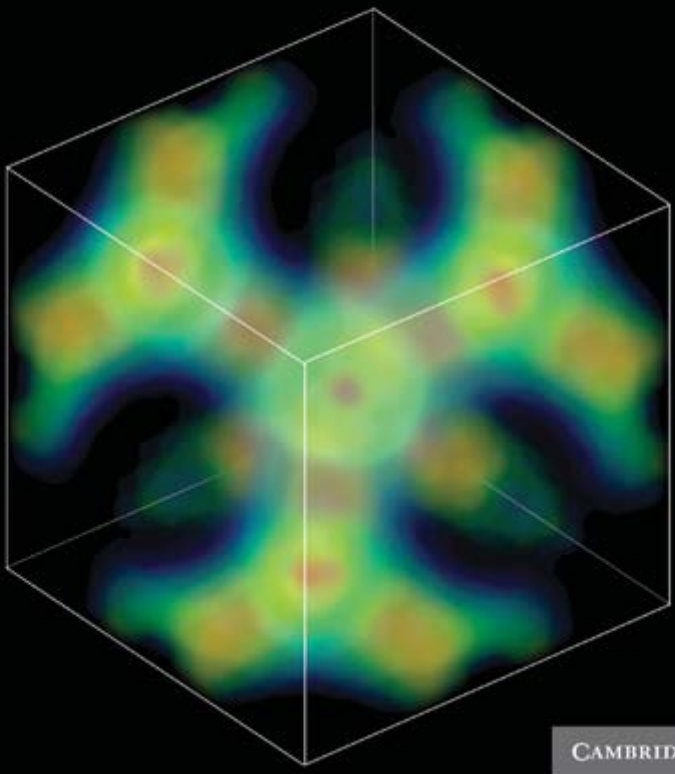
Richard M. Martin

Electronic Structure

Basic Theory and Practical Methods



CAMBRIDGE

CAMBRIDGE

more information – www.cambridge.org/9780521782852

ELECTRONIC STRUCTURE

The study of the electronic structure of materials is at a momentous stage, with new algorithms and computational methods, and rapid advances in basic theory. Many properties of materials can now be determined directly from the fundamental equations for the electrons, providing new insights into critical problems in physics, chemistry, and materials science. This book is the first of two volumes that provide a unified exposition of the basic theory and methods of electronic structure, together with instructive examples of practical computational methods and real-world applications. These books are appropriate for both graduate students and practicing scientists. This volume describes the approach most widely used today – density functional theory – with emphasis upon understanding the ideas, practical methods, and limitations. Many references are provided to original papers, pertinent reviews, and widely available books. Included in each chapter is a short list of the most relevant references and a set of exercises that reveal salient points and challenge the reader.

Richard M. Martin received his Ph.D. from the University of Chicago in 1969, followed by post-doctoral research at Bell Laboratories. In 1971 he joined the Xerox Palo Alto Research Center in California where he became Principal Scientist and Consulting Professor at Stanford University. Since 1987 he has been Professor of Physics at the University of Illinois at Urbana-Champaign, where he has organized courses, workshops, and schools on electronic structure as well as founding the Materials Computation Center. In 2008 he was appointed again as Consulting Professor at Stanford University. He has made important contributions to many areas of modern electronic structure, with over 200 published papers. He is a fellow of the American Physical Society and the American Association for the Advancement of Science, and he is a recipient of the Alexander von Humboldt Senior Scientist Award. He has served on editorial boards of the American Physical Society, including *Physical Review*, *Physical Review Letters*, and *Reviews of Modern Physics* where he was associate editor for condensed matter theory.

Electronic Structure

Basic Theory and Practical Methods

Richard M. Martin

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo,
Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521534406

First published 2004
Reprinted 2005
First paperback edition with corrections 2008
Reprinted 2010

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Martin, Richard M., 1942–
Electronic structure: basic theory and practical methods / Richard M. Martin.
p. cm.
Includes bibliographical references and index.
ISBN 978 0 521 78285 2
1. Electronic structure. I. Title.
QC176.8.E4M368 2003
530.4′11–dc21 2003044028

ISBN 978-0-521-78285-2 hardback
ISBN 978-0-521-53440-6 paperback

To Beverly

Contents

Preface

The field of electronic structure is at a momentous stage, with rapid advances in basic theory, new algorithms, and computational methods. It is now feasible to determine many properties of materials directly from the fundamental equations for the electrons and to provide new insights into vital problems in physics, chemistry, and materials science. Increasingly, electronic structure calculations are becoming tools used by both experimentalists and theorists to understand characteristic properties of matter and to make specific predictions for real materials and experimentally observable phenomena. There is a need for coherent, instructive material that provides an introduction to the field and a resource describing the conceptual structure, the capabilities of the methods, limitations of current approaches, and challenges for the future.

The purpose of this and a second volume in progress is to provide a unified exposition of the basic theory and methods of electronic structure, together with instructive examples of practical computational methods and actual applications. The aim is to serve graduate students and scientists involved in research, to provide a text for courses on electronic structure, and to serve as supplementary material for courses on condensed matter physics and materials science. Many references are provided to original papers, pertinent reviews, and books that are widely available. Problems are included in each chapter to bring out salient points and to challenge the reader.

The printed material is complemented by expanded information available on-line at a site maintained by the Electronic Structure Group at the University of Illinois (see Ch. 24). There one can find codes for widely used algorithms, more complete descriptions of many methods, and links to the increasing number of sites around the world providing codes and information. The on-line material is coordinated with descriptions in this book and will contain future updates, corrections, additions, and convenient feedback forms.

The content of this work is determined by the conviction that "electronic structure" should be placed in the context of fundamental issues in physics, while at the same time emphasizing its role in providing useful information and understanding of the properties of materials. At its heart, electronic structure is an interacting many-body problem that ranks among the most pervasive and important in physics. Furthermore, these are problems that must be solved with great accuracy in a vast array of situations to address issues relevant to materials. Indeed, many-body methods, such as quantum Monte Carlo and many-body perturbation

theory, are an increasing part of electronic structure theory for realistic problems. These methods are the subject of the second volume.

The subjects of this volume are fundamental ideas and the most useful approaches at present are based upon independent-particle approximations. *These methods address directly and quantitatively the full many-body problem because of the ingenious formulation of density functional theory and the Kohn–Sham auxiliary system*. This approach provides a way to approach the many-body problem, whereby certain properties can be calculated, in principle exactly, and in practice very accurately for many materials using feasible approximations and independent-particle methods. This volume is devoted to independent-particle methods, with emphasis on their usefulness and their limitations when applied to real problems of electrons in materials. In addition, these methods provide the starting point for much of the work described in the planned second volume. Indeed, new ideas that build upon the construction of an auxiliary system and actual independent-particle calculations are critical aspects of modern many-body theory and computational methods that can provide quantitative description of important properties of condensed matter and molecular systems.

It is a humbling experience to attempt to bring together the vast range of excellent work in this field. Many relevant ideas and examples are omitted (or given short shrift) due to lack of space, and others not covered because of the speed of progress in the field. Feedback on omissions, corrections, suggestions, examples, and ideas are welcome in person, by e-mail, or on-line.

Outline

Part I consists of the first five chapters, which include introductory material. Chapter 1 provides historical background and early developments of the theoretical methods that are foundations for more recent developments. Chapter 2 is a short summary of characteristic properties of materials and modern understanding in terms of the electronic structure. Examples are chosen to illustrate the goals of electronic structure theory and a few of the achievements of the last decades. Further details and applications are included in later chapters. Chapters 3–5 present background theoretical material: Ch. 3 summarizes basic expressions in quantum mechanics needed later; Ch. 4 provides the formal basis for the properties of crystals and establishes notation needed in the following chapters; and Ch. 5 is devoted to the homogeneous electron gas, the idealized system that sets the stage for electronic structure of condensed matter.

Part II, Chs. 6–9, is devoted to density functional theory upon which is based much of the present-day work in the theory of electronic structure. Chapter 6 presents the basic existence theorems of Hohenberg, Kohn, and others; and Ch. 7 describes the Kohn–Sham approach, which is the theoretical basis for approximate inclusion of many-body effects in practical independent-particle equations. This approach has proven to be very successful in many problems and is by far the most widely used technique for quantitative calculations. Chapter 8 covers examples of functionals; although the primary emphasis here is the use of the functionals, selected material is included on the many-body effects implicitly incorporated

into the functionals. This is required for appreciation of the limitations of widely used approximate functionals and avenues for possible improvements. Finally, general aspects of the solution of the Kohn–Sham equations are in Ch. 9, with further details and specific applications given in later chapters.

Part III, Chs. 10 and 11, addresses the solution of mean-field Hartree–Fock and Kohn–Sham equations in the simplest case, the spherical geometry of an atom, and the generation of pseudopotentials. Atomic calculations illustrate the theory and are used directly as essential parts of the methods described later. Pseudopotentials are widely used in actual calculations on real materials and, in addition, their derivation brings out beautiful theoretical issues.

Part IV, Chs. 12–17, is devoted to the three core methods for solution of independent-particle equations in solids. The goal is to describe the methods in enough detail to show key ideas, their relationships, and relative advantages in various cases. But it is not the goal to give all details needed to construct working algorithms fully. Many noteworthy aspects are placed in appendices.

Part V, Chs. 18–23, represents the culmination of present-day electronic structure, which has flowered to produce ideas and methods that enable prediction of many properties of real materials. Probably the most important single development in recent years is the "Car–Parrinello" method (Ch. 18) that has revolutionized the field of electronic structure, making possible calculations on previously intractable problems such as solids at finite temperature, liquids, molecular reactions in solvents, etc. New developments in the understanding and use of response functions and time-dependent density functional theory have proved practical methods for computing spectra for phonons and spin excitations (Ch. 19) and optical excitations (Ch. 20). New developments in the understanding and use of Wannier functions and the theory of polarization and localization in solids (Chs. 21 and 22) have led to new understanding of issues resolved only in the last decade. Finally, satisfying local descriptions of electronic properties and potentially useful linear-scaling, "order-N" methods are described in Ch. 23.

The short chapter, Ch. 24, "Where to find more" replaces a summary; instead of attempting to summarize, it is more appropriate to point to further developments in a way that will be updated in the future, namely an online site where there is further information coordinated with this volume, computer codes, and links to many other sites.

The appendices are devoted to topics that are too detailed to include in the main text and to subjects from different fields that have an important role in electronic structure.

Acknowledgments

Four people and four institutions have played the greatest role in shaping the author and this work: the University of Chicago and my advisor Morrel H. Cohen, who planted the ideas and set the level for aspirations; Bell Labs, where the theory group and interactions with experimentalists provided diversity and demanded excellence; Xerox Palo Alto Research Center (PARC), in particular, my stimulating collaborator J. W. (Jim) Allen and my second mentor W. Conyers Herring; and the University of Illinois at Urbana-Champaign, especially my close collaborator David M. Ceperley. I am indebted to the excellent colleagues and students in the Department of Physics, the Frederick Seitz Materials Research Laboratory, and the Beckman Institute.

The actual writing of this book started at the Max Planck Institut für Festkorperforschung in Stuttgart, partially funded by the Alexander von Humboldt Foundation, and continued at the University of Illinois, the Aspen Center for Physics, Lawrence Livermore National Laboratory, and Stanford University. Their support is greatly appreciated.

Funding from the National Science Foundation, the Department of Energy, the Office of Naval Research, and the Army Research Office during the writing of this book is gratefully acknowledged.

Appreciation is due to countless people who cannot all be named. Many colleagues who provided figures are specifically acknowledged in the text. Special thanks are due to David Drabold, Beverly Martin, and Richard Needs for many comments and criticisms on the entire volume. Others who contributed directly in clarifying the arguments presented here, correcting errors, and critical reading of the manuscript are: V. Akkinseni, O. K. Andersen, V. P. Antropov, E. Artacho, S. Baroni, P. Blöchl, M. Boero, J. Chelikowsky, X. Cheng, T. Chiang, S. Chiesa, M. A. Crocker, D. Das, K. Delaney, C. Elliott, G. Galli, O. E. Gunnarsson, D. R. Hamann, V. Heine, L. Hoddeson, V. Hudson, D. D. Johnson, J. Junquera, J. Kim, Y.-H. Kim, E. Koch, J. Kübler, K. Kunc, B. Lee, X. Luo, T. Martinez, J. L. Martins, N. Marzari, W. D. Mattson, I. I. Mazin, A. K. McMahan, V. Natoli, O. H. Nielsen, J. E. Northrup, P. Ordejon, J. Perdew, W. E. Pickett, G. Qian, N. Romero, D. Sanchez-Portal, S. Satpathy, S. Savrosov, E. Schwegler, G. Scuseria, E. L. Shirley, L. Shulenburger, J. Soler, I. Souza, V. Tota, N. Trivedi, A. Tsolakidis, D. H. Vanderbilt, C. G. Van de Walle, M. van Schilfgaarde, I. Vasiliev, J. Vincent, T. J. Wilkens. For corrections in 2008, I am indebted to K. Belashchenko, E. K. U. Gross, I. Souza, A. Torralba, C. G. Van de Walle, and J.-X. Zhu.

Notation

Abbreviations

BZ	first Brillouin zone
wrt	with respect to
+c.c.	denotes adding the complex conjugate of the preceding quantity

General physical quantities

E	energy
Ω	volume (to avoid confusion with V used for potential)
$P = -(\mathrm{d}E/\mathrm{d}\Omega)$	pressure
$B = \Omega(\mathrm{d}^2E/\mathrm{d}\Omega^2)$	bulk modulus (inverse of compressibility)
$H = E + P\Omega$	enthalpy
$u_{\alpha\beta}$	strain tensor (symmetrized form of $\epsilon_{\alpha\beta}$)
$\sigma_{\alpha\beta} = -(1/\Omega)(\partial E/\partial u_{\alpha\beta})$	stress tensor (note the sign convention)
$\mathbf{F}_I = -(\mathrm{d}E/\mathrm{d}\mathbf{R}_I)$	force on nucleus I
$C_{IJ} = \mathrm{d}^2E/\mathrm{d}\mathbf{R}_I\mathrm{d}\mathbf{R}_J$	force constant matrix
$n(\mathbf{r})$	density of electrons

Notation for crystals

Ω_{cell}	volume of primitive cell
$\mathbf{a}_i$	primitive translation vectors
$\mathbf{T}$ or $\mathbf{T}(\mathbf{n}) \equiv \mathbf{T}(n_1, n_2, n_3) = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$	lattice translations
$\tau_s, s = 1, \ldots, S$	positions of atoms in the basis
$\mathbf{b}_i$	primitive vectors of reciprocal lattice
$\mathbf{G}$ or $\mathbf{G}(\mathbf{m}) \equiv \mathbf{G}(m_1, m_2, m_3) = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + m_3\mathbf{b}_3$	reciprocal lattice vectors
$\mathbf{k}$	wavevector in first Brillouin zone (BZ)
$\mathbf{q}$	general wavevector ($\mathbf{q} = \mathbf{k} + \mathbf{G}$)

Hamiltonian and eigenstates

$\hat{H}$	hamiltonian for either many particles or a single particle
$\Psi(\{\mathbf{r}_i\})$	Many-body wavefunction of a set of particle positions $\mathbf{r}_i,\ i = 1, N_{\text{particle}}$; spin is assumed to be included in the argument $\mathbf{r}_i$ unless otherwise specified
E_i	energy of many-body state
$\Phi(\{\mathbf{r}_i\})$	single determinant uncorrelated wavefunction
$H_{m,m'}$	matrix element of hamiltonian between states m and m'
$S_{m,m'}$	overlap matrix elements of states m and m'
$\psi_i(\mathbf{r})$	independent-particle wavefunction or "orbital," $i = 1, \ldots, N_{\text{states}}$
ε_i	independent-particle eigenvalue, $i = 1, \ldots, N_{\text{states}}$
$f_i = f(\varepsilon_i)$	occupation of state i where f is the Fermi function
$\psi_i^\sigma(\mathbf{r})$, ε_i^σ	used when spin is explicitly indicated
$\alpha_i(\sigma_j)$	spin wavefunction for particle j; $i = 1, 2$
$\phi_i(\mathbf{r}_j, \sigma_j)$	single particle "spin-orbitals" $(= \psi_i^\sigma(\mathbf{r}_j) \times \alpha_i(\sigma_j))$
$\psi_l(r)$	single-body radial wavefunction $(\psi_{l,m}(\mathbf{r}) = \psi_l(r) Y_{lm}(\theta, \phi))$
$\phi_l(r)$	single-body radial wavefunction $\phi_l(r) = r\psi_l(r)$
$\eta_l(\varepsilon)$	phase shift
$\psi_{i,\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{\mathbf{i}\mathbf{k}\cdot\mathbf{r}} u_{i,\mathbf{k}}(\mathbf{r})$	Bloch function in crystal, with $u_{i,\mathbf{k}}(\mathbf{r})$ periodic
$\varepsilon_{i,\mathbf{k}}$	eigenvalues that define bands as a function of $\mathbf{k}$
$\hat{H}(\mathbf{k})$	"gauge transformed" hamiltonian given by Eq. (4.37); eigenvectors are the periodic parts of the Bloch functions $u_{i,\mathbf{k}}(\mathbf{r})$
$\chi_\alpha(\mathbf{r})$	single-body basis function, $\alpha = 1, \ldots, N_{\text{basis}}$. Orbital i is expanded in basis functions α, i.e. $\psi_i(\mathbf{r}) = \sum_\alpha c_{i\alpha} \chi_\alpha(\mathbf{r})$
$\chi_\alpha(\mathbf{r} - (\tau + \mathbf{T}))$	localized orbital basis function on atom at position τ in cell labelled by translation vector $\mathbf{T}$
$\chi^{\text{OPW}}(\mathbf{r})$, $\chi^{\text{APW}}(\mathbf{r})$, $\chi^{\text{LMTO}}(\mathbf{r})$	Basis function for orthogonalized, augmented or muffin-tin orbital basis functions
$w_i(\mathbf{r} - \mathbf{T})$	Wannier function i associated with band i and cell $\mathbf{T}$
$\tilde{w}_i(\mathbf{r} - \mathbf{T})$	Non-orthogonal transformation of Wannier functions

Density functional theory

$F[f]$	General notational for F a functional of the function f
$E_{\text{xc}}[n]$	exchange–correlation energy in Kohn–Sham theory
$\epsilon_{\text{xc}}(\mathbf{r})$	exchange–correlation energy per electron
$V_{\text{xc}}(\mathbf{r})$	exchange–correlation potential in Kohn–Sham theory
$V_{\text{xc}}^\sigma(\mathbf{r})$	exchange–correlation potential for spin σ
$f_{\text{xc}}(\mathbf{r}, \mathbf{r}')$	Response $\delta^2 E_{\text{xc}}[n]/\delta n(\mathbf{r})\delta n(\mathbf{r}')$

Response function and correlation functions

$\chi(\omega)$	general response function
$\chi_0(\omega)$	general response function for independent particles
$K(\omega)$	Kernel in self-consistent response function $\chi^{-1} = [\chi^0]^{-1} - K$
$\epsilon(\omega)$	frequency dependent dielectric function
$n(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$	pair distribution
$g(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$	normalized pair distribution (often omitting the spin indices)
$G(z, \mathbf{r}, \mathbf{r}')$ or $G_{m,m'}(z)$	Green's function of complex frequency z
$\rho(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$	density matrix
$\rho_\sigma(\mathbf{r}, \mathbf{r}')$	density matrix diagonal in spin for independent-particles

PART I

OVERVIEW AND BACKGROUND TOPICS

1

Introduction

Without physics there is no life

Taxi driver in Minneapolis

Summary

Since the discovery of the electron in 1896–1897, the theory of electrons in matter has ranked among the great challenges of theoretical physics. The fundamental basis for understanding materials and phenomena ultimately rests upon understanding electronic structure. This chapter provides a brief outline with original references to early developments of quantum mechanics and the pioneering quantitative theories that foreshadowed most of the methods in use today.

Electrons and nuclei are the fundamental particles that determine the nature of the matter of our everyday world: atoms, molecules, condensed matter, and man-made structures. Not only do electrons form the "quantum glue" that holds together the nuclei in solid, liquid, and molecular states, but also electron excitations determine the vast array of electrical, optical, and magnetic properties of materials. The theory of electrons in matter ranks among the great challenges of theoretical physics: to develop theoretical approaches and computational methods that can accurately treat the interacting system of many electrons and nuclei found in condensed matter and molecules.

1.1 Quantum theory and the origins of electronic structure

Although *electric* phenomena have been known for centuries, the story of *electronic structure* begins in the 1890s with the discovery of the electron as a particle – a fundamental constituent of matter. Of particular note, Hendrik A. Lorentz[1] modified Maxwell's theory of electromagnetism to interpret the electric and magnetic properties of matter in terms of the motion of charged particles. In 1896, Pieter Zeeman, a student of Lorentz in Leiden, discovered [4] the splitting of spectral lines by a magnetic field, which Lorentz explained with his electron theory, concluding that radiation from atoms was due to negatively charged particles with a very small mass. The discovery of the electron in experiments on ionized

[1] The work of Lorentz and many other references can be found in a reprint volume of lectures given in 1906 [3].

gases by J. J. Thomson at the Cavendish Laboratory in Cambridge in 1897 [5, 6] also led to the conclusion that the electron is negatively charged, with a charge to mass ratio similar to that found by Lorentz and Zeeman. For this work, the Nobel prize was awarded to Lorentz and Zeeman in 1902 and to Thomson in 1906.

The compensating positive charge is composed of small massive nuclei, as was demonstrated by experiments in the laboratory of Rutherford at Manchester in 1911 [7]. This presented a major problem for classical physics: how can matter be stable? What prevents electrons and nuclei from collapsing due to attraction? The defining moment occurred when Niels Bohr (at the Cavendish Laboratory for post-doctoral work after finishing his dissertation in 1911), met Rutherford and moved to Manchester to work on this problem. There he made the celebrated proposal that quantum mechanics could explain the stability and observed spectra of atoms in terms of a discrete set of allowed levels for electrons [8]. Although Bohr's model was fundamentally incorrect, it set the stage for the discovery of the laws of quantum mechanics, which emerged in 1923–1925, most notably through the work of de Broglie, Schrödinger, and Heisenberg.[2]

Electrons were also the testing ground for the new quantum theory. The famous Stern–Gerlach experiments [15, 16] in 1921 on the deflection of atoms in a magnetic field were formulated as tests of the applicability of quantum theory to particles in a magnetic field. Simultaneously, Compton [17] proposed that the electron possesses an intrinsic moment, a "magnetic doublet," based upon observations of convergence of beams of rays. Coupling of orbital angular momentum and an intrinsic electron spin of $\frac{1}{2}$ was formulated by Goudschmidt and Uhlenbeck [18], who noted the earlier hypothesis of Compton.

One of the triumphs of the new quantum mechanics, in 1925, was the explanation of the periodic table of elements in terms of electrons obeying the exclusion principle proposed by Pauli [19] that no two electrons can be in the same quantum state.[3] In work published early in 1926, Fermi [21] extended the consequences of the exclusion principle to the general formula for the statistics of non-interacting particles (see Eq. (1.1)) and noted the correspondence to the analogous formula for Bose–Einstein statistics [22, 23].[4] The general principle that the wavefunction for many identical particles must be either symmetric or antisymmetric when two particles are exchanged was apparently first discussed by Heisenberg [24] and, independently, by Dirac [25] in 1926.[5] Together with the later work [27] of Dirac formulating

[2] The development of quantum mechanics is discussed, for example, in the books by Jammer [9] and Waerden [10]. Early references and a short history are given by Messiah [11], Ch. 1. Historical development of the theory of metals is presented in the reviews by Hoddeson and Baym [12, 13] and the book *Out of the Crystal Maze* [14], especially the chapter "The development of the quantum mechanical electron theory of metals, 1926–1933" by Hoddeson, Baym, and Eckert.

[3] This was a time of intense activity by many people [14] and Pauli referred to earlier related work of E. C. Stoner [20].

[4] Note similarity of the title of Fermi's 1926 paper, "Zur Quantelung des Idealen Einatomigen Gases" with the title of Einstein's 1924 paper, "Quantentheorie des Idealen Einatomigen Gases."

[5] According to [14], Heisenberg learned of the ideas of statistics from Fermi in early 1926, but Dirac's work was apparently independent. In his 1926 paper, Dirac also explicitly pointed out that the wavefunction for non-interacting electrons of a given spin (up or down) can be written as a determinant of one-electron orbitals. However, it was only in 1929 that Slater showed that the wavefunction including spin can be written as a determinant of "spin orbitals" [26].

relativistic quantum mechanics, and the laws of statistical mechanics, the great advances of the 1920s form the basis of all modern theories of electronic structure of matter, from atoms and molecules to condensed matter.

Further progress quickly led to improved understanding of electrons in molecules and solids. The most fundamental notions of chemical bonding in molecules (rules for which had already been formulated by Lewis [28] and others before 1920) were placed upon a firm theoretical basis by quantum mechanics in terms of the way atomic wavefunctions are modified as molecules are formed (see, for example, Heitler and London in 1927 [29]). The rules for the number of bonds made by atoms were provided by quantum mechanics, which allows the electrons to be delocalized on more than one atom, lowering the kinetic energy and taking advantage of the attraction of electrons to each of the nuclei.

The theory of electrons in condensed matter is a many-body problem in which one must use statistical concepts to describe the intrinsic properties of materials in the large system thermodynamic limit. Progress toward quantitative theories requires approximations, of which the most widely used – still today – is the independent-electron approximation. Within this approximation each electron moves independently of the others, except that the electrons obey the exclusion principle and each moves in some average effective potential which may be determined by the other electrons. Then the state of the system is specified by independent-particle eigenstates, labeled by i, with occupation numbers f_i, which in thermal equilibrium are given by

$$f_i^\sigma = \frac{1}{e^{\beta(\epsilon_i^\sigma - \mu)} \pm 1}, \tag{1.1}$$

where the minus sign is for Bose–Einstein [22, 23] and the plus sign is for Fermi–Dirac statistics [21, 25]. Among the first accomplishments of the new quantum theory was the realization in 1926–1928 by Wolfgang Pauli and Arnold Sommerfeld [30, 31], that it resolved the major problems of the classical Drude–Lorentz theory.[6] The first step was the paper [30] of Pauli, submitted late in 1926, in which he showed that weak paramagnetism is explained by spin polarization of electrons obeying Fermi–Dirac statistics. At zero temperature and magnetic field, the electrons are spin paired and fill the lowest energy states up to a Fermi energy, leaving empty the states above this energy. For temperature or magnetic field non-zero, but low compared to the characteristic electronic energies, only the electron states near the Fermi energy are able to participate in electrical conduction, heat capacity, paramagnetism, and other phenomena.[7] Pauli and Sommerfeld based their successful theory of metals upon the model of a homogeneous free-electron gas, which resolved the major

[6] Simultaneous to Lorentz' development [3] of the theory of electric and magnetic properties of matter in terms of the motion of charged particles, Paul K. L. Drude developed a theory of optical properties of matter [32, 33] in a more phenomenological manner in terms of the motion of particles. Their work formed the basis of the purely classical theory that remains highly successful today, reinterpreted in the light of quantum mechanics.

[7] Sommerfeld learned of the ideas from Pauli in early 1927 and the development of the theory was the main subject of Sommerfeld's research seminars in Munich during 1927, which included participants such as Bethe, Eckhart, Houston, Pauling, and Peierls [12]. Both Pauli and Heisenberg were students of Sommerfeld, who went on to found the active centers of research in quantum theory, respectively in Zurich and Leipzig. The three centers were the hotbeds of activity in quantum theory, with visitors at Leipzig such as Slater, Peierls, and Wilson.

mysteries that beset the Drude–Lorentz theory. However, at the time it was not clear what would be the consequences of including the nuclei and crystal structure in the theory, both of which would be expected to perturb the electrons strongly.

Band theory for independent electrons

The critical next step toward understanding electrons in crystals was the realization of the nature of independent non-interacting electrons in a periodic potential. This was elucidated most clearly[8] in the thesis of Felix Bloch, the first student of Heisenberg in Leipzig. Bloch [36] formulated the concept of electron bands in crystals based upon what has come to be known as the "Bloch theorem" (see Chs. 4 and 12), i.e. that the wavefunction in a perfect crystal is an eigenstate of the "crystal momentum." This resolved one of the key problems in the Pauli–Sommerfeld theory of conductivity of metals: electrons can move freely through the perfect lattice, scattered only by imperfections and displacements of the atoms due to thermal vibrations.

It was only later, however, that the full consequences of band theory were recognized. Based upon band theory and the Pauli exclusion principle, the allowed states for each spin can each hold one electron per unit cell of the crystal. Rudolf Peierls, in Heisenberg's group at Leipzig, recognized the importance of filled bands and "holes" (i.e. missing electrons in otherwise filled bands) in the explanation of the Hall effect and other properties of metals [37,38]. However, it was only with the work of A. H. Wilson [39,40], also at Leipzig in the 1930s, that the foundation was laid for the classification of all crystals into metals, semiconductors, and insulators.[9]

Development of the bands, as the atoms are brought together is illustrated in Fig. 1.1, which is based upon a well-known figure by G. E. Kimball in 1935 [34]. Kimball considered diamond-structure crystals, which were difficult to study at the time because the electron states change qualitatively from those in the atom. In his words:

> Although not much of a quantitative nature can be concluded from these results, the essential differences between diamond and the metals are apparent.

The classification of materials is based upon the filling of the bonds illustrated in Fig. 1.1, which depends upon the number of electrons:

- Insulators have filled bands with a large energy gap of forbidden energies separating the ground state from all excited states of the electrons.
- Semiconductors have only a small gap, so that thermal energies are sufficient to excite the electrons to a degree that allows important conduction phenomena.
- Metals have partially filled bands with no excitation gaps, so that electrons can conduct electricity at zero temperature.

[8] Closely related work was done simultaneously in the thesis research of Hans Bethe [35] in 1928 (student of Sommerfeld in Munich), who studied the scattering of electrons from the periodic array of atoms in a crystal.

[9] Seitz [1] further divided insulators into ionic, valence, and molecular, as done in Fig. 2.1.

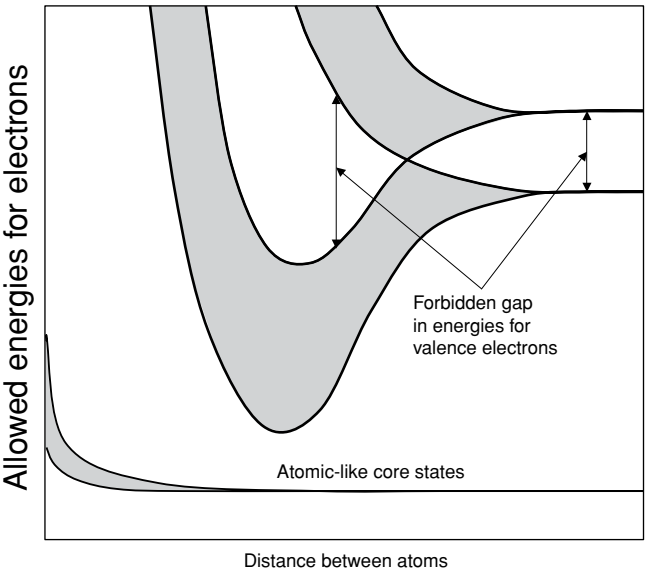


Figure 1.1. Schematic illustration of energy levels for electrons, showing the evolution from discrete atomic energies to bands of allowed states separated by forbidden gaps, as the atoms are brought together. Within the independent-particle approximation this leads to the basic division of solids into insulators, where the bands are filled with a gap to the empty states, and metals, where the bands are partially filled with no gap. Following G. Kimball [34].

1.2 Emergence of quantitative calculations

The first quantitative calculations undertaken on multi-electron systems were for atoms, most notably by D. R. Hartree[10] [43] and Hylleraas [44, 45]. Hartree's work pioneered the self-consistent field method, in which one solves the equation numerically for each electron moving in a central potential due to the nucleus and other electrons, and set the stage for many of the numerical methods still in use today. However, the approach was somewhat heuristic, and it was in 1930 that Fock [46] published the first calculations using properly antisymmetrized determinant wavefunctions, the first example of what is now known as the Hartree–Fock method. Many of the approaches used today in perturbation theory (e.g. Sec. 3.7 and Ch. 19) originated in the work of Hylleraas, which provided accurate solutions for the ground state of two-electron systems as early as 1930 [45].

The 1930s witnessed the initial formulations of most of the major theoretical methods for electronic structure of solids still in use today.[11] Among the first quantitative calculations of electronic states was the work on Na metal by Wigner and Seitz [49, 50] published in 1933

[10] D. R. Hartree was aided by his father W. R. Hartree, a businessman with an interest in mathematics who carried out calculations on a desk calculator [41]. Together they published numerous calculations on atoms. D. R. went on to become one of the pioneers of computer science and the use of electronic computers, and he published a book on calculation of electronic structure of atoms [42].

[11] The status of band theory in the early 1930s can be found in the reviews by Sommerfeld and Bethe [47] and Slater [48], and in the book *Out of the Crystal Maze* [14], especially the chapter "The development of the band theory of solids, 1933–1960" by P. Hoch.

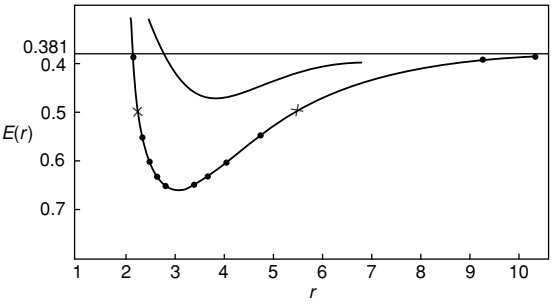


Figure 1.2. Energy versus radius for Na calculated by Wigner and Seitz [49]. Bottom curve: energy of lowest electron state calculated by the cellular method. Top curve: total energy, including an estimate of the additional kinetic energy from the homogeneous electron gas as given in Tab. 5.3. From [49].

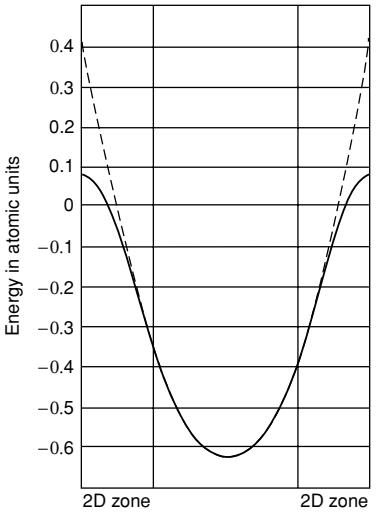


Figure 1.3. Energy bands in Na calculated in 1934 by Slater [51] using the cellular method of Wigner and Seitz [49]. The bands clearly demonstrate the nearly-free-electron character, even though the wavefunction has atomic character near each nucleus. From [51].

and 1934. They used the cellular method, a forerunner of the atomic sphere approximation, which allows the needed calculations to be done in atomic-like spherical geometry. Even with that simplification, the effort required at the time can be gleaned from their description:

> The calculation of a wavefunction took about two afternoons, and five wavefunctions were calculated on the whole, giving ten points of the figure.

The original figure, reproduced in Fig. 1.2, shows the energy of the lowest electronic state (lower curve) and the total energy of the crystal (upper curve), which are in remarkable agreement with experiment.

The electron energy bands in Na were calculated in 1934 by Slater [51] and Wigner and Seitz [50], each using the cellular method. The results of Slater are shown in Fig. 1.3;

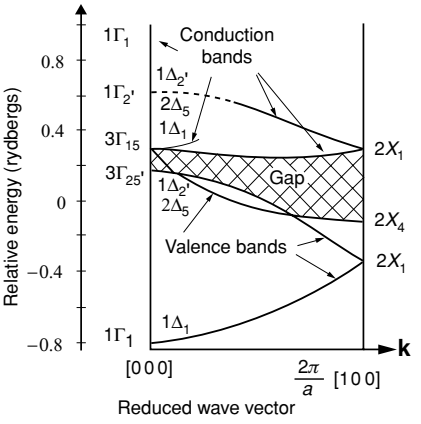


Figure 1.4. Energy bands in Ge calculated by Herman and Callaway [61] using the orthogonalized plane wave (OPW) method (Sec. 11.2). The results are for $\mathbf{k}$ along the line from Γ ($\mathbf{k} = 0$) to the X point ($\mathbf{k} = \pi/a$) as defined in Fig. 4.10. These were among the first calculations capable of reasonably accurate predictions for a diamond-structure semiconductor, and can be compared with recent calculations such in shown in Fig. 2.25. From [61].

very similar bands were found by Wigner and Seitz. Although the wavefunction has atomic character near each nucleus, nevertheless the bands are very free-electron-like, a result that has formed the basis of much of our understanding of sp-bonded metals. Many calculations were done in the 1930s and 1940s for high-symmetry metals (e.g. copper bands calculated by Krutter [52]) and ionic solids (e.g. NaCl studied by Shockley [53]) using the cellular method.

The difficulty in a general solid is to deal accurately with the electrons both near the nucleus and in the smoother bonding regions. Augmented plane waves (Ch. 16), pioneered by Slater [54] in 1937 and developed[12] in the 1950s [55,56], accomplish this with different basis sets that are matched at the boundaries. Orthogonalized plane waves (Ch. 11) were originated by Herring [57] in 1940 to take into account effects of the cores upon valence electrons. Effective potentials (forerunners of pseudopotentials, Chs. 11 and 12) were introduced in many fields of physics, e.g. by Fermi [58] in 1934 to describe scattering of electrons from atoms and neutrons from nuclei. Perhaps the original application to solids was by H. Hellmann [59,60] in 1935–1936, who developed a theory for valence electrons in metals remarkably like a modern pseudopotential calculation. Although quantitative calculations were not feasible for general classes of solids, the development of the concepts – together with experimental studies – led to many important developments, most notably the transistor.[13]

The first quantitatively accurate calculations of bands in difficult cases like semiconductors, where the electronic states are completely changed from atomic states as shown in Fig. 1.1, were done in the early 1950s, as reviewed by Herman [62, 63].[14] For example, Fig. 1.4 shows the bands of Ge calculated by Herman and Callaway [61] in 1953. They

[12] Apparently, the first published use of the term "augmented plane waves" was in the 1953 paper by Slater [55].

[13] Two of the inventors of the transistor, J. Bardeen (student of Wigner) and W. Shockley (student of Slater), did major original work in electronic structure as the topics of their theses.

[14] In his readable account in *Physics Today* [63], Herman recounts that many of the calculations were done by his mother (cf., the role of D. R. Hartree's father).

pointed out that their gap was larger than the experimental value. It turns out that this is correct: the gap in the direction studied should be larger than the lowest gap, which is in a different direction in the Brillouin zone – harder to calculate at the time. Comparison with recent calculations, e.g. in Fig. 2.25, shows that the results were basically correct.

1.3 The greatest challenge: electron correlation

Even though band theory was extremely successful in describing electrons in solids, as correlated with one another only through the exclusion principle and interacting only via the effects of some average potential, the great question was: what are the consequences of electron–electron interactions? One of the most important effects of this interaction was established early in the history of electronic structure: the underlying cause of magnetism was identified by Heisenberg [64] and Dirac [65] in terms of the "exchange energy" of interacting electrons, which depends upon the spin state and the fact that the wavefunction must change sign when two electrons are exchanged.[15] In atomic physics and chemistry, it was quickly realized that accurate descriptions must go beyond the effective independent-electron approximations because of strong correlations in localized systems and characteristic bonds in molecules [69].

In condensed matter, the great issues associated with electron–electron interactions were posed succinctly in terms of metal–insulator transitions described by Eugene Wigner [70] and Sir Nevill Mott [71–73], upon which was built much of the research on many-body effects in the 1950s to the present.[16] One way to pose the issues is to contrast the formation of bands shown in Fig. 1.1 with the effects of interactions, which are strongest in localized systems, e.g. in the atomic limit. If the atom is an open-shell system, it is well-known that Coulomb interactions lead to splitting of the independent-particle electron states into multiplets, with the ground state given by Hund's rules [74–76]. In general, there is competition between banding effects, expected to be dominant at high densities, and many-body atomic-like effects, expected to be dominant at low densities. The most challenging issues occur at intermediate densities where there are competing mechanisms. Characteristic examples are summarized in Ch. 2, especially Sec. 2.13.

The role of correlation among electrons stands out as defining the great questions and challenges of the field of electronic structure today. Experimental discoveries, such as the high-temperature superconductors and colossal-magneto-resistance materials, have stimulated yet new experimental techniques and brought to the fore issues of the theory of strongly correlated electrons. Perhaps the ideas are expressed most eloquently by P. W. Anderson in his book *Basic Notions of Condensed Matter Physics* [77] and in a paper [78] entitled "More is different," where it is emphasized that interactions may lead to phase transitions to states with broken symmetry, long-range order, and other collective behavior that emerge in systems of many particles. The lasting character of these notions are brought out in the

[15] This was another milestone for quantum mechanics, since it follows from very general theorems [66–68] that in classical mechanics, the energy of a system of charges cannot depend upon the magnetic field.

[16] The 1977 Physics Nobel Prize was awarded to P. W. Anderson, N. F. Mott, and J. H. van Vleck "for their fundamental theoretical investigations of the electronic structure of magnetic and disordered systems."

proceedings *More is Different: Fifty Years of Condensed Matter Physics* [79]. It is vital to be mindful of the "big picture," i.e. of the possible consequences of many-body electron–electron interactions, as the community of scientists progresses toward practical, efficient, theoretical approaches that can provide ever more realistic description of the electronic structure of matter.

1.4 Recent developments

In the last decades of the 1900s many developments have set the stage for new understanding and opportunities in condensed matter physics. Certainly the most important are experimental advances: discoveries of new materials such as the fullerenes and high-temperature superconductors; discoveries of new phenomena such as superconductivity and the quantum Hall effect; and new techniques for measurements that have opened doors unimagined before, such as the scanning tunnelling microscope (STM), high-resolution photoemission, and many others. A survey of experiments is completely beyond the scope of this book, but it is essential to mention certain important experimental probes, with references to specific experiments on some occasions.

With regard to the theory, perhaps the single most influential advance was the theory of superconductivity by Bardeen, Cooper, and Schrieffer (BCS) [80], which has influenced all fields of physics by providing the basis for emergence of entirely new phenomena from cooperative motions of many particles. In the broad sense, superconductivity is "electronic structure;" however, the aspects of macroscopic coherence, applications, etc., are a field unto themselves – the subject of many volumes – and only the underlying Fermi surface properties and electron–phonon interactions are considered part of present-day electronic structure theory. Indeed, Fermi surfaces are considered in many chapters, and electron–phonon interactions is an intrinsic part of modern electronic structure theory, treated especially in Ch. 19.

In a different sense, a set of theoretical developments taken together has created a new direction of research that influences all of physics and other sciences. These are the recent advances in concepts and computational algorithms that have made it possible to treat real systems, as found in nature, as well as idealized model problems. Four developments have occurred in recent years and are now the basis for most current research in theory and computational methods for electronic structure of matter:

- density functional theory for the electronic ground state and its extensions for excited states;
- quantum Monte Carlo methods, which can deal directly with the interacting many-body system of electrons and nuclei;
- many-body perturbation methods for the spectra of excitations of the electronic system;
- computational advances that make realistic calculations feasible and in turn influence the very development of the field.

The remainder of this volume is devoted to independent-particle approaches that are largely based upon density functional theory, which is the theoretical basis for approximate

inclusion of many-body effects in the independent-particle equations. These methods have proven to be very successful in many problems and are by far the most widely used approach for quantitative calculations on realistic problems.[17] Although the primary emphasis here is on the use of the functionals, selected material is included on the many-body effects implicitly incorporated into the functionals; this is required for appreciation of the limitations of widely used approximate functionals and avenues for possible improvements.

Explicit many-body methods, such as quantum Monte Carlo [81], many-body perturbation theory [82], and dynamical mean-field theory [83] are of increasing importance and should be included in a complete exposition of electronic structure. Due to lack of space, however, the present volume is limited to approaches that involve mean-field independent-particle methods. Nevertheless, it is relevant to note that the growing use of explicit many-body methods only makes it more essential to understand independent-particle methods, because inevitably they are built upon input from independent-particle calculations.

SELECT FURTHER READING

Seitz, F. *The Modern Theory of Solids* (McGraw-Hill Book Company, New York, 1940), reprinted in paperback by Dover Press, New York, 1987. A landmark for the early development of the quantum theory of solids.

Slater, J. C. *Quantum Theory of Electronic Structure*, vols. 1–4 (McGraw Hill Book Company, New York, 1960–1972). A set of volumes containing many references to original works.

[17] A tribute to the progress in crossing the boundaries between physics, chemistry, and other disciplines is the fact that the 1998 Nobel Prize in Chemistry was shared by Walter Kohn "for his development of the density-functional theory" – originally developed in the context of solids with slowly varying densities – and by John A. Pople "for his development of computational methods in quantum chemistry."

2

Overview

Summary

Theoretical analysis of the electronic structure of matter provides understanding and quantitative methods that describe the great variety of phenomena observed. A list of these phenomena reads like the contents of a textbook on condensed matter physics, which naturally divides into ground state and excited state electronic properties. The aim of this chapter is to provide an introduction to electronic structure without recourse to mathematical formulas; the purpose is to lay out the role of electrons in determining the properties of matter and to present an overview of the challenges for electronic structure theory.

The properties of matter naturally fall into two categories determined, respectively, by the *electronic ground state* and by *electronic excited states*. This distinction is evident in the physical properties of materials and also determines the framework for theoretical understanding and development of the entire field of electronic structure. In essence, the list of ground state and excited state electronic properties is the same in most textbooks [84, 86, 88] on condensed matter physics:

- Ground state: cohesive energy, equilibrium crystal structure, phase transitions between structures, elastic constants, charge density, magnetic order, static dielectric and magnetic susceptibilities, nuclear vibrations and motion (in the adiabatic approximation), and many other properties.
- Excited states: low-energy excitations in metals involved in specific heat, Pauli spin susceptibility, transport, etc; higher energy excitations that determine insulating gaps in insulators, optical properties, spectra for adding or removing electrons, and many other properties.

The reason for this division is that materials are composed of nuclei bound together by electrons. Since typical energy scales for electrons are much greater than those associated with the degrees of freedom of the more massive nuclei, the lowest energy ground state of the electrons determines the structure and low-energy motions of the nuclei. The vast array of forms of matter – from the hardest material known, diamond carbon, to the soft lubricant, graphite carbon, to the many complex crystals and molecules formed by the elements of the

periodic table – are largely manifestations of the ground state of the electrons. Motion of the nuclei, e.g. in lattice vibrations, in most materials is on a time scale much longer than typical electronic scales, so that the electrons may be considered to be in their instantaneous ground state as the nuclei move. This is the well-known adiabatic or Born–Oppenheimer approximation [89,90] (see App. C).

Since the ground state of the electrons is an important part of electronic structure, a large part of current theoretical effort is devoted to finding accurate, robust methods to treat the ground state. To build up the essential features required in a theory, we will need to understand the typical energies involved in materials. To be able to make accurate theoretical predictions, we will need to have very accurate methods that can distinguish small energy differences between very different phases of matter. By far the most widespread approach for "first principles" quantitative calculations of solids is density functional theory [91–93], which is therefore a central topic of this book. In addition, the most accurate many-body method known at the present time, quantum Monte Carlo [81,94,95], is explicitly designed to find the properties of the ground state or thermal equilibrium.

On the other hand, for given structures formed by nuclei, electronic excitations are the essence of the "electronic properties" of matter – including electrical conductivity, optical properties, thermal excitation of electrons, phenomena caused by extrinsic electrons in semiconductors, etc. These properties are governed by the spectra of excitation energies and the nature of the excited states. There are two primary types of excitation: addition or subtraction of single electrons, and excitations keeping the number of electrons constant. Since the excitations can be rigorously regarded as a perturbation upon the ground state, the methods of perturbation theory are often key to theoretical understanding and calculation of such properties [96].

Electronic excitations also couple to nuclear motion, which leads to effects such as electron–phonon interaction. This caused broadening of electronic states and to potentially large effects in metals, since normal metals *always* have excitation energies at arbitrarily low energies, which therefore mix with low-energy nuclear excitations. The coupling can lead to phase transitions and qualitative new states of matter, such as the superconducting state. Here, we will consider the theory that allows us to understand and calculate electron–phonon interactions (for example in Ch. 19), but we will not deal explicitly with the resulting phase transitions or superconductivity itself.

2.1 Electronic ground state: bonding and characteristic structures

The challenge for electronic structure theory is to provide universal methods that accurately describe real systems in nature. They must not be limited to any particular type of bonding, since otherwise they would not be successful. Nevertheless, we want the theory to provide understanding, and we seek methods that will not only provide numerical analysis (e.g. binding energies) but that will also make possible analysis of general problems using simple pictures that describe the dominant mechanisms governing the stable structures of matter.

The stable structures of solids are most naturally classified in terms of the electronic ground state, which determines the bonding of nuclei. More extensive discussion of bonding

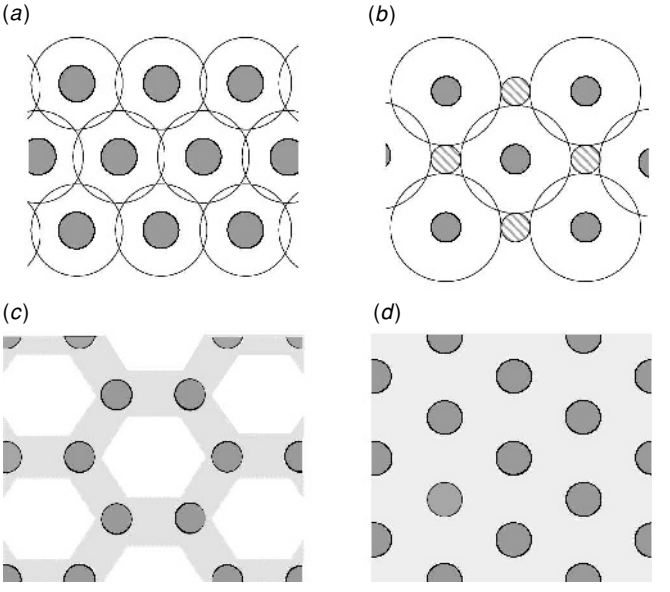


Figure 2.1. The four primary types of bonding in solids. (*a*) Closed-shell systems, typified by rare gases, remain atom-like, with weak bonding, and tend to form close-packed solids, such as fcc, hcp, and bcc. (*b*) Ionic crystals can often be considered as closed-shell systems with large negative anions and small cations in simple structures, such as NaCl, that maximize Coulomb attraction. (*c*) Covalent bonds result from the formation of electronic states with pairs of electrons forming directional bonds, leading to open structures such as diamond or graphite. (*d*) Metallic bonding is typified by itinerant conduction electrons spread among the ion cores, forming close-packed structures that are ductile and easily form alloys.

can be found in many other references [69, 84, 86, 88]; the key points for our purposes are that the lowest energy state of the electrons determines the spatial structure of the nuclei and, conversely, the spatial structure of the nuclei provides the external potential that determines the Schrödinger equation for the electrons. We will not go into detail here; the main conclusions can be reached by considering the general nature of electronic *bands*, shown in Fig. 1.1, and *bonds*, illustrated by the ground state densities shown schematically in Fig. 2.1.

The five characteristic types of bonding are listed below, four of which are illustrated in Fig. 2.1:

1. Closed-shell systems, typified by rare gases and molecular solids, have electronic states qualitatively similar to those in the atom (or molecule), with only small broadening of the bands. Characteristic structures are close-packed solids, for the rare gases, and complex structures, for solids formed from non-spherical molecules. The binding is often described as due to weak van der Waals attraction balanced by repulsion due to overlap; however, more complete analysis reveals that other mechanisms are also important.

2. Ionic crystals are compounds formed from elements with a large difference in electronegativity. They can be characterized by charge transfer to form closed-shell ions, leading to structures with large anions in a close-packed arrangement (hcp, fcc, or bcc) and small cations in positions to maximize Coulomb attraction. However, quantitative experiments and theory, as discussed in Ch. 22, show that it is not possible to identify charges uniquely associated with ions; the key point is that the system is an insulator with an energy gap.
3. Metallic systems are conductors because there is no energy gap for electronic excitation, as illustrated in Fig. 1.1 when the bands are partially filled. Then the bands can easily accept different numbers of electrons, leading to the ability of metals to form alloys among atoms with different valency, and to the tendency for metals to adopt close-packed structures, such as fcc, hcp, and bcc (see Ch. 4). Because the homogeneous electron gas is the epitome of such behavior, it deserves special attention (Ch. 5) as an informative starting point for understanding condensed matter, especially the sp-bonded metals which are often called "simple metals." Other metals, most notably the transition series, are particularly important for their mechanical and magnetic properties, as well as providing examples of many-body effects that are a challenge to theory.
4. Covalent bonding involves a complete change of the electronic states, from those of isolated atoms or ions to well-defined bonding states in solids, illustrated by the crossover in Fig. 1.1. This involves filling of electron bands up to the energy gap: the same criterion has been recognized for the formation of covalent chemical bonds [69]. The electronic density for covalent bonding, illustrated in Fig. 2.1, has been definitively identified experimentally (e.g. [97]). Experimental densities are in good agreement with theoretical results, as illustrated in Fig. 2.2 for Si. Directional covalent bonds lead to open

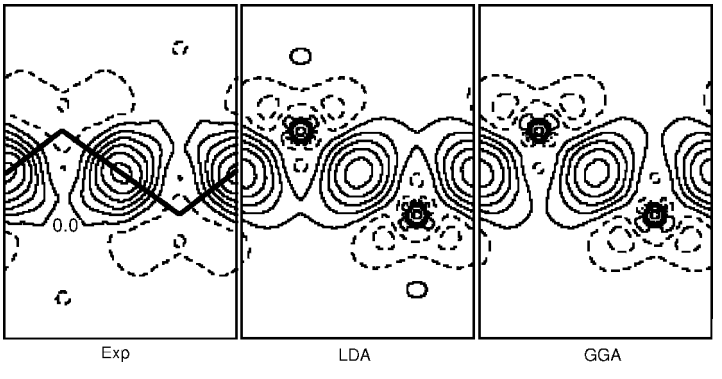


Figure 2.2. Electron density in Si shown as the *difference* of the total density from the sum of spherical atomic densities. The left-hand figure shows a contour plot of experimental (exp) measurements using electron scattering. The theoretical results were found using the linear augmented plane wave (LAPW) method (Ch. 17) and different density functionals (see text and Ch. 8). The difference density is in very good agreement with experimental measurements, with the differences in the figures due primarily to the thermal motion of the atoms in the experiment. LDA, local density approximation; GGA, generalized-gradient approximation. Provided by J. M. Zuo; similar to figure in [100].

structures, very different from the close packing typical of other types of bonding. A major success of quantitative theory is the description of semiconductors and the transition to more close-packed metallic systems under pressure.

5. Hydrogen bonding (not shown in Fig. 2.1) is often identified as another type of bonding [87]. Hydrogen is a special case, because it is the only chemically active element with no core electrons; the proton is attracted to the electrons, with none of the repulsive terms that occur for other elements due to their cores. For example, H can be stable at the bond center in silicon, as shown in Fig. 2.17. The properties of water and ice are greatly affected by the fact that protons can be shared among different molecules (see, for example, [87], [98], and *H_2O: A Biography of Water* by Philip Ball [99]). Of course, hydrogen bonding is crucial in many other molecules and is especially important for biological activity. This is one of the great challenges for ultimate application to complex materials [87]; however, we shall not deal with the complexities of structures caused by hydrogen bonding. For our purposes, it is sufficient to consider the ability of electronic structure methods to describe the strength of the hydrogen bond in selected cases.

The bonding in a real material is, in general, a combination of the above types. For example, in a metal there can be directional covalent bonding as well as contributions of ionic bonding due to local charge transfer. Molecular crystals involve strong covalent and ionic intramolecular bonds and weak intermolecular van der Waals bonding. All crystals with van der Waals bonding also have some degree of other types of bonding in which the ground state energy is lowered due to the admixture of various electronic orbitals on different atoms or molecules. Heteropolar covalent-bonded systems, such as BN, SiC, GaAs, etc., all have some ionic bonding, which goes hand-in-hand with a reduction of covalent bonding.

Electron density in the ground state

The electron density $n(\mathbf{r})$ plays a fundamental role in the theory and understanding of the system of electrons and nuclei. The density can be measured by scattering of X-rays [101] and high-energy electrons [97, 100] giving direct evidence supporting the pictures shown schematically in Fig. 2.1. Except for the lightest atoms, the total density is dominated by the core. Therefore, determination of the density reveals several features of a material: (1) the core density, which is essentially atomic-like, (2) the Debye–Waller factor, which describes smearing of the average density due to thermal and zero-point motion (dominated by the cores), and (3) the change in density due to bonding and charge transfer.

A thorough study [100] of Si has compared experimental and theoretical results calculated using the LAPW method (Ch. 17) and different density functional approximations. The total density can be compared to the theoretical value using an approximate Debye–Waller factor, which yields information about the core density. The primary conclusion is that, compared to the local density approximation (LDA), the generalized-gradient approximation (GGA, see Ch. 8) improves the description of the core density where there are large gradients;

however, there is little change in the valence region where gradients are small. In fact, nonlocal Hartree–Fock exchange is the most accurate method for determining the core density. This is a general trend that is relevant for accurate theoretical description of materials, and is discussed further in Chs. 8 and 10.

The covalent bonds are revealed by the *difference* between the crystal density and that of a sum of superimposed neutral spherical atoms.[1] This is shown in Fig. 2.2, which presents a comparison of experimental (left) with theoretical results for the LDA (middle) and GGA (right). From the figure it is apparent that the basic features are reproduced by both functionals; in addition, other calculations using LAPW [102] and pseudopotential [103, 104] methods are in good agreement. The conclusion is that the density can be measured and calculated accurately, with agreement in such detail that differences are at the level of the effects of anharmonic thermal vibrations ([100] and references cited there).

Examples of theoretically calculated valence densities [105] for the series Ge, GaAs, and ZnSe are shown later in Fig. 12.3. Theory also allows one to break the density into contributions due to each band and to a decomposition in terms of localized Wannier functions (Ch. 21) that provide much more information than the density alone.

2.2 Volume or pressure as the most fundamental variable

The equation of state as a function of pressure and temperature is perhaps the most fundamental property of condensed matter. The stable structure at a given P and T determines all the other properties of the material. The total energy E at $T = 0$ as a function of volume Ω is the most convenient quantity for theoretical analysis because it is more straightforward to carry out electronic structure calculations at fixed volume. In essence, volume is a convenient "knob" that can be tuned to control the system theoretically. Comparison of theory and experiment is one of the touchstones of "*ab initio*" electronic structure research. Because direct comparison can be made with experiment, this is one of the most important tests of the state of the theory, in particular, the approximations made to treat electron–electron interactions.

The fundamental quantities are energy E, pressure P, bulk modulus B,

$$\begin{aligned} E &= E(\Omega) \equiv E_{\text{total}}(\Omega), \\ P &= -\frac{\mathrm{d}E}{\mathrm{d}\Omega}, \\ B &= -\Omega\frac{\mathrm{d}P}{\mathrm{d}\Omega} = \Omega\frac{\mathrm{d}^2E}{\mathrm{d}\Omega^2}, \end{aligned} \tag{2.1}$$

and higher derivatives of the energy. All quantities are for a fixed number of particles, e.g. in a crystal, E is the energy per cell of volume $\Omega = \Omega_{\text{cell}}$.

The first test is to determine the theoretical prediction for the equilibrium volume Ω^0, where E is minimum or $P = 0$, and bulk modulus B for the known zero pressure crystal structure. Since Ω^0 and B can be measured with great accuracy (and extrapolated to $T = 0$),

[1] Covalent bonding is *not* readily apparent in the total density; even a superposition of atomic densities leads to a total density peaked between the atoms.

this is a rigorous test for the theory. One procedure is to calculate the energy E for several values of the volume Ω, and fit to an analytic form, e.g. the Murnaghan equation of state [108]. The minimum gives the predicted volume Ω^0 and total energy, and the second derivative is the bulk modulus B. Alternatively, P can be calculated directly from the virial theorem or its generalization (Sec. 3.3), and B from response functions (Ch. 19).

During the 1960s and 1970s computational power and algorithms, using mainly atomic orbital bases (Ch. 15) or the augmented plane wave (APW) method (Ch. 16), made possible the first reliable self-consistent calculations of total energy as a function of volume for high-symmetry solids. Examples of calculations include ones for KCl [109, 110], alkali metals [111, 112], and Cu [113].[2] A turning point was the work of Janak, Moruzzi, and Williams [106, 114], who established the efficacy of the Kohn–Sham density functional theory as a practical approach to computation of the properties of solids. They used the Koringa–Kohn–Rostocker (KKR) method (Sec. 16.3) to calculate the equilibrium volume and bulk modulus for the entire series of transition metals using the local approximation, with results shown in Fig. 2.3. Except for a few cases where magnetic effects are essential, the calculated values are remarkably accurate – within a few percent of experiment. The overall shape of the curves has a very simple interpretation: the bonding is maximum at half-filling, leading to the maximum density, binding energy, and bulk modulus. Such comparison of the predicted equilibrium properties with experimental values are now one of the routine tests of modern calculations.

Phase transitions under pressure

Pressure very different from zero is no problem for the theorist (positive or negative!), since the volume "knob" is easily turned to smaller or larger values. Here also there are excellent comparisons with experiment because there have been advances in experimental methods which have made it possible to study matter over large ranges of pressures, sufficient to change the properties of ordinary materials completely [115, 116]. In general, as the distance between the atoms is decreased, there is a tendency for all materials to transform to metallic structures, which are close packed at the highest pressures. Thus many interesting examples involve materials that have large-volume open structures at ordinary pressure, and which transform to more close-packed structures under pressure. Even though experiments are limited to structures that can actually be formed, theory has no such restrictions: understanding is gained by studying structures to quantify the reasons why they are unfavorable and to find new structures that might be metastable. This is a double-edged sword: it is very difficult for the theorist to truly "predict" new structures because of the difficulty in considering all possible structures. (Simulation techniques (Ch. 18) are beginning to reach a point where favorable structures can be found automatically, but often the time scales involved are prohibitive.) Most "predictions" have the caveat that there may be other more favorable structures not considered.

[2] A theme of the work was comparison of Slater average exchange with the Kohn–Sham formula (a factor of $2/3$ smaller). For example, Snow [113] made a careful comparison for Cu, and found the lattice constant and other properties agreed with experiment best for a factor 0.7225 instead of $2/3$.

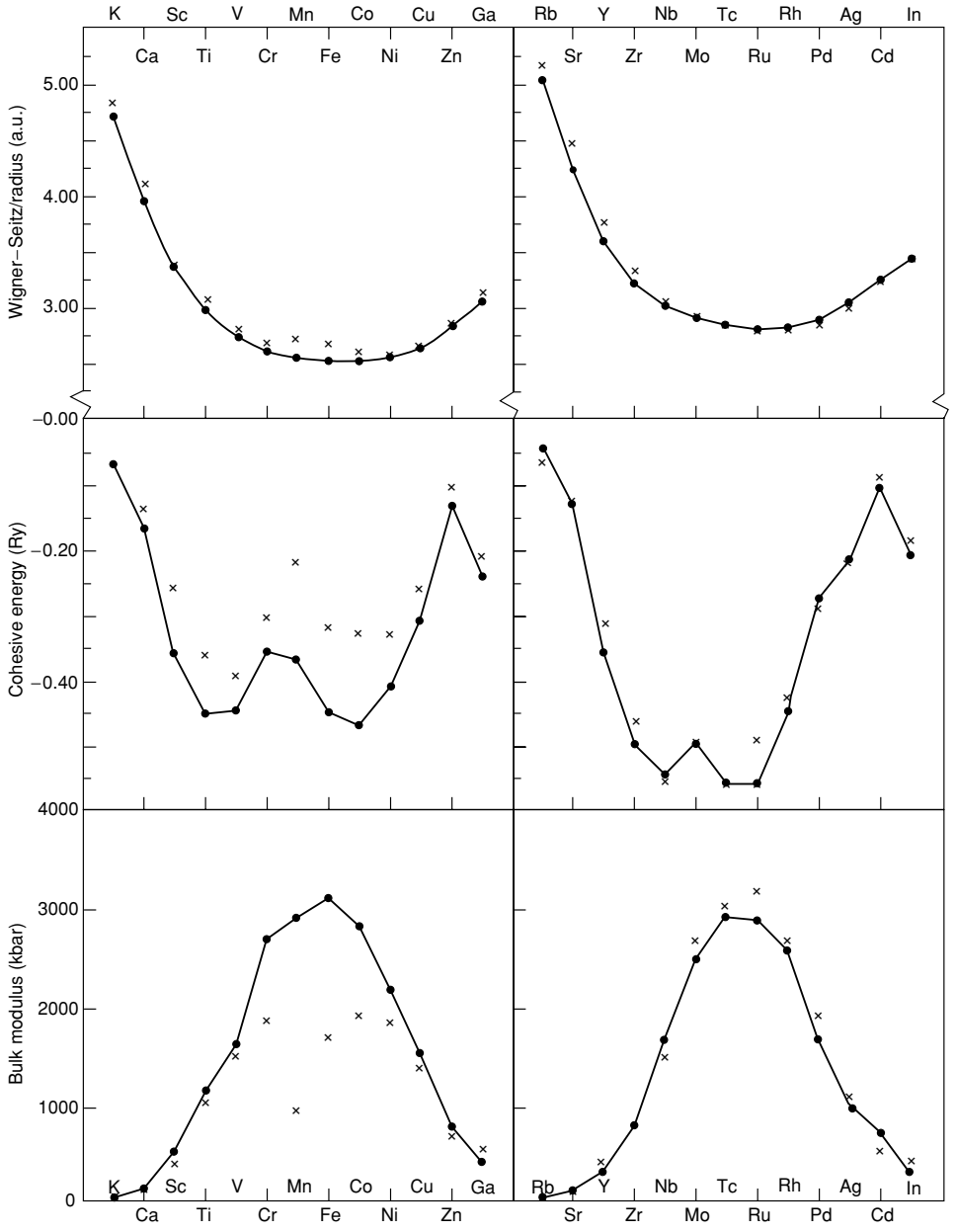


Figure 2.3. Calculated lattice constants and bulk moduli for the 3d and 4d series of transition metals, compared to experimental values denoted by x. From Moruzzi, Janak, and Williams [106] (see also [107]).

The basic quantities involved are the free energy $F(\Omega, T) = E(\Omega, T) - TS(\Omega, T)$, where the volume Ω and temperature T are the independent variables, or the Gibbs free energy $G(P,\ T) = H(P,\ T) - TS(P,\ T)$, where the pressure P and T are the independent variables. The enthalpy H is given by

$$H = E + P\Omega. \tag{2.2}$$

At temperature $T = 0$, the condition for the stable structure at constant pressure P is that enthalpy be minimum. One can also determine transition pressures by calculating $E(\Omega)$ and using the Gibbs construction of tangent lines between the $E(\Omega)$ curves for two phases, the slope of which is the pressure for the transition between the phases.

Molecular crystals and semiconductors are ideal examples to illustrate qualitative changes under pressure. The structures undergo various structural transitions and transformation from covalent open structures to metallic or ionic close-packed phases at high pressures [119, 120]. An extreme example is nitrogen, which only occurs in molecular N_2 solids and liquid at ordinary P and T, and a great challenge for many years has been to create non-molecular solid N. In Sec. 13.3 the results are given of calculations [121] that predict a new structure never before observed in any material – a real prediction.

Figure 2.4 shows the energy versus volume for Si calculated [103] using the *ab initio* plane wave pseudopotential method and the local density approximation (LDA). This approach

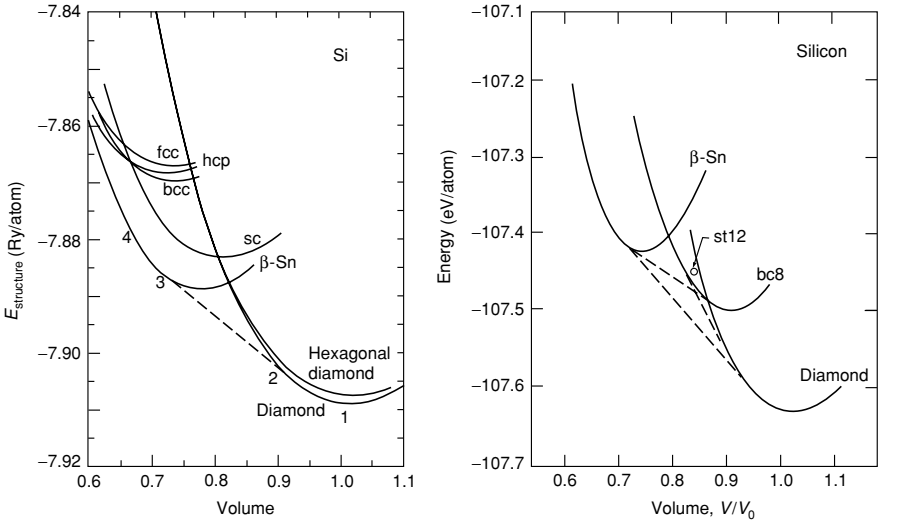


Figure 2.4. Energy versus volume for various structures of Si found using *ab initio* plane wave pseudopotential calculations. Transition pressures are given by the slopes of the tangent lines to the phases. The left-hand figure from the work of Yin and Cohen [103], is the first such fully self-consistent calculation, the success of which greatly stimulated the field of research. The tangent construction is indicated by the dashed line, the slope of which is the transition pressure. More recent calculations (e.g. Fig. 13.3) are very similar for these phases and show that improved functionals tend to lead to higher transition pressures closer to experiment. The right-hand figure is an independent calculation [117] which includes the dense tetrahedral phases bc8 and st12. The calculations find the phases to be metastable in Si but stable in C (see Fig. 2.10); similar results were found by Yin [118].

was pioneered by Yin and Cohen [103], a work that was instrumental in establishing the viability of theoretical predictions for stable structures of solids. The stable structure at $P = 0$ is cubic diamond (cd) as expected, and Si is predicted to transform to the β-Sn phase at the pressure indicated by the slope of the tangent line, $\approx$8 GPa. The right-hand figure includes phases labeled bc8 and st12, dense distorted metastable tetrahedral phases that are predicted to be almost stable; indeed they are well-known forms of Si produced upon release of pressure from the high-pressure metallic phases [122]. Many calculations have confirmed the general results and have considered many other structures [120, 123] including the simple hexagonal (sh) structure that was discovered experimentally [124, 125] and is predicted to be stable over a wide pressure range. Improved functionals increase the transition pressure by moderate amounts, as shown in Sec. 13.3, in better agreement with experimental pressure $\approx$11 GPa.

Similar calculations for carbon [117, 118, 126] correctly find the sign of the small energy difference between graphite and diamond at zero pressure, and predict that diamond will undergo phase transitions similar to those in Si and Ge, but at much higher pressures, $\approx$3,000 GPa. Interestingly, the dense tetrahedral phases are predicted to be stable above $\approx$1,200 GPa, as is indicated in the phase diagram Fig. 2.10.

There have been many such calculations and much work combining theory and experiment for the whole range of semiconductors [119, 120]. Figure 2.5 compares the lowest pressure transitions obtained experimentally with those predicted by the widely used local density approximation (LDA). *The most important conclusion to be drawn, is the agreement that depends upon the delicate energy balance between the more open covalent and more close-packed structures. This is an impressive achievement of "first principles" theory with*

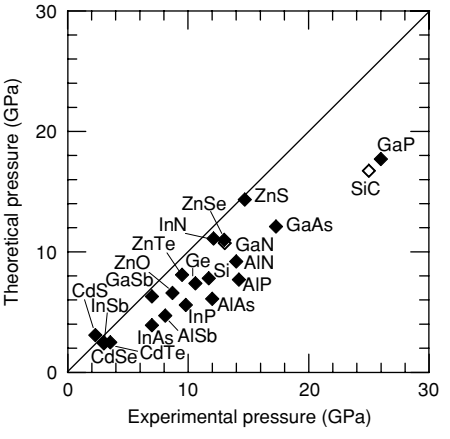


Figure 2.5. Lowest transition pressure for tetrahedral semiconductors to a high-pressure phase. Experimental pressures are plotted on the horizontal axis and theoretical pressures calculated with local density approximation (LDA) are plotted vertically. The line at 45° indicates agreement between theory and experiment. Although the agreement is impressive (see text) there is a tendency to underestimate the reported pressure. See Fig. 13.3 for the change in Si with improved functionals. Figure provided by A. Mujica, similar to plot in [120].

no adjustable parameters. The second conclusion is that the agreement is not perfect: as discussed in Ch. 8, improved functionals tend to favor more open structures. This has been tested for Si using other functionals, with results [127] much closer to experiment as shown later in Fig. 13.3. A combination of theory and experiment has led to a new picture of the systematics of the transition from tetrahedral to high-pressure structures [119, 120], with a number of competing distorted structures with low symmetries indicating interesting structural instabilities, followed by the well-known phases (β-Sn or simple hexagonal for less ionic materials, or NaCl structures for more ionic materials) and, finally, the close-packed structures.

2.3 Elasticity: stress–strain relations

The venerable subject of stress and strain in materials has also been brought into the fold of electronic structure. This means that the origins of the stress–strain relations are traced back to the fundamental definition of stress in quantum mechanics, and practical equations have been derived that are now routine tools in electronic structure theory [104, 129]. Because the development of the theory has occurred in recent years, the subject is discussed in more detail in App. G.

The basic definition of the stress tensor $\sigma_{\alpha\beta}$ is the generalization of (2.2) to anisotropic strain,

$$\sigma_{\alpha\beta} = -\frac{1}{\Omega}\frac{\partial E_{\text{total}}}{\partial u_{\alpha\beta}}, \tag{2.3}$$

where $u_{\alpha\beta}$ is the symmetric strain tensor defined in (G.2). Likewise, the theoretical expressions are the generalization of the virial theorem for pressure to anisotropic stress [104, 129].

Figure 2.6 shows the first reported calculation [128] of stress in a solid from the electronic structure, which illustrates the basic ideas. This figure shows stress as a function of uniaxial strain in Si. The linear slopes yield two independent elastic constants, and non-linear variations can be used to determine non-linear constants. For non-linear strains, it is most convenient to use Lagrange stress and strain, $t_{\alpha\beta}$ and $\eta_{\alpha\beta}$ (see [128]), which reduce to the usual expressions in the linear regime. Linear elastic constants have been calculated for many materials, with generally very good agreement with experiment, typically ≈5–10%. Non-linear constants are much harder to measure, so that the theoretical values are often predictions.

As an example of predictions of theory, Nielsen [130] has calculated the properties of diamond for general uniaxial and hydrostatic stresses, including second-, third-, and fourth-order elastic constants, internal strains, and other properties. The second-order constants agree with experiment to within ≈6% and the higher order terms are predictions. At extremely large uniaxial stresses (4 Mb) the electronic band gap collapses, and a phonon instability of the metallic diamond structure is found for compressions along the [1 1 0] and [1 1 1] crystal axes. This is relevant for ultimate stability of diamond anvils in high-pressure experiments.

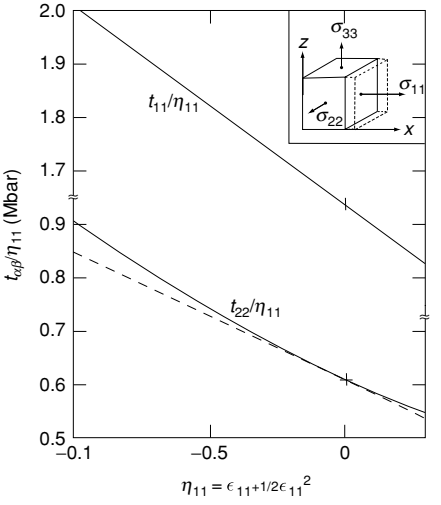


Figure 2.6. Stress in Si calculated as a function of strain in the (100) direction [128]. Two elastic constants can be found from the linear slopes; the non-linear variation determines non-linear constants. For non-linear strains it is most convenient to use Lagrange stress and strain, $t_{\alpha\beta}$ and $\eta_{\alpha\beta}$ defined in [128]. From [128].

2.4 Magnetism and electron–electron interactions

Magnetic systems are ones in which the ground state has a broken symmetry with spin and/or orbital moments of the electrons. In ferromagnetic systems there is a net moment and in antiferromagnetic systems there are spatially varying moments which average to zero by symmetry. The existence of a magnetic ground state is intrinsically a many-body effect caused by electron–electron interactions.[3] Before the advent of quantum mechanics, it was recognized that the existence of magnetic materials was one of the key problems in physics, since it can be shown that within classical physics it is impossible for the energy of the system to be affected by an external magnetic field [66–68]. The solution was recognized in the earliest days of quantum mechanics, since a single electron has half-integral spin and interacting electrons can have net spin and orbital moments in the ground state. In open-shell atoms, this is summarized in Hund's rules that the ground state has maximum total spin and maximum orbital momentum consistent with the possible occupations of electrons in the states in the shell.

In condensed matter, the key problem is the ordering of moments into long-range ordered magnetic states. In many cases this is a matter of ordering of atomic-like spins, i.e. the problem breaks into two parts: the formation of atomic moments due to electron–electron interactions and the ordering of these localized moments, which can be described by the models of Heisenberg or Ising [131]. On the other hand, many materials are magnetic even

[3] In a finite system with an odd number of electrons, there must be some unpaired spin moment. This is a relatively trivial case which can be considered separately.

though the electronic states are greatly modified from those of the atom, a situation that is termed "band magnetism" or "itinerant electron magnetism" [132].

A qualitative picture of magnetism emerges from the band picture, in which the effects of exchange and correlation among the electrons is replaced by an effective Zeeman field H_{Zeeman} represented by an added term in the hamiltonian $m(\mathbf{r})V_m(\mathbf{r})$, where m is the spin magnetization $m = n^{\uparrow} - n^{\downarrow}$ and $V_m = \mu H_{\text{Zeeman}}$.[4] In analogy to the considerations of energy versus volume in Sec. 2.2, it is most convenient to find energy for fixed field V_m and the problem is to find the minimum of the energy and the susceptibility. The basic equations are:

$$\begin{aligned} E &= E(V_m) \equiv E_{\text{total}}(V_m), \\ m(\mathbf{r}) &= -\frac{\mathrm{d}E}{\mathrm{d}V_m(\mathbf{r})}, \\ \chi(\mathbf{r},\mathbf{r}') &= -\frac{\mathrm{d}m(\mathbf{r})}{\mathrm{d}V_m(\mathbf{r}')} = \frac{\mathrm{d}^2E}{\mathrm{d}V_m(\mathbf{r})\mathrm{d}V_m(\mathbf{r}')}. \end{aligned} \tag{2.4}$$

If the electrons did not interact, the curvature of the energy χ would be positive with a minimum at zero magnetization, corresponding to bands filled with paired spins. However, exchange tends to favor aligned spins, so that $V_m(\mathbf{r})$ itself depends upon $m(\mathbf{r}')$ and can lead to a maximum at $m(\mathbf{r}') = 0$ and a minimum at non-zero magnetization; ferromagnetic if the average value $\bar{m}$ is non-zero, antiferromagnetic otherwise. In general, $V_m(\mathbf{r})$ and $m(\mathbf{r}')$ must be found self-consistently.

The mean-field treatment of magnetism is a prototype for many problems in the theory of electronic structure. Magnetic susceptibility is an example of response functions described in App. D where self-consistency leads to the mean-field theory expression (D.11); for magnetism the expressions for the magnetic susceptibility have the form first derived by Stoner [131, 133],

$$\chi = \frac{N(0)}{1 - I\ N(0)}, \tag{2.5}$$

where $N(0)$ is the density of states at the Fermi energy and the effective field has been expanded at linear order in the magnetization $V_m = V_m^{\text{ext}} + I\ m$ for the effective interaction.[5] The denominator in (2.5) indicates a renormalization of the independent-particle susceptibility $\chi^0 = N(0)$, and an instability to magnetism is heralded by the divergence when the Stoner parameter $I\ N(0)$ equals unity. Figure 2.7 shows a compilation by Moruzzi et al. [107] of $I\ N(0)$ from separate calculations of the two factors I and $N(0)$ using density functional theory. Clearly, the theory is quite successful since the Stoner parameter exceeds unity only for the actual ferromagnetic metals Fe, Co, and Ni, and it is near unity for well-known enhanced paramagnetic cases like Pd.

[4] Density functional theory, discussed in the following chapters, shows that there exists a unique mean-field potential; however, no way is known of finding it exactly and there are only approximate forms at present.

[5] The Stoner parameter $I\ N(0)$ can be understood simply as the product of the second derivative of the exchange–correlation energy with respect to the magnetization (the average effect per electron) times the density of independent-particle electronic states at the Fermi energy (the number of electrons able to participate). This idea contains the essential physics of all mean-field response functions as exemplified in App. D and Chs. 19 and 20.

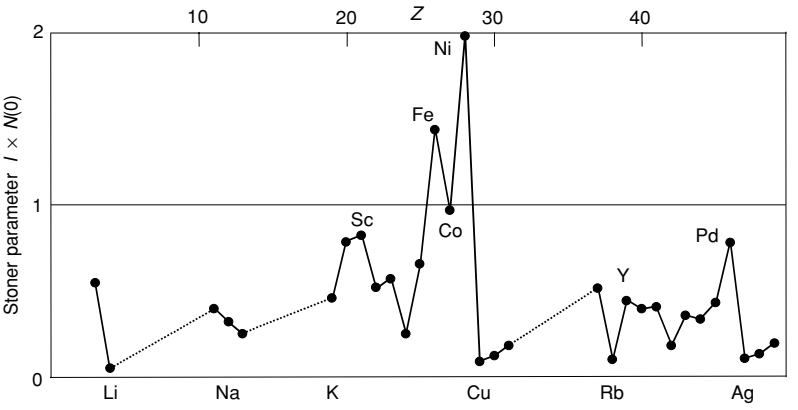


Figure 2.7. Stoner parameter for the elemental metals derived from densities of states and spin-dependent mean-field interactions from density functional theory. This shows the basic tendency for magnetism in the 3d elements Mn, Fe, Co, and Ni, due to the large density of states and interactions, both of which are consequences of the localized atomic-like nature of the 3d states discussed in the text. Data from Kübler and Eyert [134], originally from Moruzzi, et al. [107].

Modern calculations can treat the spin susceptibility and excitations accurately within density functional theory using either "frozen" spin configurations or response functions, with forms that are the same as for phonons described in the following section, Sec. 2.5 and Ch. 19. An elegant formulation of the former approach based upon a Berry's phase [135,136] is described in Sec. 19.2. Examples of both Berry's phase and response function approaches applied to Fe are given in Figs. 19.3 and 19.5.

Magnetism is difficult to treat because real materials are in neither the atomic nor the band limit. This is one of the primary examples of competition between intra-atomic correlation effects and interatomic bonding effects (the broadening in Fig. 1.1). If spin–orbit interactions can be neglected,[6] the problem divides into spin and orbital moments. Spin is relatively easy to treat in terms of spinors referred to appropriate axes in space. However, orbital moments are notoriously difficult in all cases except spherically symmetric atoms (Ch. 10). Thus the theory of magnetism in solids is one of the central challenges in condensed matter physics, intrinsically involving many-body correlation, long-range order and phase transitions, and intricate problems of the coexistence of orbital currents and crystalline order.

2.5 Phonons and displacive phase transitions

A wealth of information about materials is provided by the vibrational spectra that are measured experimentally by infrared absorptions, light scattering, inelastic neutron scattering, and other techniques. The same holds for the response of the solid to electric fields, etc. Such properties are ultimately a part of electronic structure, since the electrons determine the changes in the energy of the material if the atoms are displaced or if external fields are applied. So long as the frequencies are low the electrons can be considered to remain in

[6] Spin–orbit interactions result from relativistic effects in the core and are important in heavy atoms (Ch. 10).

their ground state, which evolves as a function of the displacements of the nuclei. The total energy can be viewed as a function of the positions of the nuclei $E(\{\mathbf{R}_I\})$ independent of the nuclear velocities. This is the adiabatic or Born–Oppenheimer regime (see Ch. 3 and App. C), which is an excellent approximation for lattice vibrations in almost all materials. In exact analogy to (2.2), the fundamental quantities are energy $E(\{\mathbf{R}_I\})$, forces on the nuclei $\mathbf{F}_I$, force constants C_{IJ},

$$\begin{aligned} E &= E(\{\mathbf{R}_I\}) \equiv E_{\text{total}}(\{\mathbf{R}_I\}), \\ \mathbf{F}_I &= -\frac{\mathrm{d}E}{\mathrm{d}\mathbf{R}_I}, \\ C_{IJ} &= -\frac{\mathrm{d}\mathbf{F}_I}{\mathrm{d}\mathbf{R}_J} = \frac{\mathrm{d}^2 E}{\mathrm{d}\mathbf{R}_I \mathrm{d}\mathbf{R}_J}, \end{aligned} \tag{2.6}$$

and higher derivatives of the energy.

Quantitatively reliable theoretical calculations have added new dimensions to our understanding of solids, providing information that is *not directly available from experiments*. For example, except in a few cases, only frequencies and symmetries of the vibration modes are actually measured; however, knowledge of eigenvectors is also required to reconstruct the interatomic force constants C_{IJ}. In the past this has led to a plethora of models for the force constants that all fit the same data (see, e.g., [137], [138] and references therein). Large differences in eigenvectors were predicted for certain phonons in GaAs, and the issues were resolved only when reliable theoretical calculations became possible [139]. Modern theoretical calculations provide complete information directly on the force constants, which can serve as a data base for simpler models and understanding of the nature of the forces. Furthermore, the same theory provides much more information: static dielectric constants, piezoelectric constants, effective charges, stress–strain relations, electron–phonon interactions, and much more.

As illustrated in the examples given here and in Ch. 19, theoretical calculations for phonon frequencies have been done for many materials, and agreement with experimental frequencies within $\approx$5% is typical. Since there are no adjustable parameters in the theory, the agreement is a genuine measure of the success of current theoretical methods for such ground state properties. This is an example where theory and experiment can work together, with experiment providing the crucial data and new discoveries, and the theory providing solid information on eigenvectors, electron–phonon interactions, and many other properties.

There are two characteristic approaches in quantitative calculations:

- Direct calculation of the total energy as a function of the positions of the atoms. This is often called the "frozen phonon" method.
- Calculations of the derivatives of the energy explicitly at any order. This is called the "response function" or "Green's function" method.

"Frozen phonons"

The term "frozen phonons" denotes the direct approach in which the total energy and forces are calculated with the nuclei "frozen" at positions $\{\mathbf{R}_I\}$. This has the great advantage

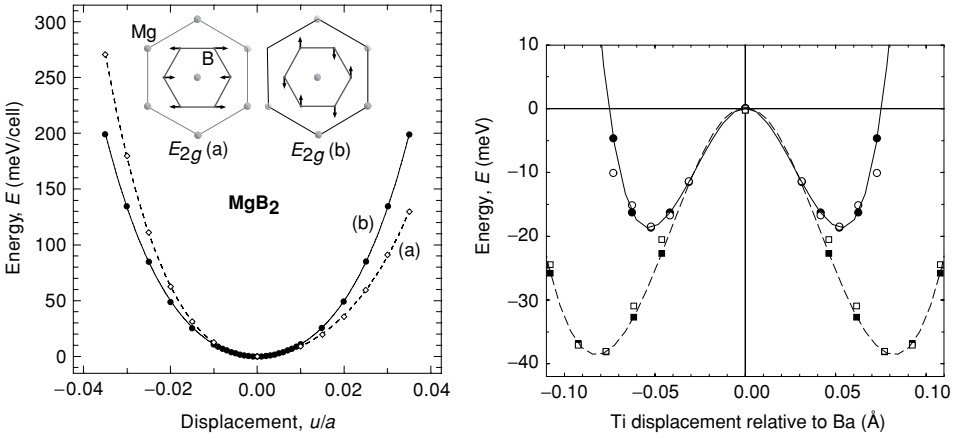


Figure 2.8. "Frozen phonon" calculations of energy versus displacement. Left: Two optic mode displacements (that are degenerate in the harmonic approximation) in the superconductor MgB_2 calculated using pseudopotentials and plane waves [140]. The deviations at large displacement illustrate their different cubic anharmonicity considered to be potentially important for superconductivity (see Sec. 2.12). Right: Two optic mode displacements of Ti atoms in $BaTiO_3$ in the tetragonal and rhombohedral directions. The points show results from two methods: dark symbols calculated in [141] using local orbitals (Ch. 15) compared with open symbols from [142] using full-potential LAPW methods (Ch. 17). The centrosymmetric position is unstable, and the most stable minimum is for the rhombohedral direction resulting in a ferroelectric phase in agreement with experiment. Figures provided by K. Kunc (left, similar to figure in [140]) and by R. Weht and J. Junquera (right, similar to figure in [141]).

that the calculations use *exactly* the same computational machinery as for other problems: for example, the same program (with only slightly different input) can be used to calculate phonon dispersion curves (Ch. 19), surface and interface structures (Ch. 13), and many other properties. Among the first calculations were phonons in semiconductors, calculated in 1976 using empirical tight-binding methods [143], and again in 1979 using density functional theory and perhaps the first use of the energy functional Eq. (9.9), to find the small changes in energy [144]. Today these are standard applications of total energy methods.

Two recent examples of energy versus displacement are shown in Fig. 2.8. On the left is shown the energy for an optic phonon displacement in MgB_2 calculated [140] using pseudopotentials and plane waves. This illustrates cubic anharmonicity, which is very sensitive to the details of the Fermi surface and may be relevant for the superconductivity recently discovered in this compound [145]. On the right-hand side is shown energy versus displacement of Ti atoms in $BaTiO_3$, which is a ferroelectric with the perovskite structure shown in Fig. 4.8. The negative curvature at the centrosymmetric position indicates the instability of this structure. Displacements in the tetragonal and rhombohedral directions are shown; the latter has the lowest energy minimum and is the predicted ferroelectric phase in agreement with experiment. The points shown are calculated using two different methods, the LAPW approach (Ch. 17) calculated in [142], and a numerical local orbital method (Ch. 15) as reported in [141]. Such calculated energies can also be used as the basis for statistical

mechanics models (see, e.g. [146]) to describe the ferroelectric phase transition as a function of temperature.

"Frozen polarization" and ferroelectricity

Incredible as it may seem, the problem of calculation of electric polarization from electron wavefunctions was only solved in the 1990s, despite the fact that expressions for the energy and forces have been known since the 1920s. As described in Ch. 22, the advance in recent years [147, 148] relates a *change* in polarization to a "Berry's phase" [149] involving the *change* in *phases* of the electron wavefunctions. The theory provides practical methods for calculation of polarization in pyroelectrics, and for effective charges and piezoelectric effects to all orders in lattice displacements and strains. It is especially important that the formulation allows calculation of the intrinsic polarization of a ferroelectric from the intrinsic wavefunctions in the bulk crystal. Examples of results are given in Ch. 22.

Linear (and non-linear) response

Response function approaches denote methods in which the force constants are calculated based upon expansions in powers of the displacements from equilibrium positions. This has the great advantage that it builds upon the theory of response functions (App. D), which can be measured in experiments and was formulated [150–152] in the 1960s. Recent developments (see the review [153] and Ch. 19) have cast the expressions in forms much more useful for computation, so that it is now possible to calculate phonon dispersion curves on an almost routine basis.

Figure 2.9 shows a comparison between experimental and theoretical phonon dispersion curves for GaAs [154]; such near-perfect agreement with experiment is found for many semiconductors using plane wave pseudopotential methods. Another example, this time for MgB_2, is shown in Fig. 2.32 using the linear muffin-tin orbital (LMTO) approach. In

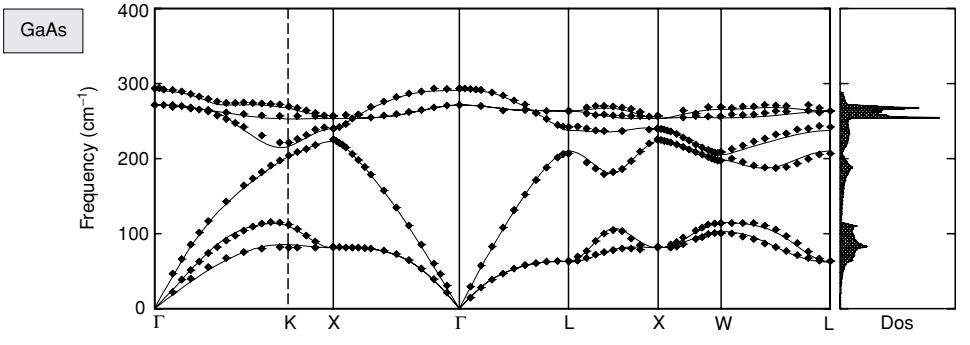


Figure 2.9. Phonon dispersion curves calculated for the semiconductor GaAs [154]. The points are from experiment and the curves from density functional theory using the response function method (Ch. 19). Similar agreement is found for the entire family of semiconductors. Calculations for many types of materials, e.g. in Figs. 19.4 and 19.5, have shown the wide applicability of this approach.

Ch. 19 examples of results for metals are shown, where the results are also impressive, but the agreement with experiment is not as good for the transition metals. Similar results are found for many materials, agreeing to within ≈5% with experimental frequencies. The response function approach is also especially efficient for calculations of dielectric functions, effective charges, electron–phonon matrix elements, and other properties, as discussed in Ch. 19.

2.6 Thermal properties: solids, liquids, and phase diagrams

One of the most important advances in electronic structure theory of recent decades is "quantum molecular dynamics" (QMD) pioneered by Car and Parrinello in 1985 [156] and often called "Car–Parrinello" simulations. As described in Ch. 18, QMD denotes classical molecular dynamics simulations for the nuclei, with the forces on the nuclei determined by solution of electronic equations as the nuclei move. By treating the *entire problem of electronic structure and motion of nuclei together*, this has opened the way for electronic structure to study an entire range of problems far beyond previous capabilities, including liquids and solids as a function of temperature beyond the harmonic approximation, thermal phase transitions such as melting, chemical reactions including molecules in solution, and many other problems. This work has stimulated many advances that are now embedded in electronic structure methods, so that calculations on molecules and complex crystals routinely optimize the structure in addition to calculating the electronic structure.

Because of advances in QMD simulations, it is now possible to determine equilibrium thermodynamic phases and dynamics of the nuclei as a function of temperature and pressure. As an example, Fig. 2.10 shows the prediction for the phase diagram of solid and liquid

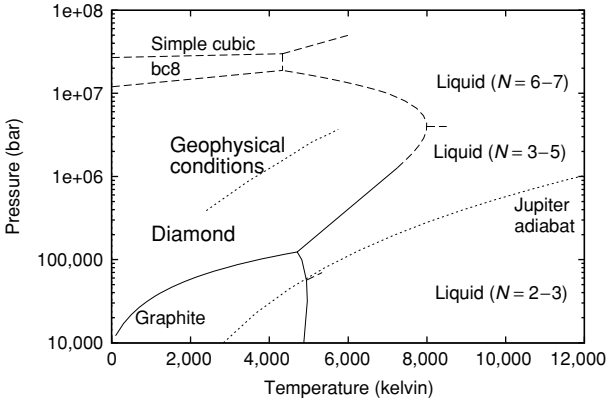


Figure 2.10. Predicted phase diagram of carbon [157] based upon experimental data at low P, T and *ab initio* simulations of various solid and liquid phases at high P, T beyond experimental capabilities. The line indicated in the liquid denotes a rapid change of average coordination N from <4 (graphite-diamond-like) to ≈ 6 (similar to liquid silicon) around 5 Mbar of pressure and $T > 10,000$ K. Although this is apparently not a phase transition, it signifies the change of slope of the liquid–solid phase boundary. Such calculations have led to revised understanding of the carbon phase diagram which previously was based upon analogies to other group IV elements [158].

carbon. This is of great interest in many fields of science, with technical, geological, and astrophysical implications, as indicated by the conditions expected inside the Earth and the planet Jupiter; however, previously there had been various wildly divergent proposals for the phase diagram [158]. As shown in Fig. 2.10, there is a predicted [159, 160] increase in melting temperature of diamond with pressure (opposite to Si and Ge) that has been confirmed by experiments [158, 161]. The higher pressure regions, however, are beyond current experimental capabilities, so that the results shown are predictions. Prominent features are that above $P > \approx 5$ Mbar, C is predicted to act like the other group VI elements, with T_{melt} decreasing with P. The diamond melting curve at high pressure is *not* directly observed in the simulation, but it can be inferred from the Claussius–Clapyron equation that relates the slope $\mathrm{d}P_{\text{melt}}/\mathrm{d}T_{\text{melt}}$ to the change in specific volume at the transition. The finding of the simulation is that at low P the liquid is less dense than diamond, whereas for $P > \approx 5$ Mbar, the nature of molten carbon changes to a higher coordination (>4) dense phase like that known to occur in molten Si and Ge at $P = 0$. At still higher pressure, static total energy calculations [117, 118] have found transitions to dense tetrahedrally coordinated structures (bc8, st12, etc.) and to the simple cubic metallic phase at $P \approx 30$ Mbar. This last prediction is confirmed by the QMD simulations, where the phase boundary shown in Fig. 2.10 has been determined directly by melting and solidification as the temperature is raised or lowered [157]. Further results are described in Sec. 18.6, especially in the lower pressure range where calculations have been done with both Car–Parrinello [159] and tight-binding [162] QMD methods.

Among the foremost challenges in geophysics is to understand the nature of the core of the Earth, which is made up primarily of Fe with other elements in solid and liquid phases. This is a case where first principles QMD can provide crucial information, complementing experiments that are very difficult at the appropriate temperature and pressure conditions found deep in the Earth. Recent work has made great progress, and examples of full thermal QMD simulations [163, 164] of Fe using the projector augmented plane wave (PAW) method (Sec. 13.2) are given in Sec. 18.6.

Water and aqueous solutions

Certainly, water is the liquid most important for life [99]. As a liquid, or in ice crystalline forms, it exemplifies the myriad of complex features due to hydrogen bonding [87, 98, 99]. QMD has opened up the possibilities for new understanding of water, ice, and aqueous solutions of ions and molecules. Of course, isolated molecules of H_2O and small clusters can be understood very well by the methods of quantum chemistry, and many properties of the condensed states are described extremely well by fitted potentials. QMD can play a special role in determining cases that are not understood from current experimental information. Examples include the actual atomic-scale nature of diffusion processes, which involves rearrangements of many molecules, the behavior of water under extreme conditions of pressure and temperature, and high pressure phases of ice.

The first step is to describe hydrogen bonding accurately within density functional theory. Tests for liquid water [165, 166] and ice [167] have shown that the results are very sensitive to both exchange and correlation. The local approximation gives bonding that is much

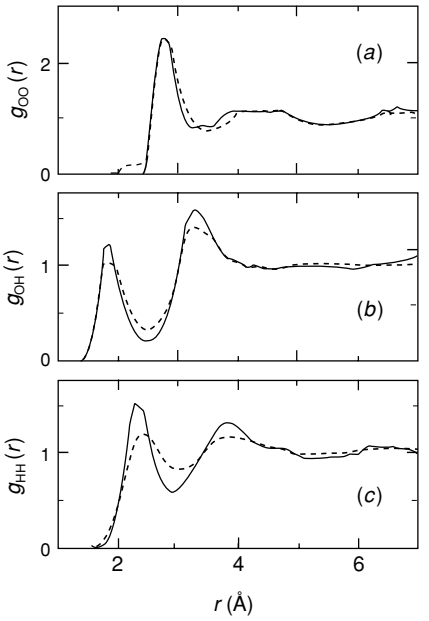


Figure 2.11. Radial density distributions $g(r)$ for O-O, O-H, and H-H distances calculated [165] with the Car–Parrinello QMD method (Ch. 18) using plane waves, pseudopotentials and the "BLYP" functional (Ch. 8) compared with experimental results. The hydrogen bonding is very sensitive to the functional. Two widely used forms (see text) appear to be good approximations [165, 166] for water, but the theoretical basis is not well understood. From [165].

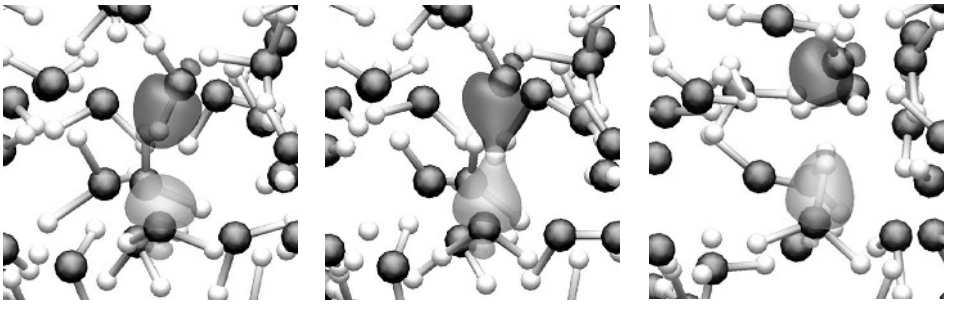


Figure 2.12. "Snapshots" of the motion of H and O atoms in a Car–Parrinello QMD simulation of water under high-pressure, high-temperature conditions [166]. The motion of the atoms is shown from left to right, particularly one proton that transfers, and the Wannier function for the electronic state that transfers to form H^+ and $(H_3O)^-$. (The Wannier functions are defined by the "maximal localization" condition of Ch. 21.) Provided by E. Schwegler; essentially the same as Fig. 2 of [166].

too strong and some correlation functionals give bonding much too weak. Two widely used generalized gradient approximation (GGA) functionals (PBE and BLYP, see Ch. 8) appear to improve the description of water [165, 166]. For example, the radial density distributions for O-O, O-H, and H-H are compared with experimental results in Fig. 2.11, taken from [165]; similar agreement is found in [166].

QMD simulations have been applied to many properties involving water, including dissociation under standard conditions (very rare events) [168], at high pressures where the molecules are breaking up [166, 169], supercritical water [170], ions in solution ([171] and references given there), and many other properties. For example, the transfer of protons and concomitant transfer of electronic orbitals are shown in Fig. 2.12, which also illustrates the use of maximally localized Wannier functions (Ch. 21) to describe the electronic states. A different approach for identification of the nature of the electronic states is the so-called "electron localization function" (ELF) described in App. H.

2.7 Atomic motion: diffusion, reactions, and catalysis

A greater challenge yet is to describe chemical reactions catalyzed in solutions or on surfaces also in solution. As an example of the important role of theoretical calculations in understanding materials science and chemistry, QMD simulations [172, 173] have apparently explained a long-standing controversy in the Ziegler–Nata reaction that is a key step in the formation of polymers from the common alpha olefins, ethylene and propylene. This is the basis for the huge chemical industry of polyethylene manufacture, used for "plastic" cups, grocery bags, the covers for CDs, etc. Since propylene is not as symmetric as ethylene, special care must be paid to produce a stereoregular molecular chain, where each monomer is bound to the next with a constant orientation. These high-quality polymers are intended for special use, e.g. for biomedical and space applications. The Ziegler–Nata process allows these cheap, harmless polymers to be made from common commercial gas, without strong acids, high temperatures, or other expensive procedures.

The process involves molecular reactions to form polymers at Ti catalytic sites on $MgCl_2$ supports; an example of a good choice for the support is made by cleaving $MgCl_2$ to form a (110) plane, as shown in Fig. 2.13. On this surface $TiCl_4$ sticks efficiently, giving a high density of active sites. The QMD simulations find that the relevant energetics,

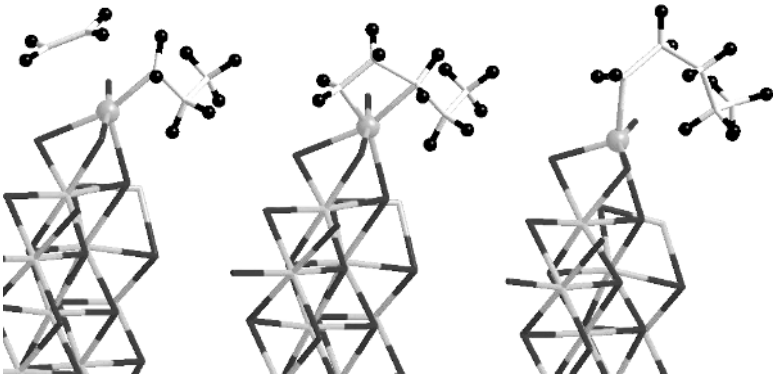


Figure 2.13. Simulation of the Ziegler–Nata reaction, which is of great importance for the production of polyethylene. Predicted steps in the main phases of the second insertion leading to the chain propagation: the π-complex (left), the transition state (middle), and insertion (right) of the ethylene molecule lengthening the polymer. Figure provided by M. Boero; essentially the same as Fig. 11 of Ref. [172].

as well as the reactivity in the alkyl chain formation process, strongly depend on the local geometry. The dynamical approach follows the reaction pathway in an unbiased way during deposition of $TiCl_4$ and complex formation, which are energetically downhill. Constrained dynamics can then be used to determine free-energy profiles and estimate activation barriers in the alkene insertion processes. Steps in the insertion of a second ethylene molecule shown in the sequence in Fig. 2.13 offers insight into the chain growth process and the stereochemical character of the polymer, providing a complete picture of the reaction mechanism.

2.8 Surfaces, interfaces, and defects

Surfaces

There is no infinite crystal in nature: every solid has a surface; all "bulk" experiments proceed with interactions through the surface. More and more experiments can probe the details of the surface on an atomic scale, e.g. the scanning tunnelling microscope and vastly improved X-ray and electron diffraction from surfaces. In no sense can this vast subject be covered here; a few examples are selected, especially semiconductors where the disruption of the strong covalent bonding leads to a variety of reconstructions of the surface. The image Fig. 2.14, from [174], illustrates essentially all aspects of the problem: atomic-scale structure, steps, and the core of the spiral, which is the end of a dislocation that continues into the bulk and controls the growth mechanism. In addition, reactions at surfaces of catalysts are particularly interesting and challenging for theory; an example is the Ziegler–Nata reaction described in Sec. 2.7.

As illustrative of the great number of theoretical studies of surfaces, ionic semiconductors, such as III–V and II–VI crystals, present challenging issues for electronic structure calculations. There is the possibility of anion- and cation-terminations with varying stoichiometry: not only must the total energy be compared for various possible reconstructions with the same numbers of atoms, but also one must compare structures with different numbers of atoms of each type, i.e. different stoichiometries. As an illuminating first step, the stoichiometry of the atoms for different types of structures can be predicted from simple electron counting rules [175, 176], i.e. charge compensation at the surface by the filling of all anion dangling bonds and the emptying of cation dangling bonds. The full analysis, however, requires that the surface energy be determined with reference to the chemical potential μ_I for each type of atom I, which can be controlled by varying its partial pressure in the gas (or other phase) in contact with the surface [177, 178], thereby allowing experimental control of the surface stoichiometry. The quantity to be minimized is not the free energy $E - T\ S$, but the grand potential [179],

$$\Omega = E - T\ S - \sum_I \mu_I N_I. \tag{2.7}$$

How can this be properly included in the theory? Fortunately, there are simplifications in a binary AB compound [179]:

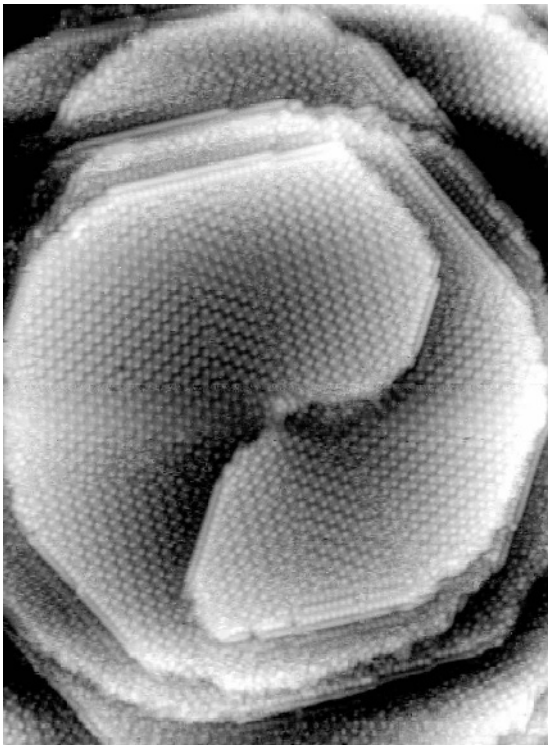


Figure 2.14. Scanning tunnelling microscope (STM) image of a GaN (0 0 0 −1) surface, illustrating many features that are challenges for theory: the atomic scale structure to perfect flat surfaces, bulk line defects that terminate the surface, and steps that form in spirals around the defect that are the sites for growth of the crystal. From Smith et al. [174].

1. The energy of the crystal E_{AB} is close to its value at $T = 0$ (corrections for finite T can be made if needed).
2. Assuming that the surface is in equilibrium with the bulk, there is only one free chemical potential, which can be taken to be μ_{A} since $\mu_{\mathrm{A}} + \mu_{\mathrm{B}} = \mu_{\mathrm{AB}} \approx E_{\mathrm{AB}}$.
3. In equilibrium, the ranges of μ_{A} and μ_{B} are limited since each can never exceed the energy of the condensed pure element, $\mu_{\mathrm{A}} \leq E_{\mathrm{A}}$, and $\mu_{\mathrm{B}} \leq E_{\mathrm{B}}$.

This is sufficient for the theory to predict reconstructions of the surface *assuming equilibrium* as a function of the real experimental conditions.

An example of recent work is the study of various ZnSe (1 0 0) surface reconstructions based upon pseudopotential plane wave calculations [180]. Examples of the structures are shown in Fig. 2.15. A $c(2 \times 2)$ reconstruction with half-monolayer coverage of two-fold coordinated Zn atoms is stable in the Zn-rich limit. Under moderately Se-rich conditions, the surface adopts a (2×1) Se-dimer phase. In the extreme Se-rich limit, the theoretical calculations predicted a new structure with one and a half monolayer coverage of Se. This

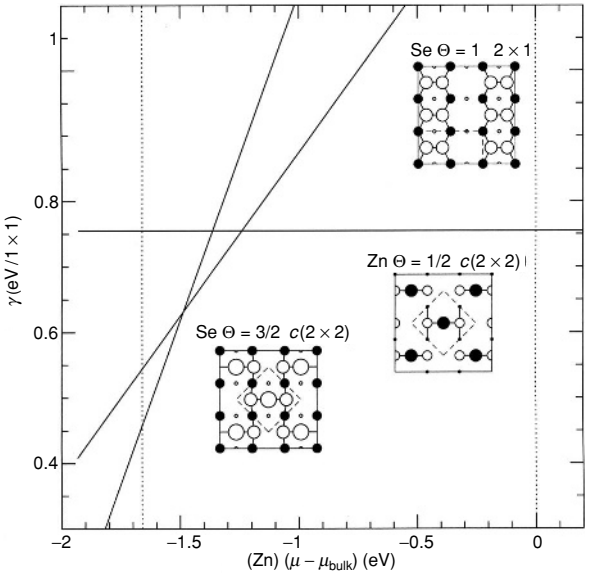


Figure 2.15. Energies of selected structures of the (1 0 0) surface of ZnSe as a function of the chemical potential (see text) calculated by the plane wave pseudopotential method. The Se-rich structure is an example of a theoretical prediction later found in experiment. Provided by A. Garcia and J. E. Northrup; essentially the same as in [180].

was proposed to account for the high growth rates observed in atomic layer epitaxy and migration enhanced epitaxy experiments at relatively low temperatures.

The structures in Fig. 2.15 also illustrate the importance of electrostatic effects in determining the pattern of a surface reconstruction that involves charge transfer within certain "building blocks." The preference for $c(2 \times 2)$ or 2×1 ordering of the building blocks is determined so that the electrostatic interaction (i.e. minimization of the surface Madelung energy – see App. F) is optimized, which was pointed out for GaAs (0 0 1) surfaces [181].

Interfaces

The surface is really just an interface between a solid and vacuum. Other interfaces include those between two materials. Of particular interest are semiconductor interfaces that have been prepared and characterized with great control. Of particular importance are the "band offsets" at the interface which confined the carriers in semiconductor quantum devices [182]. This is an issue that requires two aspects of electronic structure: establishing the proper reference energy, which depends upon calculations of the interface dipole (Sec. F.5), and the single particle energies relative to the reference. Studies of semiconductors described in Ch. 13 are an example of a triumph of theory working with experiment, with a revision of previously held rules due to theoretical calculations [183, 184].

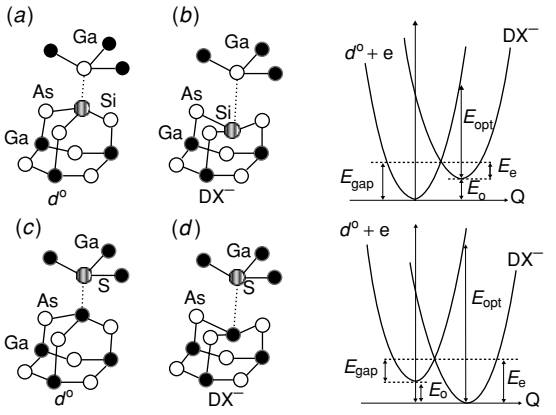


Figure 2.16. GaAs doped with Si was predicted by theory [185, 186] to have "negative U" defect centers formed by the off-center displacement of Si atoms as a function of charge. Because of the change in band structure upon alloying with Al, the stability of the center changes, as shown on the right-hand side of the figure. Adapted from [185].

Defects

Every solid has defects. Crystals are particularly interesting because there are characteristic defects: point defects, like vacancies and interstitials; one-dimensional defects, like dislocations; and two-dimensional defects, like grain boundaries, interfaces (and the surface viewed as a defect in the infinite crystal). An excellent example showing the role that modern simulations can play in revealing the nature of important defects is the saga of the DX center in GaAs [185]. Extensive experimental and theoretical work had shown that donors in III–V compounds give rise to two types of electronic states, a shallow delocalized state associated with the normal position of the donor and a deep state, called "DX," associated with lattice displacements. Chadi and Chang [185] carried out theoretical total energy calculations to show that the atomic displacements responsible for the formation of DX centers in Si- and S-doped GaAs are large bond-rupturing displacements, indicating that DX is a highly localized and negatively charged defect. They found that the defect center changes abruptly, with the atoms rebonding as shown in Fig. 2.16 and the charge state changing by two electrons to form a "negative U" center. They concluded that DX centers are an unavoidable feature of substitutional dopants and suggested alternative doping procedures.

Another case that apparently has been sorted out after many years of research is the role of interstitial H in Si (see [187] and references given there). Extensive calculations for H in different charge states and positions has shown that it forms a "negative-U" system, changing from H^+ to H^- as the Fermi energy of the electrons is increased. The neutral state H^0 is always higher energy, meaning that if one electron is added to H^+, it attracts a second electron to create H^-. How can this happen since electrons repel one another? The answer

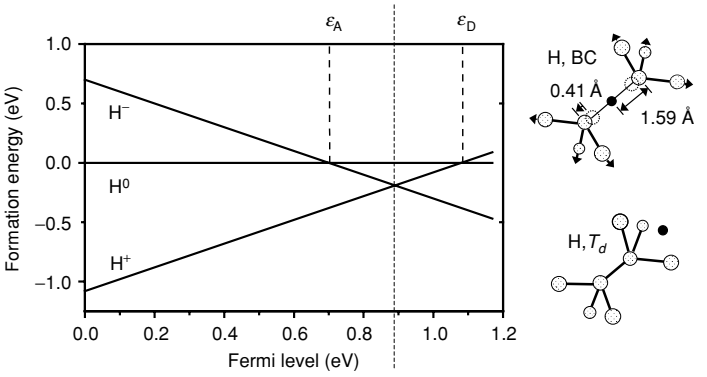


Figure 2.17. Minimum energy positions of H in Si for the three possible charge states. At the left-hand side of the figure is shown the formation energy as a function of the Fermi energy, which is the chemical potential for electrons and plays the same role as the chemical potentials in Fig. 2.15. This shows that H acts as a "negative-U" defect, with the lowest energy state changing from H^+ in p-type material to H^- in n-type. The Fermi energy is referred to the top of the valence band and the formation energy to the neutral H^0 energy. At the right-hand side of the figure are shown the minimum energy structures: H^+ (and also H^0) occupies bond-center positions, whereas H^- prefers the position in the center of the tetrahedral hole in the Si crystal. From [187].

shown in Fig. 2.17 is that H is mobile and moves to different positions in the different charge state, leading to the energy of formation of the interstitial shown on the right-hand side of the figure.

2.9 Nanomaterials: between molecules and condensed matter

Among the most dynamic new areas of experimental and theoretical research are nanoclusters and nanostructures. In some ways nanoclusters are just large molecules; there is no precise distinction, but nanoclusters share the property of condensed matter that clusters of varying size are made of. Yet nanoclusters are small enough that the properties can be tuned by varying the size. This is exemplified by metallic clusters, the size of which can be varied from a few atoms to macroscopic dimensions. At intermediate sizes the properties are controlled by finite size quantization effects and by the fact that a large fraction of the atoms are in surface regions. Because the structure is extremely hard to determine directly from experiment, theory has a great role to play. The observation of "magic numbers" for Na clusters can be understood on very simple grounds in terms of filling of shells in a sphere [188, 189]. The atomic-scale structures and optical spectra of such clusters are described in more detail in Ch. 20.

Semiconductor nanostructures have been of particular interest because confinement effects lead to large increases in the band gaps and efficient light emission has been observed, even in Si for which coupling to light is extremely weak in the bulk crystal. In the case of a pure semiconductor, the broken bonds lead to reconstruction of the surface, and in the smallest clusters there is little resemblance to the bulk structures, as illustrated in Fig. 2.18.

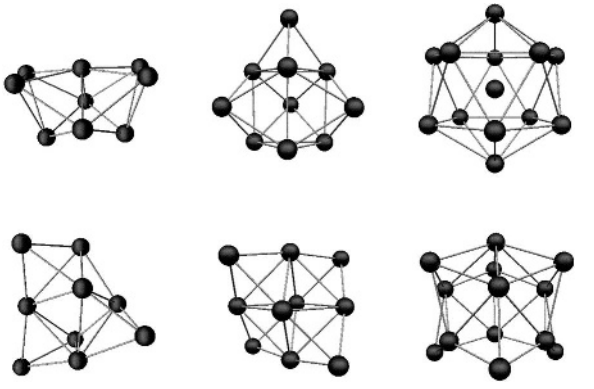


Figure 2.18. Atomic positions in competing Si_n clusters consisting of $n = 9$, 10, and 13 atoms. In each case, two very different structures are shown. The correct structure is not known from direct experiments and theory plays an important role in sorting out likely candidates. Provided by J. Grossman; from among the cases studied in [192].

For example, in Si_{13} there is competition between a symmetric structure with 12 outer atoms surrounding a central atom and the low-symmetry structure found by Car–Parrinello methods and simulated annealing [190]. The symmetric structure was argued [191] to be stabilized by correlations not accounted for in the local approximation in density functional theory; however, quantum Monte Carlo calculations [192, 193] found the low-symmetry structure to be the most stable, in agreement with Car–Parrinello simulations.

On the other hand, if the surface is terminated by atoms such as hydrogen or oxygen, which remove the dangling bonds, then the cluster is much more like a small, terminated piece of the bulk. Nanoclusters of Si have been the subject of much experimental investigation owing to their strong emission of light, in contrast to bulk Si which has very weak emission. The energy of the light emitted is increased by quantum confinement of the electron states in the cluster, and the emission strength is greatly increased by breaking of bulk selection rules due to the cluster size, shape, and detailed structure. This is an ideal case for combined theoretical and experimental work to interpret experiments and improve the desired properties. Calculations using time-dependent density functional theory (Sec. 7.6) are used as an illustration of the methods in Ch. 20, e.g. the variation of the gaps versus size, shown in Fig. 20.3.

Among the exciting discoveries of the last decades are the carbon fullerenes, C_{60}, C_{70}, ..., by Kroto et al. in 1985 [195], and nanotubes, by Iijima [196]. They are extraordinary not only because of their exceptional properties but also because of their elegant simplicity. C_{60} is the most symmetric molecule in the sense that its point group (icosahedral) with 120 symmetry operations is the largest point group of the known molecules. As shown in Fig. 2.19, a "buckyball" has the shape of a football (a soccer ball in the USA), with all 60 carbon atoms equivalent.[7] Interest in fullerenes increased dramatically when Krätschmer, et al. [197]

[7] The name for the structures derives from R. Buckminster Fuller, a visionary engineer who conceived the geodesic dome. Interestingly, he was professor at Southern Illinois University in Carbondale.

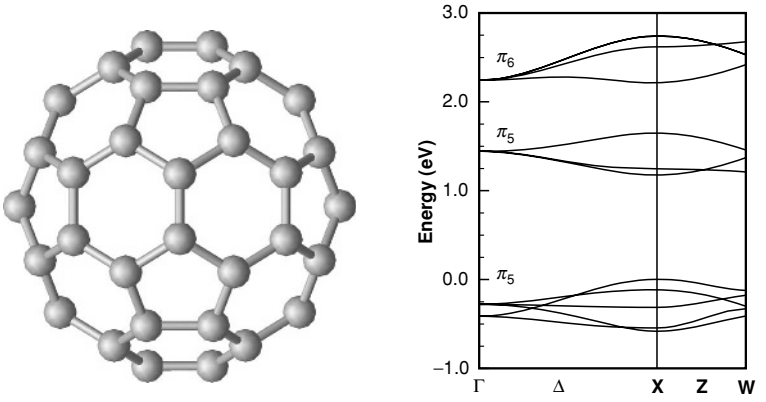


Figure 2.19. The structure of C_{60}, the most symmetric of all molecules, with the shape of a football (soccer ball) and the strength of a geodesic dome made famous by R. Buckminster Fuller. On the right are shown the calculated [194] bands of a fcc crystal of C_{60}. The highest occupied and lowest unoccupied bands are most interesting for electrical properties and are derived from molecular states designated in the text. Right figure from J. L. Martins; essentially the same as in [194].

discovered how to produce C_{60} in large enough quantities to make solids (fullerites). In rapid succession it was found that intercalation of alkali–metal atoms in solid C_{60} leads to metallic behavior [198], and that some alkali-doped compounds (fullerides) are superconductors with transition temperatures surpassed only by the cuprates. Thus the electronic bands, electron–phonon interactions, and electron–electron interactions are of great interest in these materials, as reviewed by Gunnarsson [199].

The bands of a fcc crystal of C_{60} shown in Fig. 2.19 were calculated using plane waves and norm-conserving pseudopotentials [194]. The essential results are that the bands are primarily derived from the radial π orbitals and are broad enough to lead to band-like conductivity, even though the states retain their molecular character in the sense that each set of bands is derived primarily from one set of degenerate molecular orbitals. The highest occupied and lowest unoccupied bands are derived respectively from h_u and t_{1u} molecular states that are five-fold and three-fold degenerate.

Figure 2.20 shows STM images [200] of C_{60} on Si, which is representative of productive collaboration of experiment and theory. The calculations were done using local orbital methods (Ch. 15) and a more detailed figure of atomic-scale bonding of the C_{60} molecule to the Si surface is shown in Fig. 15.5.

Nanotubes of carbon, discovered in 1991 by Iijima [196], are made from graphene-like sheets (or multiple sheets) rolled into a tube [203–205].[8] In perfect nanotube structures there are no pentagons and every carbon atom is at the vertex joining three hexagons. The various ways the sheet can be rolled lead to an enormous variety of semiconductors and metals, in some cases with helicity, such as the example shown in Fig. 14.8. These are ideal systems as the bands are beautifully described by theoretical rolling of the Brillouin zone

[8] Graphene denotes a single plane; the various structures of graphite result from different stackings of graphene planes.

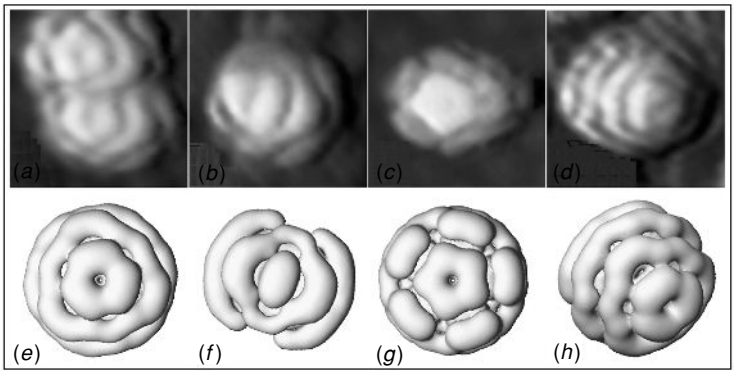


Figure 2.20. Scanning tunnelling microscope images of C_{60} "buckyballs" in Si (top) compared to calculated images from [200]. The calculations were done with the local orbital SIESTA code (Ch. 15) and using the Tersoff–Hamann [201] theoretical expressions for the STM image.

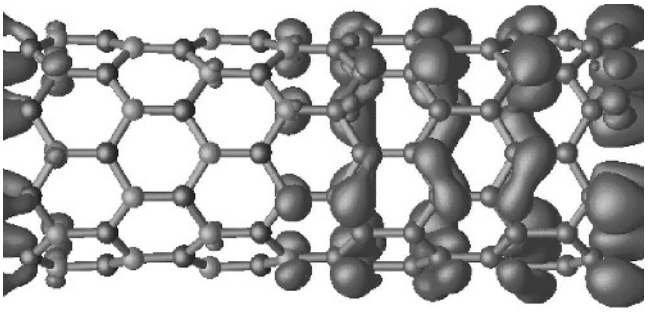


Figure 2.21. Example of a BN/C nanotube junction, illustrating the atom positions in an (8,0) tube along with the density of the highest occupied electronic state which is localized near the C portion of the tube. This can be considered as a "quantum dot" or as a supercell constructed to calculate the properties of the individual BN/C interfaces in the same spirit as illustrated for three-dimensional materials in Fig. 13.6. Provided by J. Bernholc; similar to figure in [202]. See Ch. 14 for examples of atomic and electron structures of nanotubes.

of graphene. However, the curvature adds a coupling on the σ and π bonds not present in flat graphene sheets [203, 204] and large changes in the bands can occur in tubes with very small radius, as described in Sec. 13.5 based on the work of [206] and in Sec. 14.7. Nanotubes are chosen as an elegant, instructive example of the tight-binding approach in Ch. 14.

Similar tubes of BN have been proposed theoretically [207] and latter made experimentally [208]. BN tubes always have a gap and are potential semiconductor devices. In addition they can have interesting piezolectric and pyroelectric effects [202]. Figure 2.21 shows the electron density for the highest occupied state in an example where the C nanotube is metallic, leading to metal–semiconductor junctions in the C–BN junction nanotube. The calculations [202] were done using real-space methods [209] described in Chs. 12 and 13.

2.10 Electronic excitations: bands and band gaps

Electronic excitations can be grouped into two types: excited states with the same number N of electrons as the ground state, and single particle excitations in which one electron is subtracted $N \to N-1$ or added $N \to N+1$. The former excitations determine the specific heat, linear response, optical properties, *etc*., whereas the latter are probed experimentally by tunnelling and by photoemission or inverse photoemission [85].

The most important quantity for adding and removing electrons is the *fundamental gap*, which is the minimum difference between the energy for adding and subtracting an electron. The lowest gap is *not* an approximate concept restricted to independent-particle approximation. It is defined in a general many-body system as the difference in energy between adding an electron and removing one: if the ground state has N electrons, the fundamental gap is

$$E_{\text{gap}}^{\min} = \min\{[E(N+1) - E(N)] - [E(N) - E(N-1)]\}. \tag{2.8}$$

Metals are systems in which the gap vanishes *and* the lowest energy electron states are delocalized. On the other hand if the fundamental gap is non-zero or if the states are localized (due to disorder) the system is an insulator.

Angle- and energy-resolved photoemission

The primary tool for direct observation of the spectrum of energies for removing an electron as a function of the crystal momentum $\mathbf{k}$ [85, 210] is angle-resolved photoemission, shown schematically in Fig. 2.22. Because the electrons are restricted to a surface region, photoemission is a surface probe and care must be taken to extract bulk information. The momentum of the excitation in the crystal parallel to the surface is determined by momentum

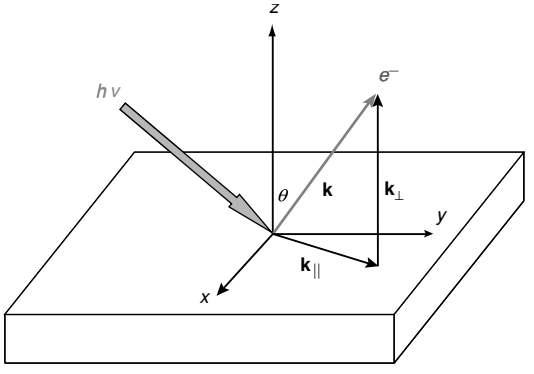


Figure 2.22. Schematic diagram of photoemission spectroscopy (PES) used to investigate the electron removal spectrum by an incident photon as indicated. The electron can escape through the surface if its energy is above the work function threshold. The momentum in the plane is fixed directly by the measured angle as indicated and the momentum perpendicular to the surface within the crystal must be calculated within a model, as discussed in the text and Fig. 2.23. The time-reversed experiment of "inverse photoemission" is a measure of the addition spectra.

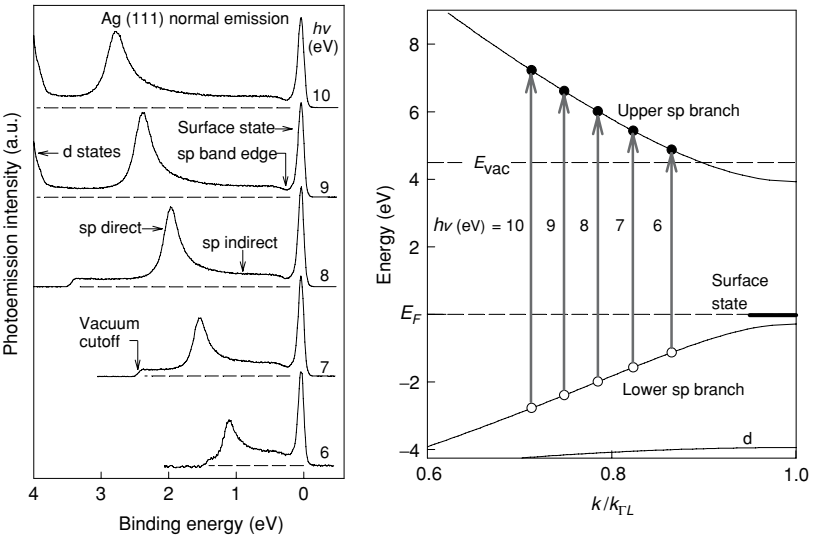


Figure 2.23. Illustration of the determination of electron momentum perpendicular to the surface by the dependence upon incident photon energy. Spectra for emission from the (111) surface of Ag at five photon energies are show in the left-hand figure [211]. (The state with no dispersion is a surface state.) Interpretation of the dispersion is indicated in the right-hand figure. The momentum perpendicular to the surface is selected by the dispersion relation of the final state, which is nearly-free-electron-like. This leads to the dispersion of the s-like band perpendicular to the surface (the Γ–L direction) shown at the bottom right. Provided by T. C. Chiang.

conservation as illustrated in Fig. 2.22. The method for determining the dispersion perpendicular to the surface is illustrated in Fig. 2.23; assuming a known dispersion for the excited electron inside the crystal (for example, if the higher bands are free-electron-like), the occupied bands $\varepsilon(\mathbf{k}_\perp)$ can be mapped out from the dependence upon the photon energy. In an independent-particle picture there are sharp peaks in the energies of the emitted electrons that are the eigenvalues or bands for the electrons. Weak interactions lead to small broadenings and shifts of the peaks, whereas strong interactions can lead to qualitative changes. The lower right part of Fig. 2.23 shows the *s* band of Ag, for which the sharp peaks observed in the experiment [211] are well described by band theory.

Angle-resolved photoemission was demonstrated as a quantitative experimental method in the late 1970s, and has become a very powerful method for studies of electrons. The main points are already shown in the earliest work illustrated by Fig. 2.24 for Cu and Fig. 12.2 for GaAs. In each case the theory came before the experiment and the agreement is a clear indication of the usefulness of independent-particle methods and band theory. The bands of Cu shown in Fig. 2.24 consist of five narrow d bands and one partially filled s band, in remarkable agreement with experiment. The theoretical calculations were done with the APW method and an approximate potential derived from an atomic calculation [213]. In fact, the agreement is *not* as good for more recent self-consistent density functional calculations. A typical result is that the d bands are too close to the Fermi energy, which

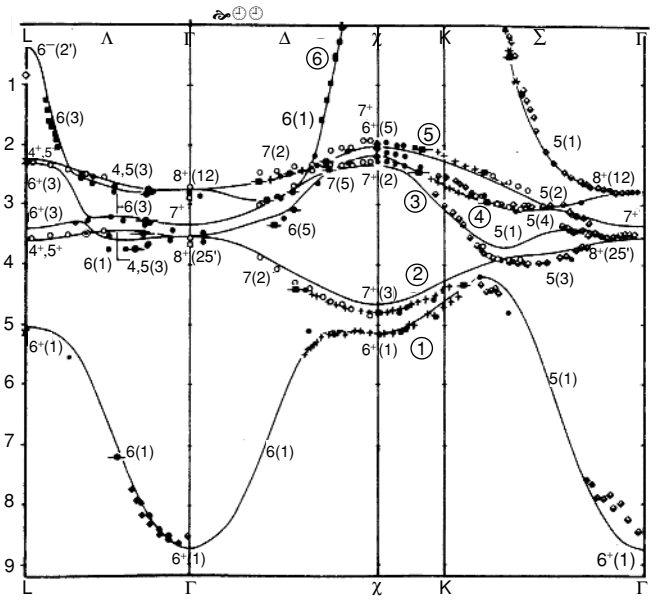


Figure 2.24. Experimental energy bands of Cu measured by angle-resolved photoemission [212] (points) compared with the classic APW calculations of Burdick [213] using the Chodorow potential, which is a sum of potentials derived from atomic calculations. (More recent band calculations are very similar.) Independent particle calculations describe Cu very well because it has an essentially filled d shell plus a wide s band (like Na shown in Fig. 5.6). Provided by S. Hufner.

is a symptom of the inaccuracies in the approximations for exchange and correlation (see Ch. 8).

The measured bands [214] for GaAs in Fig. 12.2 are in excellent agreement with the prior theoretical work of Pandey and Phillips [215] calculated with an empirical pseudopotential. As discussed in Ch. 12, the near-perfect agreement with the photoemission data is due to the fact that the pseudopotential was adjusted to fit optical data; nevertheless, the agreement shows the value of the interpretation based upon independent-particle theory.

With recent dramatic improvements in resolution [210] using synchrotron radiation, photoemission has become a powerful tool to measure the detailed dispersion and many-body effects for the one-electron removal spectrum in crystals. As an example, the spectra for MgB_2 shown in Fig. 2.30 indicate that the bands are well described by independent-particle theory. In other materials, however, there is not such good agreement. Cases that are more strongly correlated provide the most important challenges in electronic structure today [216].

Electron addition spectra: inverse photoemission

Inverse photoemission can map out the electron addition spectrum, i.e. the empty states in independent-particle theories. The process is the inverse of that shown in Fig. 2.22.

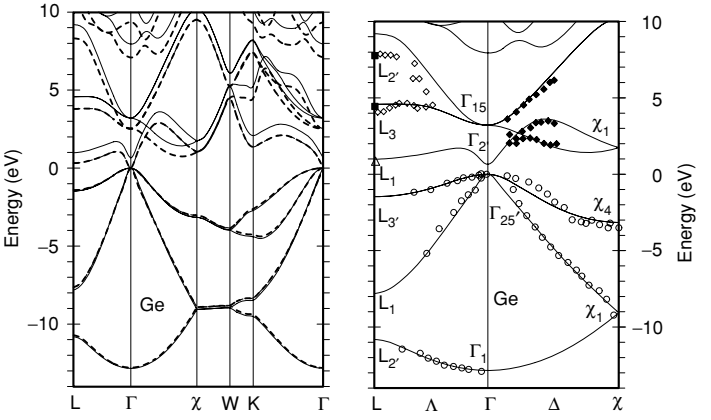


Figure 2.25. Quasiparticle bands in Ge calculated using the "GW" approximation with pseudopotentials and a gaussian basis [219] compared to experimental points from photoemission [221] and inverse photoemission [222] (right figure) and to LDA bands (dashed line in left figure). The LDA bands are essentially the same as those in Fig. 17.8 calculated by the LMTO and plane wave methods. The LDA bands illustrate the well-known "band-gap problem" that leads to a zero gap in Ge. This is improved by many-body quasiparticle methods; similar improvements are found with the exact exchange (EXX) density functional [223] (see Fig. 2.26 and Ch. 8).

Figure 2.25 illustrates the comparison of theory and experiment for Ge for both addition and removal spectra. This example is chosen because it illustrates both the success of band theory and a spectacular failure of the widely used LDA approximation. The left-hand panel compares two types of theoretical results: independent-particle bands calculated using the local density approximation (LDA, see Ch. 8) and the many-body "GW" quasiparticle theory [217–219].[9] The two methods agree closely for the filled bands but the LDA predicts a zero band gap, so that Ge would be a metal in this approximation. This is a striking example of the general result that gaps are predicted to be too small. The right-hand panel shows experimental results for both photoemission and inverse photoemission, which are in good agreement with the "GW" quasiparticle energies.

Electron addition and removal spectra: theory

Despite the impressive agreement with experiment of many density functional theory calculations for ground state properties, the same calculations for insulators often lead to mediocre (or disastrous) predictions for excitations. The fundamental gap is the key issue, and widely used approximate functionals in density functional theory lead to gaps (Eq. (2.8)) that are significantly below experimental values for essentially all materials. This is illustrated by results of calculations using the local density approximation (LDA) for a range of

[9] "GW" methods are essentially random phase approximation (RPA) calculations (see Sec. 5.4) for the quasiparticle self-energy, originally developed for jellium [220] and now being carried out on complex materials (for a review see [82]). Such methods also include exchange plus higher order diagrams.

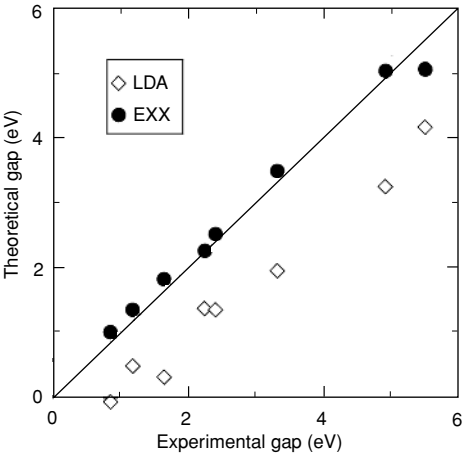


Figure 2.26. The lowest gaps of semiconductors predicted by the two different density functionals: local density approximation (LDA) and the non-local "exact exchange" (EXX; see Ch. 8). The LDA results illustrate the well-known underestimate of gaps in most DFT calculations, even a zero gap in Ge. Exact exchange theory is more difficult to treat, but it gives much improved gaps in remarkable agreement with experiment. From [223]. Provided by M. Staedele.

semiconductors, shown by the open symbols in Fig. 2.26. All gaps are too small and (as illustrated in Fig. 2.25) Ge is even predicted to be a metal. The underestimation of the gap may be caused by effects similar to the underestimate of transition pressures in Fig. 2.5; however, the effect upon the gaps is much larger and brings out much more fundamental challenges to the theory.

Improvement of the theory of excitations in insulators is a key part of current electronic structure research. The fundamental issues in the Kohn–Sham approach are brought out in Ch. 7 and actual improved functionals are presented in Secs. 8.6 and 8.7. *The low gaps are not intrinsic to the Kohn–Sham approach* and are greatly improved by better treatment of the non-local exchange, such as the orbital-dependent "exact exchange" (EXX) version of Kohn–Sham theory and "hybrid functionals," which incorporate features missing in the LDA and any GGA-type functionals. For example, the EXX functional leads to greatly improved gaps, as shown in Fig. 2.26, without destroying the accuracy of the ground state energies [223] if a local functional for correlation is included. Hybrid functionals are widely used for molecules and also lead to similar improved results for total energies and gaps [224], as illustrated in Fig. 15.3. The goal is to provide an approach to calculation of excitation spectra and energy gaps that is accurate, robust, and less computationally intensive than the many-body "GW" quasiparticle calculations.

2.11 Electronic excitations: heat capacity, conductivity and optical spectra

Excitations that conserve the number of electrons can be viewed as *electron–hole excitations* in which the added electron interacts with a "hole" left by removing an electron. The lowest

energy is less than or equal to $E_{\rm gap}^{\rm min}$,

$$E_{\rm ex}^{\rm min} < E_{\rm gap}^{\rm min}, \tag{2.9}$$

since the electron–hole interaction is attractive. Therefore, measurements, such as specific heat and optical spectra, that conserve the number of electrons can be used to establish bounds on $E_{\rm gap}^{\rm min}$ and establish that a material is a metal or insulator.

The most universal measure of excitations that conserve electron number is the heat capacity since it encompasses all possible excitations equally with no bias. The heat capacity at low temperature T is the measurable quantity that definitively separates systems with an energy gap for electronic excitations from those with no gap; in a perfect crystal this is also the distinction between metals and insulators. A normal metal is characterized by specific heat $\propto T$, which is the fundamental evidence for the Landau Fermi liquid theory [96, 225, 226]. This leads to the of idea of "quasiparticles," low-energy excitations that act like weakly interacting electrons, even though there are in fact strong interactions. Such pictures often provide an excellent description of specific heat, electrical conductivity, and Pauli paramagnetism. The density of states in the independent-particle picture is discussed in Sec. 4.7 and examples of single-particle DOS are given in Figs. 2.31 and 16.13. If the specific heat is exponentially small at low T, this is definitive evidence for a gap and no low-energy excitations.

The Fermi surface is a surface in reciprocal space that is defined in a many-body system as the locus of points where the quasiparticle lifetime is infinitely long and the quasipartcle energy equals the Fermi energy μ. In an independent-particle approximation the states have infinite lifetime at all energies, and the Fermi surface separating filled and empty states is the surface defined by $\varepsilon(\mathbf{k}) = \mu$. The surface has been mapped out in great detail in many crystals from the very low-energy excitations at low temperature, e.g. using the small periodic changes in magnetization of the electrons as a function of applied magnetic field that were first observed in 1930 by de Haas and van Alfen, and which became a key experimental tool in the 1950s. The theoretical analysis is described in many texts [84, 86, 88]. The calculated Fermi surface for MgB_2 shown in Fig. 2.31 is an example that illustrates very different portions of the Fermi surface that play distinct roles in the thermal and electrical properties and superconductivity.

Electronic conductivity and optical properties

Dielectric functions and conductivity are the most important response functions in condensed matter physics because they determine the optical properties of materials, electrical conductivity, and a host of technological applications. In addition, optical spectra are perhaps the most widespread tool for studying the electronic excitations themselves. The phenomenological formulation of Maxwell's equations in the presence of polarizable or conducting media can be cast in terms of the complex frequency-dependent dielectric function $\epsilon(\omega)$ or conductivity $\sigma(\omega)$. The relations are summarized in App. E and the formulation in terms of electronic excitations is the subject of Ch. 20. Note that the number of electrons

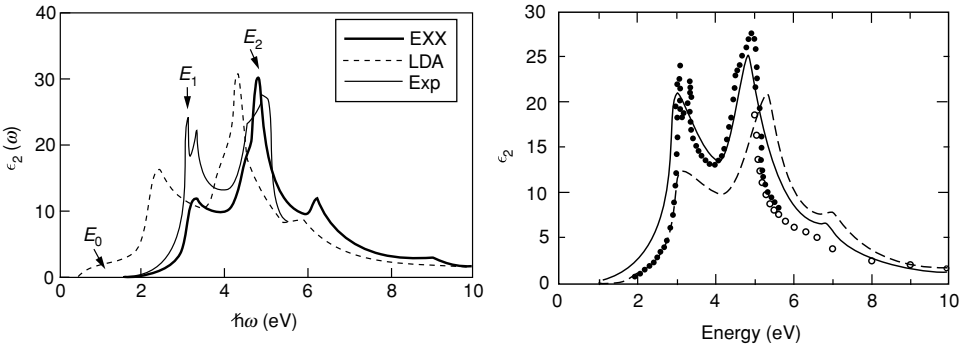


Figure 2.27. Calculated and experimental optical spectra of GaAs. Left: Comparison of experiment (light solid) with two density functional theory calculations [223], LDA (dashed) and EXX (heavy solid). Right: Spectra calculated with GW quasiparticles with (solid lines) and without (dashed) electron–hole interaction [227].

does not change, i.e. optical absorption can be viewed as the simultaneous addition of an electron and a hole, which can interact with one another.

Figure 2.27 shows two examples of calculated optical spectra of the semiconductor GaAs. The figure on the left illustrates the results from two different density functionals: the famous underestimate of the band gap using LDA, and the improvement using "exact exchange" (EXX; see Fig. 2.26 and Ch. 8). However, there is still a major discrepancy in the heights of the peaks. On the right are shown results from two many-body calculations. Using the "GW" quasiparticles without including the electron–hole interaction suffers from the same problem; however, inclusion of electron–hole interaction by solving the two-particle Bethe–Salpeter equation leads to much better overall agreement with experiment.

The effect of electron–hole interactions is much greater in wide-band-gap materials like CaF_2, as shown in Fig. 2.28. In this case there are qualitative changes with most of the absorption shifted to the bound exciton state. Similar results have been found for LiF, and other wide-band-gap insulators.

There is another approach to calculation of excitation spectra for the case where the number of electrons does not change: time-dependent density functional theory (TDDFT) [229–232] which in principle provides the *exact* solution for $n(t)$ that follows from the time-dependent Schrödinger equation (see Ch. 20). This approach has been used with approximate exchange–correlation functionals with considerable success for optical spectra confined systems such as molecules and clusters [231, 233–235] and magnetic excitations in solids [236]. An example of the energy gap in hydrogen-terminated Si clusters as a function of cluster size is shown in Fig. 20.3 calculated assuming the usual adiabatic LDA functional. However, the adiabatic functional misses important physics and the search for improved time-dependent functionals is a topic of much current research [237, 238].

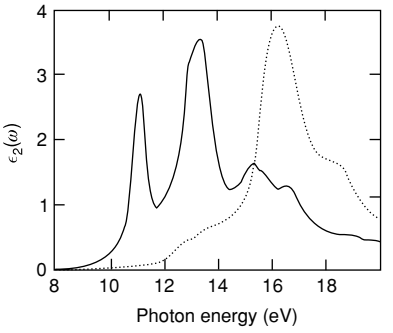


Figure 2.28. Optical spectrum of CaF_2 calculated neglecting (dashed line) and including (solid line) electron–hole interactions. This illustrates the fact that spectra can be completely modified by strong electron–hole interactions in wide-band-gap insulators, with much of the oscillator strength in the bound exciton peak around 11 eV that is completely missing in an independent electron approach. (The width of the bound exciton peak and the weak tails in the spectra at low energies are artifacts of the calculation.) From [228].

2.12 Example of MgB_2: bands, phonons, and superconductivity

Magnesium diboride MgB_2 serves as an example of many aspects of modern electronic structure theory and experiment and the interplay between the two. Although this material has been known for many years, a flurry of activity was stimulated by the discovery [145] of superconductivity at the relatively high temperature of $T_c = 39$ K. Much work has been done to elucidate the electronic states, phonons and the mechanism for superconductivity. This is an excellent example of combined theory and experiment to understand a new material with unusual electronic states. Intitially, the theoretical calculations were real predictions that have been tested by recent experiments such as angle-resolved photoemission [239] that have become feasible as sample quality has improved.

It is useful to consider MgB_2 in the light of the similarities and differences from its cousin hexagonal graphite. The structures of the two materials can be understood in terms of the honeycomb graphene plane shown in Fig. 4.5. The simple hexagonal form of graphite consists of these planes stacked with hexagons over one another in the three-dimensional simple hexagonal structure, Fig. 4.2. This is also the structure of MgB_2 which is illustrated in 4.6. The boron atoms form graphene-like planes in the simple hexagonal structure and the Mg atoms occupy sites in the centers of the hexagons between the layers. Since each Mg atom provides two valence electrons, the total electron valence count per cell is the same for graphite and MgB_2. Thus we can expect the band structures to be closely related and the bands near the Fermi level to be similar.

The bands have been calculated by many groups with the same conclusions; the results [240] presented in Fig. 2.29 are chosen because they show the comparison with hexagonal graphite, given on the right. The symmetry point notations are those for the simple hexagonal Brillouin zone in Fig. 4.10. The shading of points indicates the degree of σ bonding character of the states, i.e. the strong in-plane bonding states. The graphite bands are only slightly

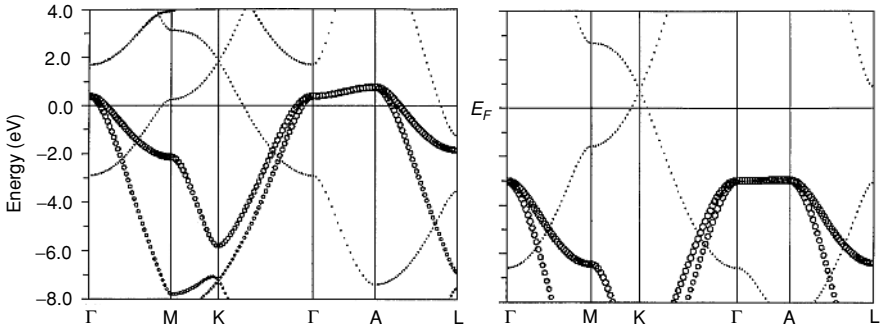


Figure 2.29. Electron bands of MgB_2 (left) and primitive hexagonal graphite calculated [240] using LAPW method (Ch. 17). The planar σ bonding states, highlighted with larger symbols, are higher in energy in MgB_2 so that they are partially unoccupied. The symmetry labels are given in Fig. 4.10 and the Fermi surface calculated using a similar method is shown in Fig. 2.31. From [240].

modified from those of a single plane of graphene, which has the Fermi energy exactly at the K points where the π bands touch to give a zero gap and a Fermi surface that is a set of points in two-dimensions (see Sec. 14.7). In graphite, the σ bonding states are shifted well below the Fermi energy, whereas in MgB_2 the bonding is weaker so that the σ bands cross the Fermi energy. In addition, there is greater dispersion perpendicular to the layers (e.g. $\Gamma \rightarrow$ A), especially for the π bands, which are rather three-dimensional in nature.

Subsequent to the calculations, measurements of the dispersion has been made by angle-resolved photoemission [239]. Since certain bands have little dispersion perpendicular to the layers, the parallel dispersion can be measured directly as shown in Fig. 2.22. Measured spectra and the plots of the peaks obtained by analysis of second derivatives of the spectra are shown in Fig. 2.30. The agreement with the bands in Fig. 2.29 is interpreted [239] as evidence that the electronic states are not strongly correlated so that an independent-particle approach captures the salient features.

The calculated Fermi surface of MgB_2 has two very different parts as shown in Fig. 2.31 [241]. The sheets of the surface near Γ are almost two-dimensional; this is also clear in Fig. 2.29, which shows the nearly flat band close to the Fermi energy in the $\Gamma \rightarrow$ A direction that is formed from the in-plane σ bonding states. Since the σ bands are degenerate at Γ, the small splitting near Γ shown in Fig. 2.29 leads to two closely spaced sheets of the Fermi surface. Although the area of the surface is small the nearly two-dimensional character leads to this part of the Fermi surface contributing $\approx$30% of the density of states at the Fermi energy [241].

The other ingredients in superconductivity are the phonons and electron–phonon interactions. Many calculations have been done for MgB_2, with the basic conclusion that the strongest coupling is provided by the E_g optic phonons at $\mathbf{k} \approx 0$. An example of a calculation of the entire phonon dispersion curves $\omega(\mathbf{k})$, density of states $F(\omega)$, and electron–phonon interaction $\alpha^2 F(\omega)$ is shown in Fig. 2.32. The black dots indicate the strength of the interaction of various phonons with exceptionally strong coupling of the E_g phonons at frequency near 600 cm^{-1}. It is also evident that the optic phonons near $\mathbf{k} = 0$ are "softened," i.e. there

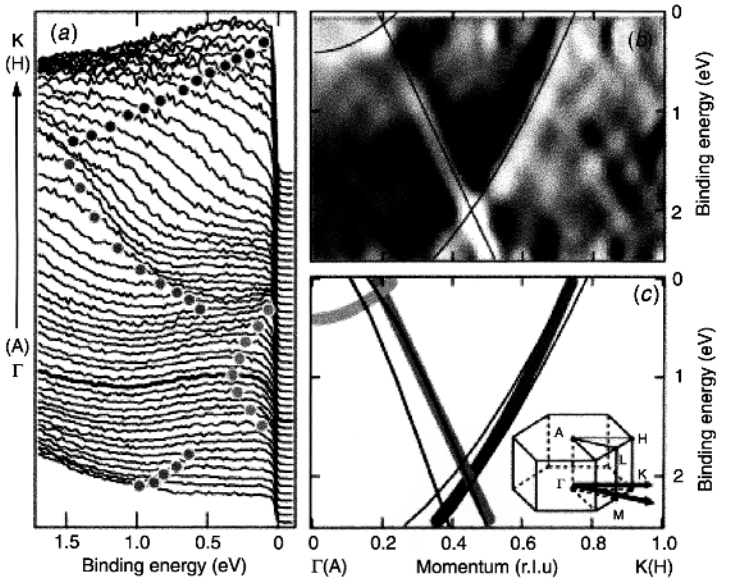


Figure 2.30. Angle-resolved photoemission!spectra of MgB_2 (left) and analysis of second derivatives of the spectra for dispersion in the plane in the $\Gamma \rightarrow$ K direction [239]. The "experimental bands" have remarkable agreement with the calculated bands in Fig. 2.29, which is interpreted as showing that the electronic states are well described by an independent-particle approach. The state near Γ does not correspond to bulk bands and is interpreted as a surface state. From [239].

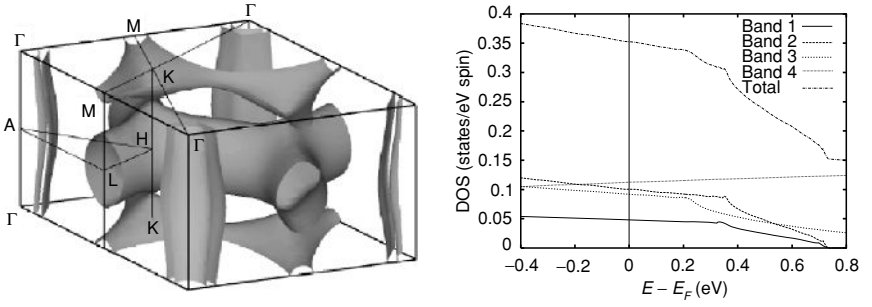


Figure 2.31. Left: Fermi surface of MgB_2 from [241]. The sheets near the Γ point are nearly two-dimensional and are composed of in-plane σ bonding states, as shown in Fig. 2.29. The other sheets are from the π bands that have larger dispersion. Right: Electronic density of states (DOS) for the different bands in MgB_2 near the Fermi energy. The partial DOS correspond to the two wide π bands and the two nearly two-dimensional σ bonding bands shown in Fig. 2.29 and the Fermi surface at the left. Although the surface around Γ is small, it is crucial in the superconducting properties: it accounts for a large fraction of the density of states at the Fermi energy and the bonding states couple strongly to phonons as shown in Fig. 2.32. Provided by J. Kortus and I. I. Mazin.

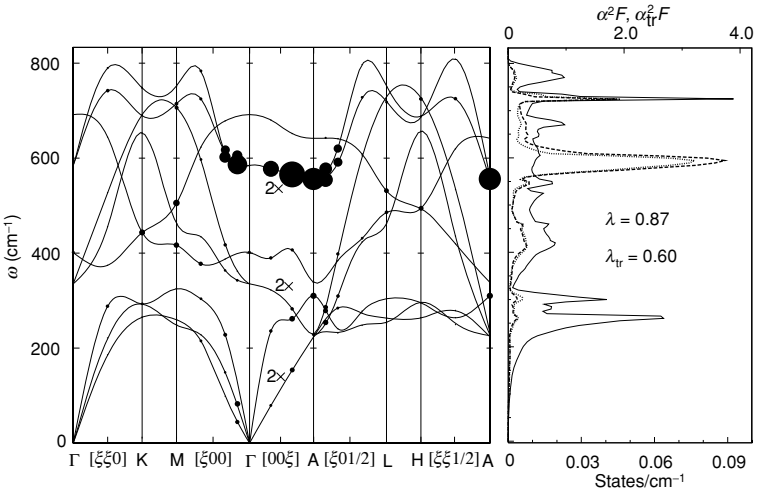


Figure 2.32. Dispersion curves, phonon density of states, and the electron-phon-coupling spectrum [242, 243] α^2F for MgB_2 calculated [155] using the LMTO method (Ch. 17). The size of the black dots indicates the strength of the electron–phonon interaction, showing strongest coupling for the E_g phonons at frequency near 600 cm^{-1}. The dip in the dispersion curves for optic phonons near $\mathbf{k} = 0$ also indicates strong electron–phonon coupling (see also Fig. 19.4.). Calculations of the transition temperature [155, 244] based upon the calculated properties are in general agreement with experiment. From [155].

is a dip in the dispersion curves that indicate strong electron–phonon coupling. (Such an effect is also present in other strong coupling superconductors, for example for Nb shown in Fig. 19.4.) In addition, there is a large cubic anharmonicity for the E_g displacement pattern, illustrated in Fig. 2.8, which is very sensitive to the details of the Fermi surface indicating large electron–phonon coupling.

The full theory of superconductivity is outside the scope of this volume. A very short summary is given in Sec. 19.8 and there are excellent reviews such as [242, 243]. The important point for our purposes is that the basic ingredients in the theory of phonon-mediated superconductivity are the bread and butter of electronic structure: the electronic bands, the Fermi surface, the single-particle densities of states, phonon dispersion, and electron–phonon interactions. For the case of MgB_2, calculations using different methods [155, 240, 241, 244] support the phonon-mediated mechanism. The surprisingly large transition temperature appears to be due to the states at the Fermi surface that are nearly two-dimensional and have σ bonding character (the cylinders of the Fermi surface in Fig. 2.31) which couple strongly to the phonons. Furthermore, there are other interesting features that emerge from the theory, including two-gap superconductivity that results from solution of the Eliashberg equations [155, 244].

2.13 The continuing challenge: electron correlation

The competition between correlation due to interactions and delocalization due to kinetic energy leads to the most challenging problems in the theory of electrons in condensed matter.

Correlations are responsible for metal–insulator transitions, the Kondo effect, heavy fermion systems, high-temperature superconductors, and many other phenomena (see, e.g., [216]). Low dimensionality leads to larger effects of correlation and new phenomena such as the quantum Hall effect. In one dimension, the Fermi liquid is replaced by a Luttinger–Tomanaga liquid in which the excitations are "holons" and "spinons."

In some cases, independent-particle methods are sufficient to include effects of correlation through effective mean-field interactions; indeed, this is the key to the success of density functional theory in quantitative description of systems with weak or moderate correlation. The term "strong correlation" is generally used to denote just those systems where mean-field approximations (at least those used at present) break down and fail to describe the important physics. Such problems are appropriate subjects for entire texts, and this section is merely a pointer to remind the reader of important issues *not solved* by present-day approximate forms of density functional theory and other independent-particle methods.

Materials systems that exhibit strong correlation effects are ones on the boundary between localized and delocalized. They often involve the latter 3d and 4d transition metals; the anomalous 4f rare earths Ce, Sm, Eu, and Tb; and the 5f actinides. These states are intermediate between the highly localized 4f rare earths and the more delocalized band-like metals, which include the early 3d and 4d elements and the non-transition metals. The 3d transition metal oxides are in the intermediate range where metal–insulator transitions occur, along with high-temperature superconductivity, which is thought to involve electron correlations, even though the understanding is still elusive [216]. The anomalous rare earth elements are ones where the addition or removal energy is close to the Fermi energy, indicating that the 4f occupation is *unstable*, leading to "mixed-valence" and "heavy-fermion" behavior. For example, $CeCu_2Si_2$ has a specific heat coefficient $\approx$1,000 times larger than expected from a band calculation [245] of "electrons."

As the theory of electronic structure becomes more powerful and more predictive, it is even more relevant to keep in mind the "big picture" of the possible consequences of many-body electron–electron interactions. Not only is a proper accounting of the effects of interactions essential for quantitative description of real materials, but also imaginative exploration of correlation can lead to exciting new phenomena qualitatively different from predictions of usual mean-field theories. As emphasized in the introduction, Sec. 1.3, the concepts are captured in the notion [78] of "More is different" with recent developments summarized in *More is Different: Fifty Years of Condensed Matter Physics* [79].

SELECT FURTHER READING

Basic references on condensed matter:

Ashcroft, N. and Mermin, N. *Solid State Physics* (W. B. Saunders Company, New York, 1976).

Chaikin, P. N. and Lubensky, T. C. *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, 1995).

Ibach, H. and Luth, H. *Solid State Physics: An Inroduction to Theory and Experiment* (Springer-Verlag, Berlin, 1991).

Kittel, C. *Introduction to Solid State Physics* (John Wiley and Sons, New York, 2000).

Marder, M. *Condensed Matter Physics* (John Wiley and Sons, New York, 2000).

3

Theoretical background

Summary

Our understanding of the electronic structure of matter is based upon theoretical methods of quantum mechanics and statistical mechanics. This chapter reviews fundamental definitions and expressions, including the most basic forms valid for many-body systems of interacting electrons and useful, simplified formulas valid for non-interacting particles. This material is the foundation for succeeding chapters, which deal with further developments of the theory and the methods to carry out calculations.

3.1 Basic equations for interacting electrons and nuclei

The subject of this book is recent progress toward describing properties of matter from theoretical methods firmly rooted in the fundamental equations. Thus our starting point is the hamiltonian for the system of electrons and nuclei,

$$\hat{H} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_{i,I}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|} + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_I \frac{\hbar^2}{2M_I}\nabla_I^2 + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{R}_I - \mathbf{R}_J|}, \tag{3.1}$$

where electrons are denoted by lower case subscripts and nuclei, with charge Z_I and mass M_I, denoted by upper case subscripts. It is essential to include the effects of difficult many-body terms, namely electron–electron Coulomb interactions and the complex structures of the nuclei that emerge from the combined effects of all the interactions. The issue central to the theory of electronic structure is the development of methods to treat electronic correlations with sufficient accuracy that one can predict the diverse array of phenomena exhibited by matter, starting from (3.1).[1] It is most informative and productive to start with the fundamental many-body theory. Many expressions, such as the force theorem, are

[1] Here relativistic effects, magnetic fields, and quantum electrodynamics are not included. These will be incorporated later to varying degrees. For example, the Dirac equation is included in Ch. 10 on atoms, and inclusion in solids is discussed in Ch. 11 on pseudopotentials and Ch. 16 on augmented methods. Magnetic fields are included explicitly only as Zeeman terms.

more easily derived in the full theory with no approximations. It is then straightforward to specialize to independent-particle approaches and the actual formulas needed for most of the following chapters.

There is only one type of term in the general hamiltonian that can be regarded as "small," the inverse mass of the nuclei $1/M_I$. A perturbation series can be defined in terms of this parameter which is expected to have general validity for the full interacting system of electrons and nuclei. If we first set the mass of the nuclei to infinity, then the kinetic energy of the nuclei can be ignored. This is the Born–Oppenheimer or adiabatic approximation [89] defined in App. C, which is an excellent approximation for many purposes, e.g. the calculation of nuclear vibration modes in most solids [90, 152]. In other cases, it forms the starting point for perturbation theory in electron-phonon interactions, which is the basis for understanding electrical transport in metals, polaron formation in insulators, certain metal–insulator transitions, and the BCS theory of superconductivity. Thus we shall focus on the hamiltonian for the electrons, in which the positions of the nuclei are parameters.

Ignoring the nuclear kinetic energy, the fundamental hamiltonian for the theory of electronic structure can be written as

$$\hat{H} = \hat{T} + \hat{V}_{\text{ext}} + \hat{V}_{\text{int}} + E_{II}. \tag{3.2}$$

If we adopt Hartree atomic units $\hbar = m_e = e = 4\pi/\epsilon_0 = 1$, then the terms may be written in the simplest form. The kinetic energy operator for the electrons $\hat{T}$ is

$$\hat{T} = \sum_i -\frac{1}{2}\nabla_i^2, \tag{3.3}$$

$\hat{V}_{\text{ext}}$ is the potential acting on the electrons due to the nuclei,

$$\hat{V}_{\text{ext}} = \sum_{i,I} V_I(|\mathbf{r}_i - \mathbf{R}_I|), \tag{3.4}$$

$\hat{V}_{\text{int}}$ is the electron–electron interaction,

$$\hat{V}_{\text{int}} = \frac{1}{2}\sum_{i\neq j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{3.5}$$

and the final term E_{II} is the classical interaction of nuclei with one another and any other terms that contribute to the total energy of the system but are not germane to the problem of describing the electrons. Here the effect of the nuclei upon the electrons is included in a fixed potential "external" to the electrons. This general form is still valid if the bare nuclear Coulomb interaction is replaced by a pseudopotential that takes into account effects of core electrons (except that the potentials are "non-local;" see Ch. 11). Also, other "external potentials," such as electric fields and Zeeman terms, can readily be included. Thus, for electrons, the hamiltonian, (3.2), is central to the theory of electronic structure.

Schrödinger equation for the many-body electron system

The fundamental equation governing a non-relativistic quantum system is the time-dependent Schrödinger equation,

$$i\hbar \frac{\mathrm{d}\Psi(\{\mathbf{r}_i\};t)}{\mathrm{d}t} = \hat{H}\Psi(\{\mathbf{r}_i\};t), \tag{3.6}$$

where the many-body wavefunction for the electrons is $\Psi(\{\mathbf{r}_i\};t) \equiv \Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N;t)$, the spin is assumed to be included in the coordinate $\mathbf{r}_i$, and, of course, the wavefunction must be antisymmetric in the coordinates of the electrons $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$. The eigenstates of (3.6) can be written as $\Psi(\{\mathbf{r}_i\};t) = \Psi(\{\mathbf{r}_i\})\mathrm{e}^{-\mathrm{i}(E/\hbar)t}$. This is the basis for understanding dynamical properties in Ch. 20 and App. D.

For an eigenstate, the time-independent expression for any observable is an expectation value of an operator $\hat{O}$, which involves an integral over all coordinates,

$$\langle \hat{O} \rangle = \frac{\langle \Psi | \hat{O} | \Psi \rangle}{\langle \Psi | \Psi \rangle}. \tag{3.7}$$

The density of particles $n(\mathbf{r})$, which plays a central role in electronic structure theory, is given by the expectation value of the density operator $\hat{n}(\mathbf{r}) = \sum_{i=1,N} \delta(\mathbf{r} - \mathbf{r}_i)$,

$$n(\mathbf{r}) = \frac{\langle \Psi | \hat{n}(\mathbf{r}) | \Psi \rangle}{\langle \Psi | \Psi \rangle} = N \frac{\int \mathrm{d}^3 r_2 \cdots \mathrm{d}^3 r_N \sum_{\sigma_1} |\Psi(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_N)|^2}{\int \mathrm{d}^3 r_1 \mathrm{d}^3 r_2 \cdots \mathrm{d}^3 r_N |\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_N)|^2}, \tag{3.8}$$

which has this form because of the symmetry of the wavefunction in all the electron coordinates. (The density for each spin results if the sum over σ_1 is omitted.) The total energy is the expectation value of the hamiltonian,

$$E = \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \equiv \langle \hat{H} \rangle = \langle \hat{T} \rangle + \langle \hat{V}_{\mathrm{int}} \rangle + \int \mathrm{d}^3 r \, V_{\mathrm{ext}}(\mathbf{r}) n(\mathbf{r}) + E_{II}, \tag{3.9}$$

where the expectation value of the external potential has been explicitly written as a simple integral over the density function. The final term E_{II} is the electrostatic nucleus–nucleus (or ion–ion) interaction, which is essential in the total energy calculation, but is only a classical additive term in the theory of electronic structure.

The eigenstates of the many-body hamiltonian are stationary points (saddle points or the minimum) of the energy expression (3.9). These may be found by varying the ratio in (3.9) or by varying the numerator subject to the constraint of orthonormality ($\langle \Psi | \Psi \rangle = 1$), which can be done using the method of Lagrange multipliers,

$$\delta[\langle \Psi | \hat{H} | \Psi \rangle - E(\langle \Psi | \Psi \rangle - 1)] = 0. \tag{3.10}$$

This is equivalent to the well-known Rayleigh–Ritz principle [245, 246] that the functional

$$\Omega_{\mathrm{RR}} = \langle \Psi | \hat{H} - E | \Psi \rangle \tag{3.11}$$

is stationary at any eigensolution $|\Psi_m\rangle$.[2] Variation of the bra $\langle\Psi|$ leads to

$$\langle\delta\Psi|\hat{H} - E|\Psi\rangle = 0. \tag{3.12}$$

Since this must hold for all possible $\langle\delta\Psi|$, this can be satisfied only if the ket $|\Psi\rangle$ satisfies the time-independent Schrödinger equation

$$\hat{H}|\Psi\rangle = E|\Psi\rangle. \tag{3.13}$$

In Exercise 3.1 it is shown that the same equations result from explicit variation of Ψ in (3.9) without Lagrange multipliers.

The ground state wavefunction Ψ_0 is the state with lowest energy, which can be determined, in principle, by minimizing the total energy with respect to all the parameters in $\Psi(\{\mathbf{r}_i\})$, with the constraint that Ψ must obey the particle symmetry and any conservation laws. Excited states are saddle points of the energy with respect to variations in Ψ.

Ground and excited electronic states

The distinction between ground and excited states pointed out in Ch. 2 is equally obvious when approached from the point of view of solving the many-body equations for the electrons. Except in special cases, the ground state must be treated by non-perturbative methods because the different terms in the energy equation are large and tend to cancel. The properties of the ground state include the total energy, electron density, and correlation functions. From the last one can derive properties that at first sight would not be considered to be ground state properties, such as whether the material is a metal or an insulator. In any case, one needs to establish which state is the ground state, often comparing states that are very different in character but similar in energy.

On the other hand, excitations in condensed matter are usually small perturbations on the entire system. These perturbations can be classified into variations of the ground electronic state (e.g. small displacements of the ions in phonon modes) or true electronic excitations, e.g. optical electronic excitations. In both cases, perturbation theory is the appropriate tool. Using perturbation techniques, one can calculate excitation spectra and the real and imaginary parts of response functions. Nevertheless, even in this case one needs to know the ground state, since the excitations are perturbations on the ground state.

These approaches apply both in independent-particle and in many-body problems. The ground state is special in both cases and it is interesting that both density functional theory and quantum Monte Carlo are primarily ground state methods. The role of perturbation theory is rather different in independent-particle and many-body problems – in the latter, it plays a key role in the basic formulation of the problem in diagrammatic perturbation series and in suggesting the key ideas for summation of appropriate diagrams.

[2] This is an example of functional derivatives described in App. A for the case where the energy functional (3.9) is linear in both the bra $\langle\Psi|$ and the ket $|\Psi\rangle$ functions. Thus one may vary either, or both at the same time, with the same result.

3.2 Coulomb interaction in condensed matter

It is helpful to clarify briefly several points that are essential to a proper definition of energies in extended systems with long-range Coulomb interaction. For a more complete analysis see App. F. The key points are:

- Any extended system must be neutral if the energy is to be finite.
- Terms in the energy must be organized in neutral groups for actual evaluation.

In (3.9) the most convenient approach is to identify and group together terms representing the classical Coulomb energies,

$$E^{\mathrm{CC}} = E_{\mathrm{Hartree}} + \int \mathrm{d}^3 r\, V_{\mathrm{ext}}(\mathbf{r}) n(\mathbf{r}) + E_{II}, \tag{3.14}$$

where E_{Hartree} is the self-interaction energy of the density $n(\mathbf{r})$ treated as a classical charge density

$$E_{\mathrm{Hartree}} = \frac{1}{2} \int \mathrm{d}^3 r \mathrm{d}^3 r' \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{3.15}$$

Since E_{II} is the interaction among the positive nuclei and $\int \mathrm{d}^3 r\, V_{\mathrm{ext}}(\mathbf{r}) n(\mathbf{r})$ is the interaction of the electrons with the nuclei, (3.14) is a neutral grouping of terms so long as the system is neutral. Evaluation of classical Coulomb energies is an intrinsic part of quantitative electronic structure calculations; methods for dealing with long-range Coulomb interaction are described in App. F.

It then follows that the total energy expression, (3.9), can be written as

$$E = \langle \hat{T} \rangle + (\langle \hat{V}_{\mathrm{int}} \rangle - E_{\mathrm{Hartree}}) + E^{\mathrm{CC}}, \tag{3.16}$$

where each of the three terms is well defined. The middle term in brackets, $\langle \hat{V}_{\mathrm{int}} \rangle - E_{\mathrm{Hartree}}$, is the difference between the Coulomb energies of interacting, correlated electrons with density $n(\mathbf{r})$ and that of a continuous classical charge distribution having the same density, which is defined to be the potential part of the exchange–correlation energy E_{xc} in density functional theory (see Secs. 6.3 and 7.3, especially the discussion related to Eq. (7.15)).[3] Thus all long-range interactions cancel in the difference, so that effects of exchange and correlation are short ranged. This is a point to which we will return in Ch. 7 and App. H.

3.3 Force and stress theorems

Force (Hellmann–Feynman) theorem

One of the beautiful theorems of physics is the "force theorem" for the force conjugate to any parameter in the hamiltonian. This is a very general idea perhaps formulated first in 1927 by Ehrenfest [251], who recognized that it is crucial for the correspondence principle of quantum and classical mechanics. He established the relevant relation by showing that the

[3] This definition differs from that given in many texts (see Sec. 3.6) as the difference from Hartree–Fock, since the Hartree–Fock density differs from the true density.

expression for force given below equals the expectation value of the operator corresponding to acceleration $\langle \mathrm{d}^2\hat{x}/\mathrm{d}t^2 \rangle$. The ideas are implicit in the 1928 work of Born and Fock [252], and the explicit formulas used today were given by Güttiger [253] in 1932. The formulas were included in the treatises of Pauli [254] and Hellmann [255], the latter reformulating them as a variational principle in a form convenient for application to molecules. In 1939, Feynman [256] derived the force theorem and explicitly pointed out that the force on a nucleus is given strictly in terms of the charge density, independent of the electron kinetic energy, exchange, and correlation. Thus as an "electrostatic theorem," it should apparently be attributed to Feynman. The nomenclature "Hellmann–Feynman theorem" has been widely used, apparently originating with Slater [41]; however, we will use the term "force theorem."

The force conjugate to any parameter describing a system, such as the position of a nucleus $\mathbf{R}_I$, can always be written

$$\mathbf{F}_I = -\frac{\partial E}{\partial \mathbf{R}_I}. \tag{3.17}$$

From the general expression for the total energy (3.9), the derivative can be written using first-order perturbation theory (the normalization does not change and we assume $\langle \Psi | \Psi \rangle = 1$ for convenience),

$$-\frac{\partial E}{\partial \mathbf{R}_I} = -\langle \Psi | \frac{\partial \hat{H}}{\partial \mathbf{R}_I} | \Psi \rangle - \langle \frac{\partial \Psi}{\partial \mathbf{R}_I} | \hat{H} | \Psi \rangle - \langle \Psi | \hat{H} | \frac{\partial \Psi}{\partial \mathbf{R}_I} \rangle - \frac{\partial E_{II}}{\partial \mathbf{R}_I}. \tag{3.18}$$

Using the fact that at the exact ground state solution the energy is extremal with respect to all possible variations of the wavefunction, it follows that the middle two terms in (3.18) vanish and the only non-zero terms come from the *explicit* dependence of the nuclear position. Furthermore, using the form of the energy in (3.9), it follows that the force depends upon only the density n of the electrons and the other nuclei,

$$\mathbf{F}_I = -\frac{\partial E}{\partial \mathbf{R}_I} = -\int \mathrm{d}^3 r n(\mathbf{r}) \frac{\partial V_{\mathrm{ext}}(\mathbf{r})}{\partial \mathbf{R}_I} - \frac{\partial E_{II}}{\partial \mathbf{R}_I}. \tag{3.19}$$

Here $n(\mathbf{r})$ is the *unperturbed* density and the other nuclei are held fixed, as shown schematically in the left-hand side of Fig. I.1. Since each nucleus interacts with the electrons and other nuclei via Coulomb interactions, the right-hand side of (3.19) can be shown (Exercise 3.3) to equal the nuclear charge times the electric field due to the electrons, which is the electrostatic theorem of Feynman. Thus even though the kinetic energy and internal interactions change as the nuclei move, all such terms cancel in the force theorem.

In the case of non-local potentials (such as pseudopotentials), the force cannot be expressed solely in terms of the electron density. However, the original expression is still valid and useful expressions can be directly derived from

$$-\frac{\partial E}{\partial \mathbf{R}_I} = -\langle \Psi | \frac{\partial \hat{H}}{\partial \mathbf{R}_I} | \Psi \rangle - \frac{\partial E_{II}}{\partial \mathbf{R}_I}. \tag{3.20}$$

Because the force theorem depends upon the requirement that the electronic states are at their variational minimum, it follows that there must be a continuum of "force theorems" that corresponds to the addition of any linear variation in Ψ or n to the above expression. Although such terms vanish in principle, they can have an enormous impact upon the

accuracy and physical interpretation of resulting formulas. The most relevant example in electronic structure is the case of core electrons: it is more physical and more accurate computationally to move the electron density in the core region along with the nucleus rather than holding the density strictly fixed. Methods to accomplish this are described in App. I and illustrated in Fig. I.1.

Finally, there are drawbacks to the fact that expressions for the force theorem depend upon the electronic wavefunction being an exact eigenstate. If the basis is not complete, or the state is approximated, then there may be additional terms. For example, if the basis is not complete and it depends upon the positions of the nuclei, then there are additional terms that must be explicitly included so that the expression for the force given by the force theorem is identical to the explicit derivative of the energy (Exercise 3.4). Explicit expressions are given for use in independent-particle Kohn–Sham calculations in Sec. 9.4.

Generalized force theorem and coupling constant integration

The derivative of the energy with respect to any parameter λ in the hamiltonian can be calculated using the variational property of the wavefunction. Furthermore, an integral expression provides a way to calculate finite energy difference between any two states connected by a continuous variation of the hamiltonian. The general expressions can be written,

$$\frac{\partial E}{\partial \lambda} = \langle \Psi_\lambda | \frac{\partial \hat{H}}{\partial \lambda} | \Psi_\lambda \rangle \tag{3.21}$$

and

$$\Delta E = \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \frac{\partial E}{\partial \lambda} = \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \langle \Psi_\lambda | \frac{\partial \hat{H}}{\partial \lambda} | \Psi_\lambda \rangle. \tag{3.22}$$

For example, if a parameter such as the charge squared of the electron e^2 in the interaction energy in the hamiltonian is scaled by $e^2 \to e^2\lambda$, then λ can be varied from 0 to 1 to vary the hamiltonian from the non-interacting limit to the fully interacting problem. Since the hamiltonian involves the charge only in the interaction term, and (3.5) is linear in e^2 (the nuclear term is treated separately as the "external potential"), it follows that the change in energy can be written

$$\Delta E = \int_0^1 \mathrm{d}\lambda \langle \Psi_\lambda | V_{\mathrm{int}} | \Psi_\lambda \rangle, \tag{3.23}$$

where V_{int} is the full interaction term (3.5) and Ψ_λ is the wavefunction for intermediate values of the interaction[4] given by $e^2 \to e^2\lambda$. The disadvantage of this approach is that it requires the wavefunction at intermediate (unphysical) values of e; nevertheless, it can be very useful, e.g. in the construction of density functionals for interacting many-body systems.

[4] The change in energy can be computed for any ground or excited state. States of different symmetry can be followed uniquely even if they cross. In many cases it is more efficient to solve a matrix equation (the size of the number of states of the same symmetry that are strongly mixed).

Stress (generalized virial) theorem

A physically different type of variation is a scaling of space, which leads to the "stress theorem" [128,129] for total stress. This is a generalization of the well-known virial theorem [254, 257–260] for pressure P which was derived in the early days of quantum mechanics. An elegant derivation was given by Fock [259] in terms of "Streckung des Grundgebietes" ("stretching of the ground state").

The stress is a generalized force for which the ideas of the force theorem can be applied. The key point is that for a system in equilibrium, the stress tensor $\sigma_{\alpha\beta}$ is minus the derivative of the energy with respect to strain $\epsilon_{\alpha\beta}$ per unit volume

$$\sigma_{\alpha\beta} = -\frac{1}{\Omega}\frac{\partial E}{\partial \epsilon_{\alpha\beta}}, \tag{3.24}$$

where α and β are the cartesian indices, and where strain is defined to be a scaling of space, $\mathbf{r}_\alpha \to (\delta_{\alpha\beta} + \epsilon_{\alpha\beta})\mathbf{r}_\beta$, where $\mathbf{r}$ is any vector in space including particle positions and translation vectors. The effect is to transform the wavefunction by scaling every particle coordinate [129],

$$\Psi_\epsilon(\{\mathbf{r}_i\}) = \det(\delta_{\alpha\beta} + \epsilon_{\alpha\beta})^{-1/2}\Psi(\{(\delta_{\alpha\beta} + \epsilon_{\alpha\beta})^{-1}\mathbf{r}_{i\beta}\}), \tag{3.25}$$

where the prefactor preserves the normalization. Since the wavefunction also depends upon the nuclear positions (either explicitly, treating the nuclei as quantum particles, or implicitly, as parameters in the Born–Oppenheimer approximation discussed after (3.1)), so also must the nuclear positions be scaled. Of course, the wavefunction and the nuclear positions actually change in other ways if the system is compressed or expanded; however, this has no effect upon the energy to first order because the wavefunction and the nuclear positions are at variational minima.

Substituting $\Psi_\epsilon(\{\mathbf{r}_i\})$ into expression (3.9) for the energy, changing variables in the integrations, and using (3.24) leads directly to the expression [129]

$$\sigma_{\alpha\beta} = -\left\langle \Psi \left| \sum_k \frac{\hbar^2}{2m_k}\nabla_{k\alpha}\nabla_{k\beta} - \frac{1}{2}\sum_{k\neq k'} \frac{(\mathbf{x}_{kk'})_\alpha (\mathbf{x}_{kk'})_\beta}{x_{kk'}} \left(\frac{\mathrm{d}}{\mathrm{d}x_{kk'}}\hat{V}\right) \right| \Psi \right\rangle, \tag{3.26}$$

where the sum over k and k' denotes a double sum over all particles, nuclei and electrons, where the interaction is a function of the distance $x_{kk'} = |\mathbf{x}_{kk'}|$. The virial theorem for pressure $P = -\sum_\alpha \sigma_{\alpha\alpha}$ is the trace of (3.26), which follows from isotropic scaling of space, $\epsilon_{\alpha\beta} = \epsilon\delta_{\alpha\beta}$. If all interactions are Coulombic and the potential energy includes all terms due to nuclei and electrons, the virial theorem leads to

$$3P\Omega = 2E_{\text{kinetic}} + E_{\text{potential}}, \tag{3.27}$$

where Ω is the volume of the system. The expression (3.26) is a general result valid in any system in equilibrium, classical or quantum, at any temperature, so long as all particles interact with central two-body forces. Explicit expressions [104, 128] used in practical calculations in Fourier space are discussed in App. G.

3.4 Statistical mechanics and the density matrix

From quantum statistical mechanics one can derive expressions for the energy U, entropy S, and free energy $F = U - TS$, at a temperature T. The general expression for F is

$$F = \mathrm{Tr}\hat{\rho}\left(\hat{H} + \frac{1}{\beta}\ln\hat{\rho}\right), \tag{3.28}$$

where $\hat{\rho}$ is the density matrix and $\beta = 1/k_B T$. Here Tr means trace over all the states of the system which have a fixed number of particles N. The final term is the entropy term, which is the log of the number of possible states of the system. A general property of the density matrix is that it is positive definite, since its diagonal terms are the density. The correct equilibrium density matrix is the positive definite matrix that minimizes the free energy,

$$\hat{\rho} = \frac{1}{Z}\mathrm{e}^{-\beta\hat{H}}, \tag{3.29}$$

with the partition function given by

$$Z = \mathrm{Tr}\,\mathrm{e}^{-\beta\hat{H}} = \mathrm{e}^{-\beta F}. \tag{3.30}$$

In a basis of eigenstates Ψ_i of $\hat{H}$, $\hat{\rho}$ has only diagonal matrix elements,

$$\rho_{ii} \equiv \langle\Psi_i|\hat{\rho}|\Psi_i\rangle = \frac{1}{Z}\mathrm{e}^{-\beta E_i};\ Z = \sum_j \mathrm{e}^{-\beta E_j}, \tag{3.31}$$

where ρ_{ii} is the probability of state i. Since the Ψ_i form a complete set, the operator $\hat{\rho}$ in (3.29) can be written

$$\hat{\rho} = \sum_i |\Psi_i\rangle\rho_{ii}\langle\Psi_i|, \tag{3.32}$$

in Dirac bra and ket notation.

In the grand canonical ensemble, in which the number of particles is allowed to vary, the expressions are modified to include the chemical potential μ and the number operator $\hat{N}$. The grand potential Ω and the grand partition function Z are given by

$$Z = \mathrm{e}^{-\beta\Omega} = \mathrm{Tr}\,\mathrm{e}^{-\beta(\hat{H}-\mu\hat{N})}, \tag{3.33}$$

where now the trace is over all states with any particle number, and the grand density matrix operator is the generalization of (3.29),

$$\hat{\rho} = \frac{1}{Z}\mathrm{e}^{-\beta(\hat{H}-\mu\hat{N})}. \tag{3.34}$$

All the equilibrium properties of the system are determined by the density matrix, just as they are determined by the ground state wavefunction at $T = 0$. In particular, any expectation value is given by

$$\langle\hat{O}\rangle = \mathrm{Tr}\,\hat{\rho}\hat{O}, \tag{3.35}$$

which reduces to a ground state expectation value of the form of (3.7) at $T = 0$. For the case of non-interacting particles, the general formulas reduce to the well-known expressions for fermions and bosons given in the next section.

3.5 Independent-electron approximations

There are two basic independent-particle approaches that may be classified as "non-interacting" and "Hartree–Fock." They are similar in that each assumes the electrons are uncorrelated except that they must obey the exclusion principle. However, they are different in that Hartree–Fock includes the electron–electron Coulomb interaction in the energy, while neglecting the correlation that is introduced in the true wavefunction due to those interactions. In general, "non-interacting" theories have some effective potential that incorporates some effect of the real interaction, but there is no interaction term explicitly included in the effective hamiltonian. This approach is often referred to as "Hartree" or "Hartree-like," after D. R. Hartree [43] who included an average Coulomb interaction in a rather heuristic way.[5] More to the point of modern calculations, *all* calculations following the Kohn–Sham method (see Chs. 7–9) involve a non-interacting hamiltonian with an effective potential chosen to incorporate exchange and correlation effects approximately.

Non-interacting (Hartree-like) electron approximation

With our broad definition, all non-interacting electron calculations involve the solution of a Schrödinger-like equation

$$\hat{H}_{\text{eff}}\psi_i^\sigma(\mathbf{r}) = \left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{\text{eff}}^\sigma(\mathbf{r})\right]\psi_i^\sigma(\mathbf{r}) = \varepsilon_i^\sigma\psi_i^\sigma(\mathbf{r}), \tag{3.36}$$

where $V_{\text{eff}}^\sigma(\mathbf{r})$ is an effective potential that acts on each electron of spin σ at point $\mathbf{r}$.[6] The ground state for many non-interacting electrons is found by occupying the lowest eigenstates of (3.36) obeying the exclusion principle. If the hamiltonian is not spin-dependent, then up and down spin states are degenerate and one can simply consider spin as a factor of two in the counting. Excited states involve occupation of higher energy eigenstates. There is no need to construct an antisymmetric wavefunction literally. Since the eigenstates of the independent-particle Schrödinger equation are automatically orthogonal, an antisymmetric wavefunction like (3.43) can be formed from a determinant of these eigenstates. It is then straightforward to show that, *if the particles are non-interacting*, the relations reduce to the expressions given below for the energy, density, etc. (Exercise 3.6).

The solution of equations having the form of (3.36) is at the heart of the methods described in this volume. The basic justification of the use of such independent-particle equations for electrons in materials is density functional theory, which is the subject of Chs. 6–9. The

[5] Historically, the first quantitative calculations on many-electron systems were carried out on atoms by D. R. Hartree [43] who solved, numerically, the equation for each electron moving in a central potential due to other electrons and the nucleus. Hartree defined a different potential for each electron because he subtracted a self-term for each electron that depended upon its orbital. However, following the later development of the Hartree–Fock method [46], it is now customary to define the effective "Hartree potential" with an unphysical self-interaction term so that the potential is orbital independent. This unphysical term has no effect since it is cancelled by the exchange term in Hartree–Fock calculations.

[6] Spin is introduced at this point because it is necessary to introduce a spin-dependent effective potential in order for the independent-particle equations to reproduce spin polarized electron states properly.

following chapters are devoted to methods for solving the equations and applications to the properties of matter, such as predictions of structures, phase transitions, magnetism, elastic constants, phonons, piezoelectric and ferroelectric moments, and many other quantities.

At finite temperature it is straightforward to apply the general formulas of statistical mechanics given in the previous section to show that the equilibrium distribution of electrons is given by the Fermi–Dirac (or Bose–Einstein) expression (1.1) for occupation numbers of states as a function of energy (Exercise 3.7). The expectation value (3.35) is a sum over many-body states Ψ_j, each of which is specified by the set of occupation numbers $\{n_i^\sigma\}$ for each of the independent particle states with energy ε_i^σ. Given that each n_i^σ can be either 0 or 1, with $\sum_i n_i^\sigma = N^\sigma$, it is straightforward (see Exercise 3.8) to show that (3.35) simplifies to

$$\langle \hat{O} \rangle = \sum_i^\sigma f_i^\sigma \langle \psi_i^\sigma | \hat{O} | \psi_i^\sigma \rangle, \tag{3.37}$$

where $\langle \psi_i^\sigma | \hat{O} | \psi_i^\sigma \rangle$ is the expectation value of the operator $\hat{O}$ for the one-particle state ψ_i^σ, and f_i^σ is the probability of finding an electron in state i, σ given in general by (1.1). The relevant case is the Fermi–Dirac distribution

$$f_i^\sigma = \frac{1}{e^{\beta(\varepsilon_i^\sigma - \mu)} + 1}, \tag{3.38}$$

where μ is the Fermi energy (or chemical potential) of the electrons. For example, the energy is the weighted sum of non-interacting particle energies ε_i^σ

$$E(T) = \langle \hat{H} \rangle = \sum_i^\sigma f_i^\sigma \varepsilon_i^\sigma. \tag{3.39}$$

Just as in the general many-body case, one can define a single-body density matrix operator

$$\hat{\rho} = \sum_i |\psi_i^\sigma\rangle f_i^\sigma \langle \psi_i^\sigma|, \tag{3.40}$$

in terms of which an expectation value (3.37) is $\langle \hat{O} \rangle = \mathrm{Tr}\, \hat{\rho} \hat{O}$ in analogy to (3.35). For example, in an explicit spin and position representation, $\hat{\rho}$ is given by

$$\rho(\mathbf{r}, \sigma; \mathbf{r}', \sigma') = \delta_{\sigma,\sigma'} \sum_i \psi_i^{\sigma *}(\mathbf{r}) f_i \psi_i^\sigma(\mathbf{r}'), \tag{3.41}$$

where the density is the diagonal part

$$n^\sigma(\mathbf{r}) = \rho(\mathbf{r}, \sigma; \mathbf{r}, \sigma) = \sum_i f_i^\sigma |\psi_i^\sigma(\mathbf{r})|^2. \tag{3.42}$$

Hartree–Fock approximation

A standard method of many-particle theory is the Hartree–Fock method that was first applied to atoms in 1930 by Fock [46]. In this approach one writes a properly antisymmetrized determinant wavefunction for a fixed number N of electrons, and finds the single determinant

that minimizes the total energy for the full interacting hamiltonian (3.2). If there is no spin–orbit interaction, the determinant wavefunction Φ can be written as a Slater determinant[7]

$$\Phi = \frac{1}{(N!)^{1/2}} \begin{vmatrix} \phi_1(\mathbf{r}_1,\sigma_1) & \phi_1(\mathbf{r}_2,\sigma_2) & \phi_1(\mathbf{r}_3,\sigma_3) & \dots \\ \phi_2(\mathbf{r}_1,\sigma_1) & \phi_2(\mathbf{r}_2,\sigma_2) & \phi_2(\mathbf{r}_3,\sigma_3) & \dots \\ \phi_3(\mathbf{r}_1,\sigma_1) & \phi_3(\mathbf{r}_2,\sigma_2) & \phi_3(\mathbf{r}_3,\sigma_3) & \dots \\ . & . & . & \dots \\ . & . & . & \dots \end{vmatrix}, \tag{3.43}$$

where the $\phi_i(\mathbf{r}_j,\sigma_j)$ are single particle "spin-orbitals" each of which is a product of a function of the position $\psi_i^\sigma(\mathbf{r}_j)$ and a function of the spin variable $\alpha_i(\sigma_j)$. (Note that $\psi_i^\sigma(\mathbf{r}_j)$ is independent of spin σ in closed-shell cases. In open-shell systems, this assumption corresponds to the "spin-restricted Hartree–Fock approximation.") The spin-orbitals must be linearly independent and if, in addition, they are orthonormal the equations simplify greatly; it is straightforward to show (Exercise 3.10) that Φ is normalized to 1. Furthermore, if the hamiltonian is independent of spin or is diagonal in the basis $\sigma = |\uparrow\rangle; |\downarrow\rangle$, the expectation value of the hamiltonian (3.2), using Hartree atomic units, with the wavefunction (3.43) is given by (Exercise 3.11)

$$\begin{aligned} \langle\Phi|\hat{H}|\Phi\rangle = &\sum_{i,\sigma} \int \mathrm{d}\mathbf{r}\psi_i^{\sigma *}(\mathbf{r}) \left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r})\right] \psi_i^\sigma(\mathbf{r}) + E_{II} \\ &+ \frac{1}{2} \sum_{i,j,\sigma_i,\sigma_j} \int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \psi_i^{\sigma_i *}(\mathbf{r}) \psi_j^{\sigma_j *}(\mathbf{r}') \frac{1}{|\mathbf{r}-\mathbf{r}'|} \psi_i^{\sigma_i}(\mathbf{r}) \psi_j^{\sigma_j}(\mathbf{r}') \\ &- \frac{1}{2} \sum_{i,j,\sigma} \int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \psi_i^{\sigma *}(\mathbf{r}) \psi_j^{\sigma *}(\mathbf{r}') \frac{1}{|\mathbf{r}-\mathbf{r}'|} \psi_j^{\sigma}(\mathbf{r}) \psi_i^{\sigma}(\mathbf{r}'). \end{aligned} \tag{3.44}$$

The first term groups together the single-body expectation values which involve a sum over orbitals, whereas the third and fourth terms are the direct and exchange interactions among electrons, which are double sums. We have followed the usual practice of including the $i = j$ "self-interaction," which is spurious but which cancels in the sum of direct and exchange terms. When this term is included, the sum over all orbitals gives the density and the direct term is simply the Hartree energy defined in (3.15). The "exchange" term, which acts only between same spin electrons since the spin parts of the orbitals are orthogonal for opposite spins, is discussed below in Sec. 3.6 and in the chapters on density functional theory.

The Hartree–Fock approach is to minimize the total energy with respect to all degrees of freedom in the wavefunction with the restriction that it has the form (3.43). Since orthonormality was used to simplify the equations, it must be maintained in the minimization, which can be done by Lagrange multipliers as in (3.10) to (3.13). If the spin functions are quantized along an axis, variation of $\psi_i^{\sigma *}(\mathbf{r})$ for each spin σ leads to the Hartree–Fock

[7] The determinant formulation had been realized by Dirac [25] before Slater's work, but the determinant of spin-orbitals is due to Slater [26], who considered this as his most popular work [41] because it replaced difficult group theoretical arguments by this simple form.

equations

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + \sum_{j,\sigma_j}\int \mathrm{d}\mathbf{r}'\psi_j^{\sigma_j *}(\mathbf{r}')\psi_j^{\sigma_j}(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right]\psi_i^{\sigma}(\mathbf{r}) - \sum_j \int \mathrm{d}\mathbf{r}'\psi_j^{\sigma *}(\mathbf{r}')\psi_i^{\sigma}(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}\psi_j^{\sigma}(\mathbf{r}) = \varepsilon_i^{\sigma}\psi_i^{\sigma}(\mathbf{r}), \tag{3.45}$$

where the exchange term is summed over all orbitals of the same spin including the self-term $i = j$ that cancels the unphysical self-term included in the direct term. If the exchange term is modified by multiplying and dividing by $\psi_i^{\sigma}(\mathbf{r})$, (3.45) can be written in a form analogous to (3.36) except that the effective hamiltonian is an operator that depends upon the state

$$\hat{H}^i_{\text{eff}}\psi_i^{\sigma}(\mathbf{r}) = \left[-\frac{\hbar^2}{2m_e}\nabla^2 + \hat{V}^{i,\sigma}_{\text{eff}}(\mathbf{r})\right]\psi_i^{\sigma}(\mathbf{r}) = \varepsilon_i^{\sigma}\psi_i^{\sigma}(\mathbf{r}), \tag{3.46}$$

with

$$\hat{V}^{i,\sigma}_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}(\mathbf{r}) + \hat{V}_x^{i,\sigma}(\mathbf{r}), \tag{3.47}$$

and the exchange term operator $\hat{V}_x$ is given by a sum over orbitals of the same spin σ

$$\hat{V}_x^{i,\sigma}(\mathbf{r}) = -\sum_j \int \mathrm{d}\mathbf{r}'\psi_j^{\sigma *}(\mathbf{r}')\psi_i^{\sigma}(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}\frac{\psi_j^{\sigma}(\mathbf{r})}{\psi_i^{\sigma}(\mathbf{r})}. \tag{3.48}$$

Note that this is a differential-integral equation for each orbital ψ_i^{σ} in terms of the exchange operator $\hat{V}_x^{i,\sigma}(\mathbf{r})$ that is an integral involving ψ_i^{σ} and all the other ψ_j^{σ} with the same spin. The term in square brackets is the Coulomb potential due to the "exchange charge density" $\sum_j \psi_j^{\sigma *}(\mathbf{r}')\psi_i^{\sigma}(\mathbf{r}')$ for the state i, σ. Furthermore, $\hat{V}_x^{i,\sigma}(\mathbf{r})$ diverges at points where $\psi_i^{\sigma}(\mathbf{r}) = 0$; this requires care in solving the equations, but is not a fundamental problem since the product $\hat{V}_x^{i,\sigma}(\mathbf{r})\psi_i^{\sigma}(\mathbf{r})$ has no singularity.

We will not discuss the solution of the Hartree–Fock equations in any detail since this is given in many texts [247, 261]. Unlike the case of independent Hartree–like equations, the Hartree–Fock equations can be solved directly only in special cases such as spherically symmetric atoms and the homogeneous electron gas. In general, one must introduce a basis, in which case the energy (3.44) can be written in terms of the expansion coefficients of the orbitals and the integrals involving the basis functions. Variation then leads to the Roothan and Pople–Nesbet equations widely used in quantum chemistry [247, 261]. In general, these are much more difficult to solve than the independent Hartree–like equations and the difficulty grows with size and accuracy since one must calculate N^4_{basis} integrals.[8]

Koopmans' theorem

What is the meaning of the eigenvalues of the Hartree–Fock equation (3.45)? Of course, Hartree–Fock is only an approximation to the energies for addition and removal of electrons,

[8] For large systems, Coulomb integrals between localized functions can be reduced to linear in N_{basis} [262].

since all effects of correlation are omitted. Nevertheless, it is very valuable to have a rigorous understanding of the eigenvalues, which is provided by Koopmans' theorem:

The eigenvalue of a filled (empty) orbital is equal to the change in the total energy (3.44), if an electron is subtracted from (added to) the system, i.e. decreasing (increasing) the size of the determinant by omitting (adding) a row and column involving a particular orbital $\phi_j(\mathbf{r}_i, \sigma_i)$, *keeping all the other orbitals the same*.

Koopmans' theorem can be derived by taking matrix elements of (3.45) with the normalized orbital $\psi_i^{\sigma *}(\mathbf{r})$ (see Exercise 3.18). For occupied states, the eigenvalues are lowered by the exchange term, which cancels the spurious repulsive self-interaction in the Hartree term. To find the energies for addition of electrons, one must compute empty orbitals of the Hartree–Fock equation (3.45). For these states there also is no spurious self-interaction since both the direct and the exchange potential terms in (3.45) involve only the occupied states. In general, the gap between addition and removal energies for electrons are greatly overestimated in the Hartree–Fock approximation because of the neglect of relaxation of the orbitals and other effects of correlation.

In finite systems, such as atoms, it is possible to improve upon the use of the eigenvalues as approximate excitation energies. Significant improvement in the addition and removal energies result from the "delta Hartree–Fock approximation," in which one calculates total energy differences directly from (3.44), allowing the orbitals to relax and taking into account the exchange of an added electron with all the others. The energy difference approach for finite systems can be used in any self-consistent field method; hence the name "ΔSCF." Illustrations are given in Sec. 10.6.

3.6 Exchange and correlation

The key problem of electronic structure is that the electrons form an interacting many-body system, with a wavefunction, in general, given by $\Psi(\{\mathbf{r}_i\}) \equiv \Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, as discussed in Sec. 3.1. Since the interactions always involve pairs of electrons, two-body correlation functions are sufficient to determine many properties, such as the energy given by (3.9). Writing out the form for a general expectation value (3.7) explicitly, the joint probability $n(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$ of finding electrons of spin σ at point $\mathbf{r}$ and of spin σ' at point $\mathbf{r}'$, is given by

$$
\begin{aligned}
& n(\mathbf{r}, \sigma; \mathbf{r}', \sigma') \\
& \quad = \left\langle \sum_{i \neq j} \delta(\mathbf{r} - \mathbf{r}_i)\delta(\sigma - \sigma_i)\delta(\mathbf{r}' - \mathbf{r}_j)\delta(\sigma' - \sigma_j) \right\rangle \qquad (3.49) \\
& \quad = N(N-1) \sum_{\sigma_3, \sigma_4, \ldots} \int \mathrm{d}\mathbf{r}_3 \cdots \mathrm{d}\mathbf{r}_N |\Psi(\mathbf{r}, \sigma; \mathbf{r}', \sigma'; \mathbf{r}_3, \sigma_3; \ldots, \mathbf{r}_N, \sigma_N)|^2, \qquad (3.50)
\end{aligned}
$$

assuming Ψ is normalized to unity. For uncorrelated particles, the joint probability is just the product of probabilities, so that the measure of correlation is $\Delta n(\mathbf{r}, \sigma; \mathbf{r}', \sigma') = n(\mathbf{r}, \sigma; \mathbf{r}', \sigma') - n(\mathbf{r}, \sigma)n(\mathbf{r}', \sigma')$, so that

$$
n(\mathbf{r}, \sigma; \mathbf{r}', \sigma') = n(\mathbf{r}, \sigma)n(\mathbf{r}', \sigma') + \Delta n(\mathbf{r}, \sigma; \mathbf{r}', \sigma'). \qquad (3.51)
$$

It is also useful to define the normalized pair distribution,

$$g(\mathbf{r},\sigma;\mathbf{r}',\sigma') = \frac{n(\mathbf{r},\sigma;\mathbf{r}',\sigma')}{n(\mathbf{r},\sigma)n(\mathbf{r}',\sigma')} = 1 + \frac{\Delta n(\mathbf{r},\sigma;\mathbf{r}',\sigma')}{n(\mathbf{r},\sigma)n(\mathbf{r}',\sigma')}, \tag{3.52}$$

which is unity for uncorrelated particles so that correlation is reflected in $g(\mathbf{r},\sigma;\mathbf{r}',\sigma') - 1$. Note that all long-range correlation is included in the average terms so that the remaining terms $\Delta n(\mathbf{r},\sigma;\mathbf{r}',\sigma')$ and $g(\mathbf{r},\sigma;\mathbf{r}',\sigma') - 1$ are short range and vanish at large $|\mathbf{r} - \mathbf{r}'|$.

Exchange in the Hartree–Fock approximation

The Hartree–Fock approximation (HFA) consists of neglecting all correlations *except* those required by the Pauli exclusion principle; however, the exchange term in (3.44) represents two effects: Pauli exclusion and the self-term that must be subtracted to cancel the spurious self-term included in the direct Coulomb Hartree energy. The effect is always to lower the energy, which may be interpreted as the interaction of each electron with a positive "exchange hole" surrounding it. The exchange hole $\Delta n_x(\mathbf{r},\sigma;\mathbf{r}',\sigma')$ is given by $\Delta n(\mathbf{r},\sigma;\mathbf{r}',\sigma')$ in the HFA, where Ψ in (3.50) is approximated by the single determinant wavefunction Φ of (3.43). If the single-particle spin-orbitals $\phi_i^\sigma = \psi_i^\sigma(\mathbf{r}_j) \times \alpha_i(\sigma_j)$ are orthonormal, it is straightforward (Exercise 3.13) to show that the pair distribution function can be written

$$n_{\mathrm{HFA}}(\mathbf{r},\sigma;\mathbf{r}',\sigma') = \frac{1}{2!}\sum_{ij}\begin{vmatrix}\phi_i(\mathbf{r},\sigma) & \phi_i(\mathbf{r}',\sigma')\\ \phi_j(\mathbf{r},\sigma) & \phi_j(\mathbf{r}',\sigma')\end{vmatrix}^2, \tag{3.53}$$

and the exchange hole takes the simple form

$$\Delta n_{\mathrm{HFA}}(\mathbf{r},\sigma;\mathbf{r}',\sigma') = \Delta n_x(\mathbf{r},\sigma;\mathbf{r}',\sigma') = -\delta_{\sigma\sigma'}\left|\sum_i \psi_i^{\sigma *}(\mathbf{r})\psi_i^\sigma(\mathbf{r}')\right|^2. \tag{3.54}$$

It is immediately clear from (3.51) and (3.54) that the exchange hole of an electron involves only electrons of the same spin and that the probability vanishes, as it must, for finding two electrons of the same spin at the same point $\mathbf{r} = \mathbf{r}'$. Note that from (3.54) and (3.41), it follows that in the HFA $\Delta n_x(\mathbf{r},\sigma;\mathbf{r}',\sigma') = -\delta_{\sigma\sigma'}|\rho_\sigma(\mathbf{r},\mathbf{r}')|^2$, where $\rho_\sigma(\mathbf{r},\mathbf{r}')$ is the density matrix, which is diagonal in spin.

This is an example of the general property [263] that indistinguishability of particles leads to correlations, which in otherwise independent-particle systems can be expressed in terms of the first-order density matrix:

$$\Delta n_{\mathrm{ip}}(\mathbf{x};\mathbf{x}') = \pm|\rho_\sigma(\mathbf{x},\mathbf{x}')|^2, \tag{3.55}$$

or

$$g_{\mathrm{ip}}(\mathbf{x};\mathbf{x}') = 1 \pm \frac{|\rho_\sigma(\mathbf{x},\mathbf{x}')|^2}{n(\mathbf{x})n(\mathbf{x}')}, \tag{3.56}$$

where the plus (minus) sign applies for bosons (fermions) and $\mathbf{x}$ incorporates all coordinates including position $\mathbf{r}$ and spin (if applicable). Thus $\Delta n_{\mathrm{ip}}(\mathbf{x};\mathbf{x}')$ is always positive for independent bosons and always negative for independent fermions.

There are stringent conditions on the exchange hole: (1) it can never be positive, $\Delta n_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma') \leq 0$ (which means that $g_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma') \leq 1$), and (2) the integral of the exchange hole density $\Delta n_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$ over all $\mathbf{r}'$ is exactly one missing electron per electron at any point $\mathbf{r}$. This is a consequence of the fact that if one electron is at $\mathbf{r}$, then that same electron cannot also be at $\mathbf{r}'$. It also follows directly from (3.54), as shown in Exercise 3.12. The exchange energy, the last term in (3.44), can be interpreted as the lowering of the energy due to each electron interacting with its positive exchange hole,

$$E_x = \left[\langle \hat{V}_{\text{int}} \rangle - E_{\text{Hartree}}(n)\right]_{\text{HFA}} = \frac{1}{2} \sum_{\sigma} \int \mathrm{d}^3 r \int \mathrm{d}^3 r' \frac{\Delta n_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma')}{|\mathbf{r} - \mathbf{r}'|}. \quad (3.57)$$

In this form it is clear that the exchange energy cancels the unphysical self-interaction term in the Hartree energy.

The simplest example of an exchange hole is a one electron problem, such as the hydrogen atom. There is, of course, no real "exchange" nor any issue of the Pauli exclusion principle, and it is easy to see that the "exchange hole" is exactly the electron density. Its integral is unity, as required by the sum rule, and the exchange energy cancels the spurious Hartree term. Because of this cancellation, the Hartree–Fock equation (3.45) correctly reduces to the usual Schrödinger equation for one electron in an external potential.

The next more complex case is a two-electron singlet such as the ground state of He. In this case (see Exercise 3.16) the two spins have identical spatial orbitals and the exchange term is minus one-half the Hartree term in the Hartree–Fock equation (3.44), so that the Hartree–Fock equation (3.45) simplifies to a Hartree–like equation of the form of (3.36) with V_{eff} a sum of the external (nuclear) potential plus one-half the Hartree potential.[9]

For systems with many electrons the exchange hole must be calculated numerically, except for special cases. The most relevant for us is the homogeneous gas considered in the following section.

Beyond Hartree–Fock: correlation

The energy of a state of many electrons in the Hartree–Fock approximation (3.44) is the best possible wavefunction made from a single determinant (or a sum of a few determinants in multi-reference Hartree–Fock [247] needed for degenerate cases). Improvement of the wavefunction to include correlation introduces extra degrees of freedom in the wavefunction and therefore always lowers the energy for any state, ground or excited, by a theorem often attributed to MacDonald [264]. The lowering of the energy is termed the "correlation energy" E_c.

This is not the only possible definition of E_c, which could also be defined as the difference from some other reference state. The definition in terms of the difference from Hartree–Fock is a well-defined choice in the sense that it leads to the smallest possible magnitude of E_c, since E_{HFA} is the lowest possible energy neglecting correlation. Another well-defined

[9] This is exactly what D. R. Hartree did in his pioneering work [43]; however, his approach of subtracting a self-term for each electron is not the same as the more proper Hartree–Fock theory for more than two electrons.

choice arises naturally in density functional theory, where E_c is also defined as the difference between the exact energy and the energy of an uncorrelated state as (3.44), but with the difference that the orbitals are required to give the *exact* density (see Sec. 3.2 and Ch. 7). In many practical cases this distinction appears not to be of great importance; nevertheless, it is essential to define the energies properly, especially as electronic structure methods become more and more powerful in their ability to calculate effects of correlation.

The effects of correlation can be cast in terms of the remaining part of the pair correlation function beyond exchange $n_c(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$ defined in terms of (3.50) and (3.51) by

$$\Delta n(\mathbf{r}, \sigma; \mathbf{r}', \sigma') \equiv n_{\rm xc}(\mathbf{r}, \sigma; \mathbf{r}', \sigma') = n_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma') + n_c(\mathbf{r}, \sigma; \mathbf{r}', \sigma'). \quad (3.58)$$

Since the entire exchange–correlation hole obeys the sum rule that it integrates to 1, the correlation hole $n_c(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$ must integrate to zero, i.e. it merely redistributes the density of the hole. In general, correlation is most important for electrons of opposite spin, since electrons of the same spin are automatically kept apart by the exclusion principle. For the ground state the correlation energy is always negative and any approximation should be negative. Excited states involve *energy differences* from the ground state, e.g. an exciton energy. Depending upon the effects of correlation in the two states, the *difference* can be positive or negative.

The correlation energy is more complicated to calculate than the exchange energy because correlation affects both kinetic and potential energies. Both effects can be taken into account by a "coupling constant integration" using the methods of Sec. 3.3. Although it is not the goal in this volume to delve into the theory of interacting systems, selected results are given in the chapters on the homogeneous gas, Ch. 5, and density functional theory (see especially Ch. 7) since present-day electronic structure theory strives to incorporate some approximation for the correlation energy in realistic calculations.

3.7 Perturbation theory and the "2*n* + 1 theorem"

Perturbation theory describes the properties of a system with hamiltonian $\hat{H}^0 + \lambda \Delta \hat{H}$ as a systematic expansion in powers of the perturbation, which is conveniently done by organizing terms in powers of λ. The first order expressions depend only upon the unperturbed wavefunctions and $\Delta \hat{H}$ to first-order and have already been given as the force or "generalized force" in Sec. 3.3. To higher order one must determine the variation in the wavefunction. The general form valid in a many-body system can be written in terms of a sum over the excited states of the unperturbed hamiltonian [11, 265, 266],

$$\Delta \Psi_i(\{\mathbf{r}_i\}) = \sum_{j \neq i} \Psi_j(\{\mathbf{r}_i\}) \frac{\langle \Psi_j | \Delta \hat{H} | \Psi_i \rangle}{E_i - E_j}. \quad (3.59)$$

The change in the expectation value of an operator $\hat{O}$ in the perturbed ground state can be cast in the form

$$\Delta \langle \hat{O} \rangle = \sum_{j \neq i} \langle \Delta \Psi_j | \hat{O} | \Psi_i \rangle + \text{c.c.} = \sum_{j \neq i} \frac{\langle \Psi_i | \hat{O} | \Psi_j \rangle \langle \Psi_j | \Delta \hat{H} | \Psi_i \rangle}{E_i - E_j} + \text{c.c.}, \quad (3.60)$$

which can readily be generalized to finite T. An advantage of writing the general many-body expression is that it shows immediately that the perturbation of the many-body ground state Ψ_0 involves only the excited states, an aspect that has to be demonstrated in the simpler independent-particle methods.

In an independent-particle approximation the states are determined by the hamiltonian $\hat{H}_{\text{eff}}$ in the effective Schrödinger equation (3.36). The change in the individual independent-particle orbitals, $\Delta\psi_i(\mathbf{r})$ to first order in perturbation theory, can be written in terms of a sum over the spectrum of the unperturbed hamiltonian $\hat{H}^0_{\text{eff}}$ as [11, 265, 266],

$$\Delta\psi_i(\mathbf{r}) = \sum_{j\neq i} \psi_j(\mathbf{r}) \frac{\langle\psi_j|\Delta\hat{H}_{\text{eff}}|\psi_i\rangle}{\varepsilon_i - \varepsilon_j}, \tag{3.61}$$

where the sum is over all the states of the system, occupied and empty, with the exception of the state being considered. Similarly, the change in the expectation value of an operator $\hat{O}$ in the perturbed ground state to lowest order in $\Delta\hat{H}_{\text{eff}}$ can be written

$$\begin{aligned}\Delta\langle\hat{O}\rangle &= \sum_{i=1}^{\text{occ}} \langle\psi_i + \delta\psi_i|\hat{O}|\psi_i + \delta\psi_i\rangle \\ &= \sum_{i=1}^{\text{occ}} \sum_{j}^{\text{empty}} \frac{\langle\psi_i|\hat{O}|\psi_j\rangle\langle\psi_j|\Delta\hat{H}_{\text{eff}}|\psi_i\rangle}{\varepsilon_i - \varepsilon_j} + \text{c.c.}\end{aligned} \tag{3.62}$$

In (3.62) the sum over j is restricted to conduction states only, which follows from the fact that the contributions of pairs of occupied states i, j and j, i cancel in (3.62) (Exercise 3.21). Expressions Eqs. (3.61) and (3.62) are the basic equations upon which is built the theory of response functions (App. D) and methods for calculating static (Ch. 19) and dynamic responses (Ch. 20) in materials.

The "2n + 1 theorem" states that knowledge of the wavefunction *to all orders* 0 *through* n *determines the energy to order* $2n + 1$. Perhaps the first example of this theorem was by Hylleraas [45] in 1930 in a study of two-electron systems, where he showed that the first-order derivative of an eigenfunction with respect to a perturbation is sufficient to find the second and third derivatives of the energy. In the same paper, Hylleraas observed that there is an expression for the second derivative that is variational (minimal) with respect to errors in $\mathrm{d}\psi/\mathrm{d}\lambda$ (see Sec. 19.5). In the intervening years there have been many works proving the full "2n + 1 theorem," which recently has been extended to density functional theory and other functionals obeying minimum energy principles [153, 267, 268].

It is instructive to write down an example of the third-order energy to see the relation to variational principles following the approach in [267]. The principles can be illustrated for a single state, and the derivation is readily extended to many states [267]. If $\hat{H}$ is expanded in powers of λ, $\hat{H} = \hat{H}^{(0)} + \lambda\hat{H}^{(1)} + \lambda^2\hat{H}^{(2)}$, and similarly for ψ, and the eigenvalue ε, then the Schrödinger equation $(\hat{H} - \varepsilon)\psi = 0$ to order m can written:

$$\sum_{k=0}^{m} (\hat{H} - \varepsilon)^{(m-k)}\psi^{(k)} = 0, \tag{3.63}$$

with the constraint

$$\sum_{j=0}^{m} \langle \psi^{(j)} | \psi^{(m-j)} \rangle = 0, \ m \neq 0. \tag{3.64}$$

Here are collected all terms of order λ^m and then λ is set to 1. Taking matrix elements of (3.63) leads to

$$\sum_{j=0}^{m} \sum_{k=0}^{m} \Theta(m-j-k) \langle \psi^{(j)} | (\hat{H} - \varepsilon)^{(m-j-k)} | \psi^{(k)} \rangle = 0, \tag{3.65}$$

where $\Theta(p) = 1, \ p \geq 0; 0, \ p < 0$.

The desired expressions can be derived by applying the condition that (3.65) be variational with respect to $\psi^{(k)}$ at each order $k = 0, \ldots, m$. This is facilitated by writing (3.65) in the form of an array, illustrated here for $m = 3$,

$$\begin{aligned} 0 = & \langle \psi^{(3)} | \bar{H}^{(0)} | \psi^{(0)} \rangle \\ & + \langle \psi^{(2)} | \bar{H}^{(1)} | \psi^{(0)} \rangle + \langle \psi^{(2)} | \bar{H}^{(0)} | \psi^{(1)} \rangle \\ & + \langle \psi^{(1)} | \bar{H}^{(2)} | \psi^{(0)} \rangle + \langle \psi^{(1)} | \bar{H}^{(1)} | \psi^{(1)} \rangle + \langle \psi^{(1)} | \bar{H}^{(0)} | \psi^{(2)} \rangle \\ & + \langle \psi^{(0)} | \bar{H}^{(3)} | \psi^{(0)} \rangle + \langle \psi^{(0)} | \bar{H}^{(2)} | \psi^{(1)} \rangle + \langle \psi^{(0)} | \bar{H}^{(1)} | \psi^{(2)} \rangle + \langle \psi^{(0)} | \bar{H}^{(0)} | \psi^{(3)} \rangle. \end{aligned} \tag{3.66}$$

Variation of each $|\psi^{(k)}\rangle$ ($\langle \psi^{(k)}|$) in turn means that the sum of elements in each row (column) of (3.66) vanishes. From this it follows that one can eliminate the higher order $\psi^{(k)}, \ k = 2, 3$, with the result (Exercise 3.24)

$$\begin{aligned} \varepsilon^{(3)} = & \langle \psi^{(0)} | \hat{H}^{(3)} | \psi^{(0)} \rangle + \langle \psi^{(1)} | \hat{H}^{(2)} | \psi^{(0)} \rangle + \text{c.c.} \\ & + \langle \psi^{(1)} | \hat{H}^{(1)} - \varepsilon^{(1)} | \psi^{(1)} \rangle. \end{aligned} \tag{3.67}$$

Such expressions are used in electronic structure theory to derive accurate energies from approximate wavefunctions, e.g. in certain expressions in the linearized methods of Ch. 17.

SELECT FURTHER READING

Texts and extensive monographs:

Ashcroft, N. and Mermin, N. *Solid State Physics* (W. B. Saunders Company, New York, 1976).

Jones, W. and March, N. H. *Theoretical Solid State Physics*, *Vol. I* (John Wiley and Sons, New York, 1976).

Kittel, C. *Introduction to Solid State Physics* (John Wiley and Sons, New York, 1996).

Marder, M. *Condensed Matter Physics* (John Wiley and Sons, New York, 2000).

Slater, J. C. *Quantum Theory of Atomic Structures, Vols. 1–4* (McGraw-Hill, New York, 1960–1972).

Szabo, A. and Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (Unabridged reprinting of 1989 version: Dover, Mineola, New York, 1996).

Problems and solutions:

Mihaly, L. and Martin, M. C. *Solid State Physics: Problems and Solutions* (John Wiley and Sons, New York, 1996).

Exercises

3.1 Show that the many-body Schrödinger equation (3.13) also results from explicit variation of the energy in (3.9) without use of Lagrange multipliers.

3.2 Show that the independent-particle Schrödinger equation (3.36) is a special case of the many-body solution. First show this for one particle; then for many non-interacting particles.

3.3 As part of his undergraduate thesis, Feynman showed that the force theorem applied to a nucleus leads to the force being exactly the electric field at the given nucleus due to the charge density of the rest of the system (electrons and other nuclei) times the charge of the given nucleus. Derive this result from (3.19).

3.4 Derive the additional terms that must be included so that the expression for the force given by the force theorem is identical to the explicit derivative of the energy, if the basis depends explicitly upon the positions for the nuclei. Show that the contribution of these terms vanishes if the basis is complete.

3.5 Derive the stress theorem (3.26). Show that this equation reduces to the well-known virial theorem (3.27) in the case of isotropic pressure and Coulomb interactions.

3.6 Show that the relations for non-interacting particles given in the equations following (3.36) remain valid, if a fully antisymmetric determinant wavefunction like (3.43) is created from the orbitals. Note that this holds *only if the particles are non-interacting*.

3.7 Derive the Fermi–Dirac distribution (3.38) for non-interacting particles from the general definition of the density matrix (3.32) using the fact that the sum over many-body states in (3.32) can be reduced to a sum over all possible occupation numbers $\{n_i^\sigma\}$ for each of the independent particle states, subject to the conditions that each n_i^σ can be either 0 or 1, and $\sum_i n_i^\sigma = N^\sigma$.

3.8 Following Exercise 3.7, show that (3.35)) simplifies to (3.37) for any operator in the independent-particle approximation.

3.9 Why is the independent particle density matrix (3.41) diagonal in spin? Is this always the case?

3.10 Show that the Hartree–Fock wavefunction (3.43) is normalized if the independent-particle orbitals are orthonormal.

3.11 Show that the Hartree–Fock wavefunction (3.43) leads to the exchange term in (3.44) and that the variational equation leads to the Hartree–Fock equation (3.45) if the independent-particle orbitals are orthonormal. Explain why the forms are more complicated if the independent-particle orbitals are not orthonormal.

3.12 Show explicitly from the definition (3.54) that the exchange hole around each electron always integrates to one missing electron. Show that, as stated in the text, this is directly related to the fact that "exchange" includes a self-term that cancels the unphysical self-interaction in the Hartree energy.

3.13 Derive the formulas for the pair distribution (3.53) and the exchange hole (3.54) for non-interacting fermions by inserting the Hartree–Fock wavefunction (3.43) into the general definition (3.50).

3.14 By expanding the 2×2 determinant in (3.53): (a) show that

$$\sum_{\sigma'} \int \mathrm{d}r' \Delta n_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma') = (N-1) n(\mathbf{r}, \sigma), \tag{3.68}$$

where $n(\mathbf{r}, \sigma)$ is the density; and (b) derive the formula (3.54) for the exchange hole.

3.15 The relation (3.55) is a general property of non-interacting identical particles [263]. As shown in (3.54), $\Delta n_x(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$ is always negative for fermions. Show that for bosons with a symmetric wavefunction, the corresponding exchange term is always positive.

3.16 Derive the results stated after (3.57) that: (a) for a one-electron problem like hydrogen, the exchange term exactly cancels the Hartree term as it should; and (b) for the ground state of two electrons in a spin singlet state, e.g. in helium, the Hartree–Fock approximation leads to a V_{eff} sum of the external (nuclear) potential plus one-half the Hartree potential.

3.17 Following the exercise above, consider two electrons in a spin triplet state. Show that the situation is not so simple as for the singlet case, i.e. that in the Hartree–Fock approximation there must be two different functions V_{eff} for two different orbitals.

3.18 Derive Koopmans' theorem by explicitly taking matrix elements of the hamiltonian with an orbital to show that the eigenvalue is the same as the energy difference if that orbital is removed.

3.19 For adding electrons one must compute empty orbitals of the Hartree–Fock equation (3.45). There is no self effect of the empty state since only occupied orbitals are included in the sum. Show that the same result for the addition energy is found if one explicitly includes the state in a calculation with one added electron, but keeps the original orbitals unchanged.

3.20 In a finite system Hartree–Fock eigenfunctions have the (surprising) property that the form of the long-range decay of *all* bound states is the same, independent of binding energy. For example, core states have the same exponential decay as valence states, although the prefactor is smaller. Show that this follows from (3.45).

3.21 Show that all contributions involving i and j both occupied vanish in the expectation value (3.62).

3.22 Show that the correlation hole always integrates to zero, i.e. it rearranges the charge correlation. This does not require complex calculations beyond Hartree–Fock theory; all that is needed is to show that conservation laws must lead to this result.

3.23 As an example of the force theorem consider a one-dimensional harmonic oscillator with hamiltonian given by $-\frac{1}{2}(\mathrm{d}^2/\mathrm{d}x^2) + \frac{1}{2}Ax^2$, where A is the spring constant and the mass is set to unity. Using the exact solution for the energy and wavefunction, calculate the generalized force $\mathrm{d}E/\mathrm{d}A$ by direct differentiation and by the force theorem.

3.24 Derive the formula (3.67) for energy to third order from the preceding equations.

3.25 Exercise 19.11 considers the variational principle in perturbation theory applied to a system composed of two springs. Let each spring have a non-linear term $\frac{1}{2}\gamma_1(x_1 - x_0)^3$ and similarly for spring 2. Find an explicit expression for the change in energy to third order due to the applied force.

4

Periodic solids and electron bands

Summary

Classification of crystals and their excitations by symmetry is a general approach applicable to electronic states, vibrational states, and other properties. The first part of this chapter deals with translational symmetry which has the same universal form in all crystals, and which leads to the Bloch theorem that rigorously classifies excitations by their crystal momentum. (The discussion here follows Ashcroft and Mermin, [84], Chs. 4–8.) The other relevant symmetries are time reversal and point symmetries. The latter depend upon a specific crystal structure and are treated only briefly. Detailed classification can be found in many texts, and computer programs that deal with the symmetries can be found on-line at sites listed in Ch. 24.

4.1 Structures of crystals: lattice + basis

A crystal is an ordered state of matter in which the positions of the nuclei (and consequently all properties) are repeated periodically in space. It is completely specified by the types and positions of the nuclei in one repeat unit (primitive unit cell), and the rules that describe the repetition (translations).

- The positions and types of atoms in the primitive cell are called the basis. The set of translations, which generates the entire periodic crystal by repeating the basis, is a lattice of points in space called the Bravais lattice. Specification of the crystal can be summarized as:

$$\text{Crystal structure} = \text{Bravais lattice} + \text{basis}.$$

- The crystalline order is described by its symmetry operations. The set of translations forms a group because the sum of any two translations is another translation.[1] In addition there may be other point operations that leave the crystal the same, such as rotations,

[1] A group is defined by the condition that the application of any two operations leads to a result that is another operation in the group. We will illustrate this with the translation group. The reader is referred to other sources for the general theory and the specific set of groups possible in crystals, e.g. books on group theory, see [270–272], and the comprehensive work by Slater [269].

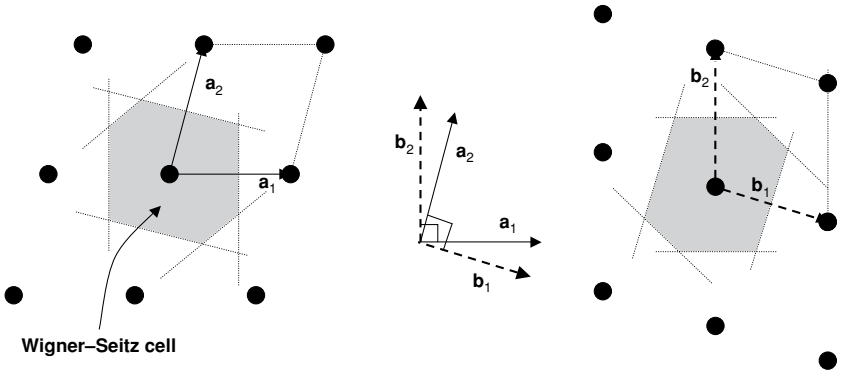


Figure 4.1. Real and reciprocal lattices for a general case in two dimensions. In the middle are shown possible choices for primitive vectors for the Bravais lattice in real space, $\mathbf{a}_1$ and $\mathbf{a}_2$, and the corresponding reciprocal lattice vectors, $\mathbf{b}_1$ and $\mathbf{b}_2$. In each case two types of primitive cells are shown, which when translated fill the two-dimensional space. The parallelepiped cells are simple to construct but are not unique. The Wigner–Seitz cell in real space is uniquely defined as the most compact cell that is symmetric about the origin; the first Brillouin zone is the Wigner–Seitz cell of the reciprocal lattice.

reflections, and inversions. This can be summarized as:

$$\text{Space group} = \text{translation group} + \text{point group.}^{2}$$

The lattice of translations

First we consider translations, since they are intrinsic to all crystals. The set of all translations forms a lattice in space, in which any translation can be written as integral multiples of primitive vectors,

$$\mathbf{T}(\mathbf{n}) \equiv \mathbf{T}(n_1, n_2, \ldots) = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + \ldots, \tag{4.1}$$

where $\mathbf{a}_i$, $i = 1, \ldots, d$ are the primitive translation vectors and d denotes the dimension of the space. For convenience we write formulas valid in any dimension whenever possible and we define $\mathbf{n} = (n_1, n_2, \ldots, n_d)$.

In one dimension, the translations are simply multiples of the periodicity length a, $T(n) = na$, where n can be any integer. The primitive cell can be any cell of length a; however, the most symmetric cell is the one chosen symmetric about the origin $(-a/2, a/2)$ so that each cell centered on lattice point n is the locus of all points closer to that lattice point than to any other point. This is an example of the construction of the Wigner–Seitz cell.

The left-hand side of Fig. 4.1 shows a portion of a general lattice in two dimensions. Space is filled by the set of all translations of any of the infinite number of possible choices

[2] In some crystals the space group can be factorized into a product of translation and point groups; in others (such as the diamond structure) there are non-symmorphic operations that can only be expressed as a combination of translation and a point operation.

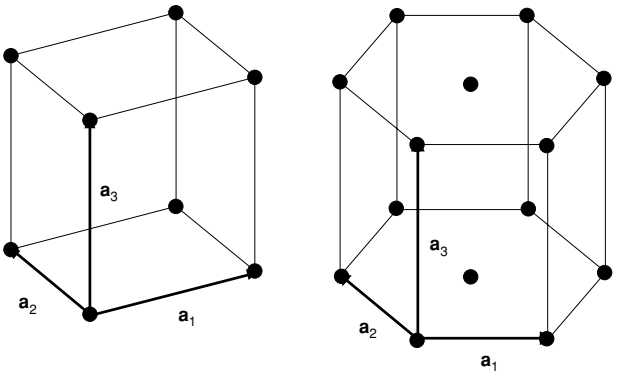


Figure 4.2. Simple cubic (left) and simple hexagonal (right) Bravais lattices. In the simple cubic case, the cell shown is the Wigner–Seitz cell and the Brillouin zone has the same shape. In the hexagonal case, the volume shown contains three atoms; the Wigner–Seitz cell is also a hexagonal prism rotated by 90° and 1/3 the volume. The reciprocal lattice is also hexagonal and rotated from the real lattice by 90°, and the Brillouin zone is shown in Fig. 4.10.

of the primitive cell. One choice of primitive cell is the parallelepiped constructed from the two primitive translation vectors $\mathbf{a}_i$. This cell is often useful for formal proofs and for simplicity of construction. However, this cell is not unique since there are an infinite number of possible choices for $\mathbf{a}_i$. A more informative choice is the Wigner–Seitz cell, which is symmetric about the origin and is the most compact cell possible. It is constructed by drawing the perpendicular bisectors of all possible lattice vectors $\mathbf{T}$ and identifying the Wigner–Seitz cell as the region around the origin bounded by those lines.

In two dimensions there are special choices of lattices that have additional symmetry when the angles between the primitive vectors are 90 or 60°. In units of the length a, the translation vectors are given by:

$$\begin{array}{lccc} & \text{square} & \text{rectangular} & \text{triangular} \\ \mathbf{a}_1 = & (1,0) & (1,0) & (1,0), \\ \mathbf{a}_2 = & (0,1) & \left(0,\frac{b}{a}\right), & \left(\frac{1}{2},\frac{\sqrt{3}}{2}\right). \end{array} \tag{4.2}$$

Examples of crystals having, respectively, square and triangular Bravais lattices are shown later in Figure 4.5.

Figures 4.2–4.4 show examples of three-dimensional lattices that occur in many crystals. The primitive vectors can be chosen to be (in units of a):

$$\begin{array}{lcccc} & \text{simple cubic} & \text{simple hex.} & \text{fcc} & \text{bcc} \\ \mathbf{a}_1 = & (1,0,0) & (1,0,0) & \left(0,\frac{1}{2},\frac{1}{2}\right) & \left(-\frac{1}{2},\frac{1}{2},\frac{1}{2}\right), \\ \mathbf{a}_2 = & (0,1,0) & \left(\frac{1}{2},\frac{\sqrt{3}}{2},0\right) & \left(\frac{1}{2},0,\frac{1}{2}\right) & \left(\frac{1}{2},-\frac{1}{2},\frac{1}{2}\right), \\ \mathbf{a}_3 = & (0,0,1) & \left(0,0,\frac{c}{a}\right) & \left(\frac{1}{2},\frac{1}{2},0\right) & \left(\frac{1}{2},\frac{1}{2},-\frac{1}{2}\right). \end{array} \tag{4.3}$$

The body centered cubic (bcc) and face centered cubic (fcc) lattices are shown, respectively, in Figs. 4.3 and 4.4, each represented in the large conventional cubic cell (indicated by the

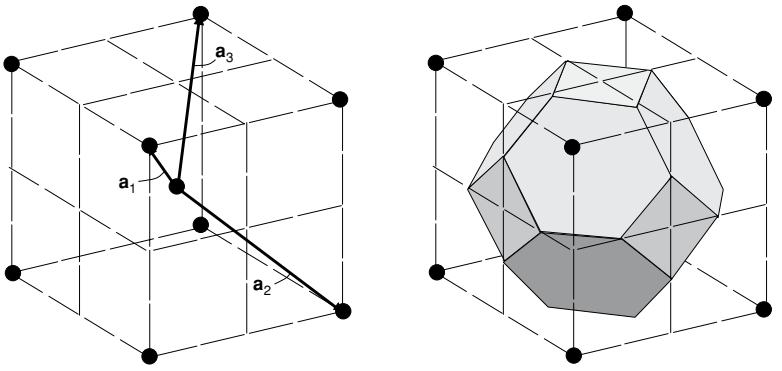


Figure 4.3. Body centered cubic (bcc) lattice, showing one choice for the three lattice vectors. The conventional cubic cell shown indicates the set of all eight nearest neighbors at a distance $\frac{\sqrt{3}}{2}a$ around the central atom. (There are six second neighbors at distance a.) On the right-hand side of the figure is shown the Wigner–Seitz cell formed by the perpendicular bisectors of the lattice vectors (this is also the Brillouin zone for the fcc lattice).

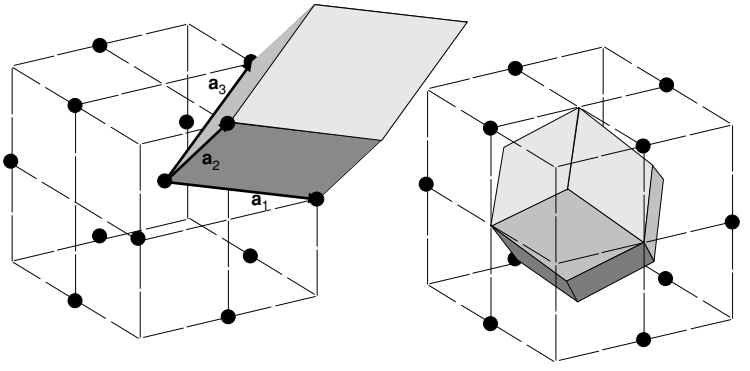


Figure 4.4. Face centered cubic (fcc) lattice, drawn to emphasize the close packing of 12 neighbors around the central site. (The location of sites at face centers is evident if the cube is drawn with a lattice site at each corner and on each face of the cube.) Left: One choice for primitive lattice vectors and the parallelepiped primitive cell, which has lower symmetry than the lattice. Right: the symmetric Wigner–Seitz cell (which is also the Brillouin zone for the bcc lattice).

dashed lines) with a lattice site at the center. All nearest neighbors of the central site are shown: eight for bcc and 12 for fcc lattices. One choice of primitive vectors is shown in each case, but clearly other equivalent vectors could be chosen, and all vectors to the equivalent neighbors are also lattice translations. In the fcc case, the left-hand side of Fig. 4.4 shows one possible primitive cell, the parallelepiped formed by the primitive vectors. This is the simplest cell to construct; however, this cell clearly does not have cubic symmetry and other choices of primitive vectors lead to different cells. The Wigner–Seitz cells for each Bravais lattice, shown respectively in Figs. 4.3 and 4.4, are bounded by planes that are perpendicular bisectors of the translation vectors from the central lattice point. The Wigner–Seitz cell is

particularly useful because it is the unique cell defined as the set of all points in space closer to the central lattice point than to any other lattice point; it is independent of the choice of primitive translations and it has the full symmetry of the Bravais lattice.

It is useful for deriving formal relations and for practical computer programs to express the set of primitive vectors as a square matrix $a_{ij} = (\mathbf{a}_i)_j$, where j denotes the cartesian component and i the primitive vector, i.e. the matrix has the same form as the arrays of vectors shown in Eqs. (4.2) and (4.3).

The volume of any primitive cell must be the same, since translations of any such cell fill all space. The most convenient choice of cell in which to express the volume is the parallelepiped defined by the primitive vectors. If we define $\Omega_{\rm cell}$ as the volume in any dimension d (i.e. it has units (length)^d), simple geometric arguments show that $\Omega_{\rm cell} = |a_1|$ $(d = 1)$; $|\mathbf{a}_1 \times \mathbf{a}_2|$, $(d = 2)$; and $|\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)|$, $(d = 3)$. In any dimension this can be written as the determinant of the $\mathbf{a}$ matrix (see Exercise 4.4),

$$\Omega_{\rm cell} = \det(\mathbf{a}) = |\mathbf{a}|. \tag{4.4}$$

The basis of atoms in a cell

The basis describes the positions of atoms in each unit cell relative to the chosen origin. If there are S atoms per primitive cell, then the basis is specified by the atomic position vectors τ_s, $s = 1, S$. Two-dimensional cases are both instructive and relevant for important problems in real materials. In particular, we consider the CuO square planar structure and the hexagonal graphene structure. These will serve as illustrative examples of simple bands in Ch. 14 and as notable examples of full calculations in Chs. 13 and 17. The square lattice for CuO_2 planes, found in the cuprate high-temperature superconductors, is shown in Fig. 4.5. The lattice vectors are given above and the atomic position vectors are conveniently chosen with the Cu atom at the origin $\tau_1 = (0, 0)$ and the other positions chosen to be $\tau_2 = (\frac{1}{2}, 0)a$ and $\tau_3 = (0, \frac{1}{2})a$. It is useful to place the Cu atom at the origin since it is the position of

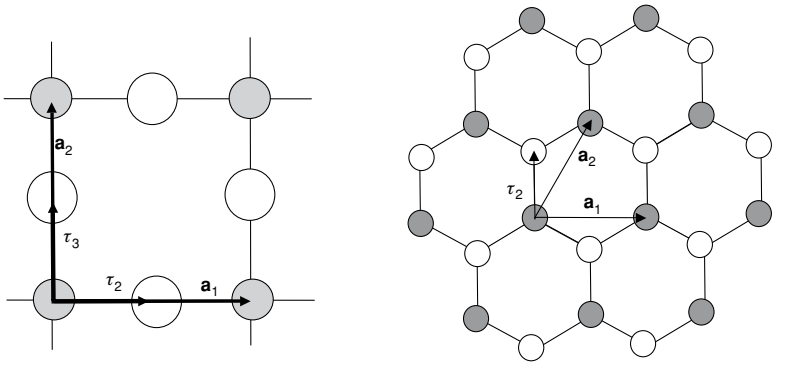


Figure 4.5. Left: Square lattice for CuO_2 planes common to the cuprate high-temperature superconductors: there are three atoms per primitive cell. Right: Honeycomb lattice for a single plane of graphite or hexagonal BN: the lattice is triangular and there are two atoms per primitive cell.

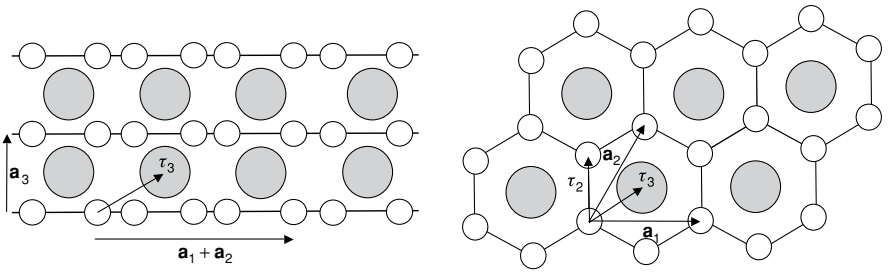


Figure 4.6. Crystal structure of MgB_2, which is an example of lattice, electronic, and superconducting properties given in Ch. 2. Right: A top view of the boron honeycomb planes with Mg atoms (shaded) in the centers of the hexagons. Left: Stacking of planes to make a simple hexagonal three-dimensional structure with Mg atoms between the boron planes. The figure shows the projection of the atoms onto a plane defined by the $\mathbf{a}_3$ and τ_3 vectors, which are in the plane of the page.

highest symmetry in the cell, with inversion, mirror planes, and four-fold rotation symmetry about this site.[3]

A second two-dimensional example is a single plane of graphite or a plane of hexagonal BN, which forms a honeycomb lattice with a triangular Bravais lattice and two atoms per primitive cell, as shown on the right-hand side of Fig. 4.5. If the two atoms are the same chemical species, the structure is that of a plane of graphite. The primitive lattice vectors are $\mathbf{a}_1 = (1, 0)a$ and $\mathbf{a}_2 = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)a$, where the nearest neighbor distance is $a/\sqrt{3}$. If one atom is at the origin, $\tau_1 = (0, 0)$, possible choices of τ_2 include $\tau_2 = (0, 1/\sqrt{3})a$ and $\tau_2 = (1, 1/\sqrt{3})a$, where the latter is symmetric with respect to the primitive vectors, as shown in Fig. 4.5. It is also useful to define the atomic positions in terms of the primitive lattice vectors by $\tau_s = \sum_{i=1}^{d} \tau_{si}^L \, \mathbf{a}_i$, where the superscript L denotes the representation in lattice vectors. In this case, one finds $\tau_2 = \frac{2}{3}(\mathbf{a}_1 + \mathbf{a}_2)$ or $\tau_1^L = [0, 0]$ and $\tau_2^L = \left[\frac{2}{3}, \frac{2}{3}\right]$.[4]

Layered three-dimensional crystals are formed by stacking two-dimensional planes, such as the actual crystals of the CuO materials and graphite. For example, the three-dimensional structure of the high-temperature superconductor $YBa_2Cu_3O_7$ is shown in Fig. 17.3. An example treated in Ch. 2 is MgB_2, which is composed of hexagonal boron planes exactly like the honeycomb graphite layers in Fig. 4.5 (with every atom equivalent) and with Mg atoms between layers in the centers of the hexagons. The three-dimensional structure is simple hexagonal as shown in Fig. 4.6, where $\mathbf{a}_3$ is perpendicular to the layers (the c-axis) and τ_3 is the vector from one boron site to the Mg site. The lattice is the same as hexagonal graphite band structures and the bands for the two crystals are compared in Fig. 2.29.

NaCl and ZnS are two examples of crystals with the fcc Bravais lattice and a basis of two atoms per cell, as shown in Fig. 4.7. The primitive translation vectors are given in the

[3] In any case where the origin can be chosen as the center of symmetry, the Fourier transforms of all properties, such as the density and potential, are real. Also, all excitations can be classified into even and odd relative to this origin, and the four-fold rotational symmetry allows the roles of the five Cu d states to be separated.

[4] Simple reasoning shows that all covalently bonded crystals are expected to have more than one atom per primitive cell (see examples of diamond and ZnS crystals and Exercise 4.8).

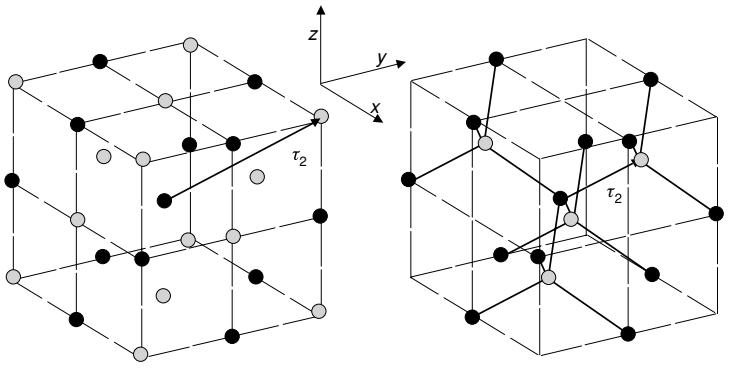


Figure 4.7. Two examples of crystals with a basis of two atoms per cell and fcc Bravais lattice. Left: Rocksalt (or NaCl) structure. Right: zinc-blende (cubic ZnS) structure. The positions of the atoms are given in the text. In the former case, a simple cubic crystal results if the two atoms are the same. In the latter case, if the two atoms are identical the resulting crystal has diamond structure.

previous section in terms of the cube edge a and illustrated in Fig. 4.4. For the case of NaCl, it is convenient to choose one atom at the origin $\tau_1 = (0, 0, 0)$, since there is inversion symmetry and cubic rotational symmetry around each atomic site, and the second basis vector is chosen to be $\tau_2 = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)a$. In terms of the primitive lattice vectors, one can see from Fig. 4.7 that $\tau_2 = \sum_{i=1}^{d} \tau_{2i}^L \, \mathbf{a}_i$, where $\tau_2^L = \left[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right]$. It is also easy to see that if the two atoms at positions τ_1 and τ_2 were the same, then the crystal would actually have a simple cubic Bravais lattice, with cube edge $a_{\text{sc}} = \frac{1}{2}a_{\text{fcc}}$.

A second example is the zinc-blende structure, which is the structure of many III–V and II–VI crystals such as GaAs and ZnS. This crystal is also fcc with two atoms per unit cell. Although there is no center of inversion in a zinc-blende structure crystal, each atom is at a center of tetrahedral symmetry; we can place the origin at one atom, $\tau_1 = (0, 0, 0)a$, and $\tau_2 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)a$, as shown in Fig. 4.7, or any of the equivalent choices. Thus this structure is the same as the NaCl structure except for the basis, which in primitive lattice vectors is simply $\tau_2^L = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]$. If the two atoms in the cell are identical, this is the diamond structure in which C, Si, Ge, and grey Sn occur. A bond center is the appropriate choice of origin for the diamond structure since this is a center of inversion symmetry. This can be accomplished by shifting the origin so that $\tau_1 = -\left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)a$, and $\tau_2 = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)a$; similarly, $\tau_1^L = -\left[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right]$ and $\tau_2^L = \left[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right]$.

The perovskite structure illustrated in Fig. 4.8 has chemical composition ABO_3 and occurs for a large number of compounds with interesting properties including ferroelectrics (e.g. $BaTiO_3$), Mott-insulator antiferromagnets (e.g. $CaMnO_3$), and alloys exhibiting metal–insulator transitions (e.g. $La_xCa_{1-x}MnO_3$). The crystal may be thought of as the CsCl structure with O on the edges. The environment of the A and B atoms is very different, with the A atoms having 12 O neighbors at a distance $a/\sqrt{2}$ and the B atoms having six O neighbors at a distance $a/2$. Thus these atoms play a very different role in the properties. Typically the A atom is a non-transition metal for which Coulomb ionic bonding favors the maximum number of O neighbors, whereas the B atom is a transition metal where the d states

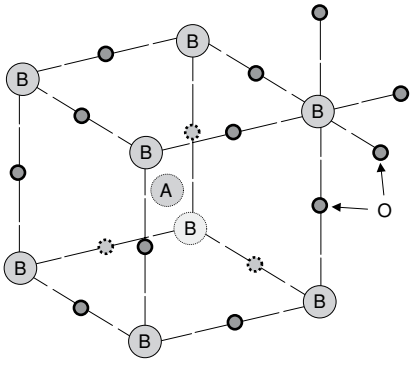


Figure 4.8. The perovskite crystal structure with the chemical composition ABO_3. This structure occurs for a large number of compounds with interesting properties, including ferroelectrics (e.g. $BaTiO_3$), antiferromagnets (e.g. $CaMnO_3$), and alloys (e.g. $Pb_xZr_{1-x}TiO_3$ and $La_xCa_{1-x}MnO_3$). The crystal may be thought of as cubes with A atoms at the center, B at the corners, and O on the edges. The environment of the A and B atoms is very different, the A atom having 12 O neighbors at a distance $a/\sqrt{2}$ and the B atoms having six O neighbors at a distance $a/2$. (The neighbors around one B atom are shown.) The Bravais lattice is simple cubic.

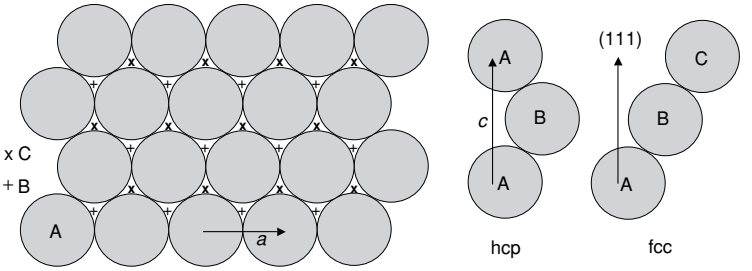


Figure 4.9. Stacking of close-packed planes to create close-packed three-dimensional lattices. Left: The only possible close packing in two dimensions is the hexagonal layer of spheres labeled A, with lattice constant a. Right: Three-dimensional stacking can have the next layers in either B or C positions. Of the infinite set of possible layer sequences, only the fcc stacking (. . . ABCABC . . .) forms a primitive lattice; hexagonal close-packed (hcp) (. . . ABABAB . . .) has two sites per primitive cell; all others have larger primitive cells.

favor bonding with the O states. Note the contrast of the planes of B and O atoms with the CuO_2 planes in Fig. 4.5: although the planes are similar, each B atom in the cubic perovskites is in three intersecting orthogonal planes, whereas in the layered structures such as La_2CuO_4, the CuO_2 planes are clearly identified, with each Cu belonging to only one plane.

Close-packed structures

In two dimensions there is only one way to make a "close-packed structure," defined as a structure in which hard spheres (or disks) can be placed with the maximum filling of space. That is the triangular lattice in Fig. 4.5 with one atom per lattice point. In the plane, each atom has six neighbors in a hexagonal arrangement, as shown in Fig. 4.9. All three-dimensional

close-packed structures consist of such close-packed planes of atoms stacked in various sequences. As shown in Fig. 4.9, the adjacent plane can be stacked in one of two ways: if the given plane is labeled A, then two possible positions for the next plane can be labeled B and C.

The face centered cubic structure (shown in Fig. 4.4) is the cubic close-packed structure, which can be viewed as the sequence of close-packed planes in the sequence . . . ABCABC . . .. It has one atom per primitive cell, as may be seen by the fact that each atom has the same relation to all its neighbors, i.e. an A atom flanked by C and B planes is equivalent to a B atom flanked by A and C planes, etc. Specifically, if the lattice is a Bravais lattice then the vector from an atom in the A plane to one of its closest neighbors in the adjacent C plane must be a lattice vector. Similarly, twice that vector is also a lattice vector, as may be verified easily. The cubic symmetry can be verified by the fact that the close-packed planes may be chosen perpendicular to any of the [111] crystal axes.

The hexagonal closed-packed structure consists of close-packed planes stacked in a sequence . . . ABABAB . . . . This is a hexagonal Bravais lattice with a basis of two atoms that are not equivalent by a translation. (This can be seen because – unlike the fcc case – twice the vector from an A atom to a neighboring B atom is *not* a vector connecting atoms. Thus the primitive cell is hexagonal as shown Fig. 4.2 with a equal to the distance between atoms in the plane and c the distance between two A planes. The ideal c/a ratio is that for packing of hard spheres, $c/a = \sqrt{8/3}$ (Exercise 4.11). (The two atoms in the primitive cell are equivalent by a combination of translation by $c/2$ and rotation by $\pi/6$, but this does not affect the analysis of the translation symmetry.)

There are an infinite number of possible stackings or "polytypes" all of which are "close packed." In particular, polytypes are actually realized in crystals with tetrahedral bonding, like ZnS. The two simplest structures are cubic (zinc-blende) and hexagonal (wurtzite), based upon the fcc and hcp lattices. In this case, each site in one of the A, B, or C planes corresponds to two atoms (Zn and S) and the fcc case is shown in Fig. 4.7.

4.2 The reciprocal lattice and Brillouin zone

Consider any function $f(\mathbf{r})$ defined for the crystal, such as the density of the electrons, which is the same in each unit cell,

$$f(\mathbf{r} + \mathbf{T}(n_1, n_2, \ldots)) = f(\mathbf{r}), \tag{4.5}$$

where $\mathbf{T}$ is any translation defined above. Such a periodic function can be represented by Fourier transforms in terms of Fourier components at wavevectors $\mathbf{q}$ defined in reciprocal space. The formulas can be written most simply in terms of a discrete set of Fourier components if we restrict the Fourier components to those that are periodic in a large volume of crystal Ω_{crystal} composed of $N_{\text{cell}} = N_1 \times N_2 \times \cdots$ cells. Then each component must satisfy the Born–Von Karmen periodic boundary conditions in each of the dimensions

$$\exp(\mathrm{i}\mathbf{q} \cdot N_1\mathbf{a}_1) = \exp(\mathrm{i}\mathbf{q} \cdot N_2\mathbf{a}_2) \ldots = 1, \tag{4.6}$$

so that $\mathbf{q}$ is restricted to the set of vectors satisfying $\mathbf{q} \cdot \mathbf{a}_i = 2\pi \frac{\text{integer}}{N_i}$ for each of the primitive vectors $\mathbf{a}_i$. In the limit of large volumes Ω_{crystal} the final results must be independent of the particular choice of boundary conditions.[5]

The Fourier transform is defined to be

$$f(\mathbf{q}) = \frac{1}{\Omega_{\text{crystal}}} \int_{\Omega_{\text{crystal}}} \mathrm{d}\mathbf{r} f(\mathbf{r}) \exp(\mathrm{i}\mathbf{q} \cdot \mathbf{r}), \tag{4.7}$$

which, for periodic functions, can be written:

$$\begin{aligned} f(\mathbf{q}) &= \frac{1}{\Omega_{\text{crystal}}} \sum_{n_1,n_2,\ldots} \int_{\Omega_{\text{cell}}} \mathrm{d}\mathbf{r}\ f(\mathbf{r}) \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}+\mathbf{T}(n_1,n_2,\ldots))} \\ &= \frac{1}{N_{\text{cell}}} \sum_{n_1,n_2,\ldots} \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{T}(n_1,n_2,\ldots)} \frac{1}{\Omega_{\text{cell}}} \times \int_{\Omega_{\text{cell}}} \mathrm{d}\mathbf{r}\ f(\mathbf{r}) \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{r}}. \end{aligned} \tag{4.8}$$

The sum over all lattice points in the middle line vanishes for all $\mathbf{q}$ except those for which $\mathbf{q} \cdot \mathbf{T}(n_1, n_2, \ldots) = 2\pi \times \textit{integer}$ for *all* translations $\mathbf{T}$. Since $\mathbf{T}(n_1, n_2, \ldots)$ is a sum of integer multiples of the primitive translations $\mathbf{a}_i$, it follows that $\mathbf{q} \cdot \mathbf{a}_i = 2\pi \times \textit{integer}$.

The set of Fourier components $\mathbf{q}$ that satisfy this condition is the "reciprocal lattice." If we define the vectors $\mathbf{b}_i$, $i = 1, d$ that are reciprocal to the primitive translations $\mathbf{a}_i$, i.e.

$$\mathbf{b}_i \cdot \mathbf{a}_j = 2\pi \delta_{ij}, \tag{4.9}$$

the only non-zero Fourier components of $f(\mathbf{r})$ are for $\mathbf{q} = \mathbf{G}$, where the $\mathbf{G}$ vectors are a lattice of points in reciprocal space defined by,

$$\mathbf{G}(m_1, m_2, \ldots) = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + \ldots, \tag{4.10}$$

where the m_i, $i = 1, d$ are integers. For each $\mathbf{G}$, the Fourier transform of the periodic function can be written,

$$f(\mathbf{G}) = \frac{1}{\Omega_{\text{cell}}} \int_{\Omega_{\text{cell}}} \mathrm{d}\mathbf{r}\ f(\mathbf{r}) \exp(\mathrm{i}\mathbf{G} \cdot \mathbf{r}). \tag{4.11}$$

The mutually reciprocal relation of the Bravais lattice in real space and the reciprocal lattice becomes apparent using matrix notation that is valid in any dimension. If we define square matrix $b_{ij} = (\mathbf{b}_i)_j$, exactly as was done for the a_{ij} matrix, then primitive vectors are related by

$$\mathbf{b}^T\mathbf{a} = 2\pi\mathbf{1} \ \rightarrow\ \mathbf{b} = 2\pi(\mathbf{a}^T)^{-1} \text{ or } \mathbf{a} = 2\pi(\mathbf{b}^T)^{-1}. \tag{4.12}$$

It is also straightforward to derive explicit expressions for the relation of the $\mathbf{a}_i$ and $\mathbf{b}_i$

[5] Of course invariance to the choice of boundary conditions in the large-system limit must be proven. For short-range forces and periodic operators the proof is straightforward, but the generalization to Coulomb forces requires care in defining the boundary conditions on the potentials. The calculation of electric polarization is especially problematic and a satisfactory theory has been developed only within the past few years, as is described in Ch. 22.

vectors; for example, in three dimensions, one can show by geometric arguments that

$$\mathbf{b}_1 = 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{|\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)|} \tag{4.13}$$

and cyclical permutations. The geometric construction of the reciprocal lattice in two dimensions is shown in Fig. 4.1.

It is easy to show that the reciprocal of a square (simple cubic) lattice is also a square (simple cubic) lattice, with dimension $\frac{2\pi}{a}$. The reciprocal of the triangular (hexagonal) lattice is also triangular (hexagonal), but rotated with respect to the crystal lattice. The bcc and fcc lattices are reciprocal to one another (Exercise 4.9). The primitive vectors of the reciprocal lattice for each of the three-dimensional lattices in Eq. (4.3) in units of $\frac{2\pi}{a}$ are given by:

$$\begin{array}{lllll} & \text{simple cubic} & \text{simple hex.} & \text{fcc} & \text{bcc} \\ \mathbf{b}_1 = & (1,0,0) & \left(1, -\frac{1}{\sqrt{3}}, 0\right) & (1,1,-1) & (0,1,1), \\ \mathbf{b}_2 = & (0,1,0) & \left(0, \frac{2}{\sqrt{3}}, 0\right) & (1,-1,1) & (1,0,1), \\ \mathbf{b}_3 = & (0,0,1) & \left(0,0,\frac{a}{c}\right) & (-1,1,1) & (1,1,0). \end{array} \tag{4.14}$$

The volume of any primitive cell of the reciprocal lattice can be found from the same reasoning as used for the Bravais in real space. This is the volume of the first Brillouin zone Ω_{BZ} (see Sec. 4.2) which can be written for any dimension d in analogy to Eq. (4.4) as

$$\Omega_{\mathrm{BZ}} = \det(\mathbf{b}) = |\mathbf{b}| = \frac{(2\pi)^d}{\Omega_{\mathrm{cell}}}. \tag{4.15}$$

This shows the mutual reciprocal relation of Ω_{BZ} and Ω_{cell}. The formulas can also be expressed in the geometric forms $\Omega_{\mathrm{BZ}} = |b_1|$ $(d = 1)$; $|\mathbf{b}_1 \times \mathbf{b}_2|$, $(d = 2)$; and $|\mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3)|$, $(d = 3)$.

The Brillouin zone

The first Brillouin zone (which we will denote as simply the "Brillouin zone" or BZ) is the Wigner–Seitz cell of the reciprocal lattice, which is defined by the planes that are the perpendicular bisectors of the vectors from the origin to the reciprocal lattice points. It is on these planes that the Bragg condition is satisfied for elastic scattering [84, 86]. For incident particles with wavevectors inside the BZ there can be no Bragg scattering. Construction of the BZ is illustrated in Figs. 4.1–4.4, and widely used notations for points in the BZ of several crystals are given in Fig. 4.10.

Useful relations

Expressions for crystals often involve the lengths of vectors in real and reciprocal space, $|\tau + \mathbf{T}|$ and $|\mathbf{k} + \mathbf{G}|$ and the scalar products $(\mathbf{k} + \mathbf{G}) \cdot (\tau + \mathbf{T})$. If the vectors are expressed in a cartesian coordinate system, the expressions simply involve sums over each cartesian

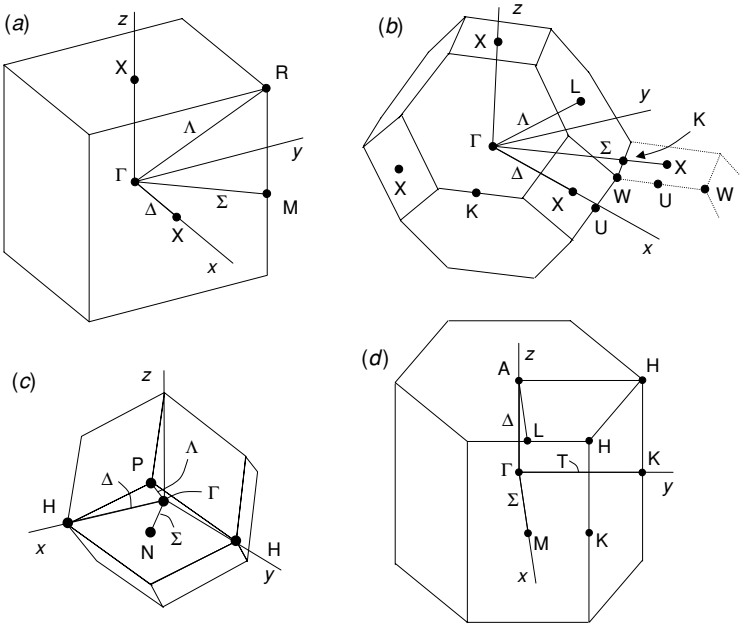


Figure 4.10. Brillouin zones for several common lattices: (*a*) simple cubic (sc), (*b*) face centered cubic (fcc), (*c*) body centered cubic (bcc), and (*d*) hexagonal (hex). High-symmetry points and lines are labeled according to Bouckaret, Smoluchowski, and Wigner; see also Slater [269]. The zone center ($\mathbf{k} = 0$) is designated Γ and interior lines by Greek letters, points on the zone boundary by Roman letters. In the case of the fcc lattice, a portion of a neighboring cell is represented by dotted lines. This shows the orientation of neighboring cells that provides useful information, for example, that the line Σ from Γ to K continues to a point outside the first BZ that is equivalent to X. This line is shown in many figures, such as Figs. 2.24 and 12.2.

component. However, it is often more convenient to represent $\mathbf{T}$ and $\mathbf{G}$ by the integer multiples of the basis vectors, and positions τ and wave vectors $\mathbf{k}$ as fractional multiples of the basis vectors. It is useful to define lengths and scalar products in this representation, i.e. to define the "metric."

The matrix formulation makes it easy to derive the desired expressions. Any position vector τ with elements $\tau_1, \tau_2, \ldots,$ in cartesian coordinates can be written in terms of the primitive vectors by $\tau = \sum_{i=1}^{d} \tau_i^L \mathbf{a}_i$, where the superscript L denotes the representation in lattice vectors and τ^L has elements $\tau_1^L, \tau_2^L, \ldots,$ that are fractions of primitive translation vectors. In matrix form this becomes (here superscript T denotes transpose)

$$\tau = \tau^L \mathbf{a}; \quad \tau^L = \tau \mathbf{a}^{-1} = \frac{1}{2\pi} \tau \mathbf{b}^T, \tag{4.16}$$

where $\mathbf{b}$ is the matrix of primitive vectors of the reciprocal lattice. Similarly, a vector $\mathbf{k}$ in reciprocal space can be expressed as $\mathbf{k} = \sum_{i=1}^{d} k_i^L \mathbf{b}_i$ with the relations

$$\mathbf{k} = \mathbf{k}^L \mathbf{b}; \quad \mathbf{k}^L = \mathbf{k}\mathbf{b}^{-1} = \frac{1}{2\pi} \mathbf{k}\mathbf{a}^T. \tag{4.17}$$

The scalar product $(\mathbf{k}+\mathbf{G})\cdot(\tau+\mathbf{T})$ is easily written in the lattice coordinates, using relation (4.9). If $\mathbf{T}(n_1, n_2, \ldots) = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + \cdots$ and $\mathbf{G}(m_1, m_2, \ldots) = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + \ldots$, then one finds the simple expression

$$(\mathbf{k}+\mathbf{G})\cdot(\tau+\mathbf{T}) = 2\pi\sum_{i=1}^{d}(k_i^L + m_i)(\tau_i^L + n_i) \equiv 2\pi(\mathbf{k}^L + \mathbf{m})\cdot(\tau^L + \mathbf{n}). \quad (4.18)$$

The relation in terms of the cartesian vectors is readily derived using (4.16) and (4.17). On the other hand, the lengths are most easily written in the cartesian system. Using (4.16) and (4.17) and the same vector notation as in (4.18), it is straightforward to show that lengths are given by,

$$|\tau+\mathbf{T}|^2 = (\tau^L + \mathbf{n})\mathbf{a}\mathbf{a}^T(\tau^L + \mathbf{n})^T; \;\; |\mathbf{k}+\mathbf{G}|^2 = (\mathbf{k}^L + \mathbf{m})\mathbf{b}\mathbf{b}^T(\mathbf{k}^L + \mathbf{m})^T, \quad (4.19)$$

i.e. $\mathbf{a}\mathbf{a}^T$ and $\mathbf{b}\mathbf{b}^T$ are the metric tensors for the vectors in real and reciprocal spaces expressed in their natural forms as multiples of the primitive translation vectors.

Finally, one often needs to find all the lattice vectors within some cutoff radius, e.g. in order to find the lowest Fourier components in reciprocal space or the nearest neighbors in real space. Consider the parallelepiped defined by all lattice points in real space $\mathbf{T}(n_1, n_2, n_3); -N_1 \le n_1 \le N_1; -N_2 \le n_2 \le N_2; -N_3 \le n_3 \le N_3$. Since the vectors $\mathbf{a}_2$ and $\mathbf{a}_3$ form a plane, the distance in space perpendicular to this plane is the projection of $\mathbf{T}$ onto the unit vector perpendicular to the plane. This unit vector is $\hat{\mathbf{b}_1} = \mathbf{b}_1/|\mathbf{b}_1|$ and, using (4.19), it is then simple to show that the maximum distance in this direction is $R_{\max} = 2\pi\frac{N_1}{|\mathbf{b}_1|}$. Similar equations hold for the other directions. The result is a simple expression (Exercise 4.15) for the boundaries of the parallelepiped that bounds a sphere of radius $R_{\max}$,

$$N_1 = \frac{|\mathbf{b}_1|}{2\pi}R_{\max}; N_2 = \frac{|\mathbf{b}_2|}{2\pi}R_{\max}; \ldots. \quad (4.20)$$

In reciprocal space the corresponding condition for the parallelepiped that bounds a sphere of radius $G_{\max}$ is,

$$M_1 = \frac{|\mathbf{a}_1|}{2\pi}G_{\max}; M_2 = \frac{|\mathbf{a}_2|}{2\pi}G_{\max}; \ldots, \quad (4.21)$$

where the vectors range from $-M_i\mathbf{b}_i$ to $+M_i\mathbf{b}_i$ in each direction.

4.3 Excitations and the Bloch theorem

The previous sections were devoted to properties of periodic functions in a crystal, such as the nuclear positions and electron density, that obey the relation (4.5), i.e. $f(\mathbf{r}+\mathbf{T}(n_1, n_2, \ldots)) = f(\mathbf{r})$ for any translation of the Bravais lattice $\mathbf{T}(\mathbf{n}) \equiv \mathbf{T}(n_1, n_2, \ldots) = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + \ldots$, as defined in (4.1). Such periodic functions have non-zero Fourier components only for reciprocal space at the reciprocal lattice vectors defined by (4.10).

Excitations of the crystal do not, in general, have the periodicity of the crystal.[6] The subject of this section is the classification of excitations according to their behavior under the translation operations of the crystal. This leads to a Bloch theorem proved, in a general way, and applicable to all types of excitations: electrons, phonons, and other excitations of the crystal.[7] We will give explicit demonstrations for independent-particle excitations; however, since the general relations apply to any system, the theorems can be generalized to correlated many-body systems.

Consider the eigenstates of any operator $\hat{O}$ defined for the periodic crystal. Any such operator must be invariant to any lattice translation $\mathbf{T}(\mathbf{n})$. For example, $\hat{O}$ could be the hamiltonian $\hat{H}$ for the Schrödinger equation for independent particles,

$$\hat{H}\psi(\mathbf{r}) = \left[-\frac{\hbar^2}{2m_e}\nabla^2 + V(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}). \tag{4.22}$$

The operator $\hat{H}$ is invariant to all lattice translations since $V_{\text{eff}}(\mathbf{r})$ has the periodicity of the crystal[8] and the derivative operator is invariant to any translation.

Similarly, we can define translation operators $\hat{T}_{\mathbf{n}}$ that act on any function by displacing the arguments, e.g.

$$\hat{T}_{\mathbf{n}}\psi(\mathbf{r}) = \psi[\mathbf{r} + \mathbf{T}(\mathbf{n})] = \psi(\mathbf{r} + n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + \ldots). \tag{4.23}$$

Since the hamiltonian is invariant to any of the translations $\mathbf{T}(\mathbf{n})$, it follows that the hamiltonian operator commutes with each of the translations operators $\hat{T}_{\mathbf{n}}$,

$$\hat{H}\hat{T}_{\mathbf{n}} = \hat{T}_{\mathbf{n}}\hat{H}. \tag{4.24}$$

From (4.24) it follows that the eigenstates of $\hat{H}$ can be chosen to be eigenstates of *all* $\hat{T}_{\mathbf{n}}$ simultaneously. Unlike the hamiltonian, the eigenstates of the translation operators can be readily determined, independent of any details of the crystal; thus they can be used to "block diagonalize" the hamiltonian, rigorously classifying the states by their eigenvalues of the translation operators, and thus leading to the "Bloch theorem" derived explicitly below.

The key point is that the translation operators form a simple group in which the product of any two translations is a third translation, so that the operators obey the relation,

$$\hat{T}_{\mathbf{n}_1}\hat{T}_{\mathbf{n}_2} = \hat{T}_{\mathbf{n}_1+\mathbf{n}_2}. \tag{4.25}$$

Thus the eigenvalues $t_{\mathbf{n}}$ and eigenstates $\psi(\mathbf{r})$ of the operators $\hat{T}_{\mathbf{n}}$

$$\hat{T}_{\mathbf{n}}\psi(\mathbf{r}) = t_{\mathbf{n}}\psi(\mathbf{r}), \tag{4.26}$$

[6] We take the Born–Von Karmen boundary conditions that the excitations are required to be periodic in the large volume Ω_{crystal} composed of $N_{\text{cell}} = N_1 \times N_2 \times \cdots$ cells, as was described previously in (4.6). See the footnote there regarding the proofs that the results are independent of the choice of boundary conditions in the thermodynamic limit of large size.

[7] The derivation here follows the "first proof" of the Bloch theorem as described by Ashcroft and Mermin [84]. Alternative proofs are given in Chs. 12 and 14.

[8] The logic also holds if the potential is a non-local operator (as in pseudopotentials, with which we will deal later).

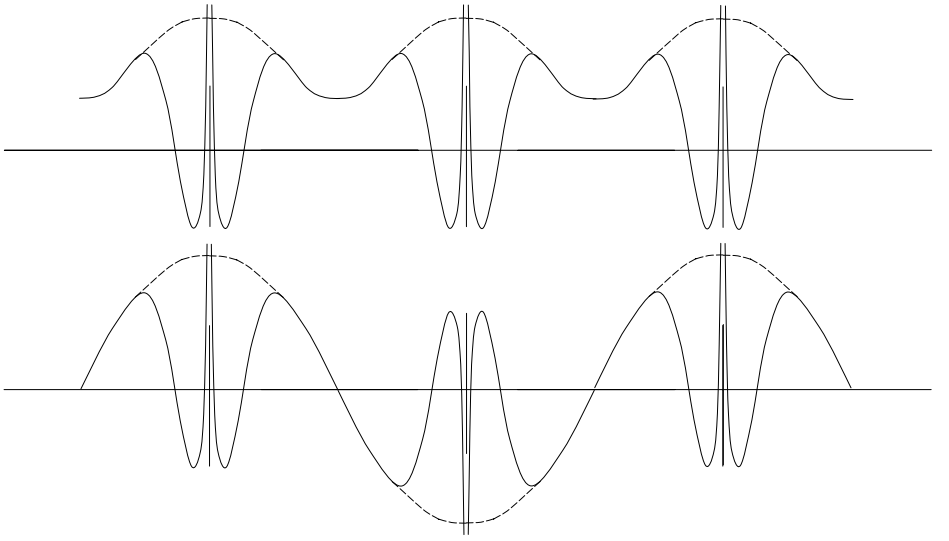


Figure 4.11. Schematic illustration of Bloch states in one dimension at $K = 0$ and at the zone boundary $k = \pi/a$. The envelope is the smooth function that multiplies a periodic array of atomic-like 3s functions, chosen to be the same as in Fig. 11.2.

must obey the relations

$$\hat{T}_{\mathbf{n}_1}\hat{T}_{\mathbf{n}_2}\psi(\mathbf{r}) = t_{(\mathbf{n}_1+\mathbf{n}_2)}\psi(\mathbf{r}) = t_{\mathbf{n}_1}t_{\mathbf{n}_2}\psi(\mathbf{r}). \tag{4.27}$$

By breaking each translation into the product of primitive translations, any $t_{\mathbf{n}}$ can be written in terms of a primitive set $t(\mathbf{a}_i)$

$$t_{\mathbf{n}} = [t(\mathbf{a}_1)]^{n_1}[t(\mathbf{a}_2)]^{n_2}\ldots. \tag{4.28}$$

Since the modulus of each $t(\mathbf{a}_i)$ must be unity (otherwise any function obeying (4.28) is not bounded), it follows that each $t(\mathbf{a}_i)$ can always be written

$$t(\mathbf{a}_i) = \mathrm{e}^{\mathrm{i}2\pi y_i}. \tag{4.29}$$

Since the eigenfunctions must satisfy periodic boundary conditions (4.6), $(t(\mathbf{a}_i))^{N_i} = 1$, so that $y_i = 1/N_i$. Finally, using the definition of the primitive reciprocal lattice vectors in (4.9), Eq. (4.28) can be written

$$t_{\mathbf{n}} = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}_{\mathbf{n}}}, \tag{4.30}$$

where

$$\mathbf{k} = \frac{n_1}{N_1}\mathbf{b}_1 + \frac{n_2}{N_2}\mathbf{b}_2 + \cdots \tag{4.31}$$

is a vector in reciprocal space. The range of $\mathbf{k}$ can be restricted to one primitive cell of the reciprocal lattice since the relation (4.30) is the same in every cell that differs by the addition of a reciprocal lattice vector $\mathbf{G}$ for which $\mathbf{G}\cdot\mathbf{T} = 2\pi \times \textit{integer}$. Note that there are exactly the same number of values of $\mathbf{k}$ as the number of cells.

This leads us directly to the desired results:

1. **The Bloch theorem.**[9] From (4.27), (4.30), and (4.31), one finds

$$\hat{T}_{\mathbf{n}}\psi(\mathbf{r}) = \psi(\mathbf{r}+\mathbf{T}_{\mathbf{n}}) = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}_{\mathbf{n}}}\psi(\mathbf{r}), \tag{4.32}$$

   which is the celebrated "Bloch theorem" that eigenstates of the translation operators vary from one cell to another in the crystal with the phase factor given in (4.32). The eigenstates of any periodic operator, such as the hamiltonian, can be chosen with definite values of $\mathbf{k}$ which can be used to classify any excitation of a periodic crystal. From (4.32) it follows that eigenfunctions with a definite $\mathbf{k}$ can also be written

$$\psi_{\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}), \tag{4.33}$$

   where $u_{\mathbf{k}}(\mathbf{r})$ is periodic ($u_{\mathbf{k}}(\mathbf{r}+\mathbf{T}_{\mathbf{n}}) = u_{\mathbf{k}}(\mathbf{r})$). Examples of the Bloch theorem for independent-particle electron states in many different representations are given in Chs. 12–17.
2. **Bands of eigenvalues.** In the limit of a large (macroscopic) crystal, the spacing of the $\mathbf{k}$ points goes to zero and $\mathbf{k}$ can be considered a continuous variable. The eigenstates of the hamiltonian may be found separately for each $\mathbf{k}$ in one primitive cell of the reciprocal lattice. For each $\mathbf{k}$ there is a discrete set of eigenstates that can be labeled by an index i. This leads to bands of eigenvalues $\varepsilon_{i,\mathbf{k}}$ and energy gaps where there can be no eigenstates for any $\mathbf{k}$.
3. **Conservation of crystal momentum.** It follows from the analysis above that in a perfect crystal the wavevector $\mathbf{k}$ is conserved *modulo any reciprocal lattice vector* $\mathbf{G}$. Thus it is analogous to ordinary momentum in free space, but it has the additional feature that it is only conserved *within one primitive cell, usually chosen to be the Brillouin zone*. Thus two excitations at vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ may have total momentum $\mathbf{k}_1 + \mathbf{k}_2$ outside the Brillouin zone at origin and their true crystal momentum should be reduced to the Brillouin zone around the origin by adding a reciprocal lattice vector. The physical process of scattering of two excitations by some perturbation is called "Umklapp scattering" [84].
4. **The role of the Brillouin zone (BZ).** All possible eigenstates are specified by $\mathbf{k}$ within any primitive cell of the periodic lattice in reciprocal space. However, the BZ is the cell of choice in which to represent excitations; its boundaries are the bisecting planes where Bragg scattering occurs and inside the Brillouin zone there are no such boundaries. Thus bands $\varepsilon_{i,\mathbf{k}}$ are analytic functions of $\mathbf{k}$ inside the BZ and non-analytic dependence upon $\mathbf{k}$ can occur only at the boundaries.

   Examples of Brillouin zones for important cases are shown in Fig. 4.10 with labels for high-symmetry points and lines using the notation of Bouckaret, Smoluchowski, and Wigner (see also Slater [269]). The labels define the directions and points used in many figures given in the present work for electron bands and phonon dispersion curves.
5. **Integrals in k space.** For many properties, such as the counting of electrons in bands, total energies, etc., it is essential to sum over the states labeled by $\mathbf{k}$. The crucial point is

[9] The properties of waves in periodic media were derived earlier by Flouquet in one dimension (see note in [84]) and is often referred to in the physics literature as the "Bloch–Flouquet theorem."

that if one chooses the eigenfunctions that obey periodic boundary conditions in a large crystal of volume Ω_{crystal} composed of $N_{\text{cell}} = N_1 \times N_2 \times \cdots$ cells, as was done in the analysis of (4.6), then there is *exactly one value of* $\mathbf{k}$ *for each cell*. Thus in a sum over states to find an intrinsic property of a crystal expressed as "per unit cell" one simply has a sum over values of $\mathbf{k}$ divided by the number of values N_k. For a general function $f_i(\mathbf{k})$, where i denotes any of the discrete set of states at each $\mathbf{k}$, the average value per cell becomes

$$\bar{f}_i = \frac{1}{N_k} \sum_{\mathbf{k}} f_i(\mathbf{k}). \tag{4.34}$$

If one converts the sum to an integral by taking the limit of a continuous variable in Fourier space with a volume per $\mathbf{k}$ point of Ω_{BZ}/N_k,

$$\bar{f}_i = \frac{1}{\Omega_{\text{BZ}}} \int_{\text{BZ}} \mathrm{d}\mathbf{k}\ f_i(\mathbf{k}) = \frac{\Omega_{\text{cell}}}{(2\pi)^d} \int_{\text{BZ}} \mathrm{d}\mathbf{k}\ f_i(\mathbf{k}), \tag{4.35}$$

where Ω_{cell} is the volume of a primitive cell in real space.

6. **Equation for the periodic part of Bloch functions.** The Bloch functions $\psi_{i,\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} u_{i,\mathbf{k}}(\mathbf{r})$ are eigenfunctions of the real hamiltonian operator $\hat{H}$. By inserting the expression for $\psi_{i,\mathbf{k}}(\mathbf{r})$ in terms of $u_{i,\mathbf{k}}(\mathbf{r})$, the equation becomes

$$\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \hat{H} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} u_{i,\mathbf{k}}(\mathbf{r}) = \varepsilon_{i,\mathbf{k}} u_{i,\mathbf{k}}(\mathbf{r}). \tag{4.36}$$

For the hamiltonian in (4.22), the equation can be written

$$\hat{H}(\mathbf{k}) u_{i,\mathbf{k}}(\mathbf{r}) = \left[-\frac{\hbar^2}{2m_e} (\boldsymbol{\nabla} + i\mathbf{k})^2 + V(\mathbf{r}) \right] u_{i,\mathbf{k}}(\mathbf{r}) = \varepsilon_{i,\mathbf{k}} u_{i,\mathbf{k}}(\mathbf{r}). \tag{4.37}$$

4.4 Time reversal and inversion symmetries

There is an additional symmetry that is present in all systems with no magnetic field. Since the hamiltonian is invariant to time reversal in the original time-dependent Schrödinger equation, it follows that the hamiltonian can always be chosen to be real. In a time-independent equation, such as (12.1), this means that if ψ is an eigenfunction, then ψ^* must also be an eigenfunction with the same real eigenvalue ε. According to the Bloch theorem, the solutions $\psi_{i,-\mathbf{k}}(\mathbf{r})$ can be classified by their wavevector $\mathbf{k}$ and a discrete band index i. If $\psi_{i,-\mathbf{k}}(\mathbf{r})$ satisfies the Bloch condition (4.32), then it follows that $\psi^*_{i,\mathbf{k}}(\mathbf{r})$ satisfies the same equation except with a phase factor corresponding to $-\mathbf{k}$. Thus there is never a need to calculate states at both $\mathbf{k}$ and $-\mathbf{k}$ in any crystal, $\psi_{i,-\mathbf{k}}(\mathbf{r})$ can always be chosen to be $\psi^*_{i,\mathbf{k}}(\mathbf{r})$, and the eigenvalues are equal $\varepsilon_{i,-\mathbf{k}} = \varepsilon_{i,\mathbf{k}}$. If in addition the crystal has inversion symmetry, then (4.37) is invariant under inversion since $V(-\mathbf{r}) = V(\mathbf{r})$ and $(\boldsymbol{\nabla} + i\mathbf{k})^2$ is the same if we replace $\mathbf{k}$ and $\mathbf{r}$ by $-\mathbf{k}$ and $-\mathbf{r}$. Thus the periodic part of the Bloch function can be chosen to satisfy $u_{i,\mathbf{k}}(\mathbf{r}) = u_{i,-\mathbf{k}}(-\mathbf{r}) = u^*_{i,\mathbf{k}}(-\mathbf{r})$.

Spin–orbit interaction

So far we have ignored spin, considering only solutions for a single electron in a non-relativistic hamiltonian. However, relativistic effects introduce a coupling of spin and spatial motion, i.e. the "spin–orbit interaction" given in Sec. 10.4. For the present purposes, the only relevant point is that time reversal leads to reversal of both spin and momentum. Thus the relation of states at $\mathbf{k}$ and $-\mathbf{k}$ is $\psi_{\uparrow,i,\mathbf{k}}(\mathbf{r}) = \psi_{\downarrow,i,-\mathbf{k}}(\mathbf{r})^*$. This is an example of the Kramers theorem which requires that all states must occur in degenerate pairs in any system with time reversal symmetry.

Symmetries in magnetic systems

If time reversal symmetry is broken, the problem is changed significantly. All effects of a magnetic field can be included by modifying the hamiltonian in two ways: $\mathbf{p} \rightarrow (\mathbf{p} - \frac{e}{c}\mathbf{A})$, where $\mathbf{A}$ is the vector potential, and $\hat{H} \rightarrow \hat{H} + \hat{H}_{\text{Zeeman}}$, with $\hat{H}_{\text{Zeeman}} = g\mu\mathbf{H} \cdot \vec{\sigma}$. The latter term is easy to add to an independent-particle calculation in which there is no spin–orbit interaction; there are simply two calculations for different spins. The first term is not hard to include in localized systems like atoms; however, it is exceptionally difficult in extended metallic systems where it leads Landau diamagnetism, very interesting effects in quantum Hall systems, *etc.* We will not treat such effects here.

In ferromagnetic systems there is a spontaneous breaking of time reversal symmetry. The ideas also apply to finite systems with a net spin, e.g. if there is an odd number of electrons. As far as symmetry is concerned, there is no difference from a material in an external magnetic field. However, the effects originate in the Coulomb interactions, which can be included in an independent-particle theory as an effective field (often a very large field). Such Zeeman-like spin-dependent terms are regularly used in independent-particle calculations to study magnetic solids such as spin-density functional theory. In Hartree–Fock calculations on finite systems, exchange induces such terms automatically; however, effects of correlation are omitted.

Antiferromagnetic solids are ones in which there is long-range order involving both space and time reversal symmetries, e.g. a Neel state is invariant to a combination of translation and time reversal. States with such a symmetry can be described in an independent-particle approach by an effective potential with this symmetry breaking form. The broken symmetry leads to a larger unit cell in real space and a translation (or "folding") of the excitations into a smaller Brillouin zone compared with the non-magnetic system. There is no corresponding exact symmetry in finite systems, only a tendency toward antiferromagnetic correlations. Antiferromagnetic solids are one of the outstanding classes of condensed matter in which many-body effects may play a crucial role. Mott insulators tend to be antiferromagnets, and metals with antiferromagnetic correlations often have large enhanced response functions. This has been brought out in Chapter 2 on qualitative description of electrons in solids, and the difficulties of the many-body problem have led to the great debate about such systems over the years.

4.5 Point symmetries

This section is a brief summary needed for group theory applications. Discussion of group theory and symmetries in different crystal classes are covered in a number of texts and monographs. For example, Ashcroft and Mermin [84] give an overview of symmetries with pictorial representation; Slater [269] gives detailed analyses for many crystals with group tables and symmetry labels; and there are many useful books on group theory [270–272]. Computer codes that automatically generate and/or apply the group operations can be found on-line, with links given in Ch. 24.

The total space group of a crystal is composed of the translation group and the point group. Point symmetries are rotations, inversions, reflections, and their combinations that leave the system invariant. In addition, there can be non-symmorphic operations that are combinations with translations or "glides" of fractions of a crystal translation vector. The set of all such operations, $\{R_n, n = 1, \ldots, N_{\text{group}}\}$ forms a group. The operation on any function $g(\mathbf{r})$ of the full symmetry system (such as the density $n(\mathbf{r})$ or the total energy E_{total}) is

$$R_n g(\mathbf{r}) = g(R_n \mathbf{r} + \mathbf{t}_n), \tag{4.38}$$

where $R_n\mathbf{r}$ denotes the rotation, inversions, or reflections of the position $\mathbf{r}$ and $\mathbf{t}_n$ is the non-symmorphic translation associated with operation n.

The two most important consequences of the symmetry operations for excitations can be demonstrated by applying the symmetry operations to the Schrödinger equation (4.22), with i replaced by the quantum numbers for a crystal, $i \to i, \mathbf{k}$. Since the hamiltonian is invariant under any symmetry operation R_n, the operation of R_n leads to a new equation with $\mathbf{r} \to R_i\mathbf{r} + \mathbf{t}_i$ and $\mathbf{k} \to R_i\mathbf{k}$ (the fractional translation has no effect on reciprocal space). It follows that the new function,

$$\psi_i^{R_i\mathbf{k}}(R_i\mathbf{r} + \mathbf{t}_i) = \psi_i^{\mathbf{k}}(\mathbf{r}); \text{ or } \psi_i^{R_i^{-1}\mathbf{k}}(\mathbf{r}) = \psi_i^{\mathbf{k}}(R_i\mathbf{r} + \mathbf{t}_i), \tag{4.39}$$

must also be an eigenfunction of the hamiltonian with the same eigenvalue $\varepsilon_i^{\mathbf{k}}$. This leads to two consequences:

- At "high symmetry" $\mathbf{k}$ points, $R_i^{-1}\mathbf{k} \equiv \mathbf{k}$, so that (4.39) leads to relations among the eigenvectors at that $\mathbf{k}$ point, i.e. they can be classified according to the group representations. For example, at $\mathbf{k} = 0$, in cubic crystals all states have degeneracy 1, 2, or 3.
- One can define the "irreducible Brillouin zone" (IBZ), which is the smallest fraction of the BZ that is sufficient to determine all the information on the excitations of the crystal. The excitations at all other $\mathbf{k}$ points outside the IBZ are related by the symmetry operations. If a group operation $R_i^{-1}\mathbf{k}$ leads to a distinguishable $\mathbf{k}$ point, then (4.39) shows that the states at $R_i^{-1}\mathbf{k}$ can be generated from those at $\mathbf{k}$ by the relations given in (4.39), apart from a phase factor that has no consequence, and the fact that the eigenvalues must be equal,

$$\varepsilon_i^{R_i^{-1}\mathbf{k}} = \varepsilon_i^{\mathbf{k}}.$$

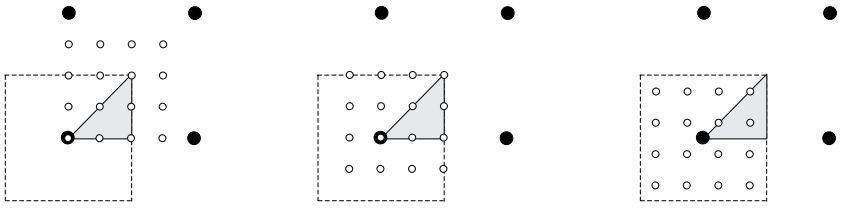


Figure 4.12. Grids for integration for a 2d square lattice, each with four times the density of the reciprocal lattice in each dimension. The left and center figures are equivalent with one point at the origin, and six inequivalent points in the irreducible BZ shown in grey. Right: A shifted special point grid of the same density but with only three inequivalent points. Additional possibilities have been given by Moreno and Soler [277], who also pointed out that different shifts and symmetrization can lead to finer grids.

In all crystals, the BZ can always be reduced by at least a factor of 2 using relation of states at $\mathbf{k}$ and $-\mathbf{k}$; in a square lattice, the IBZ is $1/8$ the BZ, as illustrated in Fig. 4.12; in the highest symmetry crystals (cubic), the IBZ is only $1/48$ the BZ.

4.6 Integration over the Brillouin zone and special points

Evaluation of many quantities, such as energy and density, require integration over the BZ. There are two separate aspects of this problem:

- Accurate integration with a discrete set of points in the BZ. This is specific to the given problem and depends upon having sufficient points in regions where the integrand varies rapidly. In this respect, the key division is between metals and insulators. Insulators have filled bands that can be integrated using only a few well-chosen points such as the "special points" discussed below. On the other hand, metals require careful integration near the Fermi surface for the those bands that cross the Fermi energy where the Fermi factor varies rapidly.
- Symmetry can be used to reduce the calculations since all independent information can be found from states with $\mathbf{k}$ in the IBZ. This is useful in all cases with high symmetry, whether metals or insulators.

Special points

The "special" property of insulators is that the integrals needed all have the form of (4.34) where *the sum is over filled bands in the full BZ.* Since the integrand $f_i(\mathbf{k})$ is some function of the eigenfunctions $\psi_{i,\mathbf{k}}$ and eigenvalues $\varepsilon_{i,\mathbf{k}}$, it is a smoothly varying,[10] periodic function of $\mathbf{k}$. Thus $f_i(\mathbf{k})$ can be expanded in Fourier components,

$$f_i(\mathbf{k}) = \sum_{\mathbf{T}} f_i(\mathbf{T}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}}, \tag{4.40}$$

[10] For an individual band the variation is not smooth at crossings with other bands; however, the relevant sums over all filled bands are smooth so long as all bands concerned as filled. This is always the case if the filled and empty bands are separated by a gap as in an insulator.

where $\mathbf{T}$ are the translation vectors of the crystal. The most important point is that the contribution of the rapidly varying terms at large $\mathbf{T}$ decreases exponentially, so that the sum in (4.40) can be truncated to a finite sum. The proof [273] is related to transformations of the expressions to traces over Wannier functions (see Ch. 21) and the observation that the range of $f_i(\mathbf{T})$ is determined by the range of the Wannier functions.

Special points are chosen for efficient integration of smooth periodic functions.[11] The single most special point is the Baldereschi point [275], where the integration reduces to a single point. The choice is based upon: (1) the fact that there is always some one "mean-value point" where the integrand equals the integral, and (2) use of crystal symmetry to find such a point approximately. The coordinates of the mean-value point for cubic lattices were found to be [275]: simple cubic, $k = (\pi/a)(1/2, 1/2, 1/2)$; body centered cubic, $k = (2\pi/a)(1/6, 1/6, 1/2)$; and face centered cubic, $k = (2\pi/a)(0.6223, 0.2953, 1/2)$. Chadi and Cohen [276] have generalized this idea and have given equations for "best" larger sets of points.

The general method proposed by Monkhorst and Pack [273] is now the most widely used method because it leads to a uniform set of points determined by a simple formula valid for any crystal (given here explicitly for three dimensions):

$$\mathbf{k}_{n_1,n_2,n_3} \equiv \sum_i^3 \frac{2n_i - N_i - 1}{2N_i}\mathbf{G}_i, \tag{4.41}$$

where $\mathbf{G}_i$ are the primitive vectors of the reciprocal lattice. The main features of the Monkhorst–Pack points are:

- A sum over the uniform set of points in (4.41), with $n_i = 1, 2, \ldots, N_i$, *exactly integrates a periodic function that has Fourier components that extend only to* $N_i\mathbf{T}_i$ *in each direction.* (See Exercise 4.21. In fact, (4.41) makes a *maximum error* for higher Fourier components.)
- The set of points defined by (4.41) is a uniform grid in $\mathbf{k}$ that is a scaled version of the reciprocal lattice and offset from $\mathbf{k} = 0$. For many lattices, especially cubic, it is preferable to choose N_i to be even [273]. Then the set does *not involve the highest symmetry points*; it omits the $\mathbf{k} = 0$ point and points on the BZ boundary.
- The $N_i = 2$ set is the Baldereschi point for a simple cubic crystal (taking into account symmetry – see below). The sets for all cubic lattices are also the same as the offset Gilat–Raubenheimer mesh (see [278]).
- An informative tabulation of grids and their efficiency, together with an illuminating description is given by Moreno and Soler [277], who emphasized the generation of different sets of regular grids using a combination of offsets and symmetry.

The logic behind the Monkhorst–Pack choice of points can be understood in one dimension, where it is easy to see that the exact value of the integral,

$$I_1 = \int_0^{2\pi} \mathrm{d}k \ \sin(k) = 0, \tag{4.42}$$

[11] In this sense, the method is analogous to Gauss–Chebyshev integration. (See [274], who found that Gauss–Chebyshev can be more efficient than the Monkhorst–Pack method for large sets of points.)

is given by the value of the integrand $f_1(k) = \sin(k)$ at the mid-point, $k = \pi$ where $\sin(k) = 0$. If one has a sum of two sin functions, $f_2(k) = A_1 \sin(k) + A_2 \sin(2k)$, then the exact value of the integral is given by a sum over two points

$$I_2 = \int_0^{2\pi} \mathrm{d}k \; f_2(k) = 0 = f_2(k = \pi/2) + f_2(k = 3\pi/2). \tag{4.43}$$

The advantage of the special point grids that do not contain the $\mathbf{k} = 0$ point is much greater in higher dimensions. As illustrated in Fig. 4.12 for a square lattice, an integration with a grid $4 \times 4 = 16$ times as dense as the reciprocal lattice can be done with only three inequivalent $\mathbf{k}$ points in the irreducible BZ (defined in the following subsection). This set is sufficient to integrate exactly any periodic function with Fourier components up to $\mathbf{T} = (4, 4) \times a$, where a is the square edge. The advantages are greater in higher dimensions.

Irreducible BZ

Integrals over the BZ can be replaced by integrals only over the IBZ. For example, the sums needed in the total energy (general expressions in Sec. 9.2 or specific ones for crystals, such as (13.1)) have the form of (4.34). Since the summand is a scalar, it must be invariant under each operation, $f_i(R_n\mathbf{k}) = f_i(\mathbf{k})$. It is convenient to define $w_\mathbf{k}$ to be the total number of *distinguishable* $\mathbf{k}$ points related by symmetry to the given $\mathbf{k}$ point in the IBZ (including the point in the IBZ) divided by the total number of points N_k. (Note that points on the BZ boundary related by $\mathbf{G}$ vectors are not distinguishable.) Then the sum (4.34) is equivalent to

$$\bar{f}_i = \sum_{\mathbf{k}}^{\mathrm{IBZ}} w_\mathbf{k} f_i(\mathbf{k}). \tag{4.44}$$

Quantities such as the density can always be written as

$$n(\mathbf{r}) = \frac{1}{N_k} \sum_{\mathbf{k}} n_\mathbf{k}(\mathbf{r}) = \frac{1}{N_{\mathrm{group}}} \sum_{R_n} \sum_{\mathbf{k}}^{\mathrm{IBZ}} w_\mathbf{k} n_\mathbf{k}(R_n\mathbf{r} + \mathbf{t}_n). \tag{4.45}$$

Here points are weighted according to $w_\mathbf{k}$, just as in (4.44), and in addition the variable $\mathbf{r}$ is transformed in each term $n_\mathbf{k}(\mathbf{r})$. Corresponding expressions for Fourier components are given in Sec. 12.7.

Symmetry operations can be used to reduce the calculations greatly. Excellent examples are the Monkhorst–Pack meshes applied to cubic crystals, where there are 48 symmetry operations so that the IBZ is 1/48 the total BZ. The set defined by $N_i = 2$ has $2^3 = 8$ points in the BZ, which reduces to 1 point in the IBZ. Similarly, $N_i = 4 \to 4^3 = 64$ points in the BZ reduces to 2 points; $N_i = 6 \to 6^3 = 216$ points in the BZ reduces to 10 points. As an example, for fcc the 2-point set is $(2\pi/a)(1/4, 1/4, 1/4)$ and $(2\pi/a)(1/4, 1/4, 3/4)$, which has been found to yield remarkably accurate results for energies of semiconductors, a fact that was very important in early calculations [143]. The 10-point set is sufficient for almost all modern calculations for such materials.

Interpolation methods

Metals present an important general class of issues for efficient sampling of the desired states in the BZ. The Fermi surface plays a special role in all properties and the integration over states must take into account the sharp variation of the Fermi function from unity to zero as a function of **k**. This plays a decisive role in all calculations of sums over occupied states for total quantities (e.g. the total electron density, energy, force, and stress in Ch. 9) and sums over both occupied and empty states for response functions and spectral functions (Ch. 19 and App. D).

In order to represent the Fermi surface, the tetrahedron method [279–282] is widely used. If the eigenvalues and vectors are known at a set of grid points, the variation between the grid points can always be approximated by an interpolation scheme using tetrahedra. This is particularly useful because tetrahedra can be used to fill all space for any grid. A simple case is illustrated on the left-hand side of Fig. 4.13, and the same construction can be used for any, e.g. an irregular grid that has more points near the Fermi surface and fewer points far from the Fermi surface where accuracy is not needed. The simplest procedure is a linear interpolation between the values known at the vertices, but higher order schemes can also be used for special grids. Tetrahedron methods are very important in calculations on transition metals, rare earths, etc., where there are exquisite details of the Fermi surfaces that must be resolved.

One example is the method proposed by Blöchl [282] in which there is a grid of **k** points and tetrahedra that reduces to a special-points method for insulators. It also provides an interpolation formula that goes beyond the linear approximation of matrix elements within the tetrahedra, which can improve the results for metals. The use of a regular grid is helpful since the irreducible **k** points and tetrahedra can be selected by an automated procedure. An example of results for Cu metal is shown on the right-hand side of Fig. 4.13. Since the

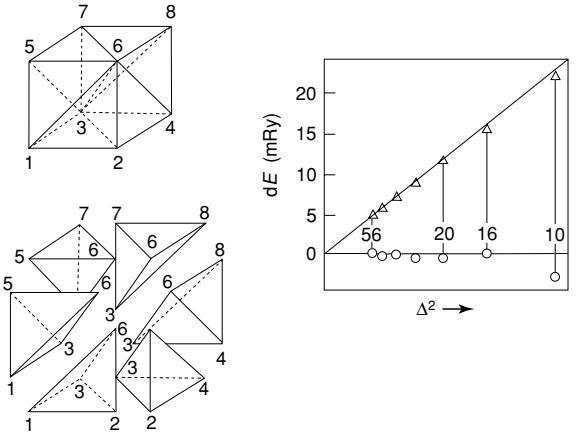


Figure 4.13. Example of generation of tetrahedra that fill the space between the grid points (left) and the results (right) of total energy calculations for Cu as a function of grid spacing Δ, comparing the linear method and the method of [282]. From [282].

Fermi surface of Cu is rather simple, the improvement over simple linear interpolation may be surprising; it is due largely to the fact that the curvature of the occupied band crossing the Fermi energy (see Fig. 2.24) is everywhere positive so that linear interpolation always leads to a systematic error.

4.7 Density of states

An important quantity for many purposes is the density of states (DOS) per unit energy E (and per unit volume Ω in extended matter),

$$\rho(E) = \frac{1}{N_k} \sum_{i,\mathbf{k}} \delta(\varepsilon_{i,\mathbf{k}} - E) = \frac{\Omega_{\text{cell}}}{(2\pi)^d} \int_{\text{BZ}} d\mathbf{k} \;\; \delta(\varepsilon_{i,\mathbf{k}} - E). \tag{4.46}$$

In the case of independent-particle states, where $\varepsilon_{i,\mathbf{k}}$ denotes the energy of an electron (or phonon), (4.46) is the number of independent-particle states per unit energy. Quantities like the specific heat involve excitations of electrons that do not change the number, i.e. an excitation from a filled to an empty state. Similarly, for independent-particle susceptibilities, such as general forms of χ^0 in App. D and the dielectric function given in (20.2), the imaginary part is given by matrix elements times a joint DOS, i.e. a double sum over bands i and j but a single sum over $\mathbf{k}$ due to momentum conservation, as a function of the energy difference $E = \varepsilon_j - \varepsilon_i$.

It is straightforward to show that the DOS has "critical points," or van Hove singularities [283], where $\rho(E)$ has analytic forms that depend only upon the space dimension. In three dimensions, each band must have square root singularities at the maxima and minima and at saddle points in the bands. A simple example is illustrated later in Fig. 14.3 for a tight-binding model in one, two, and three dimensions. Examples of single-particle electron DOS are given in Figs. 2.31 and 16.13, for optical spectra in Figs. 2.27 and 2.28, and for phonons in Figs. 2.9 and 2.32. Interestingly, the ideal of critical points can be applied to any function of a periodic variable. For example, Fig. 13.3 shows the distribution of local values of the density parameter r_s in crystalline Si [127].

SELECT FURTHER READING

Detailed analyses for many crystals with group tables and symmetry labels given in:

Ashcroft, N. and Mermin, N. *Solid State Physics*, (W. B. Saunders Company, New York, 1976).

Slater, J. C. *Symmetry and Energy Bonds in Crystals* (Collected and reprinted version of 1965 *Quantum Theory of Molecules and Solids*, Vol. 2) (Dover, New York, 1972).

Many useful books on group theory:

Heine, V. *Group Theory*, Pergamon Press, New York, 1960.

Tinkham, M. *Group Theory and Quantum Mechanics*, McGraw-HIll, New York, 1964.

Lax, M. J. *Symmetry Principles in Solid State and Molecular Physics*, John Wiley and Sons, New York, 1974.

Exercises

4.1 Derive the expression for primitive reciprocal lattice in three dimensions given in (4.13).

4.2 For a two-dimensional lattice give an expression for primitive reciprocal lattice vectors that is equivalent to the one for three dimensions given in (4.13).

4.3 Show that for the two-dimensional triangular lattice the reciprocal lattice is also triangular and is rotated by 90°.

4.4 Show that the volume of the primitive cell in any dimension is given by (4.4).

4.5 Find the Wigner–Seitz cell for the two-dimensional triangular lattice. Does it have the symmetry of a triangle or of a hexagon. Support your answer in terms of the symmetry of the triangular lattice.

4.6 Draw the Wigner–Seitz cell and the first Brillouin zone for the two-dimensional triangular lattice.

4.7 Consider a honeycomb plane of graphite in which each atom has three nearest neighbors. Give primitive translation vectors, basis vectors for the atoms in the unit cell, and reciprocal lattice primitive vectors. Show that the BZ is hexagonal.

4.8 Covalent crystals tend to form structures in which the bonds are *not* at 180°. Show that this means that the structures will have more than one atom per primitive cell.

4.9 Show that the fcc and bcc lattices are reciprocal to one another. Do this in two ways: by drawing the vectors and taking cross products and by explicit inversion of the lattice vector matrices.

4.10 Consider a body centered cubic crystal, like Na, composed of an element with one atom at each lattice site. What is the Bravais lattice in terms of the conventional cube edge a? How many nearest neighbors does each atom have? How many second neighbors? Now suppose that the crystal is changed to a diatomic crystal like CsCl with all the nearest neighbors of a Cs atom being Cl, and vice versa. Now what is the Bravais lattice in terms of the conventional cube edge a? What is the basis?

4.11 Derive the value of the ideal c/a ratio for packing of hard spheres in the hcp structure.

4.12 Derive the formulas given in (4.12), paying careful attention to the definitions of the matrices and the places where the transpose is required.

4.13 Derive the formulas given in (4.18).

4.14 Derive the formulas given in (4.19).

4.15 Derive the relations given in (4.20) and (4.21) for the parallelepiped that bounds a sphere in real and in reciprocal space. Explain the reason why the dimensions of the parallelepiped in reciprocal space involve the primitive vectors for the real lattice and vice versa.

4.16 Determine the coordinates of the points on the boundary of the Brillouin zone for fcc (X, W, K, U) and bcc (H, N, P) lattices.

4.17 Derive the formulas given in (4.20) and (4.21). Hint: Use the relations of real and reciprocal space given in the sentences before these equations.

4.18 Show that the expressions for integrals over the Brillouin zone (4.35), applied to the case of free electrons, lead to the same relations between density of one spin state n^σ and the Fermi momentum k_F^σ that was found in the section on homogeneous gas in (5.5). (From this one relation follow the other relations given after (5.5).)

4.19 In one dimension, dispersion can have singularities only at the end points where $E(k) - E_0 = A(k - k_0)^2$, with A positive or negative. Show that the singularities in the DOS form have the form $\rho(E) \propto |E - E_0|^{-1/2}$, as illustrated in the left panel of Fig. 14.3.

4.20 Show that singularities like those in Fig. 14.3 occur in three dimensions, using (4.46) and the fact that $E \propto Ak_x^2 + Bk_y^2 + Ck_z^2$ with A, B, C all positive (negative) at minima (maxima) or with different signs at saddle points.

4.21 The "special points" defined by Monkhorst and Pack are chosen to integrate periodic functions efficiently with rapidly decreasing magnitude of the Fourier components. This is a set of exercises to illustrate this property:
(a) Show that in one dimension the average of $f(k)$ at the k points $\frac{1}{4}\frac{\pi}{a}$ and $\frac{3}{4}\frac{\pi}{a}$ is exact if f is a sum of Fourier components $k + n\frac{2\pi}{a}$, with $n = 0, 1, 2, 3$, but that the error is maximum for $n = 4$.
(b) Derive the general form of (4.41).
(c) Why are uniform sets of points more efficient if they do it not include the Γ point?
(d) Derive the 2- and 10-point sets given for an fcc lattice, where symmetry has been used to reduce the points to the irreducible BZ.

4.22 The bands of any one-dimensional crystal are solutions of the Schrödinger equation (4.22) with a periodic potential $V(x + a) = V(x)$. The complete solution can be reduced to an informative analytic expression in terms of the scattering properties of a single unit cell and the Bloch theorem. This exercise follows the illuminating discussion by Ashcroft and Mermin [84], Problem 8.1, and it lays a foundation for exercises that illustrate the pseudopotential concept (Exercises 11.2, 11.6, and 11.14) and the relation to plane wave, APW, KKR, and MTO methods, respectively, in Exercises 12.6, 16.1, 16.7, and 16.13.)

An elegant approach is to consider a different problem first: an infinite line with $\tilde{V}(x) = 0$ except for a single cell in which the potential is the same as in a cell of the crystal, $\tilde{V}(x) = V(x)$ for $-a/2 < x < a/2$. At any positive energy $\varepsilon \equiv (\hbar^2/2m_e)K^2$, there are two solutions: $\psi_l(x)$ and $\psi_r(x)$ corresponding to waves incident from the left and from the right. Outside the cell, $\psi_l(x)$ is a given by $\psi_l(x) = \mathrm{e}^{iKx} + r\mathrm{e}^{-iKx}$, $x < -\frac{a}{2}$, and $\psi_l(x) = t\mathrm{e}^{iKx}$, $x > a/2$, where t and r are transmission and reflection amplitudes. There is a corresponding expression for $\psi_r(x)$. Inside the cell, the functions can be found by integration of the equation, but we can proceed without specifying the explicit solution.
(a) The transmission coefficient can be written as $t = |t|\mathrm{e}^{i\delta}$, with δ a phase shift which is related the phase shifts defined in App. J as clarified in Exercise 11.2. It is well known from scattering theory that $|t|^2 + |r|^2 = 1$ and $r = \pm i|r|\mathrm{e}^{i\delta}$, which are left as an exercise to derive.
(b) A solution $\psi(x)$ in the crystal at energy ε (if it exists) can be expressed as a linear combination of $\psi_l(x)$ and $\psi_r(x)$ evaluated at the same energy. Within the central cell all functions satisfy the same equation and $\psi(x)$ can always be written as a linear combination,

$$\psi(x) = A\psi_l(x) + B\psi_r(x), \quad -\frac{a}{2} < x < \frac{a}{2}, \tag{4.47}$$

with A and B chosen so that $\psi(x)$ satisfies the Bloch theorem for *some crystal momentum* k. Since $\psi(x)$ and $\mathrm{d}\psi(x)/\mathrm{d}x$ must be continuous, it follows that $\psi(\frac{a}{2}) = \mathrm{e}^{ika}\psi(-\frac{a}{2})$ and $\psi'(\frac{a}{2}) = \mathrm{e}^{ika}\psi'(-\frac{a}{2})$. Using this information and the forms of $\psi_l(x)$ and $\psi_r(x)$, find the 2×2 secular equation and show that the solution is given by

$$2t\cos(ka) = \mathrm{e}^{-iKa} + (t^2 - r^2)\mathrm{e}^{iKa}. \tag{4.48}$$

Verify that this is the correct solution for free electrons, $V(x) = 0$.
(c) Show that in terms of the phase shift, the solution, (4.48), can be written

$$|t|\cos(ka) = \cos(Ka + \delta), \quad \varepsilon \equiv \frac{\hbar^2}{2m_e}K^2. \tag{4.49}$$

(d) Analyse (4.49) to illustrate properties of bands and indicate which are special features of one dimension. (i) Since since $|t|$ and δ are functions of energy ε, it is most convenient to fix ε and use (4.49) to find the wavevector k; this exemplifies the "root tracing" method used in augmented methods (Ch. 16). (ii) There are necessarily band gaps where there are no solutions, except for the free electron case. (iii) There is exactly one band of allowed states $\varepsilon(k)$ between each gap. (iv) The density of states, (4.46), has the form shown in the left panel in Fig. 14.3.
(e) Finally, discuss the problems with extending this approach to higher dimensions.

5

Uniform electron gas and simple metals

Summary

The simplest model system representing condensed matter is the homogeneous electron gas, in which the nuclei are replaced by a uniform positively charged background. This system is completely specified by the density n (or r_s, which is the average distance between electrons) and the spin density $n_\uparrow - n_\downarrow$ or the polarization $\zeta = (n_\uparrow - n_\downarrow)/n$. The homogeneous gas illustrates the problems associated with interacting electrons in condensed matter and is a prelude to the electronic structure of matter, which is governed by the combined effects of nuclei and electron interaction.

The homogeneous electron gas is the simplest system for illustrating key properties of interacting electrons and characteristic magnitudes of electronic energies in condensed matter. Since all independent-particle terms can be calculated analytically, this is an ideal model system for understanding the effects of correlation. In particular, the homogeneous gas best illustrates the issues of Fermi liquid theory [225, 226], which is the basis for our understanding of the "normal" (non-superconducting) state of real metals in terms of effective independent-particle approaches.

A homogeneous system is completely specified by its density $n = N_e/\Omega$, which can be characterized by the parameter r_s, defined as the radius of a sphere containing one electron on average,

$$\frac{4\pi}{3} r_s^3 = \Omega/N_e = \frac{1}{n}; \quad \text{or} \quad r_s = \left(\frac{3}{4\pi n}\right)^{1/3}. \tag{5.1}$$

Thus r_s is a measure of the average distance between electrons. Table 5.1 gives values of r_s for valence electrons in a number of elements. The values shown are typical of characteristic electron densities in solids. For simple crystals, r_s is readily derived from the structure and lattice constant; expressions for fcc and bcc, and the VI, III–V, and II–VI semiconductors are given in Exercises 5.1 and 5.2.

Of course, density is not constant in a real solid and it is interesting to determine the variation in density. For example, Fig. 13.3 shows the distribution of local values of the density parameter r_s for valence electrons in Si [127]. In ordinary diamond-structure Si,

Table 5.1. Typical r_s values in elemental solids in units of the Bohr radius a_0. The valence is indicated by Z. The alkalis have bcc structure; Al, Cu, and Pb are fcc; the other group IV elements have diamond structure; and other elements have various structures. The values for metals are taken from [86] and [88]; precise values depend upon temperature

$Z=1$	$Z=2$	$Z=1$	$Z=2$	$Z=3$	$Z=4$
Li 3.23	Be 1.88			B	C 1.31
Na 3.93	Mg 2.65			Al 2.07	Si 2.00
K 4.86	Ca 3.27	Cu 2.67	Zn 2.31	Ga 2.19	Ge 2.08
Rb 5.20	Sr 3.56	Ag 3.02	Cd 2.59	In 2.41	Sn 2.39
Cs 5.63	Ba 3.69	Au 3.01	Hg 2.15	Tl	Pb 2.30

there is a significant volume with low density (the open parts of the diamond structure). However, in the compressed metallic phase of Si with Sn structure, the variation in r_s is only $\pm \approx 20\%$.

The hamiltonian for the homogeneous system is derived by replacing the nuclei in (3.1) with a uniform positively charged background, which leads to

$$\begin{aligned}\hat{H} &= -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \frac{1}{2}\frac{4\pi}{\epsilon_0}\left[\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i-\mathbf{r}_j|} - \int \mathrm{d}^3r\,\mathrm{d}^3r'\frac{(ne)^2}{|\mathbf{r}-\mathbf{r}'|}\right] \\ &\to -\frac{1}{2}\sum_i \nabla_i^2 + \frac{1}{2}\left[\sum_{i\neq j}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} - \int \mathrm{d}^3r\,\mathrm{d}^3r'\frac{n^2}{|\mathbf{r}-\mathbf{r}'|}\right], \end{aligned} \tag{5.2}$$

where the second expression is in Hartree atomic units $\hbar = m_e = e = 4\pi/\epsilon_0 = 1$, where lengths are given in units of the Bohr radius a_0. The last term is the average background term which must be included to cancel the divergence due to Coulomb interaction among the electrons. The total energy is given by

$$E = \langle \hat{H} \rangle = \langle \hat{T} \rangle + \langle \hat{V}_{\mathrm{int}} \rangle - \frac{1}{2}\int \mathrm{d}^3r\,\mathrm{d}^3r'\frac{n^2}{|\mathbf{r}-\mathbf{r}'|}, \tag{5.3}$$

where the first term is the kinetic energy of interacting electrons and the last two terms are the *difference* between the potential energy of the actual interacting electrons and the self-interaction of a classical uniform negative charge density, i.e. the exchange–correlation energy.[1] Note that the *difference* is well defined, since there is a cancellation of the divergent Coulomb interactions, as discussed following (3.16).

In order to understand the interacting gas as a function of density, it is useful to express the hamiltonian (5.2) in terms of scaled coordinates $\tilde{\mathbf{r}} = \mathbf{r}/r_s$ instead of atomic units ($\mathbf{r}$ in units of a_0) assumed in the second expression in (5.2). Then (5.2) becomes (see Exercise 5.3

[1] This can be derived from expression (3.16) for the energy, since in this case the total charge density (electrons + background) is everywhere zero, so that the final term in (3.16) vanishes.

for the last term that is essential for the expression to be well defined)

$$\hat{H} = \left(\frac{a_0}{r_s}\right)^2 \sum_i \left[-\frac{1}{2}\tilde{\nabla}_i^2 + \frac{1}{2}\frac{r_s}{a_0}\left(\sum_{j\neq i} \frac{1}{|\tilde{\mathbf{r}}_i - \tilde{\mathbf{r}}_j|} - \frac{3}{4\pi}\int \mathrm{d}^3\tilde{\mathbf{r}}\frac{1}{|\tilde{\mathbf{r}}|}\right)\right], \tag{5.4}$$

where energies are in atomic units. This expression shows explicitly that one can view the system in terms of a scaled unit of energy (the Hartree scaled by $(a_0/r_s)^2$) and a scaled effective interaction proportional to r_s/a_0. In other words, the properties as a function of density r_s/a_0 are completely equivalent to a system at fixed density but with scaled electron–electron interaction $e^2 \to (r_s/a_0)e^2$ at fixed density and a scaled unit of energy.

5.1 Non-interacting and Hartree–Fock approximations

In the non-interacting approximation, the solutions of (3.36) are eigenstates of the kinetic energy operator, i.e. normalized plane waves $\psi_{\mathbf{k}} = (1/\Omega^{1/2})\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$ with energy $\varepsilon_{\mathbf{k}} = \frac{\hbar^2}{2m_e}k^2$. The ground state for a given density of up and down spin electrons is the determinant function (3.43) formed from the single-electron states with wavevectors inside the Fermi surface, which is a sphere in reciprocal space of radius k_F^σ, the Fermi wavevector for each spin σ. The value of k_F^σ is readily derived, since each allowed k state in a crystal of volume Ω is associated with a volume in reciprocal space $(2\pi)^3/\Omega$ (see Exercise 5.4 and Ch. 4.) Each state can contain one electron of each spin so that

$$\frac{4\pi}{3}\left(k_F^\sigma\right)^3 = \frac{(2\pi)^3}{\Omega}N_e^\sigma; \quad \text{i.e.} \quad \left(k_F^\sigma\right)^3 = 6\pi^2 n^\sigma \quad \text{or} \quad k_F^\sigma = (6\pi^2)^{1/3}(n^\sigma)^{1/3}. \tag{5.5}$$

If the system is unpolarized, i.e. $n^\uparrow = n^\downarrow = n/2$, then $k_F = k_F^\uparrow = k_F^\downarrow$, where

$$(k_F)^3 = 3\pi^2 n; \quad \text{or} \quad k_F = (3\pi^2)^{1/3}n^{1/3} = \left(\frac{9}{4}\pi\right)^{1/3}/r_s. \tag{5.6}$$

The expression for the Fermi wavevector has the remarkable property that it also applies to interacting electron systems: the Luttinger theorem [285, 286] guarantees that the Fermi surface exists at the same k_F^σ as in the non-interacting case, so long as there is no phase transition.

In the independent-particle approximation, Fermi energy E_{F0}^σ for each spin is given by

$$E_{F0}^\sigma = \frac{\hbar^2}{2m_e}\left(k_F^\sigma\right)^2 = \frac{1}{2}\left(k_F^\sigma a_0\right)^2 \to \frac{1}{2}\left(k_F^\sigma\right)^2, \tag{5.7}$$

where the last expression is in atomic units with $a_0 = 1$. Useful relations for the Fermi wavevector and various energies are given in Tabs. 5.2 and 5.3.

The total kinetic energy per electron of a given spin in the ground state is given by integrating over the filled states

$$T_0^\sigma = \frac{\hbar^2}{2m_e}\frac{4\pi\int_0^{k_F^\sigma}\mathrm{d}k k^4}{4\pi\int_0^{k_F^\sigma}\mathrm{d}k k^2} = \frac{3}{5}E_{F0}^\sigma, \tag{5.8}$$

Table 5.2. Characteristic energies for each spin σ for the homogeneous electron gas in the Hartree–Fock approximation: the Fermi energy E^σ_{F0}; kinetic energy T^σ_0 and Hartree–Fock exchange energy per electron E^σ_x which is negative; and the increase in band width in the Hartree–Fock approximation ΔW_{HFA}.

Quantity	Expression	Atomic units
E^σ_{F0}	$\frac{\hbar^2}{2m_e}(k^\sigma_F)^2$	$\frac{1}{2}(k^\sigma_F)^2$
T^σ_0	$\frac{3}{5}E^\sigma_F$	$\frac{3}{5}E^\sigma_F$
$-E^\sigma_x$	$\frac{3e^2}{4\pi}k^\sigma_F$	$\frac{3}{4\pi}k^\sigma_F$
$\Delta W^\sigma_{\mathrm{HFA}}$	$\frac{e^2}{\pi}k^\sigma_F$	$\frac{1}{\pi}k^\sigma_F$

Table 5.3. Useful expressions for the unpolarized homogeneous electron gas in terms of r_s in units of the Bohr radius a_0. See caption of Tab. 5.2 for definitions of energies

Quantity	Expression	Atomic units	Common units
k_F	$(\frac{9}{4}\pi)^{1/3}/r_s$	$1.919{,}158/r_s$	$3.626{,}470/r_s$ (Å^{-1})
E_{F0}	$\frac{1}{2}(\frac{9}{4}\pi)^{2/3}/r_s^2$	$1.841{,}584/r_s^2$	$50.112{,}45/r_s^2$ (eV)
T_0	$\frac{3}{5}E_F$	$1.104{,}961/r_s^2$	$30.067{,}47/r_s^2$ (eV)
$-E_x$	$\frac{3}{4\pi}(\frac{9\pi}{4})^{1/3}/r_s$	$0.458{,}165{,}29/r_s$	$12.467{,}311/r_s$ (eV)
ΔW_{HFA}	$(\frac{9}{4\pi^2})^{1/3}/r_s$	$0.145{,}838{,}54/r_s$	$3.968{,}4684/r_s$ (eV)

(see Exercise 5.7 for one and two dimensions.) Since the energy is positive, the homogeneous gas is clearly unbound in this approximation. The true binding in a material is provided by the added attraction to point nuclei and the attractive exchange and correlation energies.

Density matrix

The density matrix in the homogeneous gas illustrates both the general expressions and the nature of the spatial dependence in many-electron systems. The general expression for independent fermions (3.41) simplifies for the homogeneous gas (for each spin) to

$$\rho(\mathbf{r},\mathbf{r}') = \rho(|\mathbf{r}-\mathbf{r}'|) = \frac{1}{(2\pi)^3}\int \mathrm{d}\mathbf{k}\; f(\varepsilon(k))\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')}, \tag{5.9}$$

where $\varepsilon(k) = k^2/2$, which is just a Fourier transform of the Fermi function $f(\varepsilon(k))$. To evaluate the function it is convenient to transform the expression using a partial integration [287], yielding

$$\rho(r) = \frac{\beta}{(2\pi)^2}\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\int_{-\infty}^{\infty}\mathrm{d}k\;\cos(kr)f'\left(\beta\left(\frac{1}{2}k^2-\mu\right)\right). \tag{5.10}$$

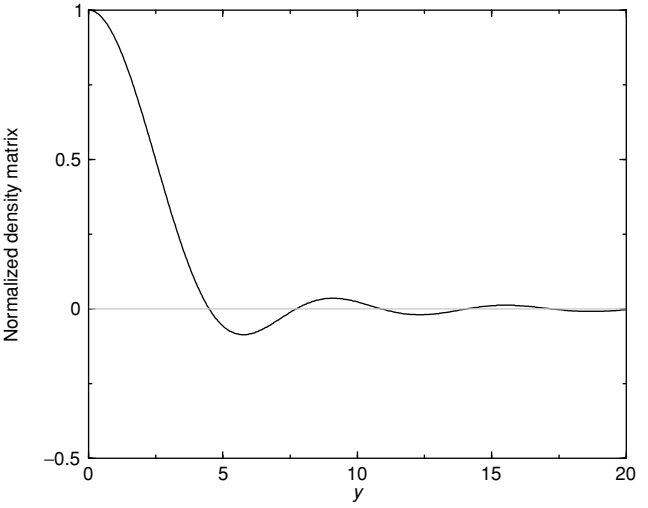


Figure 5.1. The dimensionless density matrix at $T = 0$ in the non-interacting homogeneous gas (the term in square brackets in (5.11) as a function of $y = k_F r$). The oscillations have spatial form governed by the Fermi wavevector k_F and describe charge around an impurity (Friedel oscillations) or magnetic interactions in a metal (Ruderman–Kittel–Kasuya–Yosida oscillations) [84, 86, 246]. See also Fig. 5.3 which shows the consequences for the pair correlation function.

This is a particularly revealing form that makes it clear why long-range oscillations in $r = |\mathbf{r} - \mathbf{r}'|$ must result from sharp variation in the derivative of the Fermi function $f'(\varepsilon)$, long known in Fourier transforms and attributed to Gibbs [288]. Since $f'(\varepsilon)$ approaches a delta function at low temperature, the range of $\rho(r)$ must increase as the temperature is reduced. At $T = 0$, $\rho(r)$ decays as $1/r^2$ [246],

$$\rho(r) = \frac{k_F^3}{3\pi^2}\left[3\frac{\sin(y) - y\cos(y)}{y^3}\right], \tag{5.11}$$

with $y = k_F r$. The function in square brackets is defined to be normalized (Exercise 5.6) and is plotted in Fig. 5.1, where the decaying oscillatory form is evident, often called Friedel oscillations for charge and Ruderman–Kittel–Kasuya–Yosida oscillations for magnetic interactions [84, 86, 246]. Numerical results [287, 289] and simple analytic approximations [263, 287] can be found for $T \neq 0$ which show an exponential decay constant $\propto k_B T/k_F$.

Hartree–Fock approximation

In the Hartree–Fock approximation, the one-electron orbitals are eigenstates of the non-local operator in (3.45). The solution of the Hartree–Fock equations in this case can be done analytically: the first step is to show that the eigenstates are plane waves, just as for non-interacting electrons (see Exercise 5.8). Thus the kinetic energy and the density matrix are the same as for non-interacting electrons, as they must be since the Hartree–Fock wavefunction contains no correlation beyond that required by the exclusion principle. The

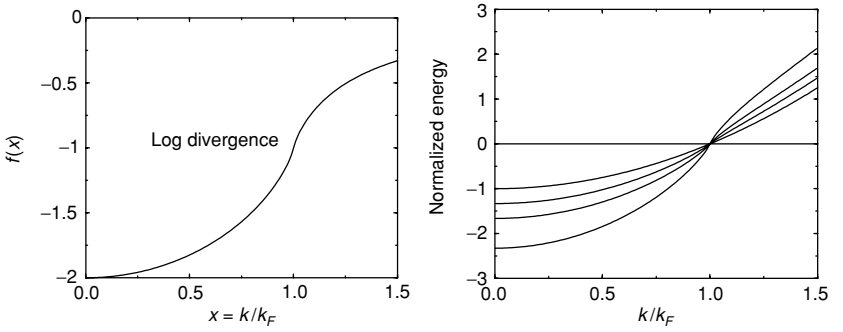


Figure 5.2. Left: The factor $f(x)$, (5.13), in the homogeneous gas that determines the dispersion $\varepsilon_{\rm HFA}(k)$ in the Hartree–Fock approximation. Right: $\varepsilon_{\rm HFA}(k)$ for three densities ($r_s = 1, 2, 4$) compared to the non-interacting case. The lowest density (largest r_s) is lowest at $k = 0$ and has the most visible singularity at the Fermi surface, $x = 1$. The normalized dimensionless eigenvalue is defined by the square brackets in (5.14), and $r_s = 0$ is the non-interacting limit $-1 + x^2$.

next step is to derive the eigenvalue for each k, which is $k^2/2$ plus the matrix element of the exchange operator (3.48). The integrals can be done analytically (the steps are outlined in Exercise 5.9 following [84, 225]), leading to

$$\varepsilon_k = \frac{1}{2}k^2 + \frac{k_F}{\pi} f(x), \tag{5.12}$$

where $x = k/k_F$ and

$$f(x) = -\left(1 + \frac{1-x^2}{2x} \ln|\frac{1+x}{1-x}|\right). \tag{5.13}$$

(Note that the expression applies to each spin separately.)

The factor $f(x)$, shown in Fig. 5.2, is negative for all x; at the bottom of the band ($x = 0$), $f(0) = -2$, and at large x it approaches zero. Near the Fermi surface ($x = 1$), $f(x)$ varies rapidly and has a divergent slope; nevertheless the limiting value at $x = 1$ is well defined, $f(x \to 1) = -1$. Thus in the Hartree–Fock approximation, exchange increases the band width W by $\Delta W = k_F/\pi$. This holds separately for each spin, and in the unpolarized case, the factor can also be written $\Delta W = \left(\frac{9}{4\pi^2}\right)^{1/3}/r_s$ (see Table 5.3 and Exercise 5.10).

The Hartree–Fock eigenvalue relative to the Fermi energy, i.e. defined with $\varepsilon_k \equiv 0$ at $k = k_F$, can be written in scaled form,

$$\varepsilon_k = \frac{1}{2}k_F^2 \left\{(x^2 - 1) + \frac{2}{\pi k_F}[f(x) + 1]\right\}. \tag{5.14}$$

The expression in curly brackets is plotted on the right-hand side of Fig. 5.2 for several values of r_s. The broadening of the filled band due to interactions in the Hartree–Fock approximation is indicated by the value at $k = 0$, which is -1 for non-interacting electrons.

The singularity at the Fermi surface, first pointed out by Bardeen, [290], is a consequence of long-range Coulomb interaction and the existence of the Fermi surface where the separation of the occupied and empty states vanishes. The velocity at the Fermi surface $\mathrm{d}\varepsilon/\mathrm{d}k$

diverges (Exercise 5.11), in blatant contradiction with experiment, where the well-defined velocities are determined by such measurements as specific heat and the de Haas–van Alfen effect [84, 86]. Thus this is an intrinsic failure of Hartree–Fock that carries over to any metal. The Hartree–Fock divergence, however, can be avoided either if there is a finite gap (i.e. in an insulator where Hartree–Fock is qualitatively correct and is widely applied in quantum chemistry) or if the Coulomb interaction is screened to be effectively short range. This is the *ansatz*, i.e. the approach, of Fermi liquid theory: that the interactions are screened for low-energy excitations, leading to weakly interacting "quasiparticles," which is commonly justified by partial summation of diagrams in the random phase approximation (RPA) [225, 226].

The exchange energy and exchange hole

The exact total energy of the homogeneous gas is given by (5.3), which can be separated into the Hartree–Fock total energy, which is the sum of kinetic energy of independent electrons plus the exchange energy, and the remainder, termed the "correlation energy." As we have seen in Sec. 3.6, the exchange energy and exchange hole can be computed directly from the wavefunctions, which can be done analytically in this case. In addition, the exchange energy per electron is simply the average of the exchange contribution to the eigenvalue $\frac{k_F}{\pi} f(x)$ in (5.12) multiplied by $1/2$ to take into account the fact that interactions should not be double-counted. Using the fact that the average value of $f(x)$ is $-3/2$ (Exercise 5.12), it follows that the exchange energy per electron is

$$\epsilon_x^\sigma = E_x^\sigma / N^\sigma = -\frac{3}{4\pi} k_F^\sigma = -\frac{3}{4} \left(\frac{6}{\pi} n^\sigma \right)^{1/3}. \tag{5.15}$$

In the unpolarized case, one finds $\epsilon_x \equiv \epsilon_x^\uparrow = \epsilon_x^\downarrow = -\frac{3}{4\pi}(\frac{9\pi}{4})^{1/3}/r_s$ and the explicit numerical relations in Table 5.3.

For partially polarized cases, the exchange energy is just a sum of terms for the two spins, which can also be expressed in an alternative form in terms of the total density $n = n^\uparrow + n^\downarrow$ and fractional polarization,

$$\zeta = \frac{n^\uparrow - n^\downarrow}{n}. \tag{5.16}$$

It is straightforward to show that exchange in a polarized system has the form

$$\epsilon_x(n, \zeta) = \epsilon_x(n, 0) + [\epsilon_x(n, 1) - \epsilon_x(n, 0)] f_x(\zeta), \tag{5.17}$$

where

$$f_x(\zeta) = \frac{1}{2} \frac{(1+\zeta)^{4/3} + (1-\zeta)^{4/3} - 2}{2^{1/3} - 1}, \tag{5.18}$$

which is readily derived [291] from (5.15).

The exchange hole g_x defined in (3.54) or (3.52) involves only electrons of the same spin and in a homogeneous system is a function only of the relative distance $|\mathbf{r}| = |\mathbf{r}_1 - \mathbf{r}_2|$, so that $g_x(\mathbf{r}_1, \sigma_1; \mathbf{r}_2, \sigma_2) = \delta_{\sigma_1,\sigma_2} g_x^{\sigma_1,\sigma_1}(|\mathbf{r}|)$. In the homogeneous gas, the form of the exchange hole can be calculated analytically in two ways (see Exercise 5.13): the definitions can

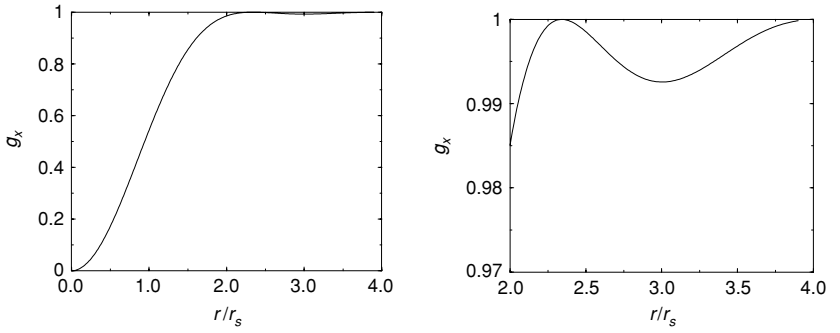


Figure 5.3. Exchange hole $g_x(r)$ in the homogeneous electron gas, (5.19) plotted as a function of r/r_s, where r_s is the average distance between electrons in an unpolarized system. The magnitude decreases rapidly with oscillation, as shown in the greatly expanded right-hand figure. Note the similarity to the calculated pair correlation function for parallel spins in Fig. 5.5.

be used directly [96] by inserting the plane wave eigenfunctions (normalized to a large volume Ω) and evaluating the resulting expression. Alternatively, $g_x(r)$ can be found from the general relation (3.56) of the pair correlation function and the density matrix in a non-interacting system[2] together with the density matrix $\rho(r)$ given by (5.11). For each spin, the hole can be given in terms of the dimensionless variable $y = k_F^\sigma r$ with the result

$$g_x^{\sigma,\sigma}(y) = 1 - \left[3\frac{\sin(y) - y\ \cos(y)}{y^3}\right]^2, \tag{5.19}$$

which is shown graphically in Fig. 5.3. The exchange hole in the homogeneous gas illustrates the principle that for fermions the hole n^x must always be negative, i.e. $g_x^{\sigma,\sigma}$ must be less than 1, and it approaches 1 as an inverse power law with the well-known Friedel oscillations due to the sharp Fermi surface.

5.2 The correlation hole and energy

"Screening" is the effect in a many-body system whereby the particles collectively correlate to reduce the net interaction among any two particles. For repulsive interactions, the hole (reduced probability of finding other particles) around each particle tends to produce a net weaker interaction strength.

Thomas–Fermi screening

The grandfather of models for screening is the Thomas–Fermi approximation for the electron gas, which is the quantum equivalent of Debye screening in a classical system. The screening is determined by analyzing the response of the gas to a static external charge density with Fourier component k. The response at wavevector k is determined by the

[2] The arguments can be applied to any non-interacting particles [263]; for bosons the result is that $g_x(r) = 1 + |\rho(r)|^2/n^2$ is always greater than 1. See Sec. 3.6 and Exercise 3.15.

change in energy of the electrons, which is a function of only density in the Thomas–Fermi approximation (Sec. 6.1). The result is that the long-range Coulomb interaction is screened to an exponentially decaying interaction, which in Fourier space can be written as

$$\frac{1}{k^2} \to \frac{1}{k^2 + k_{\rm TF}^2}, \tag{5.20}$$

where $k_{\rm TF}$ is the Thomas–Fermi screening wavevector (the inverse of the characteristic screening length). For an unpolarized system, $k_{\rm TF}$ is given by (see Exercise 5.14 and [84])

$$k_{\rm TF} = r_s^{1/2} \left(\frac{16}{3\pi^2}\right)^{1/3} k_F = \left(\frac{12}{\pi}\right)^{1/3} r_s^{-1/2}, \tag{5.21}$$

where r_s is in atomic units, i.e. in units of the Bohr radius a_0.

This is the simplest estimate for the characteristic length over which electrons are correlated, which is very useful in understanding the full results below in a homogeneous gas and estimates for real systems.

Correlation energy

It is not possible to determine the correlation hole and energy analytically. The first quantitative form for the correlation energy of a homogeneous gas was proposed in the 1930s by Wigner [70, 292], as an interpolation between low- and high-density limits.[3] At low density the electrons form a "Wigner crystal" and the correlation energy is just the electrostatic energy of point charges on the body centered cubic lattice. At the time, it was thought that the exchange energy per electron approached a constant in the high-density limit, and Wigner proposed the simple interpolation

$$\epsilon_c = -\frac{0.44}{r_s + 7.8} (\text{in a.u. } = \text{ Hartree}). \tag{5.22}$$

Correct treatment of correlation confounded many-body theory for decades until the work of Gellmann and Breuckner [293], who summed infinite series of diagrams to eliminate divergences that are present at each order and calculated the correlation energy exactly in the high-density limit, $r_s \to 0$. For an unpolarized gas ($n^\uparrow = n^\downarrow = n/2$), the result is [293, 294]

$$\epsilon_c(r_s) \to 0.311 \ \ln(r_s) - 0.048 + r_s(A \ \ln(r_s) + C) + \ldots, \tag{5.23}$$

where the ln terms are the signature of non-analyticity that causes so much difficulty. At low density the system can be considered a Wigner crystal with zero point motion leading to [226, 295]

$$\epsilon_c(r_s) \to \frac{a_1}{r_s} + \frac{a_2}{r_s^{3/2}} + \frac{a_3}{r_s^2} + \ldots, \tag{5.24}$$

[3] The first formula proposed by Wigner [70] was in error due to an incorrect expression for the low-density limit, as point out in [292].

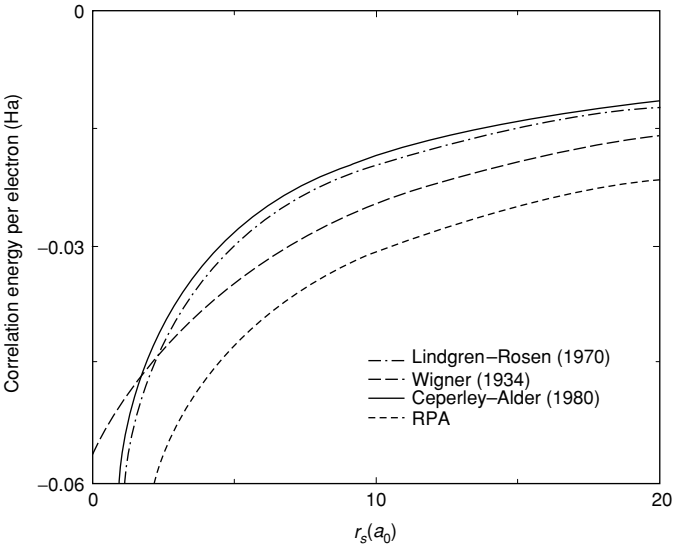


Figure 5.4. Correlation energy of an unpolarized homogeneous electron gas as a function of the density parameter r_s. The most accurate results available are quantum Monte Carlo calculations; the curve labeled "Ceperley–Alder" is the work of those authors [297] fitted to the interpolation formula of Vosko, Wilk, and Nusair (VWN) [301]; the Perdew–Zunger (PZ) fit [300] is almost identical on this scale. In comparison are shown the Wigner interpolation formula, (5.22), the RPA (see text), and an improved many-body perturbation calculation taken from Mahan [96], where it is attributed to L. Lindgren and A. Rosen. Figure provided by H. Kim.

There has been considerable work in the intervening years [96], including the well-known work of Hedin and Lundqvist [220] using the random phase approximation (RPA), which is the basis of much of our present understanding of excitations, and other recent work such as self-consistent "GW" calculations [296]. The most accurate results for ground state properties are found from quantum Monte Carlo (QMC) calculations that can treat interacting many-body systems [297–299], which are the benchmark for other methods. The QMC results for the correlation energy $\epsilon_c(r_s)$ per electron in an unpolarized gas are shown in Fig. 5.4 where they are compared with the Wigner interpolation formula, RPA, and improved many-body calculations of Lindgren and Rosen (results given in [96], p. 314). One very important result is that for materials at typical solid densities ($r_s \approx 2 - 6$) the correlation energy is much smaller than the exchange energy; however, at very low densities (large r_s) correlation becomes more important and dominates in the regime of the Wigner crystal ($r_s > \approx 80$).

The use of the QMC results in subsequent electronic structure calculations relies upon parameterized analytic forms for $E_c(r_s)$ fitted to the QMC energies calculated at many values of r_s, mainly for unpolarized and fully polarized ($n^\uparrow = n$) cases, although some calculations have been done at intermediate polarization [298]. The key point is that the formulas fit the data well at typical densities and extrapolate to the high- and low-density limits, (5.23) and (5.24). Widely used forms are due to Perdew and Zunger (PZ) [300] and

Vosko, Wilkes, and Nussair (VWN) [301], which are given in App. B and are included in subroutines for functionals referred to there and available on-line.

The simplest form for the correlation energy as a function of spin polarization is the one made by PZ [300] that correlation varies the same as exchange

$$\epsilon_c(n,\zeta) = \epsilon_c(n,0) + [\epsilon_c(n,1) - \epsilon_c(n,0)] f_x(\zeta), \tag{5.25}$$

where $f_x(\zeta)$ is given by (5.18). The slightly more complex form of VWN [301] has been found to be a slightly better fit to more recent QMC data [298].

It is also important for understanding the meaning of both exchange and correlation energies to see how they originate from the interaction of an electron with the exchange–correlation "hole" discussed in Sec. 3.6. The potential energy of interaction of each electron with its hole can be written

$$\epsilon_{\rm xc}^{\rm pot}(r_s) = E_{\rm xc}^{\rm pot}/N = \frac{1}{N}\left[\langle \hat{V}_{\rm int}\rangle - E_{\rm Hartree}(n)\right] = \frac{1}{2n}e^2 \int {\rm d}^3 r \frac{n_{\rm xc}(|\mathbf{r}|)}{|\mathbf{r}|}, \tag{5.26}$$

where the factor $1/2$ is included to avoid double-counting and we have explicitly indicated interaction strength e^2, which will be useful later. The exchange–correlation hole $n_{\rm xc}(|\mathbf{r}|)$, of course, is spherically symmetric and is a function of density, i.e. of r_s. In the ground state, $\epsilon_{\rm xc}^{\rm pot}$ is negative since exchange lowers the energy if interactions are repulsive and correlation always lowers the energy. However, this is not the total exchange–correlation energy per electron $\epsilon_{\rm xc}$, because the kinetic energy increases as the electrons correlate to lower their potential energy.

The full exchange–correlation energy including kinetic terms can be found in two ways: kinetic energy can be determined from the virial theorem [303] or from the "coupling constant integration formula" described in Ch. 3. We will consider the latter as an example of the generalized force theorem, i.e. the coupling constant integration formula (3.23) in which the coupling constant e^2 is replaced by λe^2, which is varied from $\lambda = 0$ (the non-interacting problem) to the actual value $\lambda = 1$. Just as in (3.23), the derivative of the energy with respect to λ involves only the *explicit* linear variation of $\epsilon_{\rm xc}^{\rm pot}(r_s)$ in (5.26) with λ and there is no contribution from the implicit dependence of $n_{\rm xc}^{\lambda}(|\mathbf{r}|)$ upon λ, since the energy is at a minimum with respect to such variations. This leads directly to the result that

$$\epsilon_{\rm xc}(r_s) = \frac{1}{2n}e^2 \int {\rm d}^3 r \frac{n_{\rm xc}^{\rm av}(r)}{r}, \tag{5.27}$$

where $n_{\rm xc}^{\rm av}(r)$ is the coupling-constant-averaged hole

$$n_{\rm xc}^{\rm av}(r) = \int_0^1 {\rm d}\lambda n_{\rm xc}^{\lambda}(r). \tag{5.28}$$

The exchange–correlation hole has been calculated by quantum Monte Carlo methods at full coupling strength $\lambda = 1$, with results that are shown in Fig. 5.5 for various densities labeled r_s. By comparison with the exchange hole shown in Fig. 5.3, it is apparent that correlation is much more important for antiparallel spins than for parallel spins, which are kept apart by the Pauli principle. In general, correlation tends to reduce the long-range part of the exchange hole, i.e. it tends to cause screening.

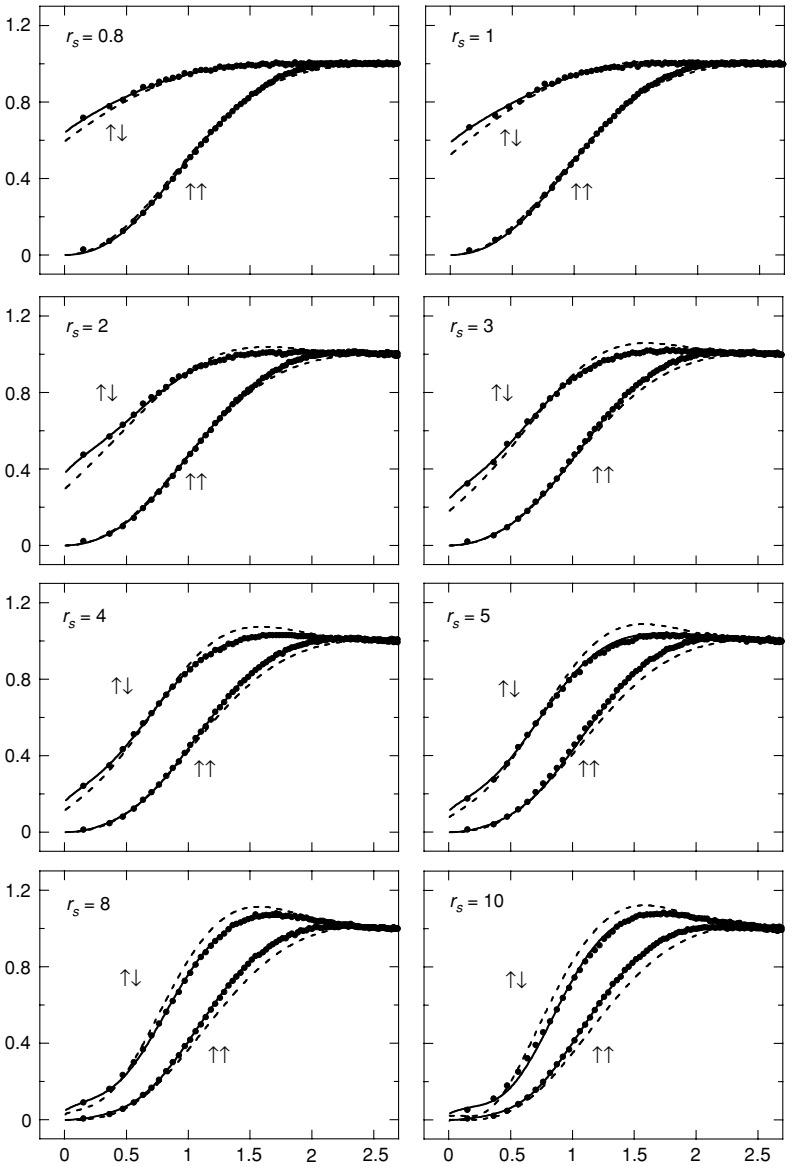


Figure 5.5. Spin-resolved normalized pair-correlation function $g_{\mathrm{xc}}(r)$ for the unpolarized homogenous electron gas as a function of scaled separation r/r_s, for r_s varying from $r_s = 0.8$ to $r_s = 10$. Dots, QMC data of [302]; dashed line, Perdew–Wang model; solid line, coupling constant integrated form of [303]. From [303].

Variation of the exchange–correlation hole with r_s can also be understood as variation with the strength of the interaction. As pointed out in (5.4), variation of e^2 from 0 to 1 at fixed density is equivalent to variation of r_s from 0 to the actual value. Working in scaled units r/r_s and $r_s \to \lambda r_s$, one finds

$$n_{\rm xc}^{\rm av}\left(\frac{r}{r_s}\right) = \int_0^1 {\rm d}\lambda n_{\rm xc}^{\lambda}\left(\frac{r}{\lambda r_s}\right). \tag{5.29}$$

Examples of the variation of the hole $n_{\rm xc}(\frac{r}{\lambda r_s})$ are shown in Fig. 5.5 for various r_s for parallel and opposite spins in an unpolarized gas. Explicit evaluation of $\epsilon_{\rm xc}(r_s)$ has been done using this approach in [303]. Note that this expression involves $\lambda r_s < r_s$ in the integrand, i.e. the hole for a system with density higher than the actual density. Exercises 5.15 and 5.16 deal with this relation, explicit shapes of the average holes for materials, and the possibility of making a relation that involves larger r_s (stronger coupling).

5.3 Binding in sp-bonded metals

The stage was set for understanding solids on a quantitative basis by Slater [304] and by Wigner and Seitz [49,50] in the early 1930s. The simplest metals, the alkalis with one weakly bound electron per atom, are represented remarkably well by the energy of a homogeneous electron gas plus the attractive interaction with the positive cores. It was recognized that the ions were effectively weak scatterers even though the actual wavefunctions must have atomic-like radial structure near the ion. This is the precursor of the pseudopotential idea (Ch. 11) and also follows from the scattering analysis of Slater's APW method and the KKR approach (Ch. 16). Treating the electrons as a homogeneous gas, and adding the energies of the ions in the uniform background, leads to the expression for total energy per electron,

$$\frac{E_{\rm total}}{N} = \frac{1.105}{r_s^2} - \frac{0.458}{r_s} + \epsilon_c - \frac{1}{2}\frac{\alpha}{r_s} + \epsilon_R, \tag{5.30}$$

where atomic units are assumed (r_s in units of a_0), and we have used the expressions in Tab. 5.3 for kinetic and exchange energies, and ϵ_c is the correlation energy per electron. The last two terms represent interaction of a uniform electron density with the ions: α is the Madelung constant for point charges in a background, and the final term is a repulsive correction due to the fact that the ion is not a point. Values of α are tabulated in Tab. F.1 for representative structures. The factor ϵ_R is due to core repulsion, which can be estimated using the effective model potentials in Fig. 11.3 that are designed to take this effect into account. This amounts to removing the attraction of the nucleus and the background in a core radius R_c around the ion

$$\epsilon_R = n2\pi \int_0^{R_c} {\rm d}r r^2 \frac{e^2}{r} = \frac{3}{4\pi r_s^3} 2\pi e^2 R_c^2 = \frac{3}{2}\frac{a_0 R_c^2}{r_s^3} = \frac{3}{2}\frac{R_c^2}{r_s^3}, \tag{5.31}$$

where the last form is in atomic units.

Expression (5.30) contains much of the essential physics for the sp-bonded metals, as discussed in basic texts on solid state physics [84,86,88]. For example, the equilibrium value

of r_s predicted by (5.30) is given by finding the extremum of (5.30). A good approximation is to neglect ϵ_c and to take $\alpha = 1.80$, the value for the Wigner–Seitz sphere that is very close to actual values in close-packed metals as shown in Tab. F.1 and (F.9). This leads to

$$\frac{r_s}{a_0} = 0.814 + \sqrt{0.899 + 3.31\left(\frac{R_c}{a_0}\right)^2}, \tag{5.32}$$

and improved expressions described in Exercise 5.17. Without the repulsive term, this leads to $r_s = 1.76$, which is much too small. However, a core radius $\approx 2a_0$ (e.g. a typical R_c in the model ion potentials shown in Fig. 11.3 and references given there) leads to a very reasonable $r_s \approx 4a_0$. The kinetic energy contribution to the bulk modulus is

$$B = \Omega\frac{\mathrm{d}^2E}{\mathrm{d}\Omega^2} = \frac{3}{4\pi r_s}\frac{1}{9}\frac{\mathrm{d}^2}{\mathrm{d}r_s^2}\frac{1.105}{r_s^2} = \frac{0.176}{r_s^5} = \frac{51.7}{r_s^5}\ \text{Mbar}, \tag{5.33}$$

where a Mbar (=100 GPa, see Tab. O) is a convenient unit. This sets a scale for understanding the bulk modulus in real materials, giving the right order of magnitude (often better) for materials ranging from sp-bonded metals to strongly bonded covalent solids.

5.4 Excitations and the Lindhard dielectric function

Excitations of a homogeneous gas can be classified into two types (see Sec. 2.10): electron addition or removal to create quasiparticles, and collective excitations in which the number of electrons does not change. The former are the bands for quasiparticles in Fermi liquid theory. How well do the non-interacting or Hartree–Fock bands shown in Fig. 5.2 agree with improved calculations and experiment? Figure 5.6 shows photoemission data for Na, which is near the homogeneous gas limit, compared to the non-interacting dispersion $k^2/2$. Interestingly, the bands are *narrower* than $k^2/2$, i.e. the opposite of what is predicted by Hartree–Fock theory. This is a field of active research in many-body perturbation theory to describe the excitations [82]. For our purposes, the important conclusion is that *the non-interacting case is a good starting point close to the measured dispersion*. This is germane to electronic structure of real solids, where it is found that Kohn–Sham eigenvalues are a reasonable starting point for describing excitations (see Sec. 7.4).

Excitations that do not change the particle number are charge density fluctuations (plasma oscillations) described by the dielectric function, and spin fluctuations that are described by spin response functions. Expressions for the response functions are given in Chs. 19 and 20 and in Apps. D and E. The point of this section is to apply the expressions to a homogeneous system where the integrals can be done analytically. The discussion here follows Pines [225], Secs. 3–5, and provides examples that help to understand the more complex behavior of real inhomogeneous systems. In a homogeneous system, the dielectric function, (E.8) and (E.11), is diagonal in the tensor indices and is an isotropic function of relative coordinates $\epsilon(|\mathbf{r} - \mathbf{r}'|, t - t')$, so that in Fourier space it is simply $\epsilon(q, \omega)$. Then, we have the simple interpretation that $\epsilon(q, \omega)$ is the response to an internal field,

$$\mathbf{D}(q, \omega) = \mathbf{E}(q, \omega) + 4\pi\mathbf{P}(q, \omega) = \epsilon(q, \omega)\mathbf{E}(q, \omega), \tag{5.34}$$

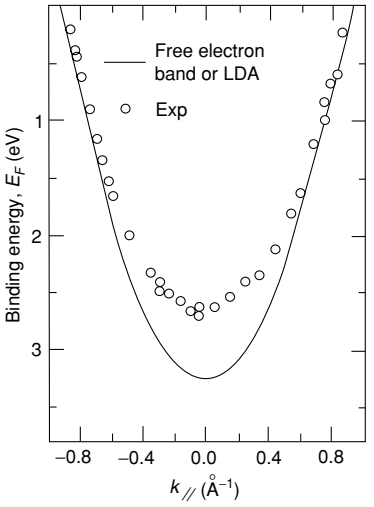


Figure 5.6. Experimental bands of Na determined from angle-resolved photoemission [305] compared to the simple $k^2/2$ dispersion for a non-interacting homogeneous electron gas at the density of Na, which is close to the actual calculated bands in Na. Such agreement is also found in other materials, providing the justification that density functional theory is a reasonable starting point for understanding electronic structure in solids such as the sp-bonded metals. From [305]; see also [306].

or, in terms of potentials,

$$\epsilon(q,\omega) = \frac{\delta V_{\rm ext}(\mathbf{q},\omega)}{\delta V_{\rm test}(\mathbf{q},\omega)} = 1 - v(q)\chi_n^*(\mathbf{q},\omega), \tag{5.35}$$

where $v(q) = \frac{4\pi e^2}{q^2}$ is the frequency-independent relation of the Coulomb potential at wavevector q to the electron density $n(q)$. No approximation has so far been made, if χ^* is the full many-body response function (called the "proper" response function) to the internal electric field.

The well-known RPA [225] is the approximation where all interactions felt by the electrons average out because of their "random phases," except for the Hartree term, in which case each electron experiences an effective potential $V_{\rm eff}$ that is the same as that for a test charge $V_{\rm test}$. Then $\chi_n^*(\mathbf{q},\omega) = \chi_n^0(\mathbf{q},\omega)$ and the RPA is an example of effective-field response functions treated in more detail in Sec. 20.2 and App. D. In a homogeneous gas, the expression for χ^0 given in Sec. 20.2 becomes an integral over states where $|\mathbf{k}| < k_F$ is occupied and $|\mathbf{k}+\mathbf{q}| > k_F$ is empty, which can be written,

$$\chi_n^0(\mathbf{q},\omega) = 4\frac{1}{\frac{4\pi}{3}k_F^3}\int^{k=k_F} \mathrm{d}\mathbf{k}\frac{1}{\varepsilon_k - \varepsilon_{|\mathbf{k}+\mathbf{q}|} - \omega + i\delta}\Theta(|\mathbf{k}+\mathbf{q}| - k_F). \tag{5.36}$$

The integral can be evaluated analytically for a homogeneous gas where $\varepsilon_k = \frac{1}{2}k^2$, leading to the Lindhard [307] dielectric function. The imaginary part can be derived as an integral over regions where the conditions are satisfied by $k < k_F$, $|\mathbf{k}+\mathbf{q}| > k_F$ and the real part

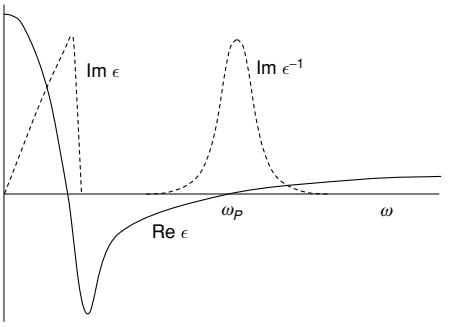


Figure 5.7. Lindhard dielectric function $\epsilon(k, \omega)$ for $k \ll k_{\rm TF}$ given in (5.38) for a homogeneous electron gas in the random phase approximation (RPA). The imaginary part of ϵ is large only for low frequency. The frequency at which the real part of $\epsilon(k, \omega)$ vanishes corresponds to a peak in the imaginary part of $\epsilon^{-1}(k, \omega)$, which denotes the plasma oscillation at $\omega = \omega_P(k)$. For low frequencies, the real part approaches $k_{\rm TF}^2/k^2$, the same as the Thomas–Fermi form (5.20).

of the energy denominator vanishes. The real part can be derived by a Kramers–Kronig transform (D.15), with the result (Exercise 5.18) [225],

$$\begin{aligned} \text{Im } \epsilon(q,\omega) &= \frac{\pi}{2}\frac{k_{\rm TF}^2}{q^2}\frac{\omega}{qv_F}, \qquad\qquad \omega < qv_F - \varepsilon_q, \\ &= \frac{\pi}{4}\frac{k_{\rm TF}^2}{q^2}\frac{k_F}{q}\left[1 - \frac{(\omega-\varepsilon_q)^2}{(qv_F)^2}\right], \qquad qv_F - \varepsilon_q < \omega < qv_F + \varepsilon_q, \\ &= 0, \qquad\qquad \omega > qv_F - \varepsilon_q, \end{aligned} \tag{5.37}$$

where v_F is the velocity at the Fermi surface, and

$$\begin{aligned} \text{Re } \epsilon(q,\omega) = 1 &+ \frac{k_{\rm TF}^2}{2q^2} \\ &+ \frac{k_F k_{\rm TF}^2}{4q^3} \times \left\{\left[1 - \frac{(\omega-\varepsilon_q)^2}{(qv_F)^2}\right] \ln\left|\frac{\omega - qv_F - \varepsilon_q}{\omega + qv_F - \varepsilon_q}\right| + \omega \to -\omega\right\}. \end{aligned} \tag{5.38}$$

The form of $\epsilon(q, \omega)$ for a homogeneous gas is shown in Fig. 5.7 for small q. The imaginary part of ϵ vanishes for $\omega > qv_F + \varepsilon_q$, so that there is no absorption above this frequency. The real part of the dielectric function vanishes at the plasmon frequency $\omega = \omega_p$, where $\omega_p^2 = 4\pi n_e e^2/m_e$, with n_e the electron density. This corresponds to a pole in the inverse dielectric function $\epsilon^{-1}(q, \omega)$. The behavior of ϵ at the plasma frequency can be derived (Exercise 5.18) from (5.38) by expanding the logarithms, but the derivation is much more easily done using the general "f sum rule" given in Sec. E.3, together with the fact that the imaginary part of $\epsilon(q, \omega)$ vanishes at $\omega = \omega_p$.

The Lindhard expression reveals many important properties that carry over qualitatively to solids. The low-frequency peak is still present in metals and is called the Drude absorption and there is generally additional broadening due to scattering [84, 86, 88]. In addition, the static screening Re $\epsilon(q, 0)$ has oscillations at twice the Fermi wavevector $q = 2k_F$, which lead to Friedel oscillations and the Kohn anomaly for phonons. Related affects carry over to

response functions of solids (App. D and Ch. 19), except that $2k_F$ is replaced by anisotropic vectors that span the Fermi surface.

The primary difference in real materials is that there are also interband transitions that give non-zero imaginary ϵ above a threshold frequency. Examples of imaginary parts of $\epsilon(q \approx 0, \omega)$ for crystals are shown in Figs. 2.27 and 2.28. Interband absorption also causes a broadening of the plasmon peak in $\epsilon^{-1}(q, \omega)$, but, nevertheless, there still tends to be a dominant peak around the plasma frequency. Examples are given in Ch. 20, where the absorption of light by nanoscale clusters exhibits clearly the plasma-like peak.

SELECT FURTHER READING

Ashcroft, N. and Mermin, N. *Solid State Physics*, (W. B. Saunders Company, New York, 1976).

Jones, W. and March, N. H. *Theoretical Solid State Physics*, Vols. I, II, (John Wiley and Sons, New York, 1976).

Mahan, G. D. *Many-Particle Physics*, 3rd Edn (Kluwer Academic/Plenum Publishers, New York, 2000).

Pines, D. *Elementary Excitations in Solids* (John Wiley and Sons, New York, 1964).

Exercises

5.1 For fcc and bcc crystals with Z valence electrons per primitive cells, show that r_s is given, respectively, by

$$r_s = \frac{a}{2}\left(\frac{3}{2\pi Z}\right)^{1/3} \text{ and } r_s = \frac{a}{2}\left(\frac{3}{\pi Z}\right)^{1/3}.$$

If r_s is in atomic units (a_0) and the cube edge a is in Å, then $r_s = 0.738Z^{-1/3}a$ and $r_s = 0.930Z^{-1/3}a$.

5.2 For semiconductors with eight valence electrons per primitive cell in diamond- or zinc-blende-structure crystals, show that $r_s = 0.369a$.

5.3 Argue that the expression for Coulomb interaction in large parentheses in (5.4) is finite due to cancellation of the two divergent terms. Show that the scaled hamiltonian given in (5.4) is indeed equivalent to the original hamiltonian (5.2).

5.4 Derive the relation (5.5) between the Fermi wavevector k_F^σ and the density n^σ for a given spin. Do this by considering a large cube of side L, and requiring the wavefunctions to be periodic in length L in each direction (Born–von Karmen boundary conditions).

5.5 Show that relation (5.6), between k_F and the density parameter r_s for an unpolarized gas, follows from the basic definition (5.5) (see also previous problem.)

5.6 Show that expression (5.10) follows from (5.9) by carrying out the indicated differentiation and partial integration. Use this form to derive the $T = 0$ form, (5.11). Also show that the factor in brackets approaches unity for $y \to 0$.

5.7 Verify expression (5.8) for the kinetic energy of the ground state of a non-interacting electron gas. Note that in (5.8), the denominator counts the number of states and the numerator is the

same integral but weighted by the kinetic energy of a state, so that this equation is independent of the number of spins. Derive the corresponding results for one and two dimensions.

5.8 Show that plane waves are eigenstates for the Hartree–Fock theory of a homogeneous electron gas – assuming the ground state is homogeneous, which may not be the case when interactions are included. Thus the kinetic energy is the same for Hartree–Fock theory as for non-interacting particles.

5.9 Derive expression (5.12) for eigenvalues in the Hartree–Fock approximation from the general definition in (3.48). Hint: The exchange integral for plane wave states has the form $-4\pi \sum_{\mathbf{k}'}^{k'<k_F} 1/|\mathbf{k}-\mathbf{k}'|^2$. This leads to the singular log form in three dimensions. For more details, see [84, 225].

5.10 Derive the broadening of the bands in the Hartree–Fock approximation from the unpolarized gas $\Delta W = (9/4\pi^2)^{1/3}/r_s$ using (5.12).

5.11 Derive analytically that the electron velocity $v = \mathrm{d}\varepsilon/\mathrm{d}k$ diverges at $k = k_F$ in the Hartree–Fock approximation. Argue that: (1) this happens in *all* metals due to the Coulomb interaction and the Hartree–Fock approximation, and (2) there is no divergence for short-range interactions.

5.12 Show that the average value of the factor $f(x)$ in (5.13) is $-3/4$, as stated before (5.15). Then, for the ground state of the homogeneous gas, verify the result for the exchange energy (5.15).

5.13 Show that (5.19) follows directly from evaluating the expressions in (3.54) or (3.52) by inserting the plane wave eigenfunctions (normalized to a large volume Ω) and evaluating the resulting expression. Alternatively, $g_x(r)$ can be found from the general relation (3.56) of the pair correlation function and the density matrix for non-interacting fermions [246,263], $g_x(r) = 1 - |\rho(r)|^2/n^2$, where n is the density and the density matrix $\rho(r)$ is given by (5.11).

5.14 Consider a point charge in an otherwise uniform gas. Use the Thomas–Fermi (TF) approximation (Ch. 6) to derive the TF screening length (5.21). (Hint: Assume of the change in the density due to the impurity is $\delta_n(r) = \exp(-k_{\mathrm{TF}}r)/r$ and determine the decay constant k_{TF} from the TF equations expanded to linear order.)

5.15 Derive the expression for the exchange–correlation hole (5.29) in terms of the hole at larger densities (smaller r_s). Would there be an analogous form that involves an integral of λ from 1 to ∞, i.e. for larger r_s?

5.16 Using Fig. 5.5 sketch the shape of the average hole, (5.29), for antiparallel-spin electrons Al, Na, and Cs.

5.17 Derive the expression for the equilibrium r_s given in (5.32) from the expression for total energy and using $\alpha = 1.80$. In which direction will the predicted r_s change if correlation is included? Find the explicit expression using the Wigner interpolation formula for ϵ_c.

5.18 Derive the Lindhard expression for the dielectric function of a homogeneous gas (5.38). This is a tedious integral and the steps are given by Pines [225], p. 144.

PART II

DENSITY FUNCTIONAL THEORY

6

Density functional theory: foundations

E Pluribus Unum

Summary

The fundamental tenet of density functional theory is that *any* property of a system of many interacting particles can be viewed as a *functional* of the ground state density $n_0(\mathbf{r})$; that is, one scalar function of position $n_0(\mathbf{r})$, in principle, determines all the information in the many-body wavefunctions for the ground state and all excited states. The existence proofs for such functionals, given in the original works of Hohenberg and Kohn and of Mermin, are disarmingly simple. However, they provide no guidance whatsoever for constructing the functionals, and no exact functionals are known for any system of more than one electron. Density functional theory (DFT) would remain a minor curiosity today if it were not for the *ansatz* made by Kohn and Sham, which has provided a way to make useful, approximate ground state functionals for real systems of many electrons. The subject of this chapter is density functional theory as a methodology for many-body systems; Ch. 7 describes the Kohn–Sham ansatz that replaces the interacting problem with an auxiliary independent-particle problem with all many-body effects included in an exchange–correlation functional; Ch. 8 deals with widely used approximations for the exchange–correlation functional; and Ch. 9 is devoted to solution of the Kohn–Sham independent-particle equations in a general form useful for all Kohn–Sham calculations. Following chapters in this volume are devoted to algorithms for actual calculations, and applications to problems in atomic, molecular, and condensed matter physics.

Density functional theory is a theory of correlated many-body systems. It is included here in close association with independent-particle methods, because it has provided the key step that has made possible development of practical, useful independent-particle approaches that incorporate effects of interactions and correlations among the particles. As such, density functional theory has become the primary tool for calculation of electronic structure in condensed matter, and is increasingly important for quantitative studies of molecules and other finite systems. The remarkable successes of the approximate local density (LDA) and generalized-gradient approximation (GGA) functionals within the Kohn–Sham approach

have led to widespread interest in density functional theory as the most promising approach for accurate, practical methods in the theory of materials.

The modern formulation of density functional theory originated in a famous paper written by P. Hohenberg and W. Kohn in 1964 [308]. These authors showed that a special role can be assigned to the density of particles in the ground state of a quantum many-body system: the density can be considered as a "basic variable," i.e. that *all* properties of the system can be considered to be unique *functionals* of the ground state density. Shortly following in 1965, Mermin [309] extended the Hohenberg–Kohn arguments to finite temperature canonical and grand canonical ensembles. Although the finite temperature extension has not been widely used, it illuminates both the generality of density functional theory and the difficulty of realizing the promise of exact density functional theory. Also in 1965 appeared the other classic work of this field by W. Kohn and L. J. Sham [92], whose formulation of density functional theory has become the basis of much of present-day methods for treating electrons in atoms, molecules, and condensed matter.

The goal of the chapters on density functional theory is to elucidate the fundamental ideas and current practices; to give the reader sufficient background to use density functional theory intelligently for real problems; and to expose potential pitfalls and possible avenues for future developments. The present chapter is concerned with the basic formulation of the theory; Ch. 7 deals with the Kohn–Sham auxiliary system that has made possible accurate, feasible approaches to the full many-body electron problem. The theory of the exchange–correlation functional and practical approximate functionals are the subject of Ch. 8, along with a few selected results. Chapter 9 deals with general aspects of the Kohn–Sham equations, with explicit algorithms and results left to later chapters.

6.1 Thomas–Fermi–Dirac approximation: example of a functional

The original density functional theory of quantum systems is the method of Thomas [316] and Fermi [317] proposed in 1927. Although their approximation is not accurate enough for present-day electronic structure calculations, the approach illustrates the way density functional theory works. In the original Thomas–Fermi method the kinetic energy of the system of electrons is approximated as an explicit functional of the density, idealized as non-interacting electrons in a homogeneous gas with density equal to the local density at any given point. Both Thomas and Fermi neglected exchange and correlation among the electrons; however, this was extended by Dirac [318] in 1930, who formulated the local approximation for exchange (see Secs. 5.1 and 8.1) still in use today. This leads to the energy functional for electrons in an external potential $V_{\rm ext}(\mathbf{r})$

$$E_{\rm TF}[n] = C_1 \int \mathrm{d}^3r\; n(\mathbf{r})^{(5/3)} + \int \mathrm{d}^3r\; V_{\rm ext}(\mathbf{r})n(\mathbf{r}) + C_2 \int \mathrm{d}^3r\; n(\mathbf{r})^{4/3} + \frac{1}{2}\int \mathrm{d}^3r\mathrm{d}^3r'\; \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}, \tag{6.1}$$

where the first term is the local approximation to the kinetic energy with $C_1 = \frac{3}{10}(3\pi^2)^{(2/3)} = 2.871$ in atomic units (see Sec. 5.1), the third term is the local exchange with $C_2 = -\frac{3}{4}(\frac{3}{\pi})^{1/3}$

(Eq. (5.15) for the case of equal up and down spins) and the last term is the classical electrostatic Hartree energy. (In App. H, improved approximations for inhomogeneous systems including gradients are given.)

The ground state density and energy can be found by minimizing the functional $E[n]$ in (6.1) for all possible $n(\mathbf{r})$ subject to the constraint on the total number of electrons

$$\int \mathrm{d}^3 r \; n(\mathbf{r}) = N. \tag{6.2}$$

Using the method of Lagrange multipliers (Exercise 6.1), the solution can be found by an unconstrained minimization of the functional

$$\Omega_{\mathrm{TF}}[n] = E_{\mathrm{TF}}[n] - \mu \left\{ \int \mathrm{d}^3 r n(\mathbf{r}) - N \right\}, \tag{6.3}$$

where the Lagrange multiplier μ is the Fermi energy. For small variations of the density $\delta n(\mathbf{r})$, the condition for a stationary point is[1]

$$\begin{aligned} &\int \mathrm{d}^3 r \left\{ \Omega_{\mathrm{TF}}[n(\mathbf{r}) + \delta n(\mathbf{r})] - \Omega_{\mathrm{TF}}[n(\mathbf{r})] \right\} \rightarrow \\ &\int \mathrm{d}^3 r \left\{ \frac{5}{3} C_1 n(\mathbf{r})^{2/3} + V(\mathbf{r}) - \mu \right\} \delta n(\mathbf{r}) = 0, \end{aligned} \tag{6.4}$$

where $V(\mathbf{r}) = V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{Hartree}}(\mathbf{r}) + V_x(\mathbf{r})$ is the total potential. Since (6.4) must be satisfied for any function $\delta n(\mathbf{r})$, it follows that the functional is stationary if and only if the density and potential satisfy the relation

$$\frac{1}{2}(3\pi^2)^{(2/3)} n(\mathbf{r})^{2/3} + V(\mathbf{r}) - \mu = 0. \tag{6.5}$$

Extensions to account for effects of inhomogeneity have been proposed by many people, the best known being the Weizsacker [319] correction, $\frac{1}{4}(\nabla n^\sigma(\mathbf{r}))^2/n^\sigma(\mathbf{r})$, but more recent work [320] has found the correction to be reduced to $\frac{1}{36}(\nabla n^\sigma(\mathbf{r}))^2/n^\sigma(\mathbf{r})$ (see [246], Sec. 2.11.6 and Appendix 2.4). Exercise 6.2 treats aspects of the equations including the gradient corrections.

The attraction of density functional theory is evident by the fact that one equation for the density is remarkably simpler than the full many-body Schrödinger equation that involves $3N$ degrees of freedom for N electrons. The Thomas-Fermi approach has been applied, for example, to equations of state of the elements [321]. However, the Thomas-Fermi-type approach starts with approximations that are too crude, missing essential physics and chemistry, such as shell structures of atoms and binding of molecules [322]. Thus it falls short of the goal of a useful description of electrons in matter.

6.2 The Hohenberg–Kohn theorems

The approach of Hohenberg and Kohn is to formulate density functional theory as an *exact theory of many-body systems*. The formulation applies to any system of interacting particles

[1] This is an example of functional equations described in App. A; see specifically (A.5).

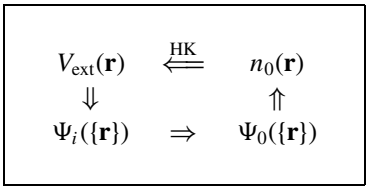


Figure 6.1. Schematic representation of Hohenberg–Kohn theorem. The smaller arrows denote the usual solution of the Schrödinger equation where the potential $V_{\rm ext}(\mathbf{r})$ determines all states of the system $\Psi_i(\{\mathbf{r}\})$, including the ground state $\Psi_0(\{\mathbf{r}\})$ and ground state density $n_0(\mathbf{r})$. The long arrow labeled "HK" denotes the Hohenberg–Kohn theorem, which completes the circle.

in an external potential $V_{\rm ext}(\mathbf{r})$, including any problem of electrons and fixed nuclei, where the hamiltonian can be written[2]

$$\hat{H} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \sum_i V_{\rm ext}(\mathbf{r}_i) + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{6.6}$$

Density functional theory is based upon two theorems first proved by Hohenberg and Kohn [308]. Here we first present the theorems and the proofs along with discussion of the consequences; Sec. 6.3 contains the alternative formulation of Levy and Lieb, which is more general and gives a more intuitive definition of the functional. The relations established by Hohenberg and Kohn are illustrated in Fig. 6.1 and can be started as follows:

- **Theorem I:** For any system of interacting particles in an external potential $V_{\rm ext}(\mathbf{r})$, the potential $V_{\rm ext}(\mathbf{r})$ is determined uniquely, except for a constant, by the ground state particle density $n_0(\mathbf{r})$.
  Corollary I: Since the hamiltonian is thus fully determined, except for a constant shift of the energy, it follows that the many-body wavefunctions for all states (ground and excited) are determined. *Therefore all properties of the system are completely determined given only the ground state density* $n_0(\mathbf{r})$.
- **Theorem II:** A *universal functional* for the energy $E[n]$ in terms of the density $n(\mathbf{r})$ can be defined, valid for any external potential $V_{\rm ext}(\mathbf{r})$. For any particular $V_{\rm ext}(\mathbf{r})$, the exact ground state energy of the system is the global minimum value of this functional, and the density $n(\mathbf{r})$ that minimizes the functional is the exact ground state density $n_0(\mathbf{r})$.
  Corollary II: The functional $E[n]$ alone is sufficient to determine the exact ground state energy and density. In general, excited states of the electrons must be determined by other means. Nevertheless, the work of Mermin (Sec. 6.4) shows that thermal equilibrium properties such as specific heat are determined directly by the free-energy functional of the density.

These assertions are so encompassing and the proofs are so simple, that it is crucial for any practitioner in the field to understand the basis of the theorems and the limits of the logical consequences.

[2] The nuclei–nuclei interaction can be added later; it is irrelevant, except that care is need to treat Coulomb interactions in extended systems (Sec. 3.2). Special considerations are required to include magnetic fields and there are subtle issues for electric fields in extended systems, see Sec. 6.4.

Proof of Theorem I: density as a basic variable

The proofs of the Hohenberg–Kohn theorems are disarmingly simple. Consider first Theorem I, using the general expressions given in (3.8) and (3.9) for the density and energy in terms of the many-body wavefunction. Suppose that there were two different external potentials $V_{\rm ext}^{(1)}(\mathbf{r})$ and $V_{\rm ext}^{(2)}(\mathbf{r})$ which differ by more than a constant and which lead to the same ground state density $n(\mathbf{r})$. The two external potentials lead to two different hamiltonians, $\hat{H}^{(1)}$ and $\hat{H}^{(2)}$, which have different ground state wavefunctions, $\Psi^{(1)}$ and $\Psi^{(2)}$, which are hypothesized to have the same ground state density $n_0(\mathbf{r})$. (It is straightforward to find different Ψs with the same density, as discussed below.) Since $\Psi^{(2)}$ is not the ground state of $\hat{H}^{(1)}$, it follows that

$$E^{(1)} = \langle\Psi^{(1)}|\hat{H}^{(1)}|\Psi^{(1)}\rangle < \langle\Psi^{(2)}|\hat{H}^{(1)}|\Psi^{(2)}\rangle. \tag{6.7}$$

The strict inequality follows if the ground state is non-degenerate, which we will assume here following the arguments of Hohenberg and Kohn.[3] The last term in (6.7) can be written

$$\langle\Psi^{(2)}|\hat{H}^{(1)}|\Psi^{(2)}\rangle = \langle\Psi^{(2)}|\hat{H}^{(2)}|\Psi^{(2)}\rangle + \langle\Psi^{(2)}|\hat{H}^{(1)} - \hat{H}^{(2)}|\Psi^{(2)}\rangle \tag{6.8}$$

$$= E^{(2)} + \int \mathrm{d}^3r \left[V_{\rm ext}^{(1)}(\mathbf{r}) - V_{\rm ext}^{(2)}(\mathbf{r})\right] n_0(\mathbf{r}), \tag{6.9}$$

so that

$$E^{(1)} < E^{(2)} + \int \mathrm{d}^3r \left[V_{\rm ext}^{(1)}(\mathbf{r}) - V_{\rm ext}^{(2)}(\mathbf{r})\right] n_0(\mathbf{r}). \tag{6.10}$$

On the other hand if we consider $E^{(2)}$ in exactly the same way, we find the same equation with superscripts (1) and (2) interchanged,

$$E^{(2)} < E^{(1)} + \int \mathrm{d}^3r \left[V_{\rm ext}^{(2)}(\mathbf{r}) - V_{\rm ext}^{(1)}(\mathbf{r})\right] n_0(\mathbf{r}). \tag{6.11}$$

Now if we add together (6.10) and (6.11), we arrive at the contradictory inequality $E^{(1)} + E^{(2)} < E^{(1)} + E^{(2)}$. This establishes the desired result: there cannot be two different external potentials differing by more than a constant which give rise to the same non-degenerate ground state charge density. The density uniquely determines the external potential to within a constant.

The corollary follows since the hamiltonian is uniquely determined (except for a constant) by the ground state density. Then, in principle, the wavefunction of any state is determined by solving the Schrödinger equation with this hamiltonian. Among all the solutions which are consistent with the given density, the unique ground state wavefunction is the one that has the lowest energy.

[3] This is not a necessary restriction. The proof can readily be extended to degenerate cases [323], which are also included in the alternative formulation by Levy [324–326] discussed in Sec. 6.3. Except in special cases the density of any one of the degenerate ground states uniquely determines the external potential. In the exercises is an example, where two degenerate states have exactly the same density so that the expectation values of general operators cannot be unique functionals of the density. Even then, the expectation value of the energy is the same for all linear combinations of the degenerate states so that the Hohenberg–Kohn theorem is recovered.

Despite the appeal of this result, it is clear from the reasoning that no prescription has been given to solve the problem. Since all that was proved is that $n_0(\mathbf{r})$ uniquely determines $V_{\rm ext}(\mathbf{r})$, we are still left with the problem of solving the many-body problem in the presence of $V_{\rm ext}(\mathbf{r})$. For example, for electrons in materials, the external potential is the Coulomb potential due to the nuclei. The theorem only requires that the electron density uniquely determines the positions and types of nuclei, which can easily be proven from elementary quantum mechanics (see Exercise 6.6). At this level we have gained nothing: we are still faced with the original problem of many interacting electrons moving in the potential due to the nuclei.

Proof of Theorem II

The second theorem is just as easily proven once one has carefully defined the meaning of a functional of the density and restricted the space of densities. The original proof of Hohenberg–Kohn is restricted to densities $n(\mathbf{r})$ that are ground state densities of the electron hamiltonian with *some* external potential $V_{\rm ext}$. Such densities are called "V-representable." This defines a space of possible densities within which we can construct *functionals* of the density. (As discussed below in Sec. 6.3 it is possible to extend the range of validity of the functional.) Since all properties such as the kinetic energy, etc., are uniquely determined if $n(\mathbf{r})$ is specified, then each such property can be viewed as a functional of $n(\mathbf{r})$, including the total energy functional

$$\begin{aligned} E_{\rm HK}[n] &= T[n] + E_{\rm int}[n] + \int {\rm d}^3 r \; V_{\rm ext}(\mathbf{r}) n(\mathbf{r}) + E_{II} \\ &\equiv F_{\rm HK}[n] + \int {\rm d}^3 r \; V_{\rm ext}(\mathbf{r}) n(\mathbf{r}) + E_{II}, \end{aligned} \tag{6.12}$$

where E_{II} is the interaction energy of the nuclei (see (3.2) and related discussion). The functional $F_{\rm HK}[n]$ defined in (6.12) includes all internal energies, kinetic and potential, of the interacting electron system,

$$F_{\rm HK}[n] = T[n] + E_{\rm int}[n], \tag{6.13}$$

which must be universal by construction since the kinetic energy and interaction energy of the particles are functionals only of the density.[4]

Now consider a system with the ground state density $n^{(1)}(\mathbf{r})$ corresponding to external potential $V_{\rm ext}^{(1)}(\mathbf{r})$. Following the discussion above, the Hohenberg–Kohn functional is equal to

[4] Note that here "universal" means *the same for all electron systems*, independent of the external potential $V_{\rm ext}(\mathbf{r})$. The Hohenberg–Kohn approach leads to different functionals for different particles depending upon their masses and interactions. In this book the functionals described are for electrons, unless explicitly indicated otherwise. In fact there is another important application of the ideas of density functional theory in the theory of electronic structure: the case of "non-interacting electrons," i.e. fermions with the electron mass but with no interactions among themselves, which are the particles that explicitly enter the Kohn–Sham equations. It is advantageous to use the general ideas of density functionals in that case as well, and we will carefully indicate the distinction between the different use of the functionals for the Kohn–Sham equations.

the expectation value of the hamiltonian in the unique ground state, which has wavefunction $\Psi^{(1)}$

$$E^{(1)} = E_{\rm HK}[n^{(1)}] = \langle \Psi^{(1)} | \hat{H}^{(1)} | \Psi^{(1)} \rangle. \tag{6.14}$$

Now consider a different density, say $n^{(2)}(\mathbf{r})$, which necessarily corresponds to a different wavefunction $\Psi^{(2)}$. It follows immediately that the energy $E^{(2)}$ of this state is greater than $E^{(1)}$, since

$$E^{(1)} = \langle \Psi^{(1)} | \hat{H}^{(1)} | \Psi^{(1)} \rangle < \langle \Psi^{(2)} | \hat{H}^{(1)} | \Psi^{(2)} \rangle = E^{(2)}. \tag{6.15}$$

Thus the energy given by (6.12) in terms of the Hohenberg–Kohn functional evaluated for the correct ground state density $n_0(\mathbf{r})$ is indeed lower than the value of this expression for any other density $n(\mathbf{r})$.

It follows that if the functional $F_{\rm HK}[n]$ was known, then by minimizing the total energy of the system, (6.12), with respect to variations in the density function $n(\mathbf{r})$, one would find the exact ground state density and energy. This establishes Corollary II. Note that the functional only determines the ground state properties; it does not provide any guidance concerning excited states.

6.3 Constrained search formulation of density functional theory

An alternative definition of a functional due to Levy [324–326] and Lieb [327–329] is very instructive, because it:

- extends the range of definition of the functional in a way that is formally more tractable and clarifies its physical meaning;
- provides an in-principle way to determine the exact functional;
- leads to the same ground state density and energy at the minimum as in the Hohenberg–Kohn analysis, and also applies for degenerate ground states.

The idea of Levy and Lieb (LL) is to define a *two-step* minimization procedure beginning with the usual general expression for the energy in terms of the many-body wavefunction Ψ given by (3.9). The ground state can be found, in principle, by minimizing the energy with respect to all the variables in Ψ. However, suppose one first considers the energy only for the class of many-body wavefunctions Ψ *that have the same density* $n(\mathbf{r})$. For any wavefunction, the total energy can be written

$$E = \langle \Psi | \hat{T} | \Psi \rangle + \langle \Psi | \hat{V}_{\rm int} | \Psi \rangle + \int {\rm d}^3 r \; V_{\rm ext}(\mathbf{r}) n(\mathbf{r}). \tag{6.16}$$

Now if one minimizes the energy (6.16) over the class of wavefunctions with the same density $n(\mathbf{r})$, then one can define a unique lowest energy for that density

$$\begin{aligned} E_{\rm LL}[n] &= \min_{\Psi \to n(\mathbf{r})} [\langle \Psi | \hat{T} | \Psi \rangle + \langle \Psi | \hat{V}_{\rm int} | \Psi \rangle] + \int {\rm d}^3 r \; V_{\rm ext}(\mathbf{r}) n(\mathbf{r}) + E_{II} \\ &\equiv F_{\rm LL}[n] + \int {\rm d}^3 r \; V_{\rm ext}(\mathbf{r}) n(\mathbf{r}) + E_{II}, \end{aligned} \tag{6.17}$$

where the Levy–Lieb functional of the density is defined by

$$F_{\rm LL}[n] = \min_{\Psi \to n(\mathbf{r})} \langle \Psi | \hat{T} + \hat{V}_{\rm int} | \Psi \rangle. \tag{6.18}$$

In this form, $E_{\rm LL}[n]$ is manifestly a functional of the density and the ground state is found by minimizing $E_{\rm LL}[n]$.

The Levy–Lieb formulation is much more than just a restatement of the Hohenberg–Kohn functional, (6.12). First, (6.18) clarifies the meaning of the functional and provides a way to make an operational definition: *the minimum of the sum of kinetic plus interaction energies for all possible wavefunctions having the given density* $n(\mathbf{r})$. The LL functional also has important formal differences from the Hohenberg–Kohn functional; in particular, the LL functional in (6.18) is defined for *any* density $n(\mathbf{r})$ derivable from a wavefunction Ψ_N for N electrons. This is termed "N-representability" and the existence of such a wavefunction Ψ_N for any density satisfying simple conditions is known [330], as discussed in Sec. 6.5. In contrast, the Hohenberg–Kohn functional is defined only for densities that can be generated by some external potential; this is called "V-representability" and the conditions for such densities are not known in general. At the minimum of the total energy of the system in a given external potential, the Levy–Leib functional $F_{\rm LL}[n]$ must equal the Hohenberg–Kohn functional defined in (6.13), since the minimum is a density which can be generated by an external potential. In addition, the LL form eliminates the restriction in the original proof of Hohenberg–Kohn to non-degenerate ground states; now one can do the search in the space of any one of a set of degenerate states.

Thus it has been established that a functional can be defined for any density (subject to certain conditions given below), and that by minimizing this functional one would find the exact density and energy of the true interacting many-body system. Just as for the original Hohenberg–Kohn proofs, however, we are faced with the cold fact that no method has been given to find the functional other than the original definition in terms of many-body wavefunctions. Nevertheless, as we shall see in the following chapter, the dependence of the functional upon the kinetic and potential energies of the full, correlated many-body wavefunction points the way toward constructing approximate functionals that are of great utility in practical calculations and in understanding the effects of exchange and correlation among the electrons.

6.4 Extensions of Hohenberg–Kohn theorems

Spin density functional theory

The above analysis also shows how the Hohenberg–Kohn theorems can be generalized to several types of particles. The reason for the special role of the density and the external potential in the Hohenberg–Kohn theorems, rather than some other properties of the particles, is simply that these quantities enter the total energy (3.9) explicitly only through the simple bilinear integral term $\int \mathrm{d}^3 r\, V_{\rm ext}(\mathbf{r}) n(\mathbf{r})$. If there are other terms in the hamiltonian having this form, then each such pair of external potential and particle density will obey a Hohenberg–Kohn theorem.

The most relevant example for our purposes is a Zeeman term that is different for up and down spin fermions (i.e. a magnetic field that acts only on the spins, not on the orbital motion). This is in fact one of the important effects of an external magnetic field, so that this can be considered as a physically realistic approximation. Within this model, one can rigorously generalize all the above arguments to include two types of densities, the particle density $n(\mathbf{r}) = n(\mathbf{r}, \sigma = \uparrow) + n(\mathbf{r}, \sigma = \downarrow)$ and the spin density $s(\mathbf{r}) = n(\mathbf{r}, \sigma = \uparrow) - n(\mathbf{r}, \sigma = \downarrow)$. This leads to an energy functional

$$E = E_{\mathrm{HK}}[n, s] \equiv E'_{\mathrm{HK}}[n], \tag{6.19}$$

where in the last form it is assumed that $[n]$ denotes a functional of the density which depends upon both position in space $\mathbf{r}$ and spin σ. "Spin density functional theory" is essential in the theory of atoms and molecules with net spins, as well as solids with magnetic order [291, 314, 331]. (Note that this does *not* include effects of a magnetic field upon the orbital motion, which requires an extension to current functional theory [332–335].)

In the absence of external Zeeman fields, the lowest energy solution may be spin polarized, i.e. $n(\mathbf{r}, \uparrow) \neq n(\mathbf{r}, \downarrow)$, which is analogous to the broken symmetry solution of unrestricted Hartree–Fock theory. (This must happen in a finite system with an odd number of electrons, and also occurs in some atoms polarized to Hund's rules and in magnetic solids.) The spin functional is useful in these cases as well; however, the original Hohenberg–Kohn theorem remains valid and the ground state, in principle, is determined by the total ground state density $n(\mathbf{r}) = n(\mathbf{r}, \uparrow) + n(\mathbf{r}, \downarrow)$ for any system where there is no spin-dependent external potential (see Exercise 6.9). The only modification of the statements of the theorems is to take into account the fact that the broken symmetry solution is necessarily degenerate.

Mermin finite temperature and ensemble density functional theory

The theorems of Hohenberg and Kohn for the ground state carry over to the equilibrium thermal distribution by constructing the density corresponding to the thermal ensemble. For each of the conclusions of Hohenberg and Kohn for the ground state, there exists a corresponding argument for a system in thermal equilibrium, as was shown by Mermin [309] shortly after the Hohenberg–Kohn paper. To show this, Mermin constructed a grand potential functional of the trial density matrices $\hat{\rho}$,

$$\Omega[\hat{\rho}] = \mathrm{Tr}\hat{\rho}\left[(\hat{H} - \mu\hat{N}) + \frac{1}{\beta}\ln\hat{\rho}\right], \tag{6.20}$$

whose minimum is the equilibrium grand potential

$$\Omega = \Omega[\hat{\rho}_0] = -\frac{1}{\beta}\ln \mathrm{Tr}\, \mathrm{e}^{-\beta(\hat{H}-\mu\hat{N})}, \tag{6.21}$$

where $\hat{\rho}_0$ is the grand canonical density matrix

$$\hat{\rho}_0 = \frac{\mathrm{e}^{-\beta(\hat{H}-\mu\hat{N})}}{\mathrm{Tr}\, \mathrm{e}^{-\beta(\hat{H}-\mu\hat{N})}}. \tag{6.22}$$

The proof is completely analogous to the Hohenberg–Kohn proofs and uses only the minimum property of $\Omega[\hat{\rho}]$ and the fact that the energy depends upon the external potential only through the term $\int V_{\rm ext}(\mathbf{r})n(\mathbf{r})$. (The independent-particle version of the Mermin functional is given in Sec. 9.2.)

The Mermin theorem leads to even more powerful conclusions than the Hohenberg–Kohn theorems, namely that not only the energy, but also the entropy, specific heat, etc., are functionals of the equilibrium density. However, the Mermin functional has not been widely applied. The simple fact is that it is much more difficult to construct useful, approximate functionals for the entropy (which involves sums over excited states) than for the ground state energy. For example, in the Fermi liquid description of a metal the specific heat coefficient at low temperature is directly related to the effective mass at the Fermi surface. Thus the Mermin functional for the free energy must correctly describe the effective mass (with all its many-body renormalization) as well as the ground state energy, whereas only the latter is required in the Hohenberg–Kohn functional.

The Hohenberg–Kohn theorems can also be generalized to other ensembles which are useful for aspects such as defining a functional of electron number as a continuous variable [336], whereas the original Hohenberg–Kohn theorems are formulated only for a ground state with a fixed integer number of electrons. The equilibrium thermal ensemble of Mermin at fixed chemical potential is an example where the number of electrons fluctuates around the average number given by the expectation value of the number operator $\hat{N}$. From ensemble theory, it also follows, that there must be discontinuities in the derivative of the energy with respect to number at integer occupations or, in the case of solids, for filled bands. These are difficult properties to build into the functional and are absent present-day approximate density functionals.

Current density and time-dependent density functional theory

The Hohenberg–Kohn theorems apply only to systems that are time reversal invariant. If there is a magnetic field or time-dependent electric field, the hamiltonian involves terms of the form $V_{\rm ext}(\mathbf{r})n(\mathbf{r})$ and $\mathbf{p} \cdot \mathbf{A}_{\rm ext}$. Thus by exactly the same logic as the original Hohenberg–Kohn arguments, the properties must depend upon both the density n and the current density $\mathbf{j} = -\frac{e}{m}\mathbf{p}$ [332, 333, 335]. However, the structure of the theory must be fundamentally different because there is no analogue of the variational principle for the ground state energy or equilibrium free energy.

The generalization of the Hohenberg–Kohn approach to time-dependent problems has been provided by Runge and Gross [230]. For a localized systems with simply connected geometry, the theory can be cast in terms of the time-dependent density since the current is determined by $\nabla \mathbf{j} = -dn/dt$. The result is that *given the initial wavefunction at one time* t', the state at later times t is a functional of the time-dependent density $n(\mathbf{r}, t'')$ for all $t' \leq t'' \leq t$. This may be viewed as the formal construction of a density functional theory for excitations. Although the time-dependent fuctional must be quite

intricate, there has been considerable progress within the Kohn–Sham approach, as described in Sec. 7.6.

In general, however, the theory must involve the current density. In particular, in system with no boundaries, *the evolution is not a functional only of the density*. For example, in a uniform ring of charge the density is unchanged if there is a net current, and the state is determined only if the current is specified [337]. Thus there is an essential link with current functionals and to properties such as the static electric polarization.

Electric fields and polarization

The issue of electric fields and polarization comes into play in extended systems. In infinite space, the potential due to an electric field $V(x) = Ex$ is unbounded; there is no lower bound to the energy and therefore there is no ground state. This is a famous problem [338, 339] in the theory of the dielectric properties of materials. However, if the ground state does not exist, the Hohenberg–Kohn theorems on the ground state do not apply [340].

Is there any way to include an electric field in density functional theory? This is a very subtle problem and the answer is that in the presence of an electric field, one must apply some constraint, within which there is a stable ground state. In the case of molecules, this is routinely done simply by constraining the electrons to remain near the molecule. In a solid, however, the constraint is not so obvious. To the knowledge of the author, all proposals involve constraining the electrons to be in localized Wannier functions (Ch. 21) or equivalent conditions on Bloch functions. Since the energy contains a term $\mathbf{E} \cdot \mathbf{P}$, where $\mathbf{P}$ is the macroscopic polarization, the theory must become a "density polarization theory"(see [341, 342] and references cited there). An interesting point is that in a system with a net polarization at zero field $\mathbf{E} = 0$ (e.g. a ferroelectric) *the polarization is determined by the density alone* [341], i.e. the original Hohenberg–Kohn theorem applies. (But see Chs. 7 and 22 for the opposite conclusion in the Kohn–Sham approach.)

6.5 Intricacies of exact density functional theory

The challenge posed by the Hohenberg–Kohn theorems is how to make use of the reformulation of many-body theory in terms of functionals of the density. The theorems are in terms of unknown functionals of the density, and it is easy to show that these must be non-local functionals, depending simultaneously upon $n(\mathbf{r})$ at different positions $\mathbf{r}$, which are difficult to cast in any simple form.

Allowed densities for electrons

There are a number of general questions related to the nature of the possible densities that are allowed for fermions, given only that they must integrate to the correct number of particles:

- Can one readily construct different wavefunctions Ψ that have the same density $n(\mathbf{r})$?
  Yes. An illuminating example is the homogeneous electron gas. All plane waves have the same uniform density, but only the choice of the lowest kinetic energy states gives the lowest energy ground state for the non-interacting case. Interacting electrons also have the same uniform density even though the wavefunctions are correlated and thus quite different from a single determinant. The same logic can be applied to inhomogeneous cases, such as discussed in Exercise 6.6.
- Is it possible to construct an antisymmetric wavefunction for fermions that can describe any possible density ("N-representability")?
  Yes, given a few restrictions on the density. As shown by Gilbert [330], it is possible to construct *any* density integrating to N total electrons of a given spin from a single Slater determinant of N one-electron orbitals, subject only to the condition that $n(\mathbf{r}) \geq 0$, and $\int |\nabla n(\mathbf{r})^{1/2}|^2$ is finite. In certain cases, explicit techniques exist for constructing such wavefunctions [93, 343], as described in Exercise 6.7.
- Is it possible to generate any such density as the ground state of some local external potential ("V-representability")?
  No. A number of "reasonable" looking densities have been shown to be impossible to be the ground state for any V [325,327]. Such densities are termed "non-V-representable." This applies to any linear combination of densities of a set of degenerate states; although the densities look "reasonable" they are not the ground state for the given number of electrons and any potential. An example is the spherically averaged density of an open-shell atom. If one weakens the question to ask if there are densities that cannot be generated by any smooth potential (one without delta functions) then one can find many counterexamples, e.g. any excited state density for single particles in finite systems. (The density of one electron in a 2s state in H is discussed in Exercise 6.7.)

Properties obeyed by the exact density functional theory

The Hohenberg–Kohn arguments are very general for properties of interacting particle systems, yet special emphasis is on the ground state. Thus questions arise as to what properties of a material should be given correctly by the minimization of the Hohenberg–Kohn functional, if it were known exactly. *These examples make it clear how difficult it is to fulfill all the properties guaranteed by the Hohenberg–Kohn and Mermin theorems!*

- Are excitation energies given correctly by the exact density functional theory?
  Yes. In principle, all properties are determined since the entire hamiltonian is determined.
- Are excitation energies given correctly by minimization of the exact Hohenberg–Kohn or Levy–Lieb functionals?
  No. The functional evaluated near the minimum provides no information about excitations, which are associated with *saddle points* at higher energies.
- Is the exact specific heat versus temperature given correctly by the exact finite temperature Mermin functional?

Yes. Even though the specific heat involves excitations from the ground state, nevertheless the thermal averages over these excitations must be a unique functional of the density and the temperature.

- Are static susceptibilities given correctly by the ground state functional?
  Yes. All static susceptibilities are second derivatives of ground state energies with respect to external fields. Thus they must be given correctly by the variation of the ground state Hohenberg–Kohn functional as functions of external fields.[5]
- Is the exact Fermi surface of a metal given by the exact ground state density functional theory?
  Yes. This is not a trivial question for two reasons. First, for the question to be meaningful, the many-body metal must have a well-defined Fermi surface; for the present purposes we assume this. Second, it is not a priori obvious that the Fermi surface is a ground state property. One way to see that the Fermi surface is determined by ground state properties is to consider susceptibilities to static perturbations. The exact density functional theory must lead to the correct Kohn anomalies and Friedel oscillations of the density far from an impurity, which depend in detail on the shape of the Fermi surface of the unperturbed metal.
- Must a Mott insulator (an insulator due to correlations among the electrons) be predicted correctly by the exact density functional theory?
  Yes. This follows from the above arguments on a metal in the special case where the Fermi surface vanishes.

6.6 Difficulties in proceeding from the density

The purpose of this section is to emphasize that density functional theory does *not* provide a way to understand the properties of a material merely by looking at the form of the density. Although the density is *in principle* sufficient, the relation is very subtle and no one has found a way to extract directly from the density any general set of properties, e.g. whether the material is a metal or an insulator. The key point is that the density is an allowed density of a quantum mechanical system; it is this fact that builds in the quantum effects.

The difficulty can be illustrated by considering a case where the exact solution can be found – N non-interacting electrons in an external potential. This is the central problem in the Kohn–Sham approach to density functional theory which is discussed in Ch. 7. In that case the *exact* Hohenberg–Kohn functional given by (6.12) is nothing other than the kinetic energy. In order to evaluate the kinetic energy exactly, the only way known is to revert to the usual expression in terms of a set of N wavefunctions. There is no known way to go directly from the density to the kinetic energy. The kinetic energy expressed in terms of wavefunctions has derivatives as a function of the number of electrons that are discontinuous at integer occupation numbers (see Exercise 6.12). From the virial theorem

[5] The dielectric susceptibility is a special case because the ground state is not strictly defined in the presence of an electric field. In an infinite system is it essential to consider the *polarization* in addition to the bulk density in order to have a well-defined thermodynamic limit see Ch. 22.

that relates kinetic and potential energies, it follows immediately that all parts of the exact functional (kinetic and potential) will vary in a *non-analytic* manner as a function of the number of electrons. This is a property of a global integral of the density and is not simply determined from any aspect of the density only in some local region.

In the case of solids, the density is remarkably similar to sums of overlapping atom densities. For example, Fig. 2.2 shows the difference in density of electrons in Si from superposed atoms, which is much smaller than the total density. In fact, the covalent bond is hard to distinguish in the total density. An ionic crystal is often considered as a sum of ions, but it is also well represented as the sum of neutral atoms [344]. This is possible because the negative anion is so large that its density extends around the positive cation, making the density similar to that of neutral atoms. Thus, even for well-known ionic crystals, it is not obvious how to extract pertinent information from the electron density. It is yet more difficult to distinguish metals from insulators (see Exercise 7.15 for an example).

This leads us to the Kohn–Sham approach, the success of which is based upon the fact that it includes the kinetic energy of non-interacting electrons in terms of independent-particle wavefunctions, in addition to interaction terms explicitly modelled as functionals of the density. Because the kinetic energy is treated in terms of orbitals – *not* as an *explicit* functional of the density – it builds in the quantum properties that have no simple relation to the density. In the example of an ionic crystal, the key point is that the density is made up of fermions that obey the exclusion principle. It is this fact that leads to filling of four bands per cell and an insulating gap, which is the essence of this ionic crystal. So long as the true many-body solution is sufficiently close to the independent-particle formulation, e.g. the states must have the same symmetry, then the Kohn–Sham approach provides insightful guidance and powerful methods for electronic structure theory.

SELECT FURTHER READING

Original papers:

Hohenberg, P. and Kohn, W., "Inhomogeneous electron gas," *Phys. Rev.* 136:B864–871, 1964.

Kohn, W. and Sham, L. J., "Self-consistent equations including exchange and correlation effects," *Phys. Rev.* 140:A1133–1138, 1965.

Mermin, N. D., "Thermal properties of the inhomogeneous electron gas," *Phys. Rev.* 137:A1441–1443, 1965.

Book with extensive exposition:

Parr, R. G. and Yang, W. *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

1998 Nobel Prize lecture:

Kohn, W., "Nobel lecture: electronic structure of matter wave functions and density functionals," *Rev. Mod. Phys.* 71:1253, 1999.

Exercises

6.1 Derive the Thomas-Fermi equation, (6.5), from the variational of the functional. Use the method of Lagrange multipliers as given in (3.10) and used in (6.3).

6.2 Derive the Thomas–Fermi–Weizsacker equation which is the generalization of (6.5) when the Weizsacker gradient term is included. The gradient expression is given following (6.5). Variations of a functional of the gradient of the density are discussed in Ch. 8.

6.3 See Exercise 5.14.

6.4 The simplest example of the Mermin theorem is the homogeneous gas. For a gas held at fixed volume, as the temperature is varied the density does not change. Describe the meaning of the Mermin functional in this case.

6.5 Theorem I of Hohenberg–Kohn shows that $n_0(\mathbf{r})$, *in principle*, uniquely determines all properties of the many-body system of electrons, including ground and excited states. We have argued that, for example, the electron density uniquely determines the positions and types of nuclei, which then defines the complete hamiltonian and therefore, *in principle*, determines all properties. Show explicitly that only the density and its derivatives near the nuclei are sufficient to establish the proof in this case.

6.6 In one dimension it is possible to construct orthonormal independent-particle orbitals that describe *any* density that satisfies simple positivity and continuity conditions. See Exercise 7.9.

6.7 Following the approach of Sec. 6.5, show that it is not possible to construct the density of the 2s state of hydrogen (one electron in the potential of a proton) as the ground state density of any smooth potential, i.e. one without delta functions.

6.8 Consider the lowest energy state of Li with three electrons, which may be $1s^2 2s$, or one of the degenerate states $(1s)^2 2p^0$, $1s^2 2p^-$, or $1s^2 2p^+$. The densities of the last two states are identical, so that the density does not determine the state. Show that, nevertheless, the energy is the same for any combination of these states so that the energy is still a functional of the density as needed for the Hohenberg–Kohn functional.

6.9 In this problem you are asked to show that in the absence of an external magnetic field the total density is, in principle, enough to determine all the properties of the system even if it is spin polarized. To do this, consider the system in a Zeeman field $\mathbf{h} \cdot \sigma$ that distinguishes between σ parallel and antiparallel to $\mathbf{h}$. Show that if $\mathbf{h}$ is reversed, the new solution will have exactly the same density, but with σ reversed. Using this fact show that you can reach the desired conclusion.

6.10 Suppose particles can be divided into two types (e.g. spins) of density n_1 and n_2 with internal energy $E_{\text{int}}[n_1, n_2]$. If the external potential acts on n_1 and n_2 equally, the total energy can be written $E_{\text{total}} = E[n_1, n_2] + \int V_{\text{ext}} n$, where $n = n_1 + n_2$. Show that E_{total} is a functional only of n. Do this in three ways: (a) using arguments similar to the original arguments of Hohenberg and Kohn; (b) the Levy–Lieb constrained search method; and (c) formal solution by variational equations in terms of n and $\sigma = n_1 - n_2$.

6.11 Consider a many-body hamiltonian $\hat{H} = \hat{H}_{\text{int}} + V_{\text{ext}}$, where $\hat{H}_{\text{int}}$ denotes all intrinsic internal kinetic and interaction terms and V_{ext} is the external potential. Show that the external potential $V_{\text{ext}}(\mathbf{r})$ is determined to within a constant, given $\hat{H}_{\text{int}}$ and *any* eigenfunction Ψ_i. Hint: solve for $V_{\text{ext}}(\mathbf{r})$ using the Schrödinger equation. (Note a specific example of a determinant wavefunction is considered as an exercise in Ch. 7.)

6.12 Show that in a finite system the kinetic energy must be a non-analytic function of the density n with derivatives that are discontinuous at integer occupations. Hint: It is sufficient to show the result in an independent-particle example (see Exercise 7.5) with an argument that the result must also apply to many-body cases. Generalize this argument to all properties of the system and to solids with an insulating gap.

7

The Kohn–Sham auxiliary system

If you don't like the answer, change the question.

Summary

Density functional theory is the most widely used method today for electronic structure calculations because of the approach proposed by Kohn and Sham in 1965: *to replace the original many-body problem by an auxiliary independent-particle problem*. This is an *ansatz*[1] that, in principle, leads to exact calculations of properties of many-body systems using independent-particle methods; in practice, it has made possible approximate formulations that have proved to be remarkably successful. As a self-consistent method, the Kohn–Sham approach involves *independent particles* but an *interacting density*, an appreciation of which clarifies the way the method is used. The present chapter is devoted to the basic formulation of the Kohn–Sham approach and the ideas behind the crucial ingredient, the exchange–correlation energy functional $E_{\rm xc}[n]$. Information on approximate functionals in widespread use is deferred to Ch. 8, and methods for solution of the Kohn–Sham equations using the functionals are the subjects of Ch. 9 and much of the remainder of this tome.

7.1 Replacing one problem with another

The Kohn–Sham approach is to replace the difficult interacting many-body system obeying the hamiltonian (3.1) with a different *auxiliary system* that can be solved more easily. Since there is no unique prescription for choosing the simpler auxiliary system, this is an *ansatz* that rephrases the issues. The *ansatz* of Kohn and Sham *assumes* that the ground state density of the original interacting system is equal to that of some chosen non-interacting system. This leads to independent-particle equations for the non-interacting system that can be considered exactly soluble (in practice by numerical means) with all the difficult many-body terms incorporated into an *exchange–correlation functional of the density*. By solving the equations one finds the ground state density and energy of the original interacting

[1] Ansatz: attempt, approach. A mathematical assumption, especially about the form of an unknown function, which is made in order to facilitate solution of an equation or other problem [Oxford English Dictionary].

system with the accuracy limited only by the approximations in the exchange–correlation functional.

Indeed, the Kohn–Sham approach has led to very useful approximations that are now the basis of most calculations that attempt to make "first-principles" or "*ab initio*" predictions for the properties of condensed matter and large molecular systems. The local density approximation (LDA) or various generalized-gradient approximations (GGAs) described below, are remarkably accurate, most notably for "wide-band" systems, such as the group IV and II–V semiconductors, sp-bonded metals like Na and Al, insulators like diamond, NaCl, and molecules with covalent and/or ionic bonding. It also appears to be successful for many cases in which the electrons have stronger effects of correlations, such as transition metals. However, these approximations fail for many strongly correlated cases including the copper oxide planar materials which are antiferromagnetic insulators for exactly half-filled bands, whereas the LDA or present GGA functionals find them to be metals [216]. This leads to the present situation in which there is great interest in utilizing and improving the density functional approach: to build upon the many successes of current approximations and to overcome the known deficiencies and failures in strongly correlated electron systems.

Here we will consider the Kohn–Sham *ansatz* for the ground state, which is by far the most widespread way in which the theory has been applied. However, in the big picture *this is only the first step.* The fundamental theorems of density functional theory (Chapter 6) show that *in principle* the ground state density determines *everything*. A great challenge in present theoretical work is to develop methods for calculating excited state properties. We will return to these issues at the end of this chapter, but for the moment we will be concerned only with the theory of the ground state.

The Kohn–Sham construction of an auxiliary system rests upon two assumptions:

1. The exact ground state density can be represented by the ground state density of an auxiliary system of non-interacting particles. This is called *"non-interacting-V-representability*;" although there are no rigorous proofs for real systems of interest, we will proceed assuming its validity. This leads to the relation of the actual and auxiliary systems shown in Fig. 7.1.
2. The auxiliary hamiltonian is chosen to have the usual kinetic operator and an effective *local* potential $V^{\sigma}_{\rm eff}(\mathbf{r})$ acting on an electron of spin σ at point $\mathbf{r}$. The local form is not essential,[2] but it is an extremely useful simplification that is often taken as the defining characteristic of the Kohn–Sham approach. As in Ch. 6, we assume that the external potential $\hat{V}_{\rm ext}$ is spin independent;[3] nevertheless, except in cases that are spin symmetric, the auxiliary effective potential $V^{\sigma}_{\rm eff}(\mathbf{r})$ must depend upon spin in order give the correct density for each spin.

[2] The original paper of Kohn and Sham also proposes an alternative Hartree–Fock-like approach with a non-local orbital-dependent operator for exchange, as in (3.45), to which effects of correlation are added.

[3] Here spin–orbit interactions are ignored.

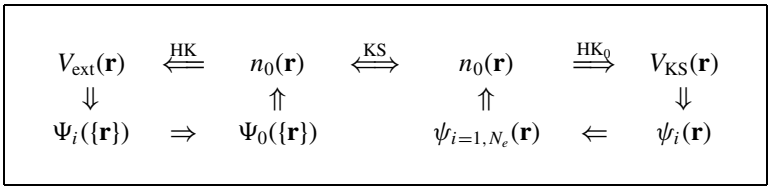


Figure 7.1. Schematic representation of Kohn–Sham *ansatz*. (Compare to Fig. 6.1.) The notation $\rm HK_0$ denotes the Hohenberg–Kohn theorem applied to the non-interacting problem. The arrow labeled KS provides the connection in both directions between the many-body and independent-particle systems, so that the arrows connect any point to any other point. Therefore, in principle, solution of the independent-particle Kohn–Sham problem determines *all properties* of the full many-body system.

The actual calculations are performed on the auxiliary independent-particle system defined by the auxiliary hamiltonian (using Hartree atomic units $\hbar = m_e = e = 4\pi/\epsilon_0 = 1$)

$$\hat{H}^\sigma_{\rm aux} = -\frac{1}{2}\nabla^2 + V^\sigma(\mathbf{r}). \tag{7.1}$$

At this point the form of $V^\sigma(\mathbf{r})$ is not specified and the expressions must apply for all $V^\sigma(\mathbf{r})$ in some range, in order to define functionals for a range of densities. For a system of $N = N^\uparrow + N^\downarrow$ independent electrons obeying this hamiltonian, the ground state has one electron in each of the N^σ orbitals $\psi^\sigma_i(\mathbf{r})$ with the lowest eigenvalues ϵ^σ_i of the hamiltonian (7.1). The density of the auxiliary system is given by sums of squares of the orbitals for each spin

$$n(\mathbf{r}) = \sum_\sigma n(\mathbf{r},\sigma) = \sum_\sigma \sum_{i=1}^{N^\sigma} |\psi^\sigma_i(\mathbf{r})|^2, \tag{7.2}$$

the independent-particle kinetic energy T_s is given by

$$T_s = -\frac{1}{2}\sum_\sigma \sum_{i=1}^{N^\sigma} \langle\psi^\sigma_i|\nabla^2|\psi^\sigma_i\rangle = \frac{1}{2}\sum_\sigma \sum_{i=1}^{N^\sigma} \int \mathrm{d}^3r |\nabla\psi^\sigma_i(\mathbf{r})|^2, \tag{7.3}$$

and we define the classical Coulomb interaction energy of the electron density $n(\mathbf{r})$ interacting with itself (the Hartree energy defined in (3.15))

$$E_{\rm Hartree}[n] = \frac{1}{2}\int \mathrm{d}^3r\,\mathrm{d}^3r' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{7.4}$$

The Kohn–Sham approach to the full interacting many-body problem is to rewrite the Hohenberg–Kohn expression for the ground state energy functional (6.12) in the form

$$E_{\rm KS} = T_s[n] + \int \mathrm{d}\mathbf{r} V_{\rm ext}(\mathbf{r})n(\mathbf{r}) + E_{\rm Hartree}[n] + E_{II} + E_{\rm xc}[n]. \tag{7.5}$$

Here $V_{\rm ext}(\mathbf{r})$ is the external potential due to the nuclei and any other external fields (assumed to be independent of spin) and E_{II} is the interaction between the nuclei (see (3.2)). Thus the sum of the terms involving $V_{\rm ext}$, $E_{\rm Hartree}$, and E_{II} forms a neutral grouping that is

well defined (see Sec. 3.2). The independent-particle kinetic energy T_s is given explicitly as a functional of the orbitals; however, T_s for each spin σ must be a unique functional of the density $n(\mathbf{r}, \sigma)$ by application of the Hohenberg–Kohn arguments applied to the independent-particle hamiltonian (7.1); see Exercise 7.4.

All many-body effects of exchange and correlation are grouped into the exchange–correlation energy $E_{\rm xc}$. Comparing the Hohenberg–Kohn, (6.12) and (6.19), and Kohn–Sham, (7.5), expressions for the total energy (recall that the auxiliary density $n(\mathbf{r}, \sigma)$ of (7.2) is required to equal the true density for each spin σ) shows that $E_{\rm xc}$ can be written in terms of the Hohenberg–Kohn functional (6.13) as

$$E_{\rm xc}[n] = F_{\rm HK}[n] - (T_s[n] + E_{\rm Hartree}[n])\,, \tag{7.6}$$

or in the more revealing form

$$E_{\rm xc}[n] = \langle \hat{T} \rangle - T_s[n] + \langle \hat{V}_{\rm int} \rangle - E_{\rm Hartree}[n]. \tag{7.7}$$

Here $[n]$ denotes a functional of the density $n(\mathbf{r}, \sigma)$ which depends upon both position in space $\mathbf{r}$ and spin σ. One can see that $E_{\rm xc}[n]$ must be a functional since the right-hand sides of the equations are functionals. The latter equation shows explicitly that $E_{\rm xc}$ is just the difference of the kinetic and the internal interaction energies of the true interacting many-body system from those of the fictitious independent-particle system with electron–electron interactions replaced by the Hartree energy.

If the universal functional $E_{\rm xc}[n]$ defined in (7.7), (or $\epsilon_{\rm xc}([n], \mathbf{r})$ in (7.14)), were known, then the exact ground state energy and density of the many-body electron problem could be found by solving the Kohn–Sham equations for independent-particles. To the extent that an approximate form for $E_{\rm xc}[n]$ describes the true exchange–correlation energy, the Kohn–Sham method provides a feasible approach to calculating the ground state properties of the many-body electron system.

7.2 The Kohn–Sham variational equations

Solution of the Kohn–Sham auxiliary system for the ground state can be viewed as the problem of minimization with respect to either the density $n(\mathbf{r}, \sigma)$ or the effective potential $V^{\sigma}_{\rm eff}(\mathbf{r})$ (see Sec. 8.7). Since T_s (7.3) is explicitly expressed as a functional of the orbitals but all other terms are considered to be functionals of the density, one can vary the wavefunctions and use the chain rule to derive the variational equation[4]

$$\frac{\delta E_{\rm KS}}{\delta \psi_i^{\sigma *}(\mathbf{r})} = \frac{\delta T_s}{\delta \psi_i^{\sigma *}(\mathbf{r})} + \left[\frac{\delta E_{\rm ext}}{\delta n(\mathbf{r}, \sigma)} + \frac{\delta E_{\rm Hartree}}{\delta n(\mathbf{r}, \sigma)} + \frac{\delta E_{\rm xc}}{\delta n(\mathbf{r}, \sigma)} \right] \frac{\delta n(\mathbf{r}, \sigma)}{\delta \psi_i^{\sigma *}(\mathbf{r})} = 0, \tag{7.8}$$

subject to the orthonormalization constraints

$$\langle \psi_i^{\sigma} | \psi_j^{\sigma'} \rangle = \delta_{i,j} \delta_{\sigma,\sigma'}. \tag{7.9}$$

[4] Note that even if $E_{\rm xc}$ is explicitly represented as a functional of the wavefunctions (as in the optimized effective potential OEP method, Sec. 8.7), one does *not* use $\delta E_{\rm xc}/(\delta \psi_i^{\sigma *}(\mathbf{r}))$, which would lead to non-local potential operators.

This is equivalent to the Rayleigh–Ritz principle [249, 250] and the general derivation of the Schrödinger equation in (3.10)–(3.12), except for the explicit dependence of $E_{\rm Hartree}$ and $E_{\rm xc}$ on n.

Using expressions (7.2) and (7.3) for $n^\sigma(\mathbf{r})$ and T_s, which give

$$\frac{\delta T_s}{\delta \psi_i^{\sigma *}(\mathbf{r})} = -\frac{1}{2}\nabla^2 \psi_i^\sigma(\mathbf{r}); \quad \frac{\delta n^\sigma(\mathbf{r})}{\delta \psi_i^{\sigma *}(\mathbf{r})} = \psi_i^\sigma(\mathbf{r}), \tag{7.10}$$

and the Lagrange multiplier method for handling the constraints (3.10)–(3.13), this leads to the Kohn–Sham Schrödinger-like equations:

$$(H_{\rm KS}^\sigma - \varepsilon_i^\sigma)\psi_i^\sigma(\mathbf{r}) = 0, \tag{7.11}$$

where the ε_i are the eigenvalues, and $H_{\rm KS}$ is the effective hamiltonian (in Hartree atomic units)

$$H_{\rm KS}^\sigma(\mathbf{r}) = -\frac{1}{2}\nabla^2 + V_{\rm KS}^\sigma(\mathbf{r}), \tag{7.12}$$

with

$$\begin{aligned} V_{\rm KS}^\sigma(\mathbf{r}) &= V_{\rm ext}(\mathbf{r}) + \frac{\delta E_{\rm Hartree}}{\delta n(\mathbf{r},\sigma)} + \frac{\delta E_{\rm xc}}{\delta n(\mathbf{r},\sigma)} \\ &= V_{\rm ext}(\mathbf{r}) + V_{\rm Hartree}(\mathbf{r}) + V_{\rm xc}^\sigma(\mathbf{r}). \end{aligned} \tag{7.13}$$

The meaning of the functional derivatives in the definitions of the Kohn–Sham potential, (7.8) and (7.13), is described in App. A along with illustrative examples. The physical interpretation of $E_{\rm xc}[n]$ and $V_{\rm xc}^\sigma(\mathbf{r})$ is the subject of the following section.

Equations (7.11)–(7.13) are the well-known Kohn–Sham equations, with the resulting density $n(\mathbf{r},\sigma)$ and total energy $E_{\rm KS}$ given by (7.2) and (7.5). The equations have the form of independent-particle equations with a potential that must be found self-consistently with the resulting density. These equations are independent of any approximation to the functional $E_{\rm xc}[n]$, and would lead to the exact ground state density and energy for the interacting system, if the exact functional $E_{\rm xc}[n]$ were known. Furthermore, it follows from the Hohenberg–Kohn theorems (see Exercise 7.3) that the ground state density uniquely determines the potential at the minimum (except for a trivial constant), so that there is a unique Kohn–Sham potential $V_{\rm eff}^\sigma(\mathbf{r})|_{\rm min} \equiv V_{\rm KS}^\sigma(\mathbf{r})$ associated with any given interacting electron system.

Solution of the equations is deferred to later chapters. General aspects of the solution of the self-consistent equations are the subject of Ch. 9. Specific approaches and results are the subject of much of the rest of this volume.

7.3 $E_{\rm xc}$, $V_{\rm xc}$, and the exchange–correlation hole

The genius of the Kohn–Sham approach is that by explicitly separating out the independent-particle kinetic energy and the long-range Hartree terms, the remaining exchange–correlation functional $E_{\rm xc}[n]$ can reasonably be approximated as a local or nearly local

functional of the density. This means that the energy $E_{\rm xc}$ can be expressed in the form

$$E_{\rm xc}[n] = \int {\rm d}\mathbf{r}\; n(\mathbf{r})\epsilon_{\rm xc}([n], \mathbf{r}), \tag{7.14}$$

where $\epsilon_{\rm xc}([n], \mathbf{r})$ is an energy per electron at point $\mathbf{r}$ that depends only upon the density $n(\mathbf{r}, \sigma)$ in some neighborhood of point $\mathbf{r}$.[5] Only the total density appears in (7.14) because the Coulomb interaction is independent of spin; in a spin polarized system, $\epsilon_{\rm xc}([n], \mathbf{r})$ incorporates the information on the spin densities.

Although the energy density $\epsilon_{\rm xc}([n], \mathbf{r})$ is *not uniquely defined by the integral* (7.14), a physically motivated definition of $\epsilon_{\rm xc}([n], \mathbf{r})$ follows from the analysis of the exchange–correlation hole described in Secs. 3.6, 5.1, and 5.2. An informative relation of $\epsilon_{\rm xc}([n], \mathbf{r})$ to the exchange–correlation hole can be found using the "coupling constant integration formula" described in the theoretical background, Chapter 3, which was called "adiabatic connection" by Harris [345].[6] In this case the electronic charge is varied from zero (the non-interacting case) to the actual value (1 in atomic units used here), with the added constraint that the density must be kept constant during this variation. Then all other terms remain constant and the change in energy is given by

$$E_{\rm xc}[n] = \int_0^{e^2} {\rm d}\lambda \langle \Psi_\lambda | \frac{{\rm d}V_{\rm int}}{{\rm d}\lambda} | \Psi_\lambda \rangle - E_{\rm Hartree}[n] = \frac{1}{2} \int {\rm d}^3 r n(\mathbf{r}) \int {\rm d}^3 r' \frac{\bar{n}_{\rm xc}(\mathbf{r}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \tag{7.15}$$

where $\bar{n}_{\rm xc}(\mathbf{r}, \mathbf{r}')$ is the coupling-constant-averaged hole

$$\bar{n}_{\rm xc}(\mathbf{r}, \mathbf{r}') = \int_0^1 {\rm d}\lambda n_{\rm xc}^{\lambda}(\mathbf{r}, \mathbf{r}'). \tag{7.16}$$

Here $n_{\rm xc}(\mathbf{r}, \mathbf{r}')$ is the hole described in Sec. 3.6 summed over parallel ($\sigma = \sigma'$) and antiparallel ($\sigma \neq \sigma'$) spins. Furthermore, the integral in (7.15) involves only the spherical average of the hole density.

Together with (7.14), Eq. (7.15) shows that the exchange–correlation density $\epsilon_{\rm xc}([n], \mathbf{r})$ can be written as

$$\epsilon_{\rm xc}([n], \mathbf{r}) = \frac{1}{2} \int {\rm d}^3 r' \frac{\bar{n}_{\rm xc}(\mathbf{r}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{7.17}$$

This is an important result which shows that the exact exchange–correlation energy can be understood in terms of the potential energy due to the exchange–correlation hole averaged over the interaction from $e^2 = 0$ to $e^2 = 1$. For $e^2 = 0$ the wavefunction is just the independent-particle Kohn–Sham wavefunction so that $n_{\rm xc}^0(\mathbf{r}, \sigma, \mathbf{r}', \sigma') = n_x(\mathbf{r}, \sigma, \mathbf{r}', \sigma')$, where the exchange hole is known from (3.54). Since the density everywhere is required to remain constant as λ is varied, clearly $\epsilon_{\rm xc}([n], \mathbf{r})$ is implicitly a functional of the density in all space. Thus $E_{\rm xc}[n]$ can be considered as an interpolation between the exchange-only and the full correlated energies at the given density $n(\mathbf{r}, \sigma)$.

[5] A polarized insulator is a case where $\epsilon_{\rm xc}(\mathbf{r})$ is *not* a function of the density only in a nearby region. In addition to the density, it is must be a functional of the polarization in the neighborhood of point $\mathbf{r}$. See Sec. 22.1.

[6] An extensive description is given by Parr and Yang [93].

Analysis of the nature of the averaged hole $\bar{n}_{xc}(\mathbf{r}, \mathbf{r}')$ is one of the primary approaches for developing improved approximations for $E_{xc}[n]$. In particular, the exchange–correlation hole obeys a sum rule that its integral must be unity, as shown in Chapter 3. The sum rule is satisfied for any case that is derived from an actual electron hamiltonian and it places constraints on any approximate forms that may be proposed [346]. This and other sum rules [347] are among the primary guidelines for systematic improvement of functionals.

Homogeneous gas

The exchange–correlation hole in the homogeneous electron gas has been presented in Ch. 5. The results are relevant here because they present representative cases from weak to strong correlation and they are the basis for the local density approximation. In the non-interacting limit there is no correlation between electrons of different spin and the hole is purely the exchange hole given by (5.19) and shown in Fig. 5.3. At full coupling strength the hole has been calculated by quantum Monte Carlo methods, with results shown in Fig. 5.5. The average hole is some mean between the two, which can also be found by an appropriate average of the holes from high density (where correlation is negligible) to the actual density. The key point is that Fig. 5.5 allows one to have a feeling for the radial shapes and the characteristic extent of the exchange–correlation hole.

Atoms

The holes have also been calculated in other systems. In general, of course, the hole is dependent upon the electron position and is non-spherical; however, for the energy, only spherical average is needed. In small systems such as atoms, the correlations can be calculated essentially exactly by configuration interaction methods. For example, Gunnarsson et al. [348] have found $n_x(\mathbf{r}, \mathbf{r}')$ for the neon atom as shown in Fig. 7.2, which illustrates the fact that the hole is extremely non-spherical, and yet the spherical average, shown on the right, is quite similar to the hole in the homogeneous electron gas with density equal to the local density at the point chosen.

Solids

There are very few quantitative calculations in solids; an example is shown in Figs. 7.3 and 7.4 for Si determined by a quantum Monte Carlo simulation with a chosen variational wavefunction [349]. The figures show separately the exchange and correlation holes, demonstrating the basic fact that the exchange dominates over correlation. This is a consequence of the sum rule that it integrates to 1 and the fact that its contribution to the energy is largely to remove the self-interaction term in the Hartree interactions. Despite the fact that the hole varies greatly from high-density bond-center regions to low-density interstitial regions, the spherical average is given reasonably well by the local density approximation. However, the large difference in the interstitial region shown in Fig. 7.4 indicates possible sources of inaccuracies. Since the hole obeys a sum rule, the deepening at short range due

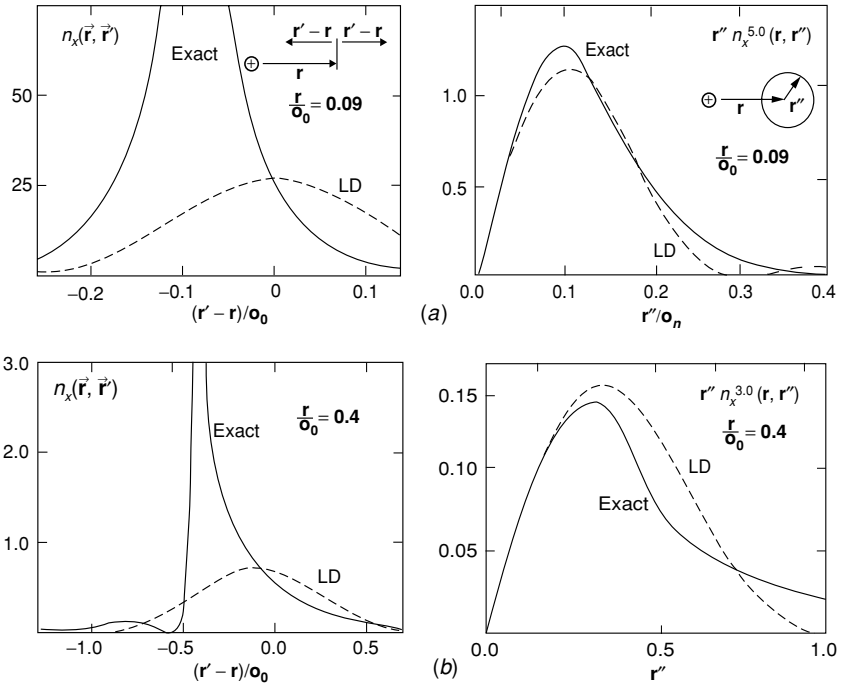


Figure 7.2. Exchange hole in a Ne atom. Left: $n(\mathbf{r}, \mathbf{r}')$ plotted for two values of $|\mathbf{r}|$ as a function of $|\mathbf{r}' - \mathbf{r}|$ along a line through the nucleus, and compared to the local density approximation. The origin is centered on an electron a distance $|r|$. All quantities are in units of the Bohr radius, a_0. Right: The spherical average, as a function of the relative distance, which shows the close resemblance to the local density approximation. From Gunnarsson et al. [348].

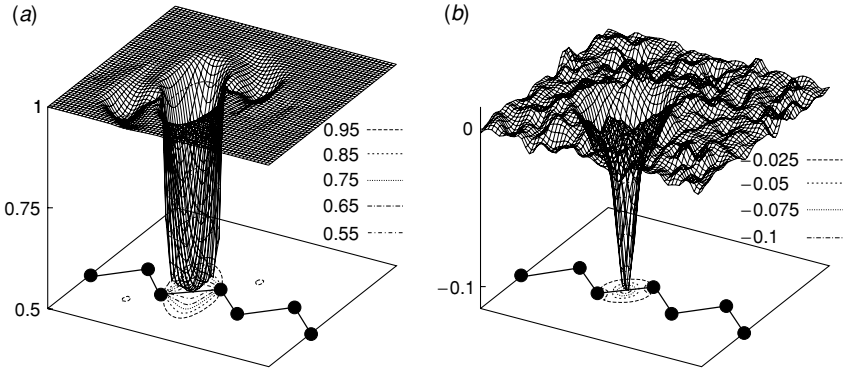


Figure 7.3. Exchange (a) and coupling-constant-averaged correlation hole (b) for an electron at the bond center in Si, calculated by a variational Monte Carlo method. Note the much smaller scale for the correlation hole. From Hood et al. [349].

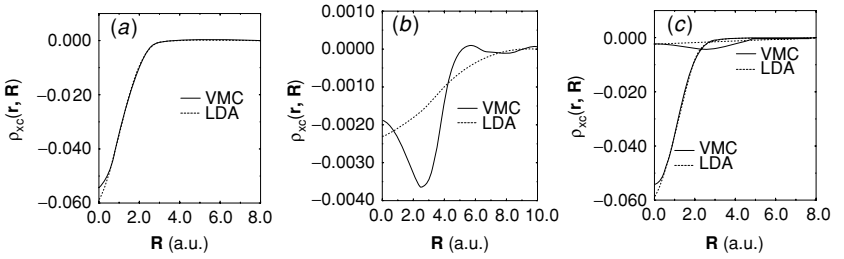


Figure 7.4. Spherical and coupling-constant-averaged exchange–correlation hole in Si, calculated as in Fig. 7.3, compared with the LDA approximation. Left: Hole around an electron at the bond center. Middle: Hole around an electron at the interstitial hole. Right: a comparison to scale. From Hood, et al. [349].

to correlation must be offset by a decrease at large range, i.e. screening that effectively decreases the range of correlation.

Exchange–correlation potential V_{xc}

The exchange–correlation potential $V_{xc}^{\sigma}(\mathbf{r})$ is the functional derivative of E_{xc}, which can be written as

$$V_{xc}^{\sigma}(\mathbf{r}) = \epsilon_{xc}([n], \mathbf{r}) + n(\mathbf{r})\frac{\delta\epsilon_{xc}([n], \mathbf{r})}{\delta n(\mathbf{r}, \sigma)}, \tag{7.18}$$

where $\epsilon_{xc}([n], \mathbf{r})$ is defined in (7.14), and is a functional of the density $n(\mathbf{r}', \sigma')$. It is instructive to examine the properties that the exact V_{xc} must satisfy. First, it is *not* a potential that can be identified with interactions between particles and it behaves in ways that seem paradoxical. Expression (7.18) illustrates such properties: the second term (sometimes called the "response potential" [350]) is due to the *change* in the exchange–correlation hole with density. In an insulator, this derivative is discontinuous at a band gap where the nature of the states changes discontinuously as a function of n. This leads to a "derivative discontinuity" whereby the Kohn–Sham potential for *all* the electrons in a crystal changes by a constant amount when a single electron is added [351, 352]. Thus even in the exact Kohn–Sham theory, the difference between the highest occupied and lowest unoccupied eigenvalues should *not* equal the actual band gap. Similarly, there can be a shift in absolute energies of states of one molecule due to the presence of another molecule far away [353].

The behavior of the Kohn–Sham potential as a function of density seems paradoxical. How can adding one electron shift the potential for all the other electrons in a solid? The answer is in the definition of the functional and the behavior can be understood from examination of the kinetic energy. The great advance of the Kohn–Sham approach over the Thomas–Fermi approximation is the incorporation of orbitals to define the kinetic energy. In terms of orbitals, it is easy to see that the kinetic energy T_s for independent particles in (7.3) changes discontinuously in going from an occupied to an empty band, since the $\psi_i^{\sigma}(\mathbf{r})$ are different for different bands. In terms of the density this means the formal density functional $T_s[n]$ has discontinuous derivatives at densities that correspond to filled bands. This is a

direct consequence of quantum mechanics and is not paradoxical; the real problem is that it is difficult to incorporate into an explicit density functional. It is likewise straightforward to see that the true exchange–correlation functional must change discontinuously. None of these properties is incorporated in any of the simple explicit functionals of the density, such as the local density or gradient approximations (Secs. 8.1 and 8.2); however, they occur naturally (and are not paradoxical) in terms of orbital-dependent formulations, such as the OEP (Sec. 8.7).

A different way to see the properties is to note that the Kohn–Sham potential $V_{\rm KS}$ is *defined* by the requirement that it yield the exact charge density. This is an exacting requirement that must be accomplished by the properties of $V_{\rm xc}$, since all the other terms in $V^\sigma_{\rm KS}(\mathbf{r}) = V_{\rm ext}(\mathbf{r}) + V_{\rm Hartree}(\mathbf{r}) + V^\sigma_{\rm xc}(\mathbf{r})$, (7.13), are known or are simple explicit functionals of the density. Thus one way to determine $V^\sigma_{\rm xc}(\mathbf{r})$ is the requirement that $V^\sigma_{\rm KS}(\mathbf{r})$ lead to the exact density. Conversely, the application of the Hohenberg–Kohn theorem to the Kohn–Sham non-interacting system implies that the exact density can be fit by only one $V^\sigma_{\rm xc}(\mathbf{r})$, which is unique except for an additive constant.

7.4 Meaning of the eigenvalues

It is often said that Kohn–Sham eigenvalues have no physical meaning. Indeed, the eigenvalues are *not* the energies to add or subtract electrons from the interacting many-body system. There is only one exception [354]: the highest eigenvalue in a finite system, which is minus the ionization energy, $-I$. The asymptotic long-range density of a bound system is governed by the occupied state with highest eigenvalue; since the density is assumed to be exact, so must the eigenvalue be exact. No other eigenvalue is guaranteed to be correct by the Kohn–Sham construction.

Nevertheless, the eigenvalues have a well-defined meaning within the theory and they can be used to construct physically meaningful quantities. One approach is the development of perturbation expressions for excitation energies starting from the Kohn–Sham eigenfunctions and eigenvalues. This can take the form of a functional [355] or it can be an operational definition, such as an explicit many-body calculation that uses the Kohn–Sham eigenfunctions and eigenvalues as input. The latter is actually done in quantum Monte Carlo and many-body perturbation approaches (for reviews see, respectively, [81] and [82]). For example, the most accurate calculations at the present time for gaps in solids are based upon fixed-node diffusion Monte Carlo, where the resulting energies are a functional only of the nodes of the many-body trial function. If the trial function is taken to be a determinant made of Kohn–Sham orbitals [81], each result is operationally a functional of the Kohn–Sham potential.

Within the Kohn–Sham formalism itself, the eigenvalues have a definite mathematical meaning, often known as the Slater–Janak theorem [356]. The eigenvalue is the derivative of the total energy with respect to occupation of a state

$$\varepsilon_i = \frac{{\rm d}E_{\rm total}}{{\rm d}n_i} = \int {\rm d}\mathbf{r} \frac{{\rm d}E_{\rm total}}{{\rm d}n(\mathbf{r})} \frac{{\rm d}n(\mathbf{r})}{{\rm d}n_i}. \tag{7.19}$$

For a non-interacting system this is trivial. However, for the Kohn–Sham problem it raises interesting points. The exchange–correlation energy is a functional of the density and the derivative of the potential terms in $\mathrm{d}E_{\mathrm{total}}/\mathrm{d}n(\mathbf{r})$ in (7.19) is the effective potential $V_{\mathrm{xc}}(\mathbf{r})$ in (7.18). As pointed out following that equation, $V_{\mathrm{xc}}(\mathbf{r})$ contains a "response part" that is the derivative of $\epsilon_{\mathrm{xc}}([n], \mathbf{r})$ with respect to $n(\mathbf{r})$. This can vary discontinuously between states giving rise to jumps in eigenvalues that are at first surprising. This is the well-known "band-gap discontinuity" [351, 352].

Thus it follows that for the critical problem of the gap in an insulator, the eigenvalues of the *ground state Kohn–Sham potential should not be the correct gap*, at least in principle. However, the magnitude of the discontinuity has not been established and there is active research especially using "optimized effective potentials" (Sec. 8.7) to clarify the issues regarding electron addition and removal energies.

7.5 Intricacies of exact Kohn–Sham theory

This section asks similar questions of Kohn–Sham theory as were asked in Sec. 6.5 of Hohenberg–Kohn density functional theory. In some cases the answers are the same and will be abbreviated here, but in other cases the difference in the answers is fundamental for understanding practical forms of density functional theory.

Allowed densities for electrons

Since the Hohenberg–Kohn theorems also apply to independent-particle problems, the reasoning of Sec. 6.5 shows that:

- One can construct different wavefunctions ψ_i that have the same density $n(\mathbf{r})$.
- An antisymmetric wavefunction for fermions can describe any possible density ("N-representability") with some analyticity conditions.
- It is *not* possible to generate any reasonable density as the ground state of some local external potential ("V-representability"). One example is a linear combination of densities of a set of degenerate states. A second is the density corresponding to an excited state of a potential, which cannot be the ground state of another potential if it is required not to have singularities. (The example of a 2s state in H is discussed in Exercise 6.7).

The new question is:

- For any ground state density of an *interacting* electron system, is it possible to reproduce the density exactly as the ground state density of a *non-interacting* electron system ("non-interacting-V-representability")?
  The answer is not known. This is the Kohn–Sham *ansatz*, which is the basis for the entire industry, but it has never been proven in general. It is obviously true for the homogeneous gas; it can be demonstrated easily for any one- or two-electron problem (see Exercises 7.2 and 7.13); and it has been shown by Kohn and Sham [92] for small deviations from

the homogeneous gas (Exercise 7.10); but, to the knowledge of the author, there are no general proofs. Nevertheless, results of calculations appear very "reasonable" and detailed tests have shown that it is possible to fit the best numerical densities in many cases. We will follow the standard practice and proceed under the *assumption* that the Kohn–Sham *ansatz* is either valid or is good enough to be worth all this effort. The definition of "exact Kohn–Sham theory" followed here is *exact – assuming that it exists*.

Properties obeyed by "exact Kohn–Sham theory"

The Kohn–Sham approach places even heavier emphasis on the ground state than the Hohenberg–Kohn theorems. The only properties guaranteed to be correct by construction in the exact Kohn–Sham theory are the density and the energy. Thus questions arise as to what properties of a material should be given correctly by Kohn–Sham theory, if the exchange correlation functional was known exactly.

- Is the spin density correct in Kohn–Sham theory?
  Yes. A spin-dependent effective potential is introduced specifically to give the correct density and spin density. Non-collinear spin functionals (Sec. 8.4) allow the proper rotation invariance, which is broken in theories that fix only the z-component of the spin.
- Are static charge and spin susceptibilities given correctly by the ground state functional?
  Yes. All static susceptibilities are second derivatives of ground state energies with respect to external fields. Thus they must be given correctly by the variation of the ground state Kohn–Sham functional as functions of external fields.[7]
- Is the macroscopic polarization in a crystal given correctly by the Kohn–Sham theory in terms of the density $n(\mathbf{r})$ in the bulk of the crystal?
  No. It has long been known that the polarization could not be derived simply from the density. Recent developments derive the polarization from the *phases* of the wavefunctions, not given correctly by the Kohn–Sham orbitals (see Ch. 22).
- Is the exact Fermi surface of a metal given by eigenvalues in the exact Kohn–Sham theory?
  No. Even though the density is reproduced, the Fermi surface may not be correct due to the requirement of a local potential [357].
- Must a Mott insulator – an insulator due to correlations among the electrons – be predicted correctly by the eigenvalues in the exact Kohn–Sham theory?
  No. This follows from the above arguments on a metal that the Fermi surface is not correct in general.
- Are excitation energies given correctly by the eigenvalues of the Kohn–Sham equations?
  No. The eigenvalues are not the true energies for adding or subtracting electrons, nor for neutral excitations (see Sec. 7.4).
- Is any excitation energy given correctly by an eigenvalue of the Kohn–Sham equations?

[7] The dielectric susceptibility is a special case and care must be taken to describe the electric polarization properly. There is a term outside the usual Kohn–Sham theory related to the following question and described in Ch. 22.

Yes. The highest eigenvalue in a finite system must be correct [354] since that state dominates the long-range tail of the density, which is defined to be correct.

- Is the exact specific heat versus temperature given correctly by the exact finite temperature Mermin functional?
  Yes. Even though the specific heat involves excitations from the ground state, nevertheless the thermal averages over these excitations must be a unique functional of the density and the temperature. However, it is more difficult to derive the exchange–correlation functional as function of temperature.
- Is it possible to determine excitation energies by any means using the Kohn–Sham theory?
  Yes. This question is in the spirit of the Hohenberg–Kohn existence proofs. Since the Kohn–Sham density is exact by construction, it follows from the Hohenberg–Kohn theorems that *all properties are determined* since the entire hamiltonian is determined. Thus there should be some way to use the Kohn–Sham potential and eigenfunctions to determine all excitations exactly, but this requires a theory beyond the naive use of Kohn–Sham eigenvalues. One approach is to use the eigenstates as the basis for a many-body calculation, which is literally done in configuration interaction, Monte Carlo [81], and many-body perturbation theory calculations [82]. In finite systems, the "ΔSCF" (Sec. 10.6) calculation of energy differences is a practical approach. Other formulations bring excitations into the fold of the Kohn–Sham approach itself, most importantly, time-dependent Kohn–Sham theory.

7.6 Time-dependent density functional theory

The Kohn–Sham *ansatz* replaces the many-body problem with an independent-particle problem, in which the effective potential depends on the density. Thus the Kohn–Sham approach involves *independent particles* but an *interacting density.* As discussed in Sec. 7.4, the eigenvalues of the Kohn–Sham equations are independent-particle eigenvalues that do not correspond to true electron removal or addition energies. Similarly, eigenvalue differences do not correspond to excitation energies.

How can the Kohn–Sham approach properly describe excitations? The answer is to return to the formulation in terms of the interacting density. In the full many-body problem, excitations are most readily described in terms of the response functions, i.e. the response of the system to external perturbations. The excitation energies in the response in (3.60) are the exact many-body excitation energies. Following the analysis of frequency-dependent dynamical response functions in App. D, the exact density response function has poles as a function of frequency ω at the exact excitation energies. Therefore, the goal is to construct a theory of the dynamical density response function within the Kohn–Sham framework.

Such a theory exists: "time-dependent Kohn–Sham density functional theory" is a remarkably simple generalization of the original static Kohn–Sham method [231, 358, 359]. The ideas are similar to the time-dependent Hartree–Fock approximation, which has a long history [318, 360], and were perhaps first used by Ando [361] for model problems in semiconductors and by Zangwill and Soven [229] for atoms. The formal theory due to Runge and Gross [230] is, in principle, exact for finite systems, as described in Sec. 6.4. The

time-dependent Kohn–Sham equations can be derived from the stationarity principle for the action [230].

$$\frac{\delta A}{\delta n(\mathbf{r}, t)} = 0, \tag{7.20}$$

where

$$A = \int_{t_0}^{t_1} \mathrm{d}t \langle \Psi(t)| \left[i \frac{\mathrm{d}}{\mathrm{d}t} - \hat{H}(t) \right] |\Psi(t)\rangle. \tag{7.21}$$

If one adds the Kohn–Sham idea of replacing the density with the density of independent particles, this leads to time-dependent Kohn–Sham density functional theory (TDDFT), in which there is a time-dependent Schrödinger-like equation

$$i\hbar \frac{\mathrm{d}\psi_i(t)}{\mathrm{d}t} = \hat{H}(t)\psi_i(t), \tag{7.22}$$

with an effective hamiltonian that depends upon time t

$$\hat{H}_{\mathrm{eff}}(t) = -\frac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\mathbf{r}, t) + \int \frac{n(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r}' + V_{\mathrm{xc}}[n](\mathbf{r}, t), \tag{7.23}$$

where $V_{\mathrm{xc}}[n](\mathbf{r}, t)$ is a function of $\mathbf{r}$ and t and a *functional* of $n(\mathbf{r}', t')$. Note that in the formally exact theory, $V_{\mathrm{xc}}[n](\mathbf{r}, t)$ is a functional of $n(\mathbf{r}', t')$ for *all earlier times* $t' \leq t$. The difficult problem is the construction of useful functionals that incorporate effects of non-locality in time. Essentially all work to date (see Ch. 20) uses the *adiabatic approximation* in which the exchange–correlation potential $V_{\mathrm{xc}}[n(t)](\mathbf{r})$ depends only upon the density at the same time, e.g. in the adiabatic LDA (ALDA), it is simply $V_{\mathrm{xc}}(\mathbf{r}, t) = V_{\mathrm{xc}}(n(\mathbf{r}, t))$.

The challenges in TDDFT are closely related to other issues. For example, the difficulty in going beyond the adiabatic approximation can be illustrated by a system driven near a resonance, where the functional should take into account the particular states involved in the transition. This is an extension of the problem in the time-independent theory of including orbital dependent effects, not easily captured in the density alone. Another issue is in extended systems where the evolution of the system must be regarded as a functional of the current density [333, 335, 337, 362]. This is an extension of the problem of polarization (Ch. 22) to the time-independent theory.

7.7 Other generalizations of the Kohn–Sham approach

The overarching guiding principle of the Kohn–Sham approach is the replacement of the full many-body problem with a simpler problem. In the usual Kohn–Sham theory of (7.1), the simpler problem is a system of non-interacting particles chosen to reproduce *only* the correct ground state density and energy. In this framework, the eigenvalues and eigenfunctions do not correspond to actual excitations, except the highest eigenvalue of a localized system. However, this is *not* essential: the density is supposed to determine *everything*. Why not

require that other properties of the Kohn–Sham system are equal to the exact values? For example, the ground state energy and density and *also the band gap*?

The essence of a "generalized Kohn–Sham theory" is that *many* possible mappings can be made of the full interacting problem onto simpler auxiliary systems. For example, the auxiliary systems could include certain interactions instead of the Kohn–Sham choice of a non-interacting system.[8] A general approach for requiring that the auxiliary system reproduce the density and some other quantity has been outlined by Jansen [363]. The simplest example is spin density theory which includes the spin density as well as number density. A recent example is the "density polarization theory" [340–342] pointed out in Ch. 22. A primary aim of on-going research is the prediction of band gaps including the derivative discontinuity; there are both formal approaches showing existence proofs [355] and practical approaches that involve approximate forms that are explicit functionals of the wavefunctions (Sec. 8.7) that are promising for description of excitations as well as ground state properties. (See, e.g., [364] and references therein.)

SELECT FURTHER READING

Original papers:

Hohenberg, P. and Kohn, W., "Inhomogeneous electron gas," *Phys. Rev.* 136:B864–871, 1964.

Kohn, W. and Sham, L. J., "Self-consistent equations including exchange and correlation effects," *Phys. Rev.* 140:A1133–1138, 1965.

Mermin, N. D., "Thermal properties of the inhomogeneous electron gas," *Phys. Rev.* 137:A1441–1443, 1965.

Books with extensive exposition:

Parr, R. G. and Yang, W. *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

Dreizler, R. M. and Gross, E. K. U., *Density Functional Theory: An Approach to the Quantum Many-Body Problem* (Springer, Berlin, 1990).

Review:

Jones, R. O. and Gunnarsson, O. "The density functional formalism, its applications and prospect," *Rev. Mod. Phys.* 61:689–746, 1989.

Edited collections:

Density Functional Methods in Physics, edited by R. M. Dreizler and J. da Providencia, Plenum, New York, 1985.

Density Functional Theory, edited by E. K. U. Gross and R. M. Dreizler, Plenum, New York, 1995.

Theory of the Inhomogeneous Electron Gas, edited by S. Lundqvist and N. H. March, Plenum, New York, 1983.

[8] The original Kohn–Sham paper [92] points out that the choice of a local potential is not the only possibility; for example, one could choose a non-local Hartree–Fock-like exchange operator.

Exercises

7.1 For any one-electron problem, one can readily determine whether or not any given density is a possible ground state density. Using the known properties of solutions of the Schrödinger equation, give a sufficient set of conditions that any function must satisfy in order to guarantee that it is the ground state density of some potential. See Exercise 7.7 for an example of an allowed density and Exercise 6.7 for a function that is not an allowed ground state density.

7.2 For any density $n(\mathbf{r})$ that is allowed (see Exercise 7.1) and integrates to one electron, show that the Kohn–Sham potential $V^{\sigma}_{\text{eff}}(\mathbf{r})|_{\text{min}} \equiv V^{\sigma}_{\text{KS}}(\mathbf{r})$ is unique, except for an arbitrary constant, and give an explicit algorithm for constructing $V^{\sigma}_{\text{KS}}(\mathbf{r})$ from $n(\mathbf{r})$. See Exercise 7.7 for an explicit example of an allowed density.

7.3 Generalize the arguments of Exercise 7.2 to show that $V^{\sigma}_{\text{KS}}(\mathbf{r})$ is unique, except for an arbitrary constant, for a non-interacting Kohn–Sham system of any integer number of electrons.

7.4 For any non-interacting Kohn–Sham system, use the result of Exercise 7.3 to show that the kinetic energy T_s for each spin σ must be a unique functional of the density $n(\mathbf{r}, \sigma)$ for that spin. Generalize the argument to show that *all* properties of the system are uniquely determined by the density.

7.5 Based upon the result of Exercise 7.4, show that in a finite system with discrete states the kinetic energy functional $T_s[n]$ must be a non-analytic function of the density n with derivatives that are discontinuous at integer occupations. Hint: Use the known solutions of the Schrödinger equation, ψ_i that are different for each i. Generalize this argument to all properties of the system and to filled bands in the case of a solids.

7.6 As an example of the fact that arbitrary densities *cannot* be constructed from the lowest eigenstates of a non-interacting hamiltonian, see Exercise 6.7. Use this example as the basis for constructing a general argument that it is not possible to construct any density from a determinant formed from the lowest N eigenvectors of a non-interacting particle problem.

7.7 As an example of the explicit construction of a potential determined by the density, find the one-dimensional potential $V(x)$ that gives the density $A\exp(-\alpha x^2)$, where normalization constant A is chosen so that the density corresponds to one electron. Express the answer in terms of α.

7.8 For a one-electron radial problem it is straightforward to find the unique Kohn–Sham potential that will lead to any radial density with no nodes. (The Schrödinger equation in radial coordinates is given in Sec. 10.1.)
(a) Find the potential $V_{\text{KS}}(r)$ that gives the hydrogen atom density.
(b) Find the potential for a gaussian density $A\exp(-\alpha r^2)$, where A is a normalization constant chosen so that the density integrates to one (See also Exercise 7.7.). Express the answer in terms of α.

7.9 This problem is an example of explicit construction of orthonormal independent-particle orbitals that describe *any* density of N particles and, furthermore, that there are many such choices for the same density. This example is for one dimension and is taken from p. 55 of [93]. For a density $n(x)$ and $s(x) \equiv n(x)/N$ given in the range $x_1 \le x \le x_2$, define the set of functions

$$\psi_k(x) = [s(x)]^{1/2} \exp\left[i2\pi k q(x)\right], \tag{7.24}$$

with $q(x) \equiv \int_{x_1}^{x} s(x')\mathrm{d}x'$ and k = integers or half-integers. Show that the orbitals satisfy the desired conditions since each has the same density $s(x)$ and the orbitals are orthonormal. Show that it follows that an infinite number of such choices can be made.

7.10 Show that, to lowest-order, small deviations from the homogeneous density can be reproduced by non-interacting fermions. Hint: Use the fact that, to lowest order, any change in the density is linear in the potential.

7.11 It is interesting to note that construction of a kinetic energy functional of the density is a "fermion problem." For non-interacting bosons, construct explicitly a practical, exact density functional theory.

7.12 Consider an independent-particle hamiltonian $\hat{H} = \hat{H}_{\text{int}} + V_{\text{ext}}$ for which the wavefunction for any state i is a single determinant Φ_i and the subscript "int" denotes all internal terms. Then the total energy can be written $E_{\text{tot}} = E_{\text{int}}[\Phi] + \int \mathrm{d}^3\mathbf{r} V_{\text{ext}}(\mathbf{r}) n(\mathbf{r})$. Show that the external potential $V_{\text{ext}}(\mathbf{r})$ is determined to within a constant given $\hat{H}_{\text{int}}$ and any eigenfunction Φ_i, not only the ground state. (Hint: Solve for $V_{\text{ext}}(\mathbf{r})$ using the Schrödinger equation.) Explain why it is more difficult numerically to find $V_{\text{ext}}(\mathbf{r})$ from the wavefunction for an excited state than for the ground state.

7.13 For a two-electron problem in a singlet state, it is straightforward to find the Kohn–Sham potential that will lead to any density with no nodes. The purpose of this exercise is to emphasize the relation to the one-electron case in Exercise 7.8 by constructing the potential $V_{\text{KS}}(r)$ for the following cases:
(a) a density that is twice that of the H atom;
(b) a gaussian density $A \exp(-\alpha r^2)$, where A is chosen so that the density integrates to two electrons.

7.14 Project: Using an atomic program (such as the one discussed in conjunction with Ch. 10) one can find the density of a closed-shell atom and the Kohn–Sham potential.
(a) This exercise is to invert the problem: construct a minimization program to find the potential $V(r)$ that will produce that density and show that it is the same potential. This is essential for the potential to be unique.
(b) Now modify the density by multiplying by a gaussian and normalizing. For this density find the potential.

7.15 Project: Use the empirical pseudopotential program (Sec. 12.6) to find the bands and charge densities of Si in the diamond structure at the lattice constant 5.431 Å. The bands should be insulating and the bonds should be visible in the charge density.
(a) Now compress the system until it is metallic (this can only be done in theory; in reality it transforms). Can you tell when the system becomes a metal just from the density? In principle, if you had the exact functional, what aspect of the density would be the signature of the insulator–metal transition?
(b) Do a similar calculation replacing the Si atoms with Al, still in the diamond structure with lattice constant 5.431 Å. (Of course this is a theoretical structure.) There are three Al electrons/atom, i.e. six electrons per cell, and it turns out to be a metal. Show that it must be metallic *without doing the calculation*. Does the density plot look a lot like Si? Can you find any feature in the density that shows it is a metal?

8

Functionals for exchange and correlation

Functional functionals

Summary

Density functional theory is the most widely used method today for electronic structure calculations because of the success of practical, approximate functionals. The crucial quantity in the Kohn–Sham approach is the exchange–correlation energy which is expressed as a functional of the density $E_{\rm xc}[n]$. This chapter is devoted to relevant approximate functionals, in particular, the local density approximation (LDA) and examples of generalized-gradient approximations (GGAs). Explicit formulas for certain widely used functionals are given in App. B. Non-local formulations are an active area of research leading to new classes of functionals, in particular, orbital-dependent functionals including the "optimized effective potential" (OEP) method and "hybrid functionals." Important features are illustrated by a few selected results on atoms and molecules.

As emphasized in the previous chapter, the genius of the Kohn–Sham approach is two-fold: first, the construction of an auxiliary system leads to tractable independent-particle equations that hold the hope of solving interacting many-body problems. The famous Kohn–Sham equations are given in (7.11)–(7.13). Second, and perhaps more important, by explicitly separating out the independent-particle kinetic energy and the long-range Hartree terms, the remaining exchange–correlation functional $E_{\rm xc}[n]$ can be reasonably approximated as a local or nearly local functional of the density. Even though the exact functional $E_{\rm xc}[n]$ must be very complex, great progress has been made with remarkably simple approximations. This chapter is devoted to those approximations.

8.1 The local spin density approximation (LSDA)

Already in their seminal paper, Kohn and Sham pointed out that solids can often be considered as close to the limit of the homogeneous electron gas. In that limit, it is known that the effects of exchange and correlation are local in character, and they proposed making the

local density approximation (LDA) (or more generally the local spin density approximation (LSDA)), in which the exchange–correlation energy is simply an integral over all space with the exchange–correlation energy density at each point assumed to be the same as in a homogeneous electron gas with that density,

$$E_{\rm xc}^{\rm LSDA}[n^\uparrow, n^\downarrow] = \int {\rm d}^3 r n(\mathbf{r}) \epsilon_{\rm xc}^{\rm hom}(n^\uparrow(\mathbf{r}), n^\downarrow(\mathbf{r})) \\ = \int {\rm d}^3 r n(\mathbf{r}) [\epsilon_x^{\rm hom}(n^\uparrow(\mathbf{r}), n^\downarrow(\mathbf{r})) + \epsilon_c^{\rm hom}(n^\uparrow(\mathbf{r}), n^\downarrow(\mathbf{r}))]. \tag{8.1}$$

(Here the axis of quantization of the spin is assumed to be the same at all points in space, but this can be generalized as in Sec. 8.4.) The LSDA can be formulated in terms of either two spin densities $n^\uparrow(\mathbf{r})$ and $n^\downarrow(\mathbf{r})$, or the total density $n(\mathbf{r})$ and the fractional spin polarization $\zeta(\mathbf{r})$ defined in (5.16),

$$\zeta(\mathbf{r}) = \frac{n^\uparrow(\mathbf{r}) - n^\downarrow(\mathbf{r})}{n(\mathbf{r})}. \tag{8.2}$$

The LSDA (with the generalization to non-collinear spins in Sec. 8.4) is the most general local approximation and is given explicitly by (5.17) and (5.18) for exchange and by approximate (or fitted) expressions given in Sec. 5.2 for correlation. For unpolarized systems the LDA is found simply by setting $n^\uparrow(\mathbf{r}) = n^\downarrow(\mathbf{r}) = n(\mathbf{r})/2$.

Once one has made the local approximation of the L(S)DA, then all the rest follows. Since the functional $E_{\rm xc}[n^\uparrow, n^\downarrow]$ is universal, it follows that it is exactly the same as for the homogeneous gas. The only information needed is the exchange–correlation energy of the homogeneous gas as a function of density; the exchange energy of the homogeneous gas is given by a simple analytic form (Ch. 5) and the correlation energy has been calculated to great accuracy with Monte Carlo methods [297]. Variations of exchange and correlation energies with density are discussed in Ch. 5 (where they are compared with insightful approximations), and explicit analytic forms fitted to the numerical results are given in App. B. As long as there are no further approximations in the calculations, the results of LDA and LSDA calculations can be considered as tests of the local approximation itself; the local approximation lives or dies depending upon how the answers agree with experiment (or with many-body calculations that can be considered essentially exact).

The rationale for the local approximation is that for the densities typical of those found in solids, the range of the effects of exchange and correlation is rather short, as discussed for the "exchange–correlation hole" described in the previous chapter. However, this is not justified by a formal expansion in some small parameter, and one must test the extent to which it works by actual applications. We expect it will be best for solids close to a homogeneous gas (like a nearly-free-electron metal) and worst for very inhomogeneous cases like atoms where the density must go continuously to zero outside the atom.

Among the most obvious faults is the spurious self-interaction term. In the Hartree–Fock approximation the unphysical self-term in the Hartree interaction is exactly cancelled by the non-local exchange interaction. However, in the local approximation to exchange, the

cancellation is only approximate and there remain spurious self-interaction terms that are negligible in the homogeneous gas but large in confined systems such as atoms. Nevertheless, even in very inhomogeneous cases, the LSDA works remarkably well. One reason for the success is that the hole obeys all the sum rules since it is the exact hole for some hamiltonian, even if it is not the correct hamiltonian [314]. Thus the hole satisfies constraints imposed by the sum rules that are difficult to satisfy if one makes arbitrary approximations. Furthermore, the detailed shape of the hole need not be correct since only the spherical average of the xc hole enters the energy.

The degree to which the LSDA is successful has made it useful in its own right, and has stimulated ideas for constructing improved functionals (such as the GGAs described in Sec. 8.2).

8.2 Generalized-gradient approximations (GGAs)

The success of the LSDA has led to the development of various Generalized-gradient approximations (GGAs) with marked improvement over LSDA for many cases. Widely used GGAs can now provide the accuracy required for density functional theory to be widely adopted by the chemistry community. In this section we briefly describe some of the physical ideas that are the foundation for construction of GGAs.

The first step beyond the local approximation is a functional of the magnitude of the gradient of the density $|\nabla n^\sigma|$ as well as the value n at each point. Such a "gradient expansion approximation" (GEA) was suggested in the original paper of Kohn and Sham, and carried out by Herman et al. [369] and others. The low-order expansion of the exchange and correlation energies is known [370]; however, the GEA does *not* lead to consistent improvement over the LSDA. It violates the sum rules and other relevant conditions [369] and, indeed, often leads to worse results. The basic problem is that gradients in real materials are so large that the expansion breaks down.

The term *generalized*-gradient expansion (GGA) denotes a variety of ways proposed for functions that modify the behavior at large gradients in such a way as to preserve desired properties. It is convenient [367] to define the functional as a generalized form of (8.1),

$$\begin{aligned} E_{\rm xc}^{\rm GGA}[n^\uparrow, n^\downarrow] &= \int {\rm d}^3 r n(\mathbf{r}) \epsilon_{\rm xc}(n^\uparrow, n^\downarrow, |\nabla n^\uparrow|, |\nabla n^\downarrow|, \ldots) \\ &\equiv \int {\rm d}^3 r n(\mathbf{r}) \epsilon_x^{\rm hom}(n) F_{\rm xc}(n^\uparrow, n^\downarrow, |\nabla n^\uparrow|, |\nabla n^\downarrow|, \ldots), \end{aligned} \tag{8.3}$$

where $F_{\rm xc}$ is dimensionless and $\epsilon_x^{\rm hom}(n)$ is the exchange energy of the unpolarized gas given in Table 5.3.

For exchange, it is straightforward to show (Exercise 8.1) that there is a "spin-scaling relation,"

$$E_x[n^\uparrow, n^\downarrow] = \frac{1}{2}\left[E_x[2n^\uparrow] + E_x[2n^\downarrow]\right], \tag{8.4}$$

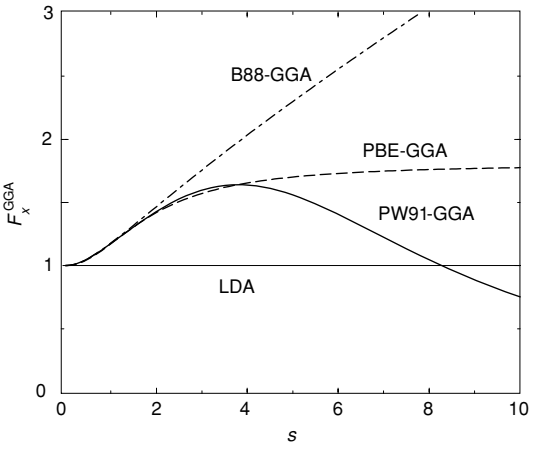


Figure 8.1. Exchange enhancement factor F_x as a function of the dimensionless density gradient s for various GGAs. (From H. Kim, similar to Fig. 1 of [367] but for a larger range of s). Note that in the relevant range for most materials, $0 < s \lesssim 3$, the magnitude of the exchange is increased by a factor $\approx$1.3–1.6. (This provides an *a posteriori* reason why there was some success in using the constant factor 4/3 in Slater's average local exchange.)

where $E_x[n]$ is the exchange energy for an unpolarized system of density $n(\mathbf{r})$. Thus for exchange we need to consider only the spin-unpolarized $F_x(n, |\nabla n|)$. It is natural to work in terms of dimensionless reduced density gradients of mth order that can be defined by

$$s_m = \frac{|\nabla^m n|}{(2k_F)^m n} = \frac{|\nabla^m n|}{2^m(3\pi^2)^{m/3}(n)^{(1+m/3)}}. \tag{8.5}$$

Since $k_F = 3(2\pi/3)^{1/3} r_s^{-1}$, s_m is proportional to the mth-order fractional variation in density normalized to the average distance between electrons r_s. The explicit expression for the first gradients can be written (Exercise 8.2)

$$s_1 \equiv s = \frac{|\nabla n|}{(2k_F)n} = \frac{|\nabla r_s|}{2(2\pi/3)^{1/3} r_s}. \tag{8.6}$$

The lowest order terms in the expansion of F_x have been calculated analytically [367, 370]

$$F_x = 1 + \frac{10}{81} s_1^2 + \frac{146}{2025} s_2^2 + \cdots \tag{8.7}$$

Numerous forms for $F_x(n, s)$, where $s = s_1$, have been proposed; these can be illustrated by the three widely used forms of Becke (B88) [371], Perdew and Wang (PW91) [372], and Perdew, Burke, and Enzerhof (PBE) [373].[1] In Fig. 8.1, we compare the factors F_x for these three approximations. Most other approximations lead to an F_x that falls between B88 and PBE, so the qualitative results obtained by employing other functionals can be appreciated from the behavior of these functionals. As shown in Fig. 8.1, one can divide the GGA into two regions: (i) small s ($0 < s \lesssim 3$) and (ii) large s($s \gtrsim 3$) regions. In region (i), which is

[1] A revised PBE form called "RPBE" has also been proposed in [374].

relevant for most physical applications, different F_xs have nearly identical shapes, which is the reason that different GGAs give similar improvement for many conventional systems with small density gradient contributions. Most importantly, $F_x \geq 1$, so all the GGAs lead to an exchange energy lower than the LDA. Typically, there are more rapidly varying density regions in atoms than in condensed matter, which leads to greater lowering of the exchange energy in atoms than in molecules and solids. This results in the reduction of binding energy, correcting the LDA overbinding, and improving agreement with experiment, which is one of the most important characteristics of present GGAs [224].

Note that in the range $0 < s \lesssim 3$ the average value of the enhancement is roughly $4/3$, making the average exchange similar to that proposed by Slater, although for very different reasons. Perhaps this accounts for the improvement that has often been found in calculations that use the factor $4/3$ or an adjustable factor called "$X\alpha$" that tends to be between 1 and $4/3$.

In region (ii), the different limiting behaviors of F_xs result from choosing different physical conditions for $s \to \infty$. In B88-GGA, $F_x^{\text{B88-GGA}}(s) \sim s/\ln(s)$ was chosen to give the correct exchange energy density ($\epsilon_x \to -1/2r$) [371]. In PW91-GGA, choosing $F_x^{\text{PW91-GGA}}(s) \sim s^{-1/2}$ satisfies the Lieb–Oxford bound (see [373]) and the non-uniform scaling condition that must be satisfied if the functional is to have the proper limit for a thin layer or a line [372]. In PBE-GGA, the non-uniform scaling condition was dropped in favor of a simplified parameterization with $F_x^{\text{PBE-GGA}}(s) \sim$ const. [373]. The fact that different physical conditions lead to very different behaviors of F_xs in region (ii) not only reflects the lack of knowledge of the large density gradient regions but also an inherent difficulty of the density gradient expansion in this region: even if one form of GGA somehow gives the correct result for a certain physical property while others fail, it is not guaranteed that the form is superior for other properties in which different physical conditions prevail.

Correlation is more difficult to cast in terms of a functional, but its contribution to the total energy is typically much smaller than the exchange. The lowest order gradient expansion at high density has been determined by Ma and Brueckner [375] (see [373]) to be

$$F_c = \frac{\epsilon_c^{\text{LDA}}(n)}{\epsilon_x^{\text{LDA}}(n)}(1 - 0.219,51 s_1^2 + \cdots). \tag{8.8}$$

For large density gradients the magnitude of correlation energy decreases and vanishes as $s_1 \to \infty$. This decrease can be qualitatively understood since large gradients are associated with strong confining potentials that increase level spacings and reduce the effect of interactions compared to the independent-electron terms. As an example of a GGA for correlation, Fig. 8.2 shows the correlation enhancement factor $F_c^{\text{PBE-GGA}}$ for the PBE functional, which is almost identical to that for the PW91-GGA. The actual analytic form for the PBE correlation is given in App. B.

There are now many GGA functionals that are used in quantitative calculations, especially in chemistry [224]. Correlation is often treated using the Lee–Yang–Parr (LYP) [376] functional, which was derived from the orbital dependent Colle–Salvetti functional [377]. That functional was in turn derived for the He atom and parameterized to fit atoms with more electrons. Krieger and coworkers [378] have constructed a functional (KCIS) based upon

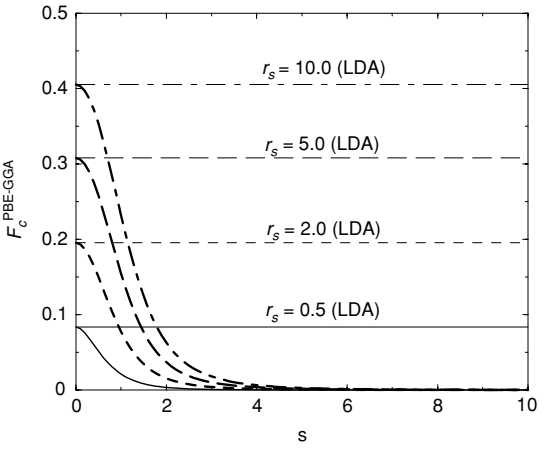


Figure 8.2. Correlation enhancement factor F_c as a function of the dimensionless density gradient s for the PBE functional. The actual form is given in App. B. Other functionals are qualitatively similar. (See caption of Fig. 8.1).

many-body calculations [379] of an artificial "jellium with a gap" problem that attempts to incorporate the effect of a gap into a functional. Most other functionals have parameters adjusted to fit to molecular data. Selected explicit forms can be found in [93, 224, 368].

8.3 LDA and GGA expressions for the potential $V_{xc}^{\sigma}(\mathbf{r})$

The part of the Kohn–Sham potential due to exchange and correlation $V_{\rm xc}^{\sigma}(\mathbf{r})$ is defined by the functional derivative in (7.13) or (7.18). The potential can be expressed more directly for LDA and GGA functionals, (8.1) and (8.3), since they are expressed in terms of functions (*not functionals*) of the local density of each spin $n(\mathbf{r}, \sigma)$ and its gradients at point $\mathbf{r}$. Explicit forms are given in App. B.

In the LDA, the form is very simple,

$$\delta E_{\rm xc}[n] = \sum_{\sigma} \int \mathrm{d}\mathbf{r} \left[\epsilon_{\rm xc}^{\rm hom} + n \frac{\partial \epsilon_{\rm xc}^{\rm hom}}{\partial n^{\sigma}} \right]_{\mathbf{r},\sigma} \delta n(\mathbf{r}, \sigma), \tag{8.9}$$

so that the potential,

$$V_{\rm xc}^{\sigma}(\mathbf{r}) = \left[\epsilon_{\rm xc}^{\rm hom} + n \frac{\partial \epsilon_{\rm xc}^{\rm hom}}{\partial n^{\sigma}} \right]_{\mathbf{r},\sigma}, \tag{8.10}$$

involves only ordinary derivatives of $\epsilon_{\rm xc}^{\rm hom}(n^{\uparrow}, n^{\downarrow})$. Here the subscript $\mathbf{r}, \sigma$ means the quantities in square brackets are evaluated for $n^{\sigma} = n(\mathbf{r}, \sigma)$. The LDA exchange terms are particularly simple: since $\epsilon_x^{\rm hom}(n^{\sigma})$ scales $(n^{\sigma})^{1/3}$ it follows that

$$V_x^{\sigma}(\mathbf{r}) = \frac{4}{3} \epsilon_x^{\rm hom}(n(\mathbf{r}, \sigma)). \tag{8.11}$$

The correlation potential depends upon the form assumed, with selected examples given in App. B.

In the GGA one can identify the potential by finding the change $\delta E_{\rm xc}[n]$ to linear order in δn and $\delta \boldsymbol{\nabla} n = \boldsymbol{\nabla} \delta n$,

$$\delta E_{\rm xc}[n] = \sum_\sigma \int {\rm d}\mathbf{r} \left[\epsilon_{\rm xc} + n \frac{\partial \epsilon_{\rm xc}}{\partial n^\sigma} + n \frac{\partial \epsilon_{\rm xc}}{\partial \boldsymbol{\nabla} n^\sigma} \boldsymbol{\nabla} \right]_{\mathbf{r},\sigma} \delta n(\mathbf{r}, \sigma). \tag{8.12}$$

The term in square brackets might be considered to be the potential; however, it does not have the form of a local potential because of the last term which is a differential operator.

There are three approaches to handling the last term. The first is to find a local $V^\sigma_{\rm xc}(\mathbf{r})$ by partial integration (see App. A) of the last term in square brackets to give

$$V^\sigma_{\rm xc}(\mathbf{r}) = \left[\epsilon_{\rm xc} + n \frac{\partial \epsilon_{\rm xc}}{\partial n^\sigma} - \boldsymbol{\nabla} \left(n \frac{\partial \epsilon_{\rm xc}}{\partial \boldsymbol{\nabla} n^\sigma} \right) \right]_{\mathbf{r},\sigma}. \tag{8.13}$$

This is the form most commonly used; however, it has the disadvantage that it requires higher derivatives of the density that can lead to pathological potentials and numerical difficulties, for example, near the nucleus or in the outer regions of atoms, where the density is rapidly varying or is very small (see Exercise 8.3).

A second approach is to use the operator form (8.12) directly by modifying the Kohn–Sham equations [380]. Using the fact that the density can be written in terms of the wavefunctions ψ_i, the matrix elements of the operator can be written (for simplicity we omit the variables $\mathbf{r}$ and σ)

$$\langle \psi_j | \hat{V}_{\rm xc} | \psi_i \rangle = \int \left[\tilde{V}_{\rm xc} \psi^*_j \psi_i + \psi^*_j \mathbf{V}_{\rm xc} \cdot \boldsymbol{\nabla} \psi_i + (\mathbf{V}_{\rm xc} \cdot \boldsymbol{\nabla} \psi^*_j) \psi_i \right], \tag{8.14}$$

where $\tilde{V}_{\rm xc} = \epsilon_{\rm xc} + n(\partial \epsilon_{\rm xc}/\partial n)$ and $\mathbf{V}_{\rm xc} = n(\partial \epsilon_{\rm xc}/\partial \boldsymbol{\nabla} n)$. This form is numerically more stable; however, it requires inclusion of the additional vector operator in the Kohn–Sham equation, which may significantly increase the computational cost; for example, in plane wave approaches four Fourier transforms are required instead of one.

Finally, a different approach proposed by White and Bird [381] is to treat $E_{\rm xc}$ strictly as a function of the density; the gradient terms are *defined* by an operational definition in terms of the density. Then (8.12) can be written using the chain rule as

$$\begin{aligned} \delta E_{\rm xc}[n] = & \sum_\sigma \int {\rm d}\mathbf{r} \left[\epsilon_{\rm xc} + n \frac{\partial \epsilon_{\rm xc}}{\partial n^\sigma} \right]_{\mathbf{r},\sigma} \delta n(\mathbf{r}, \sigma) \\ & + \sum_\sigma \int \int {\rm d}\mathbf{r} {\rm d}\mathbf{r}' n(\mathbf{r}) \left[\frac{\partial \epsilon_{\rm xc}}{\partial \boldsymbol{\nabla} n^\sigma} \right]_{\mathbf{r},\sigma} \frac{\delta \boldsymbol{\nabla} n(\mathbf{r}')}{\delta n(\mathbf{r})} \delta n(\mathbf{r}, \sigma), \end{aligned} \tag{8.15}$$

where $(\delta \boldsymbol{\nabla} n(\mathbf{r}')/\delta n(\mathbf{r}))$ denotes a functional derivative (which is independent of spin). For example, on a grid, the density for either spin is given only at grid points $n(\mathbf{r}_m)$ and the gradient at $\boldsymbol{\nabla} n(\mathbf{r}_m)$ is determined by the density by a formula of the form

$$\boldsymbol{\nabla} n(\mathbf{r}_m) = \sum_{m'} \mathbf{C}_{m-m'} n(\mathbf{r}_{m'}), \tag{8.16}$$

so that

$$\frac{\delta \nabla n(\mathbf{r}_m)}{\delta n(\mathbf{r}_{m'})} \to \frac{\partial \nabla n(\mathbf{r}_m)}{\partial n(\mathbf{r}_{m'})} = \mathbf{C}_{m-m'}. \tag{8.17}$$

(Note that each $\mathbf{C}_{m''} = \{C^x_{m''}, C^y_{m''}, C^z_{m''}\}$ is a vector in the space coordinates.) In a finite difference method, the coefficients $\mathbf{C}_{m''}$ are nonzero for some finite range; in a Fourier transform method, the $\mathbf{C}_{m''}$ follow simply by noting that

$$\nabla n(\mathbf{r}_m) = \sum_{\mathbf{G}} i\mathbf{G} n(\mathbf{G}) \mathrm{e}^{\mathrm{i}\mathbf{G}\cdot\mathbf{r}_m} = \frac{1}{N} \sum_{\mathbf{G},m'} i\mathbf{G} \mathrm{e}^{\mathrm{i}\mathbf{G}\cdot(\mathbf{r_m}-\mathbf{r_{m'}})} n(\mathbf{r}_{m'}). \tag{8.18}$$

Finally, varying $n(\mathbf{r}_m, \sigma)$ in the expression for E_{xc} and using the chain rule leads to

$$V^{\sigma}_{\mathrm{xc}}(\mathbf{r}_m) = \left[\epsilon_{\mathrm{xc}} + n\frac{\partial \epsilon_{\mathrm{xc}}}{\partial n}\right]_{\mathbf{r}_m,\sigma} + \sum_{m'} \left[n \frac{\partial \epsilon_{\mathrm{xc}}}{\partial |\nabla n|} \frac{\nabla n}{|\nabla n|}\right]_{\mathbf{r}_{m'},\sigma} \mathbf{C}_{m'-m}. \tag{8.19}$$

This form reduces the numerical problems associated with (8.13) without a vector operator as in (8.14). Note that $V^{\sigma}_{\mathrm{xc}}(\mathbf{r}_m)$ is a non-local function of $n(\mathbf{r}_{m'}, \sigma)$, the form of which depends upon the way the derivative is calculated. This is an advantage in actual calculations because it ensures consistency between E_{xc} and V_{xc}. The method can be extended to other bases by specifying the derivative in the appropriate basis.

8.4 Non-collinear spin density

In the usual (collinear) case of a spin polarized system, there are two densities $[n^{\uparrow}(\mathbf{r}), n^{\downarrow}(\mathbf{r})]$ and potentials $[V^{\uparrow}_{\mathrm{xc}}(\mathbf{r}), V^{\downarrow}_{\mathrm{xc}}(\mathbf{r})]$ for spin up and down. This is, however, not the most general form since the spin axis can vary in space. In this "non-collinear spin" case [291], the density at every point is represented by a vector giving the spin direction, or equivalently, by a local spin density matrix

$$\rho^{\alpha\beta}(\mathbf{r}) = \sum_i f_i \psi^{\alpha *}_i(\mathbf{r}) \psi^{\beta}_i(\mathbf{r}), \tag{8.20}$$

and the Kohn–Sham hamiltonian (7.12) becomes a 2×2 matrix

$$H^{\alpha\beta}_{\mathrm{KS}}(\mathbf{r}) = -\frac{\hbar^2}{2m_e}\nabla^2 + V^{\alpha\beta}_{\mathrm{KS}}(\mathbf{r}), \tag{8.21}$$

where the only part of $V^{\alpha\beta}_{\mathrm{KS}}$ that is non-diagonal in $\alpha\beta$ is $V^{\alpha\beta}_{\mathrm{xc}}$.

Although this looks like a major complication, the only real difficulty is in the nature of the functional $\epsilon^{\alpha\beta}_{\mathrm{xc}}$. In the local approximation it is given simply by finding the *local axis of spin quantization* and using the same functional form $\epsilon^{\mathrm{hom}}_{\mathrm{xc}}(n^{\uparrow}(\mathbf{r}), n^{\downarrow}(\mathbf{r}))$ given in (8.1). Examples of calculation can be found in [382–385]. Modifications of GGA expressions involve the gradient of the spin axis.

8.5 Non-local density formulations: ADA and WDA

A different approach to the generalization of the local density approximation proposed by Gunnarsson et al. [348] is to construct a non-local functional that depends on the density in some region around each point **r**. The original proposals were designed to provide a natural extension of the local functional in a way that satisfies the sum rules. This led to two approaches, the average density approximation (ADA) and the weighted density approximation (WDA) [348]. In the ADA, the exchange–correlation hole (7.16) and energy (7.17) are approximated by the corresponding quantity for a homogeneous gas of average density $\bar{n}^\sigma$ instead of the local density $n(\mathbf{r}, \sigma)$. This leads to

$$E_{\rm xc}^{\rm ADA}[n^\uparrow, n^\downarrow] = \int {\rm d}^3 r n(\mathbf{r}) \epsilon_{\rm xc}^{\rm hom}(\bar{n}^\uparrow(\mathbf{r}), \bar{n}^\downarrow(\mathbf{r})), \tag{8.22}$$

where

$$\bar{n}(\mathbf{r}) = \int {\rm d}^3 \mathbf{r}' \; w(\bar{n}(\mathbf{r}); |\mathbf{r} - \mathbf{r}'|) \; n(\mathbf{r}') \tag{8.23}$$

is a non-local functional of the density for each spin separately. The important point is the non-local nature of the ADA exchange–correlation hole whose extent depends not only upon the density at the observation point but upon a weighted average around **r**. The weight function w can be chosen in several ways. Gunnarsson et al. [348] originally proposed a form based on the linear response of the homogeneous electron gas, and given in tabular form. The WDA is related, but differs in the way the weighting is defined.

Tests have shown that there are advantages of the ADA and WDA, but there have not been extensive studies. A clear superiority over the ordinary LDA and GGAs is that in the limit where a three-dimensional system approaches two dimensional (e.g. in a confined electron gas in semiconductor quantum wells) the non-local functionals are well behaved whereas most of the LDA and GGA functionals diverge [386] (see Ex. 8.4). On the other hand, the ADA and WDA functionals suffer from the serious difficulty that core electrons distort the weighting in an unphysical way, so that any reasonable weighting must involve some shell decomposition to separate the effects on core and valence electrons.

8.6 Orbital-dependent functionals I: SIC and LDA + U

The most enduring problem with the Kohn–Sham approach is that no systematic way has been developed to improve the functionals for exchange and correlation. The problems are most severe in materials in which the electrons tend to be localized and strongly interacting, such as transition metal oxides and rare earth elements and compounds. These systems exhibit phenomena associated with correlation such as metal–insulator transitions, heavy fermion behavior, and high-temperature superconductivity (see, for example, [216]). Various methods have been developed to extend the functional approach to incorporate effects that are expected to be important on physical grounds. Two of these are SIC and LDA + U.

"SIC" denotes methods that use approximate functionals and add "self-interaction corrections" to attempt to correct for the unphysical self-interaction in many functionals for

exchange and correlation E_{xc}. The self-interaction of an electron with itself in the Hartree interaction is cancelled in exact treatments of exchange, as in Hartree–Fock and EXX discussed in Sec. 8.7. However, this is not the case for approximations to E_{xc}, and the errors can be significant since these terms involve large Coulomb interactions. There is a long history to such approaches, the first by Hartree himself [43] in his calculations on atoms. As noted in Sec. 3.5, Hartree defined a different potential for each occupied state by subtracting a self-term due to the charge density of that state. In finite systems, implementing such corrections is straightforward; however, for an extended state in a solid, the correction vanishes since the interaction scales inversely with the size of the region in which the state is localized. Thus, in extended systems there is some arbitrariness in the definition of a SIC.

An approach to extended systems has been developed in which a functional is defined with self-terms subtracted; minimization of the functional in an unrestricted manner allows the system of electrons to minimize the total energy by delocalizing the states (in a crystal, this is the usual Kohn–Sham solution with vanishing correction) or by localizing some or all of the states to produce a different solution [300,387]. This approach has an intuitive appeal in that it leads to atomic-like states in systems like transition metal oxides and rare earth systems, where the electrons are strongly interacting. This is often considered to be a better starting point for understanding such materials than the mean-field Kohn–Sham solution (see, for example, [216]). For example, studies using the SIC-LSDA have led to an improved description of the magnetic state and magnetic order in transition metal oxides [388], high T_c materials [389], and 4f occupation in rare earth compounds [390].

The quaint acronym "LDA+U" stands for methods that involve LDA- or GGA-type calculations coupled with an additional orbital-dependent interaction [366,391]. The additional interaction is usually considered only for highly localized atomic-like orbitals on the same site, i.e. of the same form as the "U" interaction in Hubbard models [392,393]. The effect of the added term is to shift the localized orbitals relative to the other orbitals, which attempts to correct errors known to be large in the usual LDA or GGA calculations. For example, the promotion energies in transition metal atoms in Fig. 10.2 illustrate the fact that the relative energies shift depending upon the approximation for exchange. Since orbital energies are shifted by occupations, the LDA+U and SIC approaches have much in common. The "U" parameter is often taken from "constrained density functional" calculations (Sec. 10.6) so that the theories do not contain adjustable parameters.

Many examples of "LDA+U" calculations are given in [366]. The prototypical examples are the magnetic oxides. For example, the usual spin density theory for NiO finds the correct spin states and an energy gap, but the value of the gap is much too small. This is corrected by a "U" term that increases the gap between the filled and empty 3d states. A much more severe case is CoO, which is an insulator with a gap of 2.4 eV, but which is a metal in density functional theory calculations unless a term involving orbital polarization is included [394]. This is an effect that can formally be considered within current density functional theory (Sec. 6.4). Within a spherical approximation around the Co atom, it leads to on-site terms for the 3d states with an additional self-consistent potential that is proportional to m_l, the component of the orbital angular momentum along the quantization axis. Since the effective field is self-consistent, the solution may be zero orbital moment or there may be an instability

to forming a moment. This was found to happen in CoO, splitting the m_l states and leading to an insulating gap [394].
Perhaps the best known cases are the parent compounds of the CuO superconductors which are found to be non-magnetic metals in the usual LDA and GGA calculations (see Ch. 17) whereas "LDA+U" calculations find the correct antiferromagnetic insulator solution [366].

8.7 Orbital-dependent functionals II: OEP and EXX

The essential advance of the Kohn–Sham approach over Thomas-Fermi-type methods is that the kinetic energy (7.3) is explicitly expressed as a functional of the independent-particle orbitals ψ_i. It is implicitly a functional of the density, since the orbitals are determined by $V_{\rm KS}$, which in turn is a functional of n; however, the functional dependence upon n must be highly non-trivial, non-analytic, and non-local. In particular, derivatives of the Kohn–Sham kinetic energy $\mathrm{d}T_s/\mathrm{d}n$ are discontinuous functions of n at densities corresponding to filled shells. This is the way in which shell structure occurs in the Kohn–Sham approach, whereas it is missing in Thomas-Fermi-type approximations.

These properties of the kinetic energy suggest a way to improve exchange–correlation functionals by expressing $E_{\rm xc}$ explicitly in terms of the independent-particle orbitals ψ_i. It is known that the true $E_{\rm xc}$ functional must have discontinuities at filled shells [351, 352, 395], which is essential for a correct description of energy gaps between filled bands. Such effects occur automatically in an orbital-dependent formulation, offering the possibility for improved description of the effect of correlation upon band gaps in density functional theory.

How does one formulate the Kohn–Sham theory with orbital-dependent functionals $E_{\rm xc}[\{\psi_i\}]$? In fact there is a long history of such methods that predates the work of Kohn and Sham, apparently first formulated in a short paper [396] by Sharp and Horton in 1953 as the problem of finding "that potential, the same for all electrons, such that when . . . given a small variation, the energy of the system remains stationary." This approach has come to be known as the optimized effective potential (OEP) method [231, 365, 397]. The key point is that if one considers orbitals ψ_i that are determined by a potential through the usual independent-particle Schrödinger equation, then it is straightforward, in principle, to define the energy functional of the potential V,

$$E_{\rm OEP}[V] = E[\{\psi_i[V]\}]. \tag{8.24}$$

The OEP method is fully within the Kohn–Sham approach since it is just the optimization of the potential V that appears in the very first Kohn–Sham equation (7.1). Furthermore, as emphasized in Sec. 9.2, the usual Kohn–Sham expressions are operationally functionals of the potential; the OEP is merely an orbital formulation of the general idea. The OEP method has been applied primarily to the Hartree–Fock exchange functional, which is straightforward to write in terms of the orbitals (the fourth term in (3.44)), which is called "exact exchange" or "EXX." However, the OEP approach is more general and applicable to orbital-dependent correlation functionals as well.

The variational equation representing the minimization of energy (8.24) can be written using the density formalism as an intermediate step. Since the potential V acts equally on all orbitals, it follows that

$$V_{\rm xc}^{\sigma,{\rm OEP}}(\mathbf{r}) = \frac{\delta E_{\rm xc}^{\rm OEP}}{\delta n^{\sigma}(\mathbf{r})}, \tag{8.25}$$

which can be written (see Exercise 8.5) using the chain rule [223, 365] as

$$\begin{aligned} V_{\rm xc}^{\sigma,{\rm OEP}}(\mathbf{r}) &= \sum_{\sigma'}\sum_{i=1}^{N^{\sigma'}}\int {\rm d}\mathbf{r}'\,\frac{\delta E_{\rm xc}^{\rm OEP}}{\psi_i^{\sigma'}(\mathbf{r}')}\frac{\psi_i^{\sigma'}(\mathbf{r}')}{\delta n^{\sigma}(\mathbf{r})} + {\rm c.c.} \\ &= \sum_{\sigma'}\sum_{i=1}^{N^{\sigma'}}\int {\rm d}\mathbf{r}'\int {\rm d}\mathbf{r}''\left[\frac{\delta E_{\rm xc}^{\rm OEP}}{\delta\psi_i^{\sigma'}(\mathbf{r}')}\frac{\delta\psi_i^{\sigma'}(\mathbf{r}')}{\delta V^{\sigma',KS}(\mathbf{r}'')} + {\rm c.c.}\right]\frac{\delta V^{\sigma',KS}(\mathbf{r}'')}{\delta n^{\sigma}(\mathbf{r})}, \end{aligned} \tag{8.26}$$

where $V^{\sigma',{\rm KS}}$ is the total potential in the independent-particle Kohn–Sham equations that determine the $\psi_i^{\sigma'}$. Each term has a clear meaning and can be evaluated from well-known expressions:

- The first term is an orbital-dependent non-local (NL) operator that can be written

$$\frac{\delta E_{\rm xc}^{\rm OEP}}{\delta\psi_i^{\sigma'}(\mathbf{r}')} \equiv V_{i,{\rm xc}}^{\sigma',{\rm NL}}(\mathbf{r}')\psi_i^{\sigma'}(\mathbf{r}'). \tag{8.27}$$

  For example, in the exchange-only approximation, $V_{i,{\rm xc}}^{\sigma',{\rm NL}}(\mathbf{r}')$ is the orbital-dependent Hartree–Fock exchange operator (3.48).
- The second term can be evaluated by perturbation theory,[2]

$$\frac{\delta\psi_i^{\sigma'}(\mathbf{r}')}{\delta V^{\sigma',{\rm KS}}(\mathbf{r}'')} = G_0^{\sigma'}(\mathbf{r}',\mathbf{r}'')\psi_i^{\sigma'}(\mathbf{r}''), \tag{8.28}$$

  where the Green's function for the non-interacting Kohn–Sham system is given by (see (D.3) which is written here with spin explicitly indicated)

$$G_0^{\sigma}(\mathbf{r},\mathbf{r}') = \sum_{j\neq i}^{\infty}\frac{\psi_j^{\sigma}(\mathbf{r})\psi_j^{\sigma *}(\mathbf{r}')}{\varepsilon_{\sigma i}-\varepsilon_{\sigma j}}. \tag{8.29}$$

- The last term is the inverse of a response function χ_0 given by

$$\chi_0^{\sigma,{\rm KS}}(\mathbf{r},\mathbf{r}') = \frac{\delta n^{\sigma}(\mathbf{r})}{\delta V^{\sigma',{\rm KS}}(\mathbf{r}'')} = \sum_{i=1}^{N^{\sigma}}\psi_i^{\sigma *}(\mathbf{r})G_0^{\sigma}(\mathbf{r},\mathbf{r}')\psi_i^{\sigma}(\mathbf{r}'), \tag{8.30}$$

  where we have used a chain rule and the fact that n is given by the sum of squares of orbitals, (7.2).

[2] Note that G_0 and the derivative in (8.28) are diagonal in spin since they involve the non-interacting Kohn–Sham system.

The integral form of the OEP equations (see Exercise 8.5) can be found by multiplying (8.27) by $\chi_0^\sigma(\mathbf{r}, \mathbf{r}')$ and integrating:

$$\sum_{i=1}^{N^\sigma} \int \mathrm{d}\mathbf{r}' \psi_i^{\sigma *}(\mathbf{r}') \left[V_{\mathrm{xc}}^{\sigma,\mathrm{OEP}}(\mathbf{r}') - V_{i,\mathrm{xc}}^{\sigma,\mathrm{NL}}(\mathbf{r}') \right] G_0^\sigma(\mathbf{r}', \mathbf{r}) \psi_i^\sigma(\mathbf{r}) + \mathrm{c.c.} = 0. \tag{8.31}$$

This form shows the physical interpretation that $V_{\mathrm{xc}}^{\sigma,\mathrm{OEP}}(\mathbf{r})$ is a particular weighted average of the non-local orbital-dependent potentials.

The integral form is the basis for useful approximations for which the potential can be given explicitly, e.g. as proposed by Krieger, Li, and Iafrate (KLI) [365,398–400]. Although KLI gave a more complete derivation, a heuristic derivation [365,396,398] is to replace the energy denominator in the Green's function by a constant $\Delta\varepsilon$. Then the value of $\Delta\varepsilon$ drops out of (8.31) and (8.29) becomes

$$G_0^\sigma(\mathbf{r}, \mathbf{r}') \to \sum_{j \neq i}^{\infty} \frac{\psi_j^\sigma(\mathbf{r}) \psi_j^{\sigma *}(\mathbf{r}')}{\Delta\varepsilon} = \frac{\delta(\mathbf{r} - \mathbf{r}') - \psi_i^\sigma(\mathbf{r}) \psi_i^{\sigma *}(\mathbf{r}')}{\Delta\varepsilon}. \tag{8.32}$$

As discussed in Exercise 8.8, the KLI approximation leads to the simple form

$$V_{\mathrm{xc}}^{\sigma,\mathrm{KLI}}(\mathbf{r}) = V_{\mathrm{xc}}^{\sigma,S}(\mathbf{r}) + \sum_{i=1}^{N^\sigma} \frac{n_i^\sigma(\mathbf{r})}{n^\sigma(\mathbf{r})} \left[\bar{V}_{i,\mathrm{xc}}^{\sigma,\mathrm{KLI}} - \bar{V}_{i,\mathrm{xc}}^{\sigma,\mathrm{NL}} \right], \tag{8.33}$$

where $V_{\mathrm{xc}}^{\sigma,S}(\mathbf{r})$ is the density-weighted average proposed by Slater [401]

$$V_{\mathrm{xc}}^{\sigma,S}(\mathbf{r}) = V_{\mathrm{xc}}^{\sigma,S}(\mathbf{r}) + \sum_{i=1}^{N^\sigma} \frac{n_i^\sigma(\mathbf{r})}{n^\sigma(\mathbf{r})} \bar{V}_{i,\mathrm{xc}}^{\sigma,\mathrm{NL}} \tag{8.34}$$

and the $\bar{V}$ are expectation values

$$\begin{aligned} \bar{V}_{i,\mathrm{xc}}^{\sigma,\mathrm{KLI}} &= \langle \psi_i^\sigma | V_{\mathrm{xc}}^{\sigma,\mathrm{KLI}} | \psi_i^\sigma \rangle, \\ \bar{V}_{i,xc}^{\sigma,\mathrm{NL}} &= \langle \psi_i^\sigma | V_{i,xc}^{\sigma,\mathrm{NL}} | \psi_i^\sigma \rangle. \end{aligned} \tag{8.35}$$

Finally, by taking matrix elements of (8.33), the equations become a set of linear equations for the matrix elements $\bar{V}_{i,\mathrm{xc}}^{\sigma,\mathrm{KLI}}$, which can be solved readily. The KLI approximation, including only exchange, has been shown to be quite accurate in many cases [365].

Slater local approximation for exchange

An interesting detour is the difference between the Kohn–Sham formula for the exchange potential (8.11) and the local form Slater had proposed earlier [401] based upon his approach of finding a local potential that is a weighted average of the non-local Hartree–Fock exchange operators (8.34). By averaging the *exchange potential* of the homogeneous gas, Slater found $V_x = 2\epsilon_x$, rather than the factor $\frac{4}{3}$ in (8.11) found by Kohn and Sham from the derivative of the exchange energy. In the context of the non-local exchange energy functional, it is not immediately clear which is the better approximation to carry over to an inhomogeneous system. Only recently has this issue been resolved [397, 399] by careful

treatment of the reference for the zero of energy in transferring the potential from the gas to an inhomogeneous system and the second factor in 8.33. The result is the Kohn–Sham form (8.11).

It has been observed that the Slater local approximation for exchange often gives eigenvalues in better agreement with experiment than does the Kohn–Sham form. This has led to the "Xα" approximation with an adjustable constant. In hindsight, this can be justified in part by the fact that gradient corrections lead to typical increases of similar magnitude, as shown in Fig. 8.1.

8.8 Hybrid functionals

The form of the coupling constant integration for the exchange–correlation energy, (7.15), is the basis for constructing a class of functionals called "hybrid" because they are a combination of orbital-dependent Hartree–Fock and an explicit density functional. These are the most accurate functionals available as far as energetics is concerned and are the method of choice in the chemistry community (see, e.g. [402] and [224]).

The hybrid formulation arises by approximating the integral in (7.15) in terms of information at the end points and the dependence as a form of the coupling constant λ. In particular, at $\lambda = 0$ the energy is just the Hartree–Fock exchange energy, which is easily expressed in terms of the exchange hole that can be calculated from the orbitals (the fourth term in (3.44)). Becke [403] has argued that the potential part of the LDA or GGA functional is most appropriate at full coupling $\lambda = 1$, and has suggested that the integral (7.15) can be approximated by assuming a linear dependence on λ leading to the "half-and-half" form

$$E_{\rm xc} = \frac{1}{2}(E_x^{\rm HF} + E_{\rm xc}^{\rm DFA}), \tag{8.36}$$

where DFA denotes an LDA or GGA functional. Later Becke presented parameterized forms that are accurate for many molecules, such as "B3P91" [403,404], a three-parameter functional that mixes Hartree–Fock exchange, the exchange functional of Becke (B88), and correlation from Perdew and Wang (PW91). Alternatively, the B3LYP form uses the LYP correlation. The definition of the exchange–correlation energy is

$$E_{\rm xc} = E_{\rm xc}^{\rm LDA} + a_0(E_x^{\rm HF} - E_x^{\rm DFA}) + a_x E_x^{\rm Becke} + a_c E_c, \tag{8.37}$$

with coefficients that are empirically adjusted to fit atomic and molecular data.

The coupling-constant integration approach has also been used to generate hybrid functionals by Perdew and coworkers, but with the idea of deriving the form theoretically. Based upon arguments on the variation of $E_{xc}(\lambda)$ as a function of λ, Perdew, Ernzerhof, and Burke [405] proposed the form

$$E_{\rm xc} = E_{\rm xc}^{\rm LDA} + \frac{1}{4}(E_x^{\rm HF} - E_{\rm xc}^{\rm DFA}), \tag{8.38}$$

that is, mixing in 1/4 the Hartree–Fock exchange energy. They have also given rationale for variations and for the values found previously by fitting. For example, the "1/4" form

has been tested using the PBE [373] form for $E_{\rm xc}^{\rm DFA}$ on a large set of molecules and found to be roughly comparable in accuracy to functionals with several fitted parameters [402].

Hybrid functionals can be used in different ways. Their use is not strictly within the usual Kohn–Sham approach if the Hartree–Fock equations are solved with a non-local exchange operator; however, it can be brought into the Kohn–Sham family by the OEP (Sec. 8.7). The most striking change due to use of the hybrid functionals is the predicted excitation energies; an example are the calculated bands of Si shown in Fig. 15.3 determined using the B3LYP functional.

8.9 Tests of functionals

It is instructive to examine the consequences of different approximations for the exchange–correlation functional in the simplest cases, where the conceptual structure is apparent and the quantitative results reveal aspects that may carry over to more complicated problems.

One-electron problems: hydrogen

For any one-electron problem, Hartree–Fock provides the exact solution for the total energy and the lowest eigenvalue of the Hartree–Fock equation (3.45) is the exact energy to remove the electron. This is because there is no correlation and the exchange potential (3.48) exactly cancels the self-interaction in the Hartree potential. However, the excited state eigenvalues of (3.45) denote the energy to add a second electron *assuming there is no correlation and no change in the occupied orbital*. There are often very large errors in the addition energies, e.g. for the H atom treated by Hartree–Fock, there are no bound states for added electrons, whereas in fact a second electron is bound by a small energy.

Exact exchange (EXX, Sec. 8.7) is Kohn–Sham density functional theory with the Hartree–Fock orbital-dependent exchange functional. For one electron, the bound state is exact, just as in Hartree–Fock; however, EXX is qualitatively different for excited states. In this case, the local Kohn–Sham potential is just the external potential, and the excited eigenvalues are the exact eigenvalues for the single electron in the external potential. For the H atom, this is the Rydberg series of excitations. Thus, excited state EXX eigenvalues correspond to excitation energies *with no change in the number of electrons* and are not energies for addition of electrons. Their interpretation as excitation energies also applies for high Rydberg states of multi-electron atoms, and suggests this interpretation for other cases as well [406].

On the other hand, one-particle problems are severe tests for approximate functionals such as the LDA and GGAs. The functionals are designed to deal with many electrons (e.g. the homogeneous gas) and their application to a one-particle problem introduces unphysical terms: (1) the unphysical self-interaction in the Hartree term is not cancelled exactly by the approximate exchange functional, and (2) it is spurious to introduce a correlation functional into a one-particle problem. The question is: how much damage is done in these cases where exact answers are known? The first row of Tabs. 8.1 and 8.2 gives the results for the value

Table 8.1. Exchange energies ($-E_x$, in Ha) for selected spherically symmetric atoms. Exact values denote "exact exchange" (EXX) in the Kohn–Sham theory. The other energies are calculated using the same EXX density. The functionals are described in the text. The last row is the mean absolute value of the error for 12 atoms given in [407].

Atom	Exact	LSDA	PBE	RPBE	BLYP	HCTH	PKZB
H	0.3125	0.2680	0.3059	0.3112	0.3098	0.3055	0.3081
He	1.0258	0.8840	1.0136	1.0313	1.0255	1.0063	1.0202
Be	2.6658	2.3124	2.6358	2.6801	2.6578	2.6114	2.6482
N	6.6044	5.908	6.5521	6.6252	6.5961	6.5145	6.5255
Ne	12.1050	11.0335	12.0667	12.1593	12.1378	12.0114	11.9514
Error, %	0	9.8	0.8	0.3	0.2	1.4	1.3%

Table 8.2. Correlation energies ($-E_c$, in Ha) for selected spherically symmetric atoms. The various functionals are evaluated for the same EXX density as in Tab. 8.1. The RPBE correlation functional is omitted since it is the same as PBE, and KCIS denotes the correlation functional derived in [378] incorporating effects of a gap. See caption of Tab. 8.1. From [407].

Atom	Exact	LSDA	PBE	BLYP	HCTH	PKZB	KCIS
H	0.0000	0.0222	0.0060	0.0000	0.0132	0.0000	0.0000
He	0.0420	0.1125	0.0420	0.0438	0.0753	0.0473	0.0408
Be	0.0950	0.2240	0.0856	0.0945	0.1505	0.0936	0.0861
N	0.1858	0.4268	0.1799	0.1919	0.2772	0.1841	0.1805
Ne	0.3939	0.7428	0.3513	0.3835	0.5046	0.3635	0.3667
Error, %	0	128.3	6.4	4.5	51.8	5.8	4.3%

of exchange and correlation energies for H resulting from various functionals [407]. The difference from the exact value indicates the accuracy. The most obvious result is that there are significant errors in the LSDA that are considerably improved by the GGAs. Note that the separate errors in exchange and correlation in the LSDA tend to cancel, so that the error in the final LSDA energy is ≈0.48 Ha. This is in surprisingly good agreement with the exact value, 0.5 Ha. Although there is no such cancellation for the GGAs, their final results for the total energy are much improved over LSDA.

Two-electron problems: He and H_2

The neutral He atom and H_2 molecule are the simplest two-electron systems which nevertheless exemplify issues related to many of the most important problems in condensed matter physics. The exchange and correlation energies for He are also given in Tabs. 8.1

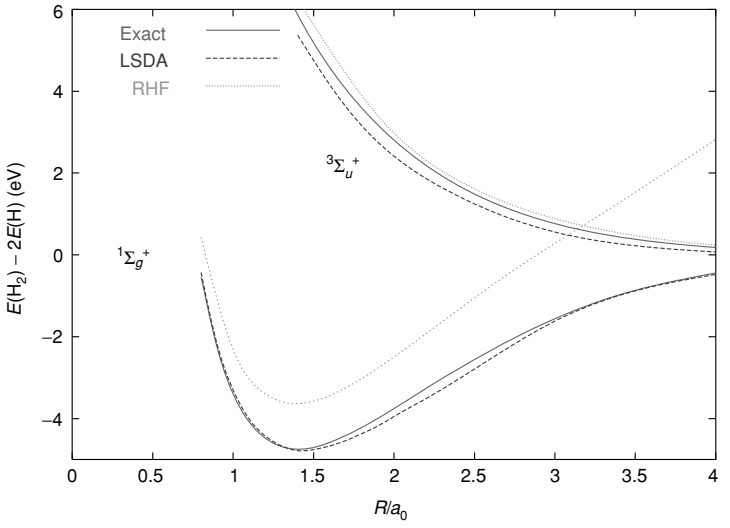


Figure 8.3. Energy versus separation R for an H_2 molecule, comparing LSDA (unrestricted) and Hartree–Fock (restricted-RHF) with the exact energies from [408]. The two sets of curves are for the spin singlet (bonding) and triplet (antibonding) states. The most remarkable result is the accuracy of the LSDA near the minimum, whereas the Hartree–Fock curve is too high since it omits correlation. At large R the unrestricted LSDA has a broken symmetry solution that approaches the usual spin-polarized isolated-atom LSDA limit. The triplet Hartree–Fock energy approaches the exact isolated atom limit, $E \equiv 0$, for large R, but the singlet approaches the wrong limit in the restricted approximation. Figure provided by O. E. Gunnarsson.

and 8.2. The good agreement for the LYP functional may not be surprising since it was constructed using He as a starting point; however, the quality of the results is impressive for functionals such as PBE constructed from information on the homogeneous gas.

The neutral H_2 molecule is a two-electron system that can be considered to be like He at very short distances R between the protons. At equilibrium lengths (and shorter) it is an excellent approximation to consider the exact system of two correlated electrons starting from the independent electron approximation, introducing correlation as a quantitative effect. The LDA is remarkably accurate, as shown in Fig. 8.3. However, at large distances there is a strong correlation between the electrons, with a greatly reduced probability of finding two electrons near the same atom at one time, compared to the probability of $\frac{1}{4}$ which would occur in the non-interacting case.

How do the eigenvalues compare with experimental removal energies? Since the highest eigenvalue is exact in an exact Kohn–Sham calculation, this is a test of approximate functionals. As illustrated in Fig. 8.4, there is a large effect in finite systems due to the long-range form of the potential [409]. The self-interaction term that occurs in approximate functionals has the effect of adding a spurious repulsive term that raises the eigenvalues and makes states too weakly bound. Proper treatment of the non-local exchange eliminates this effect and makes the states more strongly bound. Similar consequences of non-local exchange are found in calculations on many-electron atoms as illustrated in Sec. 10.5.

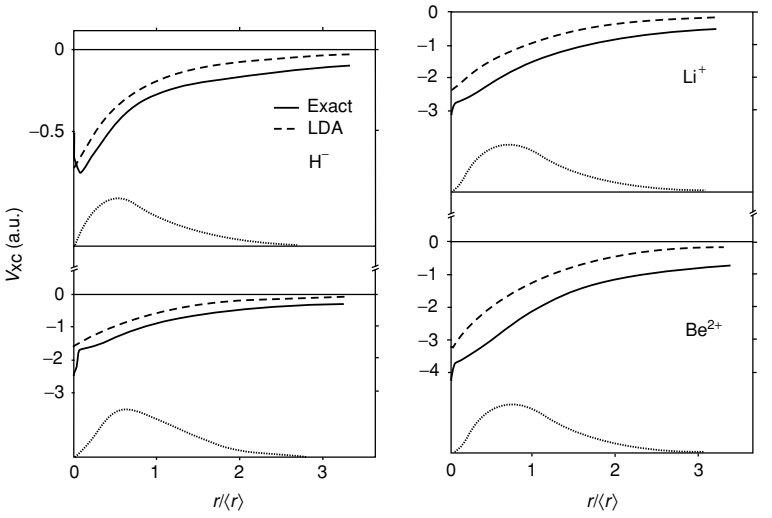


Figure 8.4. Exchange–correlation potential in two-electron ions, comparing the exact V_{xc} with the LDA. Each is derived from an essentially exact density. Note that the LDA potential is too high leading to an eigenvalue that is also too high. This can be easily understood from the fact that the exact potential has an attractive $1/r$ form at long range that is missing in the LDA. From Almbladh and Pedroza [409].

Tests for sets of atoms, molecules, and solids

The various functionals have been tested by many authors on various systems, for example extensive tests on the "extended G2" set of atoms and molecules (see [224], [410], and references cited there). The PBE [373] and hybrid PBE [405] functionals with fewer parameters have been compared to other functionals and found to be of approximately the same accuracy for the extended G2 set [402], for other molecules including transition metals [411], hydrogen bonding [167] in H_2O, for phase transitions in solids [412], and many other examples. In Tabs. 8.1 and 8.2 are given results comparing functionals for several atoms. The last line is the mean absolute error for 12 atoms tested and reported in [407], which indicates rather severe errors in LSDA and great improvement (but not always uniform) in the GGAs.

SELECT FURTHER READING

See references in "Select further reading" in Chapter 7.

References on functionals:

Anisimov, V. I., Aryasetiawan, F., and Lichtenstein, A. I., "First principles calculations of the electronic structure and spectra of strongly corelated systems: the LDA + U method," *J. Phys.: Condensed Matter* 9:767–808, 1997.

Casida, M. E., in *Recent Developments and Applications of Density Functional Theory*, edited by J. M. Seminario, Elsevier, Amsterdam, 1996, p. 391.

Grabo, T., Kreibich, T., Kurth, S., and Gross, E. K. U., in *Strong Coulomb Correlations in Electronic Structure: Beyond the Local Density Approximation*, edited by V. I. Anisimov, Gordon & Breach, Tokyo, 1998.

Koch, W., and Holthausen, M. C., *A Chemists' Guide to Density Functional Theory*, Wiley-VCH, Weinheim, 2001.

Perdew, J. P., and Burke, K., "Comparison shopping for a gradient-corrected density functional," *Int. J. Quant. Chem.* 57:309–319, 1996.

Staedele, M., Moukara, M., Majewski, J. A., Vogl, P., and Gorling, A., "Exact exchange Kohn–Sham formalism applied to semiconductors," *Phys. Rev.* B 59:10031–10043, 1999.

Towler, M. D., Zupan, A., and Causa, M., "Density functional theory in periodic systems using local gaussian basis sets," *Computer Physics Commun.* 98:181–205, 1996. (Summarizes explicit formulas for functionals.)

Exercises

8.1 Derive the "spin-scaling relation" (8.4). From this it follows that in the homogeneous gas, one needs only the exchange in the unpolarized case.

8.2 (a) Show that the expression for the dimensionless gradients $s_1 = s$ in (8.5) can be written in terms of r_s as (8.6).
(b) Find the form of the second gradient s_2 in terms of r_s.

8.3 Use the known form of the density near a nucleus to analyze the final term in (8.13) near the nucleus. Show that the term involves higher-order derivatives of the density that are singular at the nucleus.
(a) Argue that such a potential is unphysical using the facts that the exact form of the exchange potential is known and correlation is negligible compared to the divergent nuclear potential.
(b) Show that, nevertheless, the result for the total energy is correct since it is just a transformation of the equations.
(c) Finally, discuss how the singularity can lead to numerical difficulties in actual calculations.

8.4 Show that if a three-dimensional system is compressed in one direction so that the electrons are confined to a region that approaches a two-dimensional plane, the density diverges and the LDA expression for the exchange energy approaches negative infinity. Show that this is unphysical and that the exchange energy should approach a finite value that depends upon the area density. Argue that this is not necessarily the case for a GGA, but that the unphysical behavior can be avoided only by stringent conditions on the form of the GGA.

8.5 Derive the general OEP expression, (8.25), using the chain rule, and show that it leads to the compact integral expression, (8.31).

8.6 Write out explicit expressions for the inversion of the response function needed in (8.25) by expressing the response function in a basis. Consider appropriate bases for two cases: a radially symmetric atom (with the potential and density on a one-dimensional radial grid) and a periodic crystal (with all quantities represented in Fourier space).

8.7 An impediment in actual application of the OEP formula, (8.25) is the fact that the response function is singular. Show that this is the case since a constant shift in the potential causes no change in the density. Describe how such a response function can be inverted. Hint: One can define a non-singular function by projecting out the singular part. This may be most transparent in the case of a periodic crystal where trouble arises from a constant potential which is known to be undetermined.

8.8 Show that the approximation, (8.32), substituted into the integral equation, (8.31), leads to the KLI form, (8.33), and discuss the ways in which this is a much simpler expression than the integral equation, (8.31).

8.9 Problem on a diatomic molecule that demonstrates the breaking of symmetry in mean-field solutions such as LSDA.
(a) Prove that the lowest state is a singlet for two electrons in any local potential.
(b) Show this explicitly for the two-site Hubbard model with two electrons.
(c) Carry out the unrestricted HF calculation for the two-site Hubbard model with two electrons. Show that for large U the lowest energy state has broken symmetry.
(d) Carry out the same set of calculations for the hydrogen molecule in the LSDA. This can be done using programs available on-line (Ch. 24). Show that the lowest energy state changes from the correct symmetric singlet to a broken symmetry state as the atoms are pulled apart.
(e) Explain why the unrestricted solution has broken symmetry in parts (c) and (d), and discuss the extent to which it represents correct aspects of the physics even though the symmetry is not correct.
(f) Explain how to form a state with proper symmetry using the solutions of (c) and (d) and a sum of determinants.

8.10 Compute the exact exchange potential as a function of radius r in an H atom using the exact wavefunction in the ground state. This can be done with the formulas in Ch. 10 and numerical integration. Compare with the LDA approximation for the exchange potential using the exact density and expression (8.11), (note the system is full-spin polarized). Show the comparison explicitly by plotting the potentials as a function of radius. Justify the different functional forms of the potentials at large radius in the two cases.

8.11 The hydrogen atom is also a test case for correlation functionals; of course, correlation should be zero in a one-electron problem. Calculate the correlation potential using the approximate forms given in App. B (or the simpler Wigner interpolation form). Is the result close to zero? Does the correlation potential tend to cancel the errors in the local exchange approximation?

8.12 Apply the KLI approximation to H in its ground state. Is the KLI approximation exact in this case (as is the original EXX)?

9

Solving the Kohn–Sham equations

Summary

The Kohn–Sham equations provide the framework for finding the *exact* density and energy of the ground state of a many-body electron problem using standard independent-particle methods. These equations form the basis for much of the electronic structure developments described in the remainder of this volume. This chapter is devoted to the general form of the solution in terms of coupled self-consistent independent-particle Schrödinger-like equations. In order to apply the equations, one needs only the equations given in Ch. 7 (and summarized in the flow chart, Fig. 9.1) along with an explicit expression for the exchange–correlation functional $E_{\rm xc}[n]$. The reader is directed to Chs. 7 and 8 for discussion of the Kohn–Sham method itself, which follows directly from the choice of the auxiliary system, and the rationale for construction of the exchange–correlation functional $E_{\rm xc}[n]$. Examples of explicit forms for approximate functionals are given in Ch. 8 and App. B.

9.1 The self-consistent coupled Kohn–Sham equations

The Kohn–Sham equations derived in Sec. 7.2 are summarized in the flow chart in Fig. 9.1. They are a set of Schrödinger-like independent-particle equations which must be solved subject to the condition that the effective potential $V_{\rm eff}^{\sigma}(\mathbf{r})$ and the density $n(\mathbf{r}, \sigma)$ are consistent. The explicit reference to spin will be dropped, except where needed, and notation $V_{\rm eff}$ and n will be assumed to designate both space and spin dependence (of course, the potential for each spin depends upon the densities for both spins). An actual calculation utilizes a numerical procedure that successively changes $V_{\rm eff}$ and n to approach the self-consistent solution. The computationally intensive step in Fig. 9.1 is "solve KS equation" for a given potential V_{eff}. This is the subject of the following chapters. Here this step is considered a "black box" that uniquely solves the equations for a given input $V^{\rm in}$ to determine an output density $n^{\rm out}$, i.e. $V^{\rm in} \rightarrow n^{\rm out}$. Conversely, for a given form of the *xc* functional, any density n determines a potential V_{eff} as shown in the second box. (This is the same as (7.13) and examples of specific expressions are given in Sec. 8.3.)

The problem is that, except at the exact solution, the input and output potentials and densities do not agree. To arrive at the solution one defines operationally a new potential

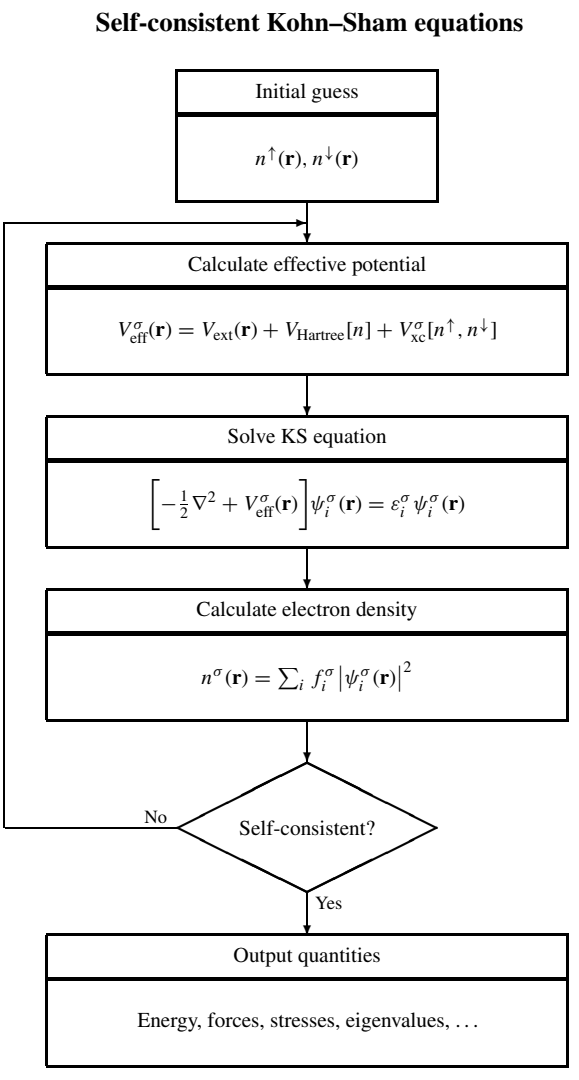


Figure 9.1. Schematic representation of the self-consistent loop for solution of Kohn–Sham equations. In general, one must iterate two such loops simultaneously for the two spins, with the potential for each spin a functional of the density of both spins.

$n^{\rm out} \rightarrow V^{\rm new}$, which can then start a new cycle with $V^{\rm new}$ as the new input potential. Clearly, the procedure shown in Fig. 9.1 can be made into the iterative progression

$$V_i \rightarrow n_i \rightarrow V_{i+1} \rightarrow n_{i+1} \rightarrow \ldots, \tag{9.1}$$

where i labels the step in the iteration. The progression converges with a judicious choice of the new potential in terms of the potential or density found at the previous step (or steps).

Methods for reaching self-consistency are described in Sec. 9.3. However, it is first best to probe the nature of various possible total energy functionals. The expressions are needed

for the final calculation of the energy and, in addition, the behavior of any of the functionals near the correct solution provides the basis for analysis of the convergence characteristics using that functional.

9.2 Total energy functionals

The subject of this section is the behavior of various functionals, all of which have the same minimum energy solution of the Kohn–Sham equations, but behave differently away from the minimum. In particular, it is not essential to regard the density as the independent variable in the equations; different functionals can be found by a Legendre transformation to change the independent and dependent variables, as is familiar in thermodynamics. In terms of the Kohn–Sham equations, this means the behavior as a functional of the difference of input and output quantities $\Delta V = V^{\text{out}} - V^{\text{in}}$ and $\Delta n = n^{\text{out}} - n^{\text{in}}$, where n^{out} is the resulting density from solving the Schrödinger-like equation with the potential V^{in}. *It is essential to utilize correct variational expressions in order to have the desired variational properties.*

The original expression for the Kohn–Sham energy functional is given by (7.5), which is repeated here, with the grouping of all the potential terms to define $E_{\text{pot}}[n]$,

$$E_{\text{KS}} = T_s[n] + E_{\text{pot}}[n], \tag{9.2}$$

$$E_{\text{pot}}[n] = \int \mathrm{d}\mathbf{r} V_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) + E_{\text{Hartree}}[n] + E_{II} + E_{\text{xc}}[n]. \tag{9.3}$$

The first three terms on the right-hand side of the second equation together form a neutral grouping equal to the classical Coulomb interaction E^{CC} in (3.14). Since the eigenvalues of the Kohn–Sham equations are given by

$$\varepsilon_i^\sigma = \langle \psi_i^\sigma | H_{KS}^\sigma | \psi_i^\sigma \rangle, \tag{9.4}$$

the kinetic energy can be expressed as

$$T_s = E_s - \sum_\sigma \int \mathrm{d}\mathbf{r}\ V^{\sigma,in}(\mathbf{r}) n^{\text{out}}(\mathbf{r}, \sigma), \tag{9.5}$$

where

$$E_s = \sum_\sigma \sum_{i=1}^{N^\sigma} \varepsilon_i^\sigma . \tag{9.6}$$

The advantages of this formulation are that the eigenvalues are available in actual calculations and, furthermore, E_s in (9.6) is itself a functional. It is the ground state energy of a non-interacting electron system, for which the Hohenberg–Kohn theorems, the force theorem, etc., all apply in a particularly simple way.

The Kohn–Sham functional of the potential $E_{\text{KS}}[V]$

Although the Kohn–Sham energy (9.2) is, *in principle*, a functional of the density, it is operationally a functional of the input potential $E_{\text{KS}}[V^{\text{in}}]$, as indicated in the flow chart,

Fig. 9.1. (Here V denotes the potential for each spin, $V^{\sigma}(\mathbf{r})$.) At any stage of a Kohn–Sham calculation when the energy is not at the minimum, V^{in} determines all the quantities in the energy. This is clearly shown if we write E_{KS} from (9.2) as

$$E_{\text{KS}}[V^{\text{in}}] = E_s[V^{\text{in}}] - \sum_{\sigma} \int \mathrm{d}\mathbf{r}\; V^{\sigma,in}(\mathbf{r}) n^{\text{out}}(\mathbf{r},\sigma) + E_{\text{pot}}[n^{\text{out}}], \tag{9.7}$$

where the first two terms on the right-hand side are a convenient way of calculating the independent-particle kinetic energy as in (9.5), and E_{pot} is the sum of potential terms given in (9.3) evaluated for $n = n^{\text{out}}$. Since E_s is the sum of eigenvalues, (9.6), and $n^{\text{out}}(\mathbf{r},\sigma)$ is the output density, each determined directly by the potential $V^{\sigma,in}(\mathbf{r})$, clearly the energy is a functional of V^{in}. Of course, E_{KS} formally can be regarded as a functional of n^{out}, since there is a one-to-one relation of the output density and the input potential (except for a trivial constant in V^{in}); however, the Kohn–Sham equations provide no way of choosing n^{out} except as an output determined by a potential.

The solution of the Kohn–Sham equations is for the potential V^{in} that minimizes the energy, (9.7). Then $V^{\text{in}} = V_{\text{KS}}$, the output density n^{out} is the ground state density n^0, and the potential and density are consistent with the relation in (7.13). The functional $E_{\text{KS}}[V^{\text{in}}]$ is variational and all other potentials lead to energies that are higher by an amount that is quadratic in the error $V^{\text{in}} - V_{\text{KS}}$. Near the minimum energy solution, the error in the energy must also be quadratic in the error in the density $\delta n = n^{\text{out}} - n^0$, so that

$$E_{\text{KS}}[V^{\text{in}}] = E_{\text{KS}}[V_{\text{KS}}] + \frac{1}{2}\sum_{\sigma,\sigma'} \int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \left[\frac{\delta^2 E_{\text{KS}}}{\delta n(\mathbf{r},\sigma)\delta n(\mathbf{r}',\sigma')}\right]_{n^0} \delta n(\mathbf{r},\sigma)\delta n(\mathbf{r}',\sigma'), \tag{9.8}$$

where the second term is always positive.

Explicit functionals of the density

As shown by Harris [415], Weinert, et al. [416], and Foulkes and Haydock [417], one can choose different expressions for the total energy functional that are given *explicitly* in terms of the density. The functional is cast in terms of the density n^{in} that, via (7.13), determines the input potential $V[n^{\text{in}}] \equiv V_{n^{\text{in}}}$, which in turn leads directly to the sum of eigenvalues, the first term on the right-hand side of (9.7). The energy is then defined by evaluating the functional $E_{\text{pot}}[n^{\text{in}}]$ in (9.3) in terms of the chosen *input* density $n^{\text{in}}(\mathbf{r},\sigma)$ (instead of the output density $n^{\text{out}}(\mathbf{r},\sigma)$ as in the Kohn–Sham functional),

$$E_{HWF}[n^{\text{in}}] \equiv E_s[V_{n^{\text{in}}}] - \sum_{\sigma} \int \mathrm{d}\mathbf{r}\; V^{\sigma}_{n^{\text{in}}}(\mathbf{r}) n^{\text{in}}(\mathbf{r},\sigma) + E_{\text{pot}}[n^{\text{in}}]. \tag{9.9}$$

The stationary properties of this functional can be understood straightforwardly following the arguments of Foulkes [417]. For a given input density n^{in} and potential $V_{n^{\text{in}}}$, the difference in the two expressions for the energy involves only the potential terms

$$\begin{aligned} E_{\text{KS}}[V^{\text{in}}] \;-\; E_{HWF}[n^{\text{in}}] = \sum_{\sigma} \int \mathrm{d}\mathbf{r}\; V^{\sigma}_{n^{\text{in}}}(\mathbf{r}) \left[n^{\text{out}}(\mathbf{r},\sigma) - n^{\text{in}}(\mathbf{r},\sigma)\right] \\ + \left[E_{\text{pot}}[n^{\text{out}}] - E_{\text{pot}}[n^{\text{in}}]\right]. \end{aligned} \tag{9.10}$$

Near the correct solution where $\Delta n = n^{\rm out} - n^{\rm in}$ is small, one can expand the difference in (9.10) in powers of Δn. The linear terms cancel (which follows from the fact that $V^{\sigma}_{n^{\rm in}}(\mathbf{r}) = [\delta E_{\rm pot}/(\delta n(\mathbf{r},\sigma))]_{n^{\rm in}}$, see Exercise 9.2), so that the lowest order terms are

$$E_{\rm KS}[V^{\rm in}] - E_{HWF}[n^{\rm in}] \approx \frac{1}{2}\sum_{\sigma,\sigma'}\int d\mathbf{r}d\mathbf{r}' K(\mathbf{r},\sigma;\mathbf{r}',\sigma')_{n^{\rm in}}\Delta n(\mathbf{r},\sigma)\Delta n(\mathbf{r}',\sigma'), \quad (9.11)$$

where the kernel K is defined to be

$$\begin{aligned} K(\mathbf{r},\sigma;\mathbf{r}',\sigma') &\equiv \frac{\delta^2 E_{\rm Hxc}[n]}{\delta n(\mathbf{r},\sigma)\delta n(\mathbf{r}',\sigma')} \\ &= \frac{1}{|\mathbf{r}-\mathbf{r}'|}\delta_{\sigma,\sigma'} + \frac{\delta^2 E_{\rm xc}[n]}{\delta n(\mathbf{r},\sigma)\delta n(\mathbf{r}',\sigma')}, \end{aligned} \quad (9.12)$$

evaluated for $n = n^{\rm in}$. (Note that K has been defined in terms of $E_{\rm Hxc}[n] \equiv E_{\rm Hartree}[n] + E_{\rm xc}[n]$; the other terms in $E_{\rm pot}[n]$ do not contribute since they are constant or linear in n.) Since the differences in the energies are quadratic in the errors in the density, it follows that at the exact solution where $\Delta n(\mathbf{r},\sigma) = 0$, the functional $E_{HWF}[n^{\rm in}]$ equals the usual Kohn–Sham energy and it is stationary. However, it is *not variational*, which can be seen from (9.11). Since the kernel K tends to be positive (see below), then $E_{HWF}[n^{\rm in}]$ is lower than $E_{\rm KS}[V^{\rm in}]$. Thus even though $E_{\rm KS}[V^{\rm in}]$ is always above the Kohn–Sham energy, $E_{HWF}[n^{\rm in}]$ may be lower by an amount that is second order in the error $\Delta n(\mathbf{r},\sigma)$.

The primary advantage of the explicit functional of the density (9.9) is that, for densities near the correct solution, it can accurately approximate the true Kohn–Sham energy. In particular, it is often an excellent approximation to stop the calculation after one calculation of eigenvalues with *no self-consistency*: in this case one does not even need to calculate the output density. This approach is remarkably successful if $n(\mathbf{r})$ is approximated by a sum of atomic densities [144, 415, 417–419]. Perhaps the first example was calculation of phonon frequencies [144]. Foulkes has used this as a conceptual basis for the success of empirical tight-binding models where the energy is given strictly by sums of eigenvalues plus additional terms that can be accounted for in this approach (see Sec. 14.4 on tight-binding). In addition, it is particularly simple to calculate the energy relative to neutral atoms in terms of the *difference in the density from a sum of neutral atoms*. This yields directly desirable physical quantities, as described in Sec. F.4.

In a full self-consistent calculation the functional (9.9) is useful at each step of the iteration in Fig. 9.1. It is now standard to calculate both energies, (9.7) and (9.9), at each step in the iteration. The KS functional of the potential is variational, but the non-variational functional of the density energy is usually closer to the true energy for reasons explained in Sec. 9.3. It is also very useful to calculate both energies and treat the difference as a measure of the lack of self-consistency during a calculation.

It is tempting to assume that the explicit density functional (9.9) is a *maximum* as a function of density. However, this is not the case in general because the second derivative functional $K(\mathbf{r},\sigma;\mathbf{r}',\sigma')$ in (9.12) is *not* guaranteed to be positive definite [420–422]. From the definition of K in (9.12), the first term is positive definite since it is due to the repulsive Hartree term. One might expect that the second attractive term would never overcome

the repulsion. However, approximations such as the LDA violate this condition since the extreme local $\delta(|\mathbf{r}-\mathbf{r}'|)$ behavior leads to large negative contributions for short wavelength density variations.

Generalized functionals of V and n, $E[V,n]$

It is also possible to define functionals of the density and potential varied independently, as pointed out by a number of authors [417,419,423,424]. We will denote n and V by n^{in} and V^{in} to emphasize that *both* are independent input functions. The expression is exactly the same as (9.9), except that V^{in} is regarded as an independent function so that the expression can be written

$$E[V^{\text{in}}, n^{\text{in}}] = E_s[V^{\text{in}}] - \sum_\sigma \int V^{\sigma,\text{in}}(\mathbf{r}) n^{\text{in}}(\mathbf{r},\sigma)\mathrm{d}\mathbf{r} + E_{\text{pot}}[n^{\text{in}}]. \tag{9.13}$$

The first term is solely a functional of V^{in}, the last term is a functional only of n^{in}, and the only coupling of V^{in} and n^{in} is through the simple bilinear second term. The properties of the functional can be seen clearly following the description by Methfessel [419]. Considering variations around any V^{in} and n^{in}, to linear order

$$\begin{aligned}\delta E[V^{\text{in}}, n^{\text{in}}] = &\sum_\sigma \int \left[V^\sigma_{n^{\text{in}}}(\mathbf{r}) - V^{\sigma,\text{in}}(\mathbf{r})\right]\delta n(\mathbf{r},\sigma)\mathrm{d}\mathbf{r} \\ &+ \sum_\sigma \int \left[n^{\text{out}}_{V^{\text{in}}}(\mathbf{r},\sigma) - n^{\text{in}}(\mathbf{r},\sigma)\right]\delta V^\sigma(\mathbf{r})\mathrm{d}\mathbf{r},\end{aligned} \tag{9.14}$$

where $V^\sigma_{n^{\text{in}}}(\mathbf{r}) = \left[\frac{\delta E_{\text{pot}}}{\delta n(\mathbf{r},\sigma)}\right]_{n^{\text{in}}}$ is the potential determined by the input density (as used in (9.9)), and $n^{\text{out}}_{V^{\text{in}}}(\mathbf{r},\sigma)$ is the output density determined by the potential V^{in} (as used in (9.7)). Since the terms in brackets vanish at self-consistency, the functional is stationary and the value equals the Kohn–Sham energy $E_{\text{KS}}[V^{\text{KS}}]$.

It is also straightforward to show [419] that for any fixed density n^{in}, the stationary point of $E[V^{\text{in}}, n^{\text{in}}]$ as a function of V^{in} is in fact a *global maximum* as a function of V^{in}, at which point the value of $E_s[V^{\max}] - \sum_\sigma \int V^{\sigma,\max}(\mathbf{r}) n^{\text{in}}(\mathbf{r},\sigma)\mathrm{d}\mathbf{r}$ equals the Kohn–Sham kinetic energy functional $T_s[n^{\text{in}}]$. Although the maximum property may seem surprising, it follows from inequalities similar to the Hohenberg–Kohn arguments and it can be understood from (9.14), which shows that

$$\frac{\delta E}{\delta V^\sigma(\mathbf{r})} = n^{\text{out}}(\mathbf{r},\sigma) - n^{\text{in}}(\mathbf{r},\sigma) \Rightarrow \frac{\delta^2 E}{\delta V^\sigma(\mathbf{r})\delta V^{\sigma'}(\mathbf{r}')} = \frac{\delta n^{\text{out}}(\mathbf{r},\sigma)}{\delta V^{\sigma'}(\mathbf{r}')}. \tag{9.15}$$

The eigenvalues of this functional are always negative since the density decreases where the potential is increased [419]. The curvature of E as a functional of n^{in} is given by the kernel (9.12), which involves only the potential terms $E_{\text{Hxc}}[n]$ since the other terms are constant or linear. As explained following (9.12), E tends to be a minimum as a functional of n^{in}; however, this is not guaranteed and only with constraints on the density variations is the solution a minimum [419].

The importance of the stationarity is that one can approximate both V^{in} and n^{in}. For example, one can choose convenient forms for the potentials, such as spherical muffin-tin-type potentials often used in augmented methods. If one carries out the Kohn–Sham calculation exactly for this potential, of course this is just a restatement of the variational property of $E_{\text{KS}}[V]$. The generalized functional shows that the errors in the energy are still quadratic if the density is also approximated using convenient functional forms. This can be used to advantage in calculations as illustrated in [419].

Thermal functionals

Introducing temperature has many potential benefits:

- Direct calculation of thermal quantities: entropy S, free energy $F = E - \text{TS}$, etc.
- The density matrix becomes shorter range as the temperature increases, which can be used to advantage, e.g. in order-N methods (Ch. 23).
- Smearing the occupation makes calculations for metals less sensitive to numerical approximations.

Expressions for the energy are given by any of the previous functionals with the sum of single particle energies $E_s \rightarrow E_s(T)$ generalized to finite T as in (3.39). The entropy is given by the single particle form of the Mermin finite temperature functional (6.20),

$$S = -\left[\sum_i f_i \ln f_i + \sum_i (1 - f_i)\ln(1 - f_i)\right], \tag{9.16}$$

where f_i denotes the occupation number $f(\varepsilon_i - \mu)$.

These formulas can be used as a clever way to calculate $E(T = 0)$. The simple idea is that $E(T)$ increases quadratically with T, whereas $F(T)$ decreases quadratically. A combination of the two, $E + F$ (see Exercise 9.4), can cancel the quadratic terms and give an expression equal to $E(T = 0)$ with only quartic corrections. For example, this been used by Gillan [425] to calculate the vacancy energy in Al using a calculation actually done at a temperature of 10,000 K. The high temperature greatly simplifies the calculations by reducing the finite size effects in the calculation.

In iterative methods (App. M), one is seeking to find the solution for both the potential and the wavefunctions at the same time, i.e. the wavefunctions are not consistent with the potential, as is assumed in the above expressions. As shown in [426], one can generalize the Fermi function f_i to a matrix f_{ij}, which is constrained to have eigenvalues in the range [0, 1]. Then the density is given by

$$n(\mathbf{r}) = \sum_{ij} f_{ij}\psi_i^*(\mathbf{r})\psi_j(\mathbf{r}), \tag{9.17}$$

and the grand energy functional (6.20) is generalized to

$$\begin{aligned}\tilde{\Omega}[V^{\text{in}}, n^{\text{in}}, T, \mu] = {} & E[V^{\text{in}}, n^{\text{in}}]_0 + \mu(N_0 - \text{Tr}[f]) \\ & + k_B T \text{Tr}\left[f \ln f + (1 - f)\ln(1 - f)\right].\end{aligned} \tag{9.18}$$

This form is particularly useful in iterative methods where it can speed the convergence in metals by effectively allowing for unitary transformations of the wavefunctions that are problematic because they correspond to low energy "slow modes" of the electronic system.

The most complete expression for a generalized functional is found by including temperature T via the Mermin functional (see Sec. 6.4) and the chemical potential μ to allow variation in particle number. Then, as shown by Nicholson et al. [424], one can define a grand functional,

$$\Omega[V^{\rm in}, n^{\rm in}, T, \mu] = E[V^{\rm in}, n^{\rm in}, T]_0 + \mu\Big(N_0 - \sum_i f_i\Big) + k_B T\Big[\sum_i f_i \ln f_i + \sum_i (1 - f_i)\ln(1 - f_i)\Big]. \tag{9.19}$$

This functional is stationary with respect to $V^{\rm in}$, $n^{\rm in}$, μ, T, and the form of the occupation function $f(\varepsilon)$.

9.3 Achieving self-consistency

A key problem is the choice of procedure for updating the potential V^σ or the density n^σ in each loop of the Kohn–Sham equations illustrated in Fig. 9.1. Obviously one can vary either V^σ or n^σ, but it is simpler to describe in terms of n^σ, which is unique, whereas V^σ is subject to shift by a constant. (The spin index σ is omitted below for simplicity.)

The simplest approach is *linear mixing*, estimating an improved density input $n^{\rm in}_{i+1}$ at step $i + 1$ as a fixed linear combination of $n^{\rm in}_i$ and $n^{\rm out}_i$ at step i,

$$n^{\rm in}_{i+1} = \alpha n^{\rm out}_i + (1 - \alpha) n^{\rm in}_i = n^{\rm in}_i + \alpha(n^{\rm out}_i - n^{\rm in}_i). \tag{9.20}$$

This is the best choice in the absence of other information and is essentially moving in an approximate "steepest descent" direction for minimizing the energy.

Why cannot one simply take the output density at one step as the input to the next? What are the limits on α? How can one do better? The answers lie in linear analysis of the behavior near the minimum [413, 427].[1] As in (9.8), let us define the deviation from the correct density to be $\delta n \equiv n - n_{\rm KS}$ at any step in the iteration. Then near the solution, the error in the output density to linear order in the error in the input is given by

$$\delta n^{\rm out}[n^{\rm in}] = n^{\rm out} - n_{\rm KS} = (\tilde{\chi} + 1)(n^{\rm in} - n_{\rm KS}), \tag{9.21}$$

where

$$\tilde{\chi} + 1 = \frac{\delta n^{\rm out}}{\delta n^{\rm in}} = \frac{\delta n^{\rm out}}{\delta V^{\rm in}} \frac{\delta V^{\rm in}}{\delta n^{\rm in}}. \tag{9.22}$$

Here $\delta n^{\rm out}/\delta V^{\rm in}$ is a response function defined to be χ^0 in (D.3) and $\delta V^{\rm in}/\delta n^{\rm in}$ is K defined in (9.12). Thus the needed function $\tilde{\chi}$ can be calculated and is closely related to other uses of response functions. The best choice for the new density is one that would make the error zero, i.e. $n^{\rm in}_{i+1} = n_{\rm KS}$. Since $n^{\rm out}_i$ and $n^{\rm in}_i$ are known from step i, if $\tilde{\chi}$ is also known, then

[1] The description here follows that of Pickett in [413].

(9.21) can be solved for $n_{\rm KS}$,

$$n_{\rm KS} = n_i^{\rm in} - \tilde{\chi}^{-1}\big(n_i^{\rm out} - n_i^{\rm in}\big). \tag{9.23}$$

If (9.23) were exact, this would be the answer and the iterations could stop; since it is not exact this gives the best input for the next iteration.

Although (9.23) is a more complex integral equation, it bears a strong resemblance to the linear-mixing equation (9.20). If we resolve the response function $\tilde{\chi}$ into eigenfunctions $\tilde{\chi}(\mathbf{r}, \mathbf{r}') = \sum_m \chi_m f_m(\mathbf{r}) f_m(\mathbf{r}')$, the eigenvalues χ_m give the optimal α for the change in density resolved into the density eigenvectors $f_m(\mathbf{r})$. Furthermore, the radius of convergence of the linear-mixing scheme is determined by the maximum eigenvalue $\tilde{\chi}_{\rm max}^{-1} = 1/\tilde{\chi}_{\rm min}$ of the matrix $\tilde{\chi}^{-1}$. If a constant α is used, it is straightforward to show [413] that the maximum error at iteration i varies as $(1 - \alpha\tilde{\chi}_{\rm max}^{-1})^i$, so that the iterations converge only if $\alpha < 2/\tilde{\chi}_{\rm max}^{-1} = 2\tilde{\chi}_{\rm min}$ (see Exercises 9.8 and 13.3).

Physically, the response of the system is a measure of the polarizability. Linear mixing with large α works well for strongly bound, rigid systems, such as regions near atom cores. However, convergence can be very difficult to achieve for "soft cases," for which metal surfaces are an especially difficult example. Convergence algorithms using the response kernel K have been proposed [428] for such cases. In these examples, it is most useful to analyze the response in Fourier space, which is done in Sec. 13.1 in the chapter on plane waves.

Numerical mixing schemes

The difficulty with the analysis in terms of the response kernel $\tilde{\chi}$ (or K) is that in real problems, it can be found only by calculations (similar to those for response functions, App. D and Ch. 19) that are more costly than many iterations of a standard minimization algorithm. It can be much more efficient to adopt methods from the numerical literature that build up the information on the Jacobian J (the second derivative matrix) of the system automatically rather than using physical arguments. In fact, the matrix $\tilde{\chi}$ is the Jacobian J, but in this section we will use the notation J to be consistent with commonly used notation (see App. L).

A general numerical approach for reaching a consistent solution[2] is the Broyden method [431] described in App. L. In this approach the desired quantity, the inverse Jacobian J^{-1} itself, is built up as the iterations proceed. Starting with an approximate form, J^{-1} is improved at each iteration in a way so that the change in density for step $i + 1$ is made in a direction orthogonal to all previous directions. (This is the general idea in all numerical methods that generate a "Krylov subspace" – see Apps. M and L.) The magnitude of the step is chosen to be such that it would give the result of step i projected onto the subspace generated thus far. (Note the similarity of this last requirement with solution (9.23) using $\tilde{\chi}^{-1}$; the difference is that in the Broyden method only partial information is known about

[2] This method was first used in solid state calculations by Bendt and Zunger [429] and described in more detail by Srivastava [430].

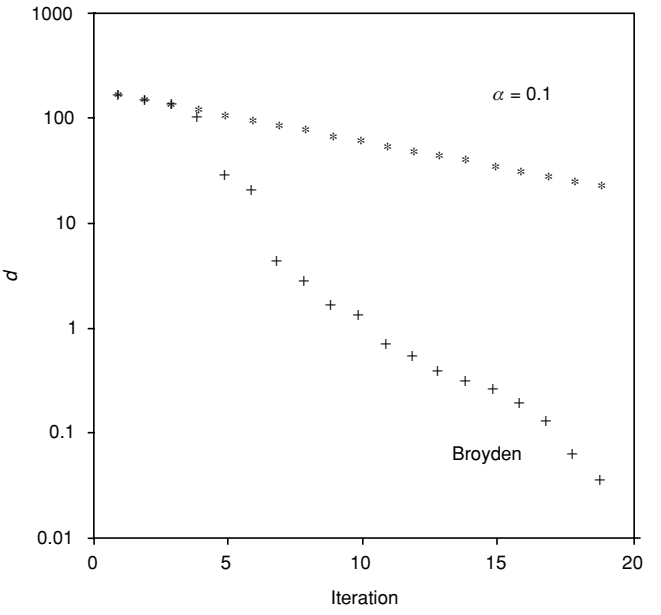


Figure 9.2. Convergence of the density for a W(1 0 0) surface (see (9.25) for the definition of d) versus iteration number for linear mixing and the Broyden method. From [432].

the Jacobian at any step i.) Thus the Broyden method combines the "best of both worlds" to make an automatic method that generates the needed parts of the Jacobian as the calculation proceeds, with essentially no added cost above that encountered in simple linear mixing.

At each iteration i the input density for the next step is given by an equation analogous to (9.23) except that $\tilde{\chi}$ is replaced by the approximate Jacobian J_i

$$n_{i+1}^{\text{in}} = n_i^{\text{in}} - J_i^{-1}\left(n_i^{\text{out}} - n_i^{\text{in}}\right), \tag{9.24}$$

and J_i^{-1} is improved at each step by expression (L.24). This can be used directly if the Jacobian matrix is small, i.e. if there are only a few components of the density for which convergence is a problem. An example is given in Sec. 13.1 in the chapter on plane waves.

Srivastava [430] has shown how to avoid storage of the Jacobian matrices by writing the predicted change $\delta n_{i+1}^{\text{in}}$ in terms of a sum over all the previous steps involving only the initial J_0^{-1}. An example of the power of the Broyden method using this approach is shown in Fig. 9.2 for the density at a (1 0 0) surface of W using an LAPW method (Ch. 17). The quantity shown is the "distance" d, which is the norm of the residual

$$d = \frac{1}{\Omega_{\text{cell}}} \int_{\Omega_{\text{cell}}} \mathrm{d}^3 r (n^{\text{out}} - n^{\text{in}})^2, \tag{9.25}$$

plotted for linear mixing with $\alpha = 0.1$ and for Broyden with $J_0 = \alpha \mathbf{1}$.

The modified Broyden method proposed by Vanderbilt and Louie [433] and adapted by Johnson [434] to also incorporate Srivistava's improvements [430] can be considered the state-of-the-art. The basic equation is given in (L.25) with discussion of the weights given

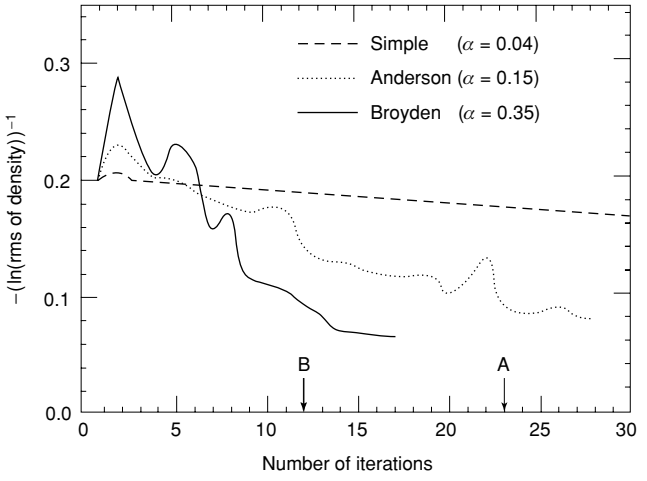


Figure 9.3. Convergence of the density for an alloy versus iteration number for linear mixing, the Anderson method [435] and the modified Broyden method. For the last case α was chosen so large that at first there is divergence until the Jacobian is improved sufficiently to lead to convergence. From [434].

in the original and subsequent papers. An example of the results for a disordered alloy $Ni_{0.35}Fe_{0.65}$ near a magnetic instability using a KKR method (Ch. 16) is given in Fig. 9.3.

9.4 Force and stress

It is straightforward to see that the usual form (Sec. 3.3) of the force theorem holds in the density functional calculations. The essential point is that – at the correct solution – the energy is at a variational minimum (or saddle point in generalized functionals) with respect to the density. Thus changes in the density as a nucleus is moved do not contribute to the first-order derivatives. The result follows from the Hohenberg–Kohn expression for the total energy, (6.12), or any of the expressions in Sec. 9.2. Since the only terms that depend *explicitly* upon the positions of the nuclei are the interaction E_{II} and the external potential, one immediately finds

$$\mathbf{F}_I = -\frac{\partial E}{\partial \mathbf{R}_I} = -\int \mathrm{d}\mathbf{r}\; n(\mathbf{r}) \frac{\partial V_{\mathrm{ext}}(\mathbf{r})}{\partial \mathbf{R}_I} - \frac{\partial E_{II}}{\partial \mathbf{R}_I}, \tag{9.26}$$

which is the "electrostatic theorem" for the forces due to Feynman [256] and given in (3.19). For non-local pseudopotentials, the force is only formally a function of the density; operationally it is defined in terms of the Kohn–Sham wavefunctions, with the general expression given in (3.20) and explicit plane wave expressions given in (13.3).

There are many possible alternative expressions for forces since any linear variation of the density can be added to (9.26) with no change in the result. The main point is very simple and is illustrated in Fig. I.1: the usual force theorem involves a nucleus moving relative to *all the electrons* as shown on the left-hand side of Fig. I.1. It is more appropriate in many

actual calculations (especially ones involving core electrons) to move part of the density along with the nucleus as illustrated in the middle part of the figure. The resulting equations can be made very simple through clever choices, as described in App. I.

In actual calculations, there are two factors that can affect the use of the force theorem (9.26): (1) explicit dependence of the basis upon the positions of the atoms, and (2) errors due to non-self-consistency. Both factors can be addressed by considering the nature of the terms omitted in going from (3.18) to (3.19). The middle terms in (3.18) that involve variations of the wavefunctions can be written in the independent-particle case as

$$\mathbf{F}_I^{(2)} = -2\mathrm{Re}\sum_i \int \mathrm{d}\mathbf{r}\frac{\partial\psi_i^*}{\partial\mathbf{R}_I}\left[\frac{1}{2}\nabla^2 + V_{\mathrm{KS}} - \varepsilon_i\right]\psi_i$$
$$- \int \mathrm{d}\mathbf{r}\,\left[V_{\mathrm{KS}} - V^{\mathrm{in}}\right]\frac{\partial n}{\partial\mathbf{R}_I}, \tag{9.27}$$

where the term involving ε_i is due to the orthonormality constraint just as in the derivation of the Kohn–Sham equation. Here V_{KS} is defined to be the self-consistent Kohn–Sham potential for the given basis set, and V^{in} is the non-self-consistent input potential that leads to the wavefunctions ψ_i.

Since the ψ_i are the eigenstates of the hamiltonian with potential V^{in}, the first term in (9.27) is zero if the changes in the ψ_i maintain orthonormality when the atom is displaced. This happens in two cases: (1) if the basis is independent of the atom positions (as in plane waves), or (2) the basis is complete. However, this term is non-zero if the basis is tied to the atoms (as in atom-centered orbitals) and the basis is incomplete. This contribution, often called the Pulay correction term [436], is straightforward – but often tedious – to include in a calculation. Only if it is included will the force be equal to the change in total energy per unit displacement. One of the great advantages of plane waves is that it is manifestly zero even if the basis is not complete.

The last term in (9.27) is the contribution due to the lack of self-consistency in the solution. This is a more serious concern for forces than for the energy, since the energy is variational (errors are second order), whereas the force expression is not. Strategies can be devised for approximate inclusion of such terms at any stage in the self-consistency iterations, even though the final potential V_{KS} is not known. These methods are based on essentially the same logic as those for achieving self-consistency discussed in Sec. 9.1, where the goal is to find the optimum choice of potential at the next step.

Stress

Stress and strain are important concepts in characterizing the states of condensed matter; however, general expressions in terms of the ground state wavefunction have been formulated only recently [104, 129]. There are a number of subtle issues and complications, so that a separate appendix, App. G, is devoted to the definition of stress and strain and to the resulting formulas that can be used in various applications.

The main results are that the stress tensor is the generalization of pressure to all the independent components of dilation and shear, and the "stress theorem" provides a way

to calculate all components of the stress tensor from the ground state wavefunction as a generalization of the virial theorem for pressure. In condensed matter, the state of the system is specified by the forces on each atom and by the macroscopic stress, which is an independent variable. The conditions for equilibrium are: (1) the total force vanishes on each atom, and (2) the macroscopic stress equals the externally applied stress. This is well established in classical simulations [437] (e.g. the Parrinello–Rahman [438] and variable metric methods [439]) and is now an integral part of electronic structure calculations [440] in which one relaxes the structure by minimizing with respect to both the positions of the atoms in a unit cell and the size and shape of the cell.

SELECT FURTHER READING

See references in "Select Further Reading" in Chapter 7.

See also:

Foulkes, W. M. C. and Haydock, R., "Tight-binding models and density-functional theory," *Phys. Rev.* B 39:12520–12536, 1989.

Jacobsen, K. W., Norskov, J. K. and Puska, M. J., "Interatomic interactions in the effective medium theory," *Phys. Rev.* B 35:7423–7442, 1987. Appendix A.

Pickett, W. E., "Pseudopotential methods in condensed matter applications," *Computer Physics Reports* 9:115, 1989.

Singh, D. J., *Planewaves, Pseudopotentials, and the APW Method*, Kluwer Academic Publishers, Boston, 1994.

Exercises

9.1 In actual calculations one can determine the energy from either of the two functionals (9.8) or (9.9). Describe how it can be useful to compute both. Which is expected to be closest to the actual converged result before convergence is reached? Which is a true variational bound? Can the difference be used as a measure of convergence?

9.2 As posed before Eq. (9.11), derive the expressions for the linear terms and thus the form of (9.11).

9.3 Fill in the steps to show that (9.13) defines a functional that is indeed extremal at the correct solution for independent variations of potential and density.

9.4 On general thermodynamic grounds, show that $E(T)$ increases quadratically with T, whereas $F(T)$ decreases quadratically. Thus a linear combination of $E(T)$ and $F(T)$ can be chosen in which the quadratic terms cancel. Using the expressions for $E(T)$ and $F(T)$ that follow from the occupation numbers, find the value of α for which $\alpha E(T) + (1-\alpha)F(T) = E(T=0)$ with corrections $\propto T^4$.

9.5 Complete the arguments to show that (9.19) is extremal at the correct solution for independent variations of all the quantities: V^{in}, n^{in}, μ, T, and the form of the occupation function $f(\varepsilon)$.

9.6 Show that the form of the electronic entropy $\sum_i f_i \ln f_i + \sum_i (1-f_i)\ln(1-f_i)$ presented in (9.19) in fact follows from the general many-body from in terms of the density matrix given by Mermin in (6.20).

9.7 Show that $\tilde{\chi}$ in (9.21) is given by

$$\tilde{\chi} + 1 = \frac{\delta n^{\text{out}}}{\delta n^{\text{in}}} = \frac{\delta n^{\text{out}}}{\delta V^{\text{in}}} \frac{\delta V^{\text{in}}}{\delta n^{\text{in}}}, \tag{9.28}$$

where $\delta n^{\text{out}}/\delta V^{\text{in}}$ is a response function defined to be χ^0 in (D.3) and the last term $\delta V^{\text{in}}/\delta n^{\text{in}}$ is K defined in (9.12). Thus the needed function $\tilde{\chi}$ can be calculated and is closely related to other uses of response functions.

9.8 Derive the constraint on the α parameter in the simple linear mixing scheme in terms of the response function; i.e. that the iterations converge only if $\alpha < 2/\tilde{\chi}_{\text{max}}^{-1} = 2\tilde{\chi}_{\text{min}}$. See also Exercise 13.3.

9.9 Derive the two terms in the corrections to the force given in (9.27) for a self-consistent independent-particle method, starting from the general form, Eq. (3.18). The self-consistency adds the second term that is not present is the general case where the hamiltonian never changes. Hint: Derive this term from the original definition of the force as a derivative of the total energy.

PART III

IMPORTANT PRELIMINARIES ON ATOMS

10

Electronic structure of atoms

Summary

This chapter is concerned with the issues of solving the self-consistent Kohn–Sham and Hartree–Fock equations in the simplest geometry. We will *not* be concerned with the intricate details of the states of many-electron atoms, but only those aspects relevant to our primary goal, the electronic structure of condensed matter and molecules. Studies of the atom illustrate the concepts and are directly relevant for following sections since they are the basis for construction of *ab initio* pseudopotentials (Ch. 11) and the augmentation functions (Ch. 16) that are at the heart of augmented plane wave (APW), linear combination of muffin-tin orbitals (LMTO), and KKR methods. Furthermore, we shall see that calculations on atoms and atomic-like radial problems are extremely useful in qualitative understanding of many aspects of condensed matter, including band widths, equilibrium volume, and bulk modulus, as discussed in Sec. 10.7.

10.1 One-electron radial Schrödinger equation

We start with the case of the hydrogenic one-electron atom. Although this is treated in many texts, it is useful to establish notation that will be used in many chapters. In the non-relativistic case, there is no spin–orbit coupling and the wavefunction can be decoupled into a product of space and spin. (The relativistic Dirac equations and spin–orbit interactions are treated in Sec. 10.4.) Since the potential acting on the electron is spherically symmetric $V_{\rm ext}(\mathbf{r}) = V_{\rm ext}(r) = -Z/r$, the spatial part of the orbital can be classified by angular momentum ($L = \{l, m_l\}$)

$$\psi_{lm}(\mathbf{r}) = \psi_l(r)Y_{lm}(\theta,\phi) = r^{-1}\phi_l(r)Y_{lm}(\theta,\phi), \tag{10.1}$$

where the normalized spherical harmonics are given by

$$Y_{lm}(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m[\cos(\theta)]\mathrm{e}^{\mathrm{i}m\phi}, \tag{10.2}$$

with $P_l^m(x)$ denoting the associated Legendre polynomials defined in Sec. K.2. The traditional labels for angular momenta are s, p, d, f, g, ..., for $l = 0, 1, 2, 3, \ldots$, and explicit formulas for the first few functions are given in (K.10).

Using the form of the Laplacian in spherical coordinates,

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left[\sin(\theta)\frac{\partial}{\partial\theta}\right] + \frac{1}{r^2\sin^2(\theta)}\left(\frac{\partial^2}{\partial\phi^2}\right), \quad (10.3)$$

the wave equation can be reduced to the radial equation (Exercise 10.1) for principal quantum number n

$$-\frac{1}{2r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left[r^2\frac{\mathrm{d}}{\mathrm{d}r}\psi_{n,l}(r)\right] + \left[\frac{l(l+1)}{2r^2} + V_{\text{ext}}(r) - \varepsilon_{n,l}\right]\psi_{n,l}(r) = 0, \quad (10.4)$$

or

$$-\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}r^2}\phi_{n,l}(r) + \left[\frac{l(l+1)}{2r^2} + V_{\text{ext}}(r) - \varepsilon_{n,l}\right]\phi_{n,l}(r) = 0. \quad (10.5)$$

The equations can be solved for bound states with the boundary conditions $\phi_{n,l}(r)$, $\psi_{n,l}(r) \to 0$ for $r \to \infty$, and $\phi_{n,l}(r) \propto r^{l+1}$ and $\psi_{n,l}(r) \propto r^l$ for $r \to 0$, and subject to the normalization condition

$$\int_0^\infty \mathrm{d}r\,\phi_{n,l}\,(r)^2 = 1. \quad (10.6)$$

For a one-electron atom, the well-known analytic solutions have eigenvalues independent of l given by

$$\varepsilon_{n,l} = -\frac{1}{2}\frac{Z^2}{n^2} \quad (10.7)$$

in Hartree atomic units.

Logarithmic grid

It is convenient to have regular grids in numerical algorithms; however, for atoms a higher density of radial points is needed near the origin and only a low density in the outer region. In the program of Herman and Skillman [445] this is accomplished by doubling the grid density several times as one proceeds toward the nucleus. This is simple in concept but leads to a complicated algorithm. Another choice is to use a logarithmic grid with $\rho \equiv \ln(r)$, which is suggested by the hydrogenic orbital which has amplitude $\propto \exp(-Zr)$, where Z is the atomic number. If we define $\tilde{\phi}_l(\rho) = r^{1/2}\psi(r)$, then the radial equation (10.5) becomes (see Fischer [441] and Exercise 10.2)

$$\left\{\frac{-\hbar^2}{2m_e}\frac{\mathrm{d}^2}{\mathrm{d}\rho^2} + \frac{l}{2}\left(l+\frac{1}{2}\right)^2 + r^2[V_{\text{ext}}(r) - \varepsilon_{n,l}]\right\}\tilde{\phi}_l(\rho) = 0, \quad (10.8)$$

This has the disadvantage of transforming the interval $0 \le r \le \infty$ to $-\infty \le \rho \le \infty$. In practice, one can treat an inner region $0 \le r \le r_1$ with a series expansion [441]

$$\phi_l(r) \propto r^{l+1}\left[1 - \frac{Zr}{l+1} + \alpha r^2 + O(r^3)\right] \quad (10.9)$$

where α is given in Sec. 6.2 of [441]. The boundary r_1 is chosen so that $\rho_1 = \ln(Zr)$ is constant for all atoms. Then the outer region $\rho_1 \leq \rho \leq \infty$ can be treated on a regular grid in the variable ρ.

The atomic equations can be solved on a regular grid in r or ρ following the flow chart for a Kohn–Sham calculation given in Fig. 9.1. The radial equations can be solved using the Numerov method described in App. L, Sec. L.1. Excellent atomic programs exist often built upon the one written originally by Herman and Skillman [445]. The ideas are described in great detail by Slater [442, 443] and by Fischer [441] in the Hartree–Fock approximation, and a simplified description is given by Koonin and Meredith [444]. Links to programs developed in the group of the author and other programs are given in Ch. 24.

10.2 Independent-particle equations: spherical potentials

The Kohn–Sham equations for a general problem have been given in (7.11)–(7.13), which are independent-particle equations with a potential that must be determined self-consistently. The same form applies to the Hartree–Fock equations (3.45) or (3.46), except that the exchange potential (3.48) is state-dependent. In each case, the solution requires solving independent-particle equations having the same form as the one-electron equation (10.5) or (10.4), except that V_{ext} is replaced by some effective potential V_{eff} that must be determined self-consistently.

For closed-shell systems, such as rare gas atoms, all the filled states are spin-paired and the charge density $n(r)$

$$n(r) = \sum_{n,l}^{\text{occupied}} (2l+1)|\psi_{n,l}(r)|^2 = \sum_{n,l}^{\text{occupied}} (2l+1)|\phi_{n,l}(r)|^2/r^2 \tag{10.10}$$

has spherical symmetry. The potential

$$V_{\text{eff}}(r) = V_{\text{ext}}(r) + V_{\text{Hartree}}(r) + V_{\text{xc}}(r). \tag{10.11}$$

is obviously spherically symmetric in the Kohn–Sham approach. In the Hartree–Fock case, the last term is the orbital-dependent exchange $\hat{V}_x^{n,l}(\mathbf{r})$, but it is not difficult to show (Exercise 10.5) that matrix elements of $\hat{V}_x$ are independent of m and σ, and lead to an effective radial potential for each n, l.

Thus the independent-particle states can be rigorously classified by the angular momentum quantum numbers $L = \{l, m_l\}$ and there is no net spin. This leads to the simplest case radial equations with eigenvectors ϕ_{l,m_l} independent of spin and eigenvalues independent of m_l. The resulting radial equation for $\phi_{n,l}(r)$, analogous to (10.5), is

$$-\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}r^2}\phi_{n,l}(r) + \left[\frac{l(l+1)}{2r^2} + V_{\text{eff}}(r) - \varepsilon_{n,l}\right]\phi_{n,l}(r) = 0, \tag{10.12}$$

which can be solved for bound states with the same boundary conditions and normalization as for the one-electron atom.

Achieving self-consistency

The general form for solution of self-consistent equations has been given in Sec. 9.3. In a closed-shell case the effective potential is spherically symmetric (Exercise 10.5). In most cases, strongly bound states pose no great problems and the linear-mixing algorithm, (9.20), is usually sufficient. (A value of $\alpha < 0.3$ will converge for most cases, but may have to be reduced for heavy atoms.) However, weakly bound "floppy" states may need the more sophisticated methods described in Sec. 9.3. and systems with near degeneracies (e.g. energies of 3d and 4s states in transition metal atoms) may present special problems, since the order of states may change during the iterations, so that filling the states according to the minimum energy principle leads to abrupt changes in the potential. Since this principle really applies only at the minimum, often there is a simple solution; however, in some cases, there is no stable solution.

An essential part of the problem is the calculation of the Hartree or Coulomb potential. There are two approaches: solution of the Poisson equation [444] or analytic formulas that can be written down for the special case of wavefunctions that are radial functions times spherical harmonics [443]. The former approach has the advantage that the Poisson equation,

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2} V_{\mathrm{Hart}}(r) = -4\pi n(r), \tag{10.13}$$

has the form of the second-order equation (L.1) and can be solved by numerical methods similar to those used for the Schrödinger equation [444]. The latter approach involves expressions that are applicable to open-shell atoms (Sec. 10.3) and special cases of the Fock integrals needed in for Hartree–Fock calculations [443].

The exchange–correlation potential $V_{\mathrm{eff}}(r)$ depends upon the type of independent-particle approximation. An explicit expressions for the state-dependent exchange potential $V_x^{n,l}(r)$ in the Hartree–Fock approximation is given in the following section. (The exchange potential is purely radial for a closed-shell atom, and is shown in Exercise 10.5.) The OEP formulation of the energy is exactly the same as for Hartree–Fock but the potential is required to be $V_x(r)$ independent of the state. In approximations such as the LDA and GGAs, explicit expressions are given in App. B. Examples of results for selected spherically symmetric atoms with various functionals are given in Tabs. 8.1 and 8.2 and are discussed further in Sec. 10.5.

10.3 Open-shell atoms: non-spherical potentials

The term "open shell" denotes cases where the spins are not paired and/or the angular momentum states $m = -l, \ldots, l$, are not completely filled for a given l. Then the proper classification is in terms of "multiplets," with given total angular momentum $J = \{j, m_j\}$ that are linear combinations of the space ($L = \{l, m_l\}$) and spin ($S = \{s, m_s\}$) variables. In general, one must deal with multiple-determinant wavefunctions. Even though the external potential

(the nuclear potential for an atom) has spherical symmetry, the effective independent-particle potential $V_{\rm eff}(\mathbf{r})$ does not, since it depends upon the occupations of the orbitals. The only simplification is that one can choose the axes of quantization so that $V_{\rm eff}(\mathbf{r}) = V_{\rm eff}(r, \theta)$ has cylindrical symmetry. Fortunately, a method due to Slater [401] shows how to reduce all the needed calculations to purely radial calculations, by using symmetry to choose appropriate multiplets. There are no general rules, but an extensive compilation of cases is given in App. 21 of [443].[1] In the atom one needs to consider only the cases where $V_{\rm eff}$ is purely radial $V_{\rm eff}(r)$ or cylindrical $V_{\rm eff}(r, \theta)$.

For open-shell problems there are a set of approximations with various degrees of accuracy:

- Restricted: treat the problem as spherical (derive $V_{\rm eff}(r)$ by a spherical average over any non-spherical terms) and independent of spin (average over spin states so that orbitals are the same for each spin state). This is the correct form for closed-shell, spin $= 0$ systems, and, with care, can be viewed as an approximation for open shells.
- Spin-unrestricted: treat as spherical but allow the potential and the orbitals to depend upon the spin. This is the correct form for half-filled shells with maximum spin so that they are closed-shell for each spin separately. This case has been treated in the section on closed shells.
- Unrestricted: treat the full problem in which only the total m_l and m_s are good quantum numbers. In this case, Slater's method [443] can be used to simplify the problem.

Equations for open-shell cases

For the fully unrestricted case, additional complications arise from the electron–electron interaction terms. If we chose an axis then the density $n(r, \theta)$ and the potential $V_{\rm xc}(r, \theta)$ have cylindrical symmetry. In the Kohn–Sham approach, the wavefunctions are expanded in spherical harmonics and angular integrals with $V_{\rm xc}(r, \theta)$ must be done numerically because the non-linear relation of $V_{\rm xc}(\mathbf{r})$ to $n(\mathbf{r})$ means that an expansion of $V_{\rm xc}(r, \theta)$ in spherical harmonics has no maximum cutoff in L. Also the Coulomb potential has multi-pole moments and solution of the Poisson equation is not as simple as in the spherical case.

For the open-shell case, the Hartree–Fock equations are actually simpler because the exchange term can be expanded in a finite sum of spherical harmonics. In order to calculate the matrix elements of the electron–electron interaction, one can use the well-known expansion [448] in terms of the spherical harmonics (App. K), which allows the factorization into terms involving $\mathbf{r}_1$ and $\mathbf{r}_2$

$$\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{1}{2l+1} \frac{r_<^l}{r_>^{l+1}} Y^*_{-lm}(\theta_2, \phi_2) Y_{-lm}(\theta_1, \phi_1), \qquad (10.14)$$

[1] In general one needs non-diagonal Lagrange multipliers $\varepsilon_{n,l;n',l'}$ [443], but these terms appear to be small [443].

where $r_<$ and $r_>$ are the lesser and the greater of r_1 and r_2. Using this expression, matrix elements involving the orbitals i, j, r, t can be written

$$\langle i\ j|\frac{1}{r_{12}}|r\ t\rangle = \delta(\sigma_i,\sigma_r)\,\delta(\sigma_j,\sigma_t)\,\delta(m_i+m_j,m_r+m_t) \times \sum_{k=0}^{k_{\max}} c^k(l_i,m_i;l_r,m_r)\,c^k(l_t,m_t;l_j,m_j)\,R^k(i,j;r,t), \quad (10.15)$$

where

$$R^k(i,j;r,t) = \int_0^\infty \int_0^\infty \phi^\dagger_{n_i,l_i}(r_1)\phi^\dagger_{n_j,l_j}(r_2)\frac{r_<^k}{r_>^{k+1}}\phi_{n_r,l_r}(r_1)\phi_{n_t,l_t}(r_2)\mathrm{d}r_1\mathrm{d}r_2. \quad (10.16)$$

The δ functions in (10.15) reflect the fact that the interaction is spin-independent and conserves the z component of the angular momentum. The angular integrals can be done analytically resulting in the Gaunt coefficients $c^k(l,m;l',m')$, which are given explicitly in [443] and App. K. Fortunately, the values of R^k are only needed for a few values of k; the maximum value is set by the vector-addition limits $|l-l'| \le k_{\max} \le |l+l'|$, and, furthermore, $c^k = 0$ except for $k+l+l' = $ even.

The radial Hartree–Fock equations are derived by functional derivatives of the energy with respect to the radial functions $\psi_{n,l,m,\sigma}(r)$. If we define a function[2]

$$Y^k(n_i,l_i;n_j,l_j;r) = \frac{1}{r^k}\int_0^r \mathrm{d}r'\,\phi^\dagger_{n_i,l_i}(r')\phi_{n_j,l_j}(r')r'^k + r^{k+1}\int_r^\infty \mathrm{d}r'\,\phi^\dagger_{n_i,l_i}(r')\phi_{n_j,l_j}(r')\frac{1}{r'^{k+1}}, \quad (10.17)$$

and use the relation between the Gaunt and Clebsch–Gordan coefficients (K.17), then the Hartree potential is given by

$$V_{\mathrm{Hartree}}(r) = \sum_{\sigma=\uparrow,\downarrow}\sum_{j=1,N_\sigma}\sum_{k=0}^{\min(2l_i,2l_j)} (-1)^{m_i+m_j}\frac{(2l_i+1)(2l_j+1)}{(2k+1)^2}\,\frac{Y^k(n_j,l_j;n_j,l_j;r)}{r} \times C^{k0}_{l_i0,l_i0}\,C^{k0}_{l_j0,l_j0}\,C^{k0}_{l_im_i,l_i-m_i}\,C^{k0}_{l_jm_j,l_j-m_j} \quad (10.18)$$

and the exchange potential acting on state n_i, l_i, σ_i can be written

$$V_x(r) = -\sum_{\sigma=\uparrow,\downarrow}\delta(\sigma,\sigma_i)\sum_{j=1,N_\sigma}\sum_{k=|l_i-l_j|}^{l_i+l_j}\frac{(2l_i+1)(2l_j+1)}{(2k+1)^2} \left[C^{k0}_{l_i0,l_j0}\,C^{km_i-m_j}_{l_im_i,l_j-m_j}\right]^2\,\frac{Y^k(n_j,l_j;n_j,l_j;r)}{r}\,\frac{\phi_{n_j,l_j}(r)}{\phi_{n_i,l_i}(r)}, \quad (10.19)$$

where $C^{j_3m_3}_{j_1m_1,j_2m_2}$ are the Clebsch–Gordan coefficients defined in App. K.

[2] This follows the definition of Slater [442], p. 180.

10.4 Relativistic Dirac equation and spin–orbit interactions

Relativistic effects are essential for heavy atoms. Fortunately, they originate deep inside the core, so that it is sufficient to solve the relativistic equations in a spherical atomic geometry. The results carry over to molecules and solids essentially unchanged. In the actual calculation on solids or molecules, relativistic effects can be included directly within the augmentation methods (Ch. 16), where the calculation is equivalent to that on an atom, or indirectly within other methods, e.g. pseudopotentials (Ch. 11). Relativistic effects can be built into pseudopotentials by generating them using relativistic atomic calculations; the pseudopotentials can then be used in a *non-relativistic Schrödinger equation* to determine the valence states *including relativistic effects* [449, 450].

The famous equation proposed by Dirac in 1928 [27, 451] generalizes the Schrödinger equation in a relativistically covariant form

$$i\hbar\frac{\partial}{\partial t}\Psi = (c\alpha\cdot\mathbf{p} + \beta mc^2)\psi = H\Psi, \tag{10.20}$$

where Ψ is a four-component single-particle wave function that describes spin-$\frac{1}{2}$ particles. Here $\mathbf{p} = -\imath\hbar\boldsymbol{\nabla}$ is the usual momentum operator, and the (4×4) matrices α_i and β are written in terms of the Pauli matrices

$$\alpha_i = \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}, \tag{10.21}$$

where σ_i are the (2×2) Pauli spin matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{10.22}$$

and the unit entries of β are (2×2) unit matrices. Solution of Eq. (10.20) in its general form can be found in many texts, such as [452] and [453], and a clear discussion of a practical solution in the case of a spherical potential can be found in [446] and [447], and in the reviews [134] and [132].

It is convenient to write the solution in the form

$$\Psi(x^\mu) = \mathrm{e}^{-\mathrm{i}\varepsilon t/\hbar}\begin{pmatrix} \phi(\mathbf{r}) \\ \chi(\mathbf{r}) \end{pmatrix}, \tag{10.23}$$

where $\phi(\mathbf{r})$ and $\chi(\mathbf{r})$ are time-independent two-component spinors describing the spatial and spin-$\frac{1}{2}$ degrees of freedom. The Dirac equation becomes a set of coupled equations for ϕ and χ,

$$\begin{aligned} c(\sigma\cdot\mathbf{p})\chi &= (\varepsilon - V - mc^2)\phi, \\ c(\sigma\cdot\mathbf{p})\phi &= (\varepsilon - V + mc^2)\chi. \end{aligned} \tag{10.24}$$

For electrons (positive energy solutions), ϕ is the large component and χ is the small component (by a factor $\propto 1/(mc^2)$).

In the case of a spherical potential $V(r)$, one can make use of conservation of parity and total angular momentum denoted by the quantum numbers jm. Then the wavefunction

for each principle quantum number n can be written in terms of radial and angular-spin functions [453],

$$\psi^l_{njm} = \begin{pmatrix} g_{nj}(r)\varphi^l_{jm} \\ i f_{nj}(r)\frac{\sigma\cdot\mathbf{r}}{r}\varphi^l_{jm} \end{pmatrix}, \tag{10.25}$$

which defines two functions with the same jm but opposite parity for the two possible values $l = j \pm \frac{1}{2}$. The two-component functions φ^l_{jm} can be written explicitly as:

$$\begin{aligned} &\text{for } j = l + \frac{1}{2}, \\ &\varphi^l_{jm} = \sqrt{\frac{l+\frac{1}{2}+m}{2l+1}}\, Y_l^{m-\frac{1}{2}}\chi_{\frac{1}{2}+\frac{1}{2}} + \sqrt{\frac{l+\frac{1}{2}-m}{2l+1}}\, Y_l^{m+\frac{1}{2}}\chi_{\frac{1}{2}-\frac{1}{2}}, \\ &\text{for } j = l - \frac{1}{2}, \\ &\varphi^l_{jm} = \sqrt{\frac{l+\frac{1}{2}-m}{2l+1}}\, Y_l^{m-\frac{1}{2}}\chi_{\frac{1}{2}+\frac{1}{2}} - \sqrt{\frac{l+\frac{1}{2}+m}{2l+1}}\, Y_l^{m+\frac{1}{2}}\chi_{\frac{1}{2}-\frac{1}{2}}. \end{aligned} \tag{10.26}$$

The resulting equations for the radial functions are simplified if we define the energy,

$$\varepsilon' = \varepsilon - mc^2, \tag{10.27}$$

a radially varying mass,

$$M(r) = m + \frac{\varepsilon' - V(r)}{2c^2}, \tag{10.28}$$

and the quantum number κ,

$$\kappa = \pm(j + \frac{1}{2}) \qquad \begin{cases} +, & \text{if } l = j + \frac{1}{2} \Rightarrow \kappa = l, \\ -, & \text{if } l = j - \frac{1}{2} \Rightarrow \kappa = -(l+1). \end{cases} \tag{10.29}$$

Note that $\kappa(\kappa+1) = l(l+1)$ in either case. Then the coupled equations can be written in the form of the radial equations [132, 134, 446, 447]

$$\begin{aligned} &-\frac{\hbar^2}{2M}\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}g_{n\kappa}}{\mathrm{d}r}\right) + \left[V + \frac{\hbar^2}{2M}\frac{l(l+1)}{r^2}\right]_{g_{n\kappa}}, \\ &-\frac{\hbar^2}{4M^2c^2}\frac{\mathrm{d}V}{\mathrm{d}r}\frac{\mathrm{d}g_{n\kappa}}{\mathrm{d}r} - \frac{\hbar^2}{4M^2c^2}\frac{\mathrm{d}V}{\mathrm{d}r}\frac{(1+\kappa)}{r}g_{n\kappa} = \varepsilon' g_{n\kappa}, \end{aligned} \tag{10.30}$$

and

$$\frac{\mathrm{d}f_{n\kappa}}{\mathrm{d}r} = \frac{1}{\hbar c}(V - \varepsilon')g_{n\kappa} + \frac{(\kappa-1)}{r}f_{n\kappa}. \tag{10.31}$$

These are the general equations for a spherical potential; no approximations have been made thus far. Equation (10.30) is the same as an ordinary Schrödinger equation except

that the mass M is a function of radius and there are two added terms on the left-hand side, which are, respectively, the Darwin term and the spin–orbit coupling. The latter can be written out explicitly in terms of the spin using the relation

$$\mathbf{L} \cdot \sigma \varphi_{\kappa m} = -\hbar(1+\kappa)\varphi_{\kappa m}, \tag{10.32}$$

where $\varphi_{\kappa m}$ is the appropriate φ^l_{jm} determined by κ.

Scalar relativistic equation and spin–orbit coupling

If we make the approximation that the spin–orbit term is small, then we can omit it in the radial equations for g and f and treat it by perturbation theory. Then (10.31)–(10.33) depend only upon the principle quantum number n and orbital angular momentum l and can be written in terms of the approximate functions, $\tilde{g}_{nl}$ and $\tilde{f}_{nl}$, leading to,

$$-\frac{\hbar^2}{2M}\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\tilde{g}_{nl}}{\mathrm{d}r}\right) + \left[V + \frac{\hbar^2}{2M}\frac{l(l+1)}{r^2}\right]\tilde{g}_{nl} - \frac{\hbar^2}{4M^2c^2}\frac{\mathrm{d}V}{\mathrm{d}r}\frac{\mathrm{d}\tilde{g}_{nl}}{\mathrm{d}r} = \varepsilon'\tilde{g}_{nl} \tag{10.33}$$

and

$$\tilde{f}_{nl} = \frac{\hbar}{2Mc}\frac{\mathrm{d}\tilde{g}_{nl}}{\mathrm{d}r}, \tag{10.34}$$

with the normalization condition

$$\int (\tilde{g}^2_{nl} + \tilde{f}^2_{nl}) r^2 \,\mathrm{d}r = 1. \tag{10.35}$$

Equation (10.33) is the scalar relativistic radial equation, which can be solved by the same techniques as the usual non-relativistic equation. The other equations can then be treated easily on the radial grid.

Finally, the spin–orbit term can be included following the approach of MacDonald et al. [447]. Together with relation (10.32) the spin–orbit hamiltonian coupling the large components of the wavefunction has the form

$$\hat{H}_{\mathrm{SO}} = \frac{\hbar^2}{2M^2c^2}\frac{1}{r}\frac{\mathrm{d}V}{\mathrm{d}r}\mathbf{L}\cdot\sigma, \tag{10.36}$$

which can often be treated as a small perturbation. Since this term originates deep in the core near the nucleus where $\frac{1}{r}\frac{\mathrm{d}V}{\mathrm{d}r}$ is large, the present spherical derivation of the spin–orbit term carries over from the atom to a solid or molecule.

10.5 Example of atomic states: transition elements

Examples for selected spherically symmetric atoms have been given in Tabs. 8.1 and 8.2 using various functionals for exchange and correlation, respectively. Three atoms shown there (He, Be, and Ne) are closed shell and the other two (H and N) are half-filled shells

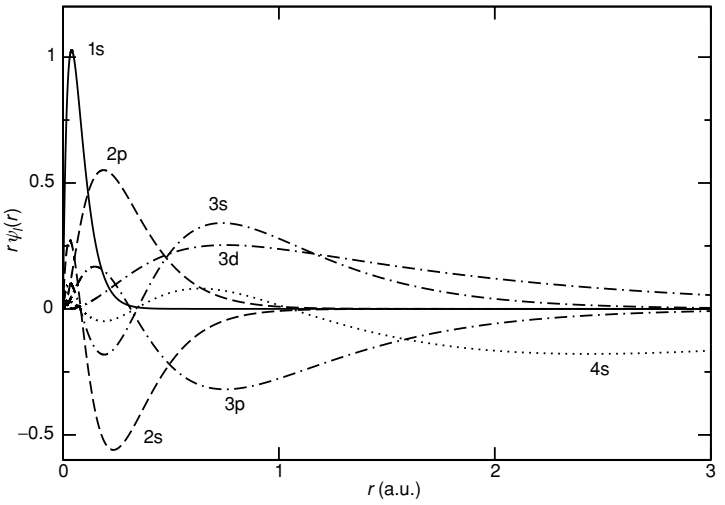


Figure 10.1. Radial wavefunction $\phi_l(r) = r\psi_l(r)$ for Mn in the $3d^{5\uparrow}4s^2$ state showing all the orbitals. Note that the 4s states are much more delocalized than the 3d states even though they have similar energies. In contrast, the maximum in the 3d is close to that of the 3s and 3p even though these are much lower in energy and are called "semi-core" states. Since the atom is spin polarized, the orbital shapes depend upon the spin. There is a clear effect on the 3d and 4s, which is not shown for simplicity.

in which the spatial wavefunction is spherically symmetric. The latter are called "spin unrestricted" and require spin functionals with separate potentials for $V^{\sigma}_{\rm eff}(r)$ for spin up and down. The results show that the local approximation works remarkably well, considering that it is derived from the homogeneous gas, and that GGAs in general improve the overall agreement with experiment.

Hydrogen is the special case where the one-electron solution for the ground state energy is exact. This is satisfied in any theory that has no self-interactions, including Hartree–Fock and exact exchange (EXX). The results in the tables indicate the error in the other functionals in this limit. The accuracy is quite remarkable, especially for functionals derived from the homogeneous gas, which supports the use of the functionals for the entire range from homogeneous solids to isolated atoms. Nevertheless, there are important errors. In particular, there are large effects upon the eigenvalues due to the long-range asymptotic form of the potential. The form is correct in Hartree–Fock and EXX calculations that take into account the non-local exchange, but is incorrect in local and GGA approximations. The effects are large in one- and two-electron cases, shown explicitly in Sec. 8.9, but are smaller in heavier atoms with many electrons.

As an example of a many-electron atom, the wavefunctions for spin polarized Mn are shown in Fig. 10.1, calculated without relativistic corrections and spin–orbit coupling. The states of this transition metal element illustrate the difference between the loosely bound outer 4s states and the more localized 3d states. The atom is in the $3d^{5\uparrow}4s^2$ state and the

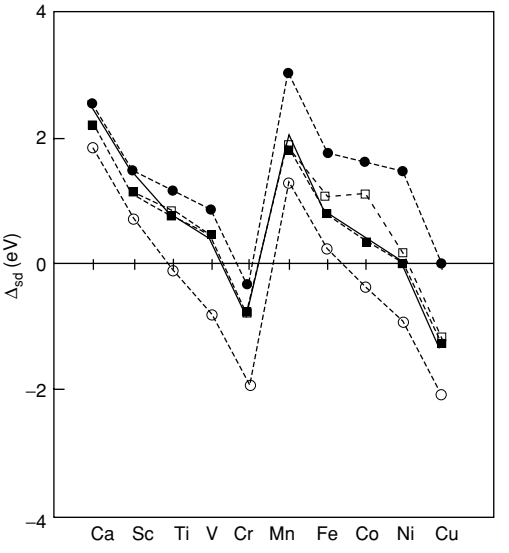


Figure 10.2. Promotion energies for d → s electrons in the 3d transition metal series. Shown are the energies $\Delta_{sd} \equiv E_{total}[3d^{n-1}4s^1] - E_{total}[3d^{n-2}4s^2]$ from experiment (solid line) and from calculated total energy differences with different functionals. This shows that LSDA (open circles), typically tends to underestimate the promotion energy, i.e. the d states are underbound, whereas full non-local exchange (solid circles) tends to have the opposite effect. Including correction for self-interaction (open squares) and screening, the exchange (solid squares) improves the results considerably. Such effects carry over to solids as well. From [454].

right-hand side of the figure shows the up and down wavefunctions for the outer states. Since the d shell is filled for one spin (Hund's first rule for maximum spin), the atom is in a spherically symmetric spatial state. Note that the 3d states are actually in the same spatial region as the strongly bound "semi-core" 3s and 3p states.

Atomic calculations can be used to gain insight into practical aspects of density functional theory and how it can be expected to work in solids. Transition metals provide an excellent example because the d states retain much of their atomic character in the solid. For example, the relative energies of the 3d and 4s states can be expected to carry over to the solid. Figure 10.2 shows the promotion energies for transferring a d electron to an s state. The energies plotted are the experimental promotion energies and the calculated values with different functionals. The calculated energies are *total energy differences* $\Delta_{sd} \equiv E_{total}[3d^{n-1}4s^1] - E_{total}[3d^{n-2}4s^2]$, not differences of eigenvalues, since Δ_{sd} is a better measure of the true energy for promotion than the eigenvalue difference. The primary point to notice is that there are opposite tendencies for the local approximation (LSDA) and for exact non-local exchange (EXX) which is very close to Hartree–Fock. The former leads to underbinding of the d relative to the state, whereas the latter leads to overbinding. The example of screened exchange is one version of hybrid functionals (Sec. 8.8) that tend to give results intermediate between Hartree–Fock and LDA.

Three conclusions can be drawn from the atomic calculations that are very important for applications in molecules and solids:

- Typical density functional theory calculations using LDA or GGA functionals can be expected to give errors in the relative positions of bands, especially bands of different character, such as localized 3d versus delocalized 4s states. The errors can be of the order of electron volts.
- Hybrid functionals (Sec. 8.8) are promising ways to improve accuracy of practical functionals applied to molecules and solids. Such functionals are widely applied in chemistry where implementation is rather simple as a mixture of Hartree–Fock and LDA or GGA calculations. Only a few applications to solids have been done, but they are promising (see, for example, tests on a large set of molecules in [402] and the band structure of Si given in Fig. 15.3.
- The results for transition metals illustrate the large difference between valence orbitals that are delocalized (s and p) and the d states that are much more localized. The effects of exchange and correlation are much more important in highly localized orbitals, leading to effects of strong correlation in transition metal systems. The essence of the "LDA+U" approach (Sec. 8.6) is to include orbital-dependent interactions that describe this large difference; however, it is rather heuristic in character and does not lead to a universal functional. The "U" term modifies the energies of selected localized orbitals as a function of their occupancy in a way that describes a correlated system better than the universal functionals. This can provide qualitatively improved descriptions of strongly correlated systems like transition metal oxides [366].

10.6 Delta-SCF: electron addition, removal, and interaction energies

In localized systems, electron excitation, addition, and removal energies can all be calculated as *energy differences* $\Delta E_{12} = E_2 - E_1$ for a transition between states 1 and 2, instead of eigenvalues calculated for state 1 or 2. This is known as "delta-SCF" and in self-consistent field methods, it produces more accurate results since the energy difference includes effects of relaxation of all the orbitals. Following the Slater transition state argument ([455], p. 51), the energy difference can be approximated by the eigenvalue calculated at the occupation *half-way between the two states.* For example, an electron removal energy is the eigenvalue when $1/2$ an electron is missing in the given state; a transition energy is the eigenvalue difference calculated when $1/2$ an electron is transferred between the two states,

$$\Delta E(N \to N-1) = E(N-1) - E(N) \approx \epsilon_i\left(N - \frac{1}{2}\right), \tag{10.37}$$

where i denotes a particular state and $N - \frac{1}{2}$ means the density is $n(\mathbf{r}) - \frac{1}{2}|\psi_i(\mathbf{r})|^2$. See Exercise 10.8 for a statement of the ideas involved in the arguments and their proof.

The delta-SCF or transition state methods can be used in atomic calculations and compared with experiment. For example, in [456] it was shown that results from both delta-SCF and transition state calculations using LDA are in good agreement with experiment for the

first and second ionization energies of 3d electrons in Cu. In addition, it is straightforward to calculate interaction energies as energy differences. An effective interaction energy that includes relaxation of the orbitals is given by the difference of first and second ionization energies, which in terms of the transition state rule can be written

$$U \equiv [E(N-1) - E(N)] - [E(N-2) - E(N-1)] \approx \epsilon_i\left(N - \frac{1}{2}\right) - \epsilon_i\left(N - \frac{3}{2}\right). \tag{10.38}$$

See Exercise 10.11 for discussion of the interpretation and suggested exercises.

In a solid it is not a obvious how to carry out such a calculation, since there is no localized state whose occupation can be varied. (Of course, there is no effect if the state is delocalized in an infinite system.) One general approach is to identify a localized state, e.g. a Wannier state or an approximation thereto, and do calculations very much like those in an atom. Another approach is the "constrained DFT" approach in which the potential is varied in a region corresponding to a localized orbital; the variation in occupation can then be used to find similar information. Examples are the calculation of addition and removal energies for 4f electrons [457] and for 3d electrons, giving results like those shown in Fig. 10.2. These calculations can also be used to find effective "U" interaction parameters. The difference between the energies needed to add or to remove electrons in the same shell is a direct measure of the interactions between electrons in that shell. This approach has been used with considerable success in many calculations for transition metal oxides, e.g. in [456] and [458] for La_2CuO_4.

10.7 Atomic sphere approximation in solids

In a solid the wavefunctions tend to be atomic-like near each atom. This results from the full calculations described later, but it is instructive to see that qualitative (sometimes quantitative) information about electronic bands, pressure, and energy of simple solids can be derived from calculations with spherical symmetry, analogous to the usual atomic calculations of Secs. 10.2 or 10.4 with only one difference: a change of boundary conditions to mimic the extreme limits of each band in a solid. Such calculations are also instructive because they are very closely related to the radial atomic-like calculations used in augmented plane wave (APW), linear combination of muffin-tin orbitals (LMTO), and KKR methods of Chs. 16 and 17.

The basic ideas are in the original work of Wigner and Seitz [49], extended by Andersen [461] to describe the width of bands formed by states with a given angular momentum. The environment of an atom in a close-packed solid is mimicked by boundary conditions on an atomic sphere, i.e. approximating the Wigner–Seitz cell as a sphere. As indicated in Fig. 10.3, for each angular momentum l the free-atom boundary conditions are replaced by the condition that the wavefunction be zero at the boundary (the highest energy antibonding-type state that corresponds to the top of the band) or have zero derivative at the boundary (the lowest energy bonding-type state that corresponds to the bottom of the band). The difference is the band width W_l for angular momentum l.

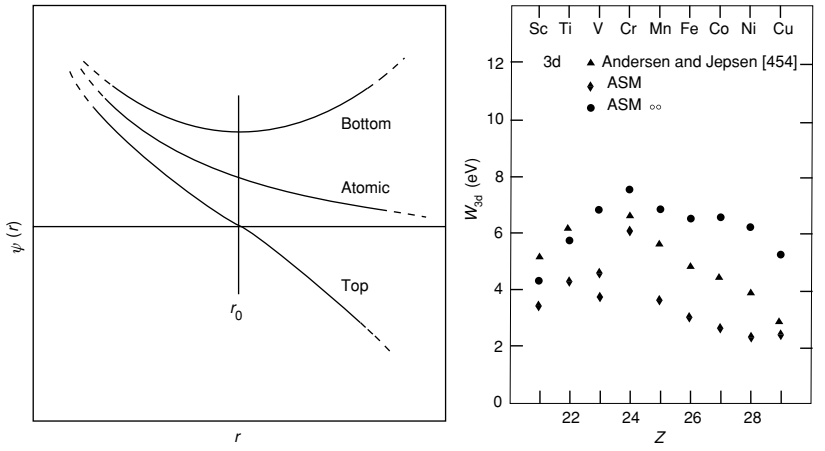


Figure 10.3. Left: Schematic figure of the radial wavefunction in the atomic sphere approximation (ASA) to a solid, where r_0 is the Wigner–Seitz radius, roughly $1/2$ the distance to a neighboring atom. The lines denote the different boundary conditions that correspond to the "bottom" lowest energy bonding-type state, the "top" highest energy antibonding-type state, and the usual free-atom state. The difference in energy from bottom to top is an estimate of the band width in a solid. Right: Estimates of the d-band width for the 3d transition metals from Eq. (10.41) (called "ASM") compared to full calculations using the LMTO method (Ch. 17) by Andersen and Jepsen [459]. The circles show an additional approximation described in [460]. From Straub and Harrison [460].

This simple picture contains the important ingredients for understanding and semiquantitative prediction of band widths in condensed matter. The width is directly related to the magnitude, the wavefunction, and its slope at the boundary. Thus the width varies from narrow-band atomic-like to wide-band delocalized states exactly as indicated in the fist figure of this book, Fig. 1.1, as there is increased overlap of the atomic states. Furthermore, it is reasonably accurate as illustrated on the right-hand side of Fig. 10.3 from the work of Straub and Harrison [455]. It gives an estimate that is particularly good in close-packed systems, but also is a good starting point for liquids and even dense plasmas [462–465] One can also calculate properties like the pressure, as discussed in Sec. I.3. This gives a great start to understanding electronic structure of condensed matter!

Explicit expressions for the band widths can be derived from the radial equation, (10.4), for the wavefunction $\psi_{n,l}(r)$. The band formed for each state n, l is considered separately, and we can drop the subscripts to simplify the equations. Consider any two solutions $\psi^1(r)$ and $\psi^2(r)$ with eigenvalues ε^1 and ε^2 obtained with two different specific boundary conditions. Let the equation for $\psi^1(r)$ be multiplied by $r^2\psi^2(r)$ and integrated from $r = 0$ to the boundary $r = r_0$, and similarly for the equation for $\psi^2(r)$. Integrating by parts and subtracting the equations leads to [460] (Exercise 10.12),

$$-\frac{1}{2}r_0^2\left(\psi^2\frac{\mathrm{d}\psi^1}{\mathrm{d}r}-\psi^1\frac{\mathrm{d}\psi^2}{\mathrm{d}r}\right)_{r-r_0}=(\varepsilon^1-\varepsilon^2)\int_0^{r_0}\mathrm{d}r r^2\psi^1\psi^2. \tag{10.39}$$

If we let $\psi^2(r)$ be the solution for the top of the band ($\psi^2(r_0) = 0$) and $\psi^1(r)$ for the bottom ($\mathrm{d}\psi^1(r)/\mathrm{d}r = 0$ at $r = r_0$), then this equation can be written

$$W \equiv \varepsilon^2 - \varepsilon^1 = -\frac{1}{2}\frac{r_0^2\left(\psi^1 \frac{\mathrm{d}\psi^2}{\mathrm{d}r}\right)_{r=r_0}}{\int_0^{r_0} \mathrm{d}r r^2 \psi^1 \psi^2}. \tag{10.40}$$

This gives the width W for each n, l in terms of the two solutions with the boundary conditions described above.

Finally, a simple, insightful expression for the band width as a function of the Wigner–Seitz radius r_0 can be derived [460] from a single atomic calculation with the usual boundary conditions. As suggested by the interpretation of the bottom and top of the band as bonding and antibonding, as illustrated in Fig. 10.3, the value of the bonding function at r_0 is approximately twice the value of the atomic function ψ^a, and similarly for the slope. Further approximating the product to be $\psi^1\psi^2 \approx [\psi^a]^2$ leads to the very simple expression (see Exercise 10.12)

$$W \approx -2\frac{r_0^2\left(\psi^a \frac{\mathrm{d}\psi^a}{\mathrm{d}r}\right)_{r=r_0}}{\int_0^{r_0} \mathrm{d}r r^2 (\psi^a)^2}. \tag{10.41}$$

The right-hand side of Fig. 10.3 shows the band widths for d-bands in transition metals calculated from this simple formula compared to full calculations using the LMTO method (Ch. 17).

SELECT FURTHER READING

Books:

Fischer, C. F., *The Hartree-Fock Method for Atoms: A Numerical Approach*, John Wiley and Sons, New York, 1977.

Koonin, S. E., and Meredith, D. C., *Computational Physics*, Addison Wesley, Menlo Park, CA, 1990.

Slater, J. C., *Quantum Theory of Atomic Structure, Vol. 1*, McGraw-Hill, New York, 1960.

Slater, J. C., *Quantum Theory of Atomic Structure, Vol. 2*, McGraw-Hill, New York, 1960.

Tinkham, M., *Group Theory and Quantum Mechanics*, McGraw-HIll, New York, 1964.

Early numerical work:

Herman, F., and Skillman, S., *Atomic Structure Calculations*, Prentice-Hall, Engelwood Cliffs, N. J., 1963.

Relativistic theory:

Koelling, D. D., and Harmon, B. N., "A technique for relativistic spin-polarized calculations," *J. Phys. C* 10:3107–3114, 1977.

Kübler, J., *Theory of Itinerant Electron Magnetism*, Oxford University Press, Oxford, 2001.

Kübler, J., and Eyert, V., in *Electronic and Magnetic Properties of Metals and Ceramics*, edited by K. H. J. Buschow, VCH-Verlag, Weinheim, Germany, 1992, p. 1.

MacDonald, A. H., Pickett, W. E., and Koelling, D., "A linearised relativistic augmented-plane-wave method utilising approximate pure spin basis functions," *J. Phys. C: Solid State Phys.* 13:2675–2683, 1980.

Exercises

10.1 Show explicitly that the wave equation can indeed be written in the form of Eq. (10.4).

10.2 Derive the form of the radial equation, (10.8) in terms of the transformed variable $\rho \equiv \ln(r)$. Give two reasons why a uniform grid in the variable ρ is an advantageous choice for an atom.

10.3 Show that the Hartree–Fock equations are exact for the states of H. Show that the change in energy computed by *energy differences* gives exact excitations; but the eigenvalues do not.

10.4 Show that the OEP equations are exact for the states of H, just like Hartree–Fock. But unlike Hartree–Fock the eigenvalues give exact excitation energies.

10.5 Show that the general Hartree–Fock equations simplify in the closed-shell case so that the exchange potential is spherically symmetric.

10.6 Show that for the ground state of He the general Hartree–Fock equations simplify to the very simple problem of one electron moving in the average potential of the other, with both electrons required to be in the same spatial orbital.

10.7 There are results that emerge in relativistic quantum mechanics that may be surprising. For example, show that there is a 2p state that has non-zero expectation value at the origin, whereas it is zero in the non-relativistic theory.

10.8 The Slater transition state argument ([455], p. 51) is based upon two facts. First, an eigenvalue is the derivative of the energy with respect to the occupation of the given state (the "Janak theorem"), and, second, that the eigenvalue varies with occupation and can be represented in a power series.
(a) Using these facts derive the "half-way" rule.
(b) Argue that one can derive the "half-way" rule based purely on the fact that one wants a result that is symmetric between the two states.
(c) Derive the explicit expression (10.37) for electron removal.

10.9 Solve the Schrödinger equation in Sec. 10.1 for a particle in a spherical box of radius R. If the boundary conditions are that $\psi = 0$ at $r = R$, show that the solutions are $\psi(r) = \sin kr/r$ and derive the eigenvalues and normalization factors for the states with the three lowest energies. Show that all energies scale as $\propto 1/R^2$.

10.10 Derive the pressure $-\mathrm{d}E/\mathrm{d}\Omega$ from the expression for the energy in the problem above. Show that this is equivalent to the expression for the pressure in a spherical geometry given in (I.8).

10.11 The expression (10.38) provides a way to calculate interactions.
(a) Show that these are "effective" in the sense that orbital relations are included and are *exact* if the energies $E(N)$, $E(N-1)$, and $E(N-2)$ are exact.
(b) Derive expression (10.38) using the same arguments as in Exercise 10.8.
(c) Use an atomic code to calculate the first and second ionization energies of 3d electrons in Cu. The difference is the effective d–d interaction. A better measure of the net effect in a solid

is to calculate the difference $E(3d^9) - E(3d^8 4s^1)$. Compare your results with those of [456]. In this case, the effective d–d interaction is decreased because the added s electron "screens" the change in charge of the d state. As argued in [457] this is close to the screening that occurs in a solid; hence, the screened interaction is the appropriate effective interaction in the solid.

10.12 Following the arguments given in conjunction with (10.39)–(10.41), derive the approximate expressions (10.40) and (10.41) for the band width. The full argument requires justifying the argument that this corresponds to the maximum band width in a solid and deriving the explicit expression using the linearized formulas for energy as a function of boundary condition.

10.13 The wavefunction for atomic hydrogen can be used to estimate hydrogen band widths at various states, using the approximate form of Eq. (10.41). Apply this approach to the H_2 molecule to calculate bonding/anti-bonding splitting and compare these with the results shown in Fig. 8.3. Use this expression to derive a general argument for the functional form of the splitting as a function of proton separation R at large R. Evaluate explicitly at the equilibrium R and compare with Fig. 8.3. Calculate the band width expected for hydrogen at high density ($r_s = 1.0$) where it is expected to be stable as a close packed crystal with 12 neighbors. (The result can be compared with the calculations in Exercises 12.13 and 13.5.)

10.14 Use an atomic code (possibly modified to have different boundary conditions) to calculate the band widths for elemental solids using the approach described in Sec. 10.7. As an example consider 3d and 4s bands in fcc Cu. Compare these with the bands shown in Fig. 2.24.

11

Pseudopotentials

Summary

The fundamental idea of a "pseudopotential" is the replacement of one problem with another. The primary application in electronic structure is to replace the strong Coulomb potential of the nucleus and the effects of the tightly bound core electrons by an effective ionic potential acting on the valence electrons. A pseudopotential can be generated in an atomic calculation and then used to compute properties of valence electrons in molecules or solids, since the core states remain almost unchanged. Furthermore, the fact that pseudopotentials are not unique allows the freedom to choose forms that simplify the calculations and the interpretation of the resulting electronic structure. The advent of "*ab initio* norm-conserving" and "ultrasoft" pseudopotentials has led to accurate calculations that are the basis for much of the current research and development of new methods in electronic structure, as described in the following chapters.

Many of the ideas originated in the orthogonalized plane wave (OPW) approach that casts the eigenvalue problem in terms of a smooth part of the valence functions plus core (or core-like) functions. The OPW method has been brought into the modern framework of total energy functionals by the projector augmented wave (PAW) approach that uses pseudopotential operators but keeps the full core wavefunctions.

11.1 Scattering amplitudes and pseudopotentials

The scattering properties of a localized spherical potential at any energy ε can be formulated concisely in terms of the phase shift $\eta_l(\varepsilon)$, which determines the cross-section and all properties of the wavefunction outside the localized region. The derivation and explicit formulas are given in Sec. J.1. This is a central concept for many phenomena in physics, such as scattering cross-sections in nuclear and particle physics, resistance in metals due to scattering from impurities, and electron states in crystals described by phase shifts in the augmented plane wave and multiple scattering KKR methods (Ch. 16). *The essential point for this chapter is that all properties of the wavefunction outside the scattering region are invariant to changes in the phase shift by any multiple of* 2π.

Pseudopotentials have a long history in such problems. The basic idea is that the scattering, i.e. phase shifts modulo $2n\pi$, can be reproduced over a range of energies by a different potential chosen to have more desirable properties. One of the early examples of this idea is illustrated in Fig. 11.1, taken from papers by Fermi and coworkers on low-energy electron scattering from atoms [58] and low-energy neutron scattering from nuclei [477]. The incident plane wave is resolved into spherical harmonics as in Fig. J.1, and the figure shows the radial wavefunction for one angular momentum l in a scattering state with a small positive energy. The closely spaced nodes in the wavefunction near the origin indicate that the kinetic energy is large, i.e. that there is a strong attractive potential. In fact, there must be lower energy bound states (with fewer nodes) to which the scattering state must be orthogonal.[1]

It is instructive to consider the changes in the wavefunction $\phi = r\psi$ outside the scattering region as a function of the scattering potential. If there were no potential, i.e. phase shift $\eta_l(\varepsilon) = 0$, then Eq. (J.4) leads to $\phi \propto r j_l(\kappa r)$, which extrapolates to zero at $r = 0$. In the presence of a potential the wavefunction outside the central region is also a free wave but phase shifted as in (J.4). A weak potential leads to a small phase shift $\eta < 2\pi$. If the potential is made more attractive, the phase shift increases with a new bound state formed for each integer multiple 2π. From the explicit solution (J.4), it is clear that the wavefunction outside the central region is exactly the same for any potential that gives the same phase shift $\eta_l(\varepsilon)$ modulo any multiple of 2π. In particular, the scattering in Fig. 11.1 can be reproduced at the given energy ε by a weak potential that has no bound states and a scattering state with no nodes. For example, one can readily find a square well with the same scattering properties at this energy (see Exercise 11.2). The aim of pseudopotential theory is to find useful pseudopotentials that faithfully represent the scattering over a desired energy range.

Perhaps the first use of pseudopotentials in solids was by Hellmann [59, 60] in 1935, who developed an effective potential for scattering of the valence electrons from the ion cores in metals and formulated a theory for binding of metals that is remarkably similar to present-day pseudopotential methods. The potentials, however, were not very weak [478], so that the calculations were not very accurate using perturbation methods available at the time.

Interest in pseudopotentials in solids was revived in the 1950s by Antoncik [479, 480] and Phillips and Kleinman [481], who showed that the orthogonalized plane wave (OPW) method of Herring [57, 482] (see Sec. 11.2) can be recast in the form of equations for the valence states only that involves a weaker effective potential. Their realization that the band structures of sp-bonded metals and semiconductors could be described accurately by a few empirical coefficients led to the basic understanding of a vast array of properties of sp-bonded metals and semiconductors. Excellent descriptions of the development of pseudopotentials before 1970 can be found in the review of Heine and Cohen [467, 469] and in the book *Pseudopotentials in the Theory of Metals* by Harrison [468].

[1] The figure also illustrates that for low-energy scattering, the phase shift is equivalent to a scattering length (Exercise 11.1); however, the linear extrapolation is not useful, in general, if the scattering wavelength becomes comparable to the size of the scatterer.

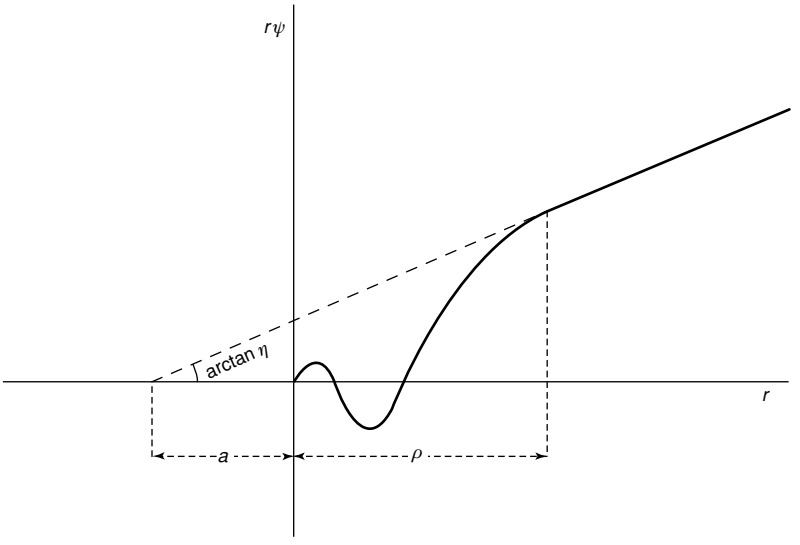


Figure 11.1. Radial wavefunction $\phi = r\psi$ for low-energy scattering as illustrated in a figure from the 1934 and 1935 papers of Fermi and coworkers for low-energy electron scattering from atoms [58] and neutron scattering from nuclei [477]. The nodes in the wavefunction near the origin show that the potential is attractive and strong enough to have bound states. The cross-section for scattering from the localized potential is determined by the phase shift (or equivalently the extrapolated scattering length as discussed in Exercise 11.1) and is the same for a weaker pseudopotential with the same phase shift modulo 2π.

Most modern pseudopotential calculations are based upon "*ab initio* norm-conserving" potentials (Secs. 11.4–11.8), which in large measure are a return to the model potential concepts of Fermi and Hellmann, but with important additions. The requirement of "norm-conservation" is the key step in making *accurate, transferable* pseudopotentials, which is essential so that a pseudopotential constructed in one environment (usually the atom) can faithfully describe the valence properties in different environments including atoms, ions, molecules, and condensed matter.[2] The basic principles are given in some detail in Sec. 11.4 because they are closely related to scattering phase shifts (App. J), the augmentation approaches of Ch. 16, and the properties of the wavefunctions needed for linearization and given explicitly in Sec. 17.1. Section 11.5 is devoted to the generation of "semilocal" potentials $V_l(r)$ that are l-dependent, i.e. act differently upon different angular momenta l. In Sec. 11.8 we describe the transformation to a separable, fully non-local operator form that is often advantageous.

This approach has been extended by Blöchl [473] and Vanderbilt [474], who showed that one can make use of auxiliary localized functions to define "ultrasoft pseudopotentials" (Sec. 11.10). By expressing the pseudofunction as the sum of a smooth part and a more

[2] Of course, there is some error due to the assumption that the cores do not change. Many tests have shown that this is an excellent approximation in atoms with small, deep cores. Errors occur in cases with shallow cores and requiring high accuracy.

rapidly varying function localized around each ion core (formally related to the original OPW construction [57] and the Phillips–Kleinman–Antoncik transformation), the accuracy of norm-conserving pseudopotentials can be improved, while at the same time making the calculations less costly (although at the expense of added complexity in the programs).

Most recently, the advent of the projector augmented wave (PAW) formulation (Sec. 11.11) has completed reformulation of the OPW method into a form that is particularly appropriate for density functional theory calculations of total energies and forces. The valence wavefunctions are expressed as a sum of smooth functions plus core functions, which leads to a generalized eigenvalue equation just as in the OPW approach. Unlike pseudopotentials, however, the PAW method retains the entire set of all-electron core functions along with smooth parts of the valence functions. Matrix elements involving core functions are treated using muffin-tin spheres as in the augmented methods (Ch. 16). As opposed to augmented methods, however, the PAW approach maintains the advantage of pseudopotentials that forces can be calculated easily.

The concept of a pseudopotential is not limited to reproducing all-electron calculations within independent-particle approximations, such as the Kohn–Sham density functional approach. In fact, the original problem of "replacing the effects of core electrons with an effective potential" presents a larger challenge: can this be accomplished in a true many-body theory taking into account the fact that all electrons are indistinguishable? Although the details are beyond the scope of the present chapter, Sec. 11.12 provides the basic issues and ideas for construction of pseudopotentials that describe the effects of the cores *beyond the independent electron approximation*.

11.2 Orthogonalized plane waves (OPWs) and pseudopotentials

Orthogonalized plane waves (OPWs), introduced by Herring [57, 482] in 1940, were the basis for the first quantitative calculations of bands in materials other than sp-bonded metals (see e.g. [61, 483, 484] and the review by Herman [62]). The calculations of Herman and Callaway [61] for Ge, done in the 1950s, is shown in Fig. 1.4; similarly, OPW calculations provided the first theoretical understanding that Si is an indirect band-gap material with the minimum of the conduction band near the X ($\mathbf{k} = (1, 0, 0)$) zone boundary point [485,486]. Combined with experimental observations [487], this work clarified the nature of the bands in these important materials. The OPW method is described in this chapter because it is the direct antecedent of modern pseudopotential and projector augmented wave (PAW) methods.

The original OPW formulation [57] is a very general approach for construction of basis functions for valence states with the form

$$\chi_{\mathbf{q}}^{\mathrm{OPW}}(\mathbf{r}) = \frac{1}{\Omega}\left\{ e^{i\mathbf{q}\cdot\mathbf{r}} - \sum_j \langle u_j|\mathbf{q}\rangle u_j(\mathbf{r}) \right\}, \tag{11.1}$$

where

$$\langle u_j|\mathbf{q}\rangle \equiv \int d\mathbf{r} u_j(\mathbf{r}) e^{i\mathbf{q}\cdot\mathbf{r}}, \tag{11.2}$$

from which it follows that $\chi_{\mathbf{q}}^{\mathrm{OPW}}$ is orthogonal to each function u_j. The functions $u_j(\mathbf{r})$ are left unspecified, but are required to be localized around each nucleus.

If the localized functions u_j are well chosen, (11.1) divides the function into a smooth part plus the localized part, as illustrated on the left-hand side of Fig. 11.2. In a crystal a smooth function can be represented conveniently by plane waves; hence the emphasis upon plane waves in the original work. In the words of Herring [57]:

> This suggests that it would be practical to try to approximate [the eigenfunction in a crystal] by a linear combination of a few plane waves, plus a linear combination of a few functions centered about each nucleus and obeying wave equations of the form[3]

$$\frac{1}{2}\nabla^2 u_j + \left(E_j - V_j\right) u_j = 0. \tag{11.3}$$

The potential $V_j = V_j(r)$ and the functions u_j are to be chosen to be optimal for the problem. With this broad definition present in the original formulation [57], the OPW approach is the prescience of all modern pseudopotential and PAW methods. As is clear in the sections below, those methods involve new ideas and clever choices for the functions and operations on the functions. This has led to important advances in electronic structure that have made many of the modern developments in the field possible.

For present purposes it is useful to consider the orthogonalized form for the valence states in an atom, where the state is labeled by angular momentum lm and, of course, the added functions must also have the same lm. Using the definitions (11.1) and (11.2), it follows immediately that the general OPW-type relation takes the form

$$\psi_{lm}^{v}(\mathbf{r}) = \tilde{\psi}_{lm}^{v}(\mathbf{r}) + \sum_j B_{lmj} u_{lmj}(\mathbf{r}), \tag{11.4}$$

where ψ_{lm}^{v} is the valence function, $\tilde{\psi}_{lm}^{v}$ is the smooth part, and all quantities can be expressed in terms of the original OPWs by Fourier transforms:

$$\psi_{lm}^{v}(\mathbf{r}) = \int \mathrm{d}\mathbf{q}\; c_{lm}(\mathbf{q})\; \chi_{\mathbf{q}}^{\mathrm{OPW}}(\mathbf{r}), \tag{11.5}$$

$$\tilde{\psi}_{lm}^{v}(\mathbf{r}) = \int \mathrm{d}\mathbf{q}\; c_{lm}(\mathbf{q})\; \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{r}}, \tag{11.6}$$

$$B_{lmj} = \int \mathrm{d}\mathbf{q}\; c_{lm}(\mathbf{q})\; \langle u_j | \mathbf{q} \rangle. \tag{11.7}$$

A schematic example of a $3s$ valence state and the corresponding smooth function is illustrated in Fig. 11.2.

It is also illuminating to express the OPW relation (11.4) as a transformation

$$|\psi_{lm}^{v}\rangle = \mathcal{T}|\tilde{\psi}_{lm}^{v}\rangle. \tag{11.8}$$

[3] This is the original equation except that the factor of $\frac{1}{2}$ was not included since Herring's equation was written in Rydberg atomic units.

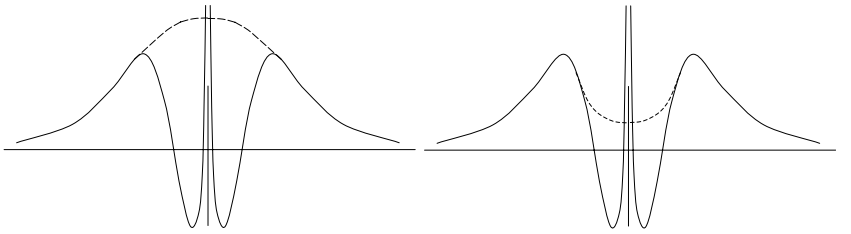


Figure 11.2. Schematic example of a valence function that has the character of a 3s orbital near the nucleus (which is properly orthogonal to the 1s and 2s core states) and two examples of smooth functions (dashed lines) that equal the full wavefunction outside the core region. Left: Represents the smooth part of the valence function $\tilde{\psi}$ defined by OPW-like equations (11.4) and (11.6). Right: Illustrates a smooth pseudofunction $\psi_l^{\rm PS}$ that satisfies the norm-conservation condition (11.21). In general, $\psi_l^{\rm PS}$ is not as smooth as $\tilde{\psi}$.

This is, of course, nothing more than a rewritten form of Eq. (11.4), but it expresses in compact form the idea that a solution for the smooth function $\tilde{\psi}_{lm}^v$ is sufficient; one can always recover the full function ψ_{lm}^v using a linear transformation denoted $\mathcal{T}$ in (11.8). This is exactly the form used in the PAW approach in Sec. 11.11.

The simplest approach is to choose the localized states to be core orbitals $u_{lmi} = \psi_{lmi}^c$, i.e. to choose the potential in (11.3) to be the actual potential (assumed to be spherical near the nucleus), so that ψ_{lmi}^c are the lowest eigenstates of the hamiltonian

$$H\psi_{lmi}^c = \varepsilon_{li}^c \psi_{lmi}^c. \tag{11.9}$$

Since the valence state ψ_{lm}^v must be orthogonal to the core states ψ_{lmi}^c, the radial part of $\psi_l^v(r)$ must have as many nodes as there are core orbitals with that angular momentum. One can show (Exercise 11.3) that the choice of $u_{li} = \psi_{li}^c$ leads to a smooth function $\tilde{\psi}_l^v(\mathbf{r})$ that has no radial nodes, i.e. it is indeed smoother than $\psi_l^v(\mathbf{r})$. Furthermore, often the core states can be assumed to be the same in the molecule or solid as in the atom. This is the basis for the actual calculations [62] in the OPW method.

There are several relevant points to note. As is illustrated on the left-hand side of Fig. 11.2, an OPW is like a smooth wave with additional structure and reduced amplitude near the nucleus. The set of OPWs is not orthonormal and each wave has a norm less than unity (Exercise 11.4)

$$\langle \chi_{\mathbf{q}}^{\rm OPW} | \chi_{\mathbf{q}}^{\rm OPW} \rangle = 1 - \sum_j |\langle u_j | \mathbf{q} \rangle|^2. \tag{11.10}$$

This means that the equations for the OPWs have the form of a generalized eigenvalue problem with an overlap matrix.

The pseudopotential transformation

The pseudopotential transformation of Phillips and Kleinman [481] and Antoncik [479,480] (PKA) results if one inserts the expression, (11.4), for $\psi_i^v(\mathbf{r})$ into the original equation for

the valence eigenfunctions

$$\hat{H}\psi_i^v(\mathbf{r}) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{r})\right]\psi_i^v(\mathbf{r}) = \varepsilon_i^v\psi_i^v(\mathbf{r}), \tag{11.11}$$

where V is the total effective potential, which leads to an equation for the smooth functions, $\tilde{\psi}_i^v(\mathbf{r})$,

$$\hat{H}^{\mathrm{PKA}}\tilde{\psi}_i^v(\mathbf{r}) \equiv \left[-\frac{1}{2}\nabla^2 + \hat{V}^{\mathrm{PKA}}\right]\tilde{\psi}_i^v(\mathbf{r}) = \varepsilon_i^v\tilde{\psi}_i^v(\mathbf{r}). \tag{11.12}$$

Here

$$\hat{V}^{\mathrm{PKA}} = V + \hat{V}^R, \tag{11.13}$$

where $\hat{V}^R$ is a non-local operator that acts upon $\tilde{\psi}_i^v(\mathbf{r})$ with the effect

$$\hat{V}^R\tilde{\psi}_i^v(\mathbf{r}) = \sum_j \left(\varepsilon_i^v - \varepsilon_j^c\right)\langle\psi_j^c|\tilde{\psi}_i^v\rangle\psi_j^c(\mathbf{r}). \tag{11.14}$$

Thus far this is nothing more than a formal transformation of the OPW expression, (11.11). The formal properties of the transformed equations suggest both advantages and disadvantages. It is clear that $\hat{V}^R$ is repulsive since (11.14) is written in terms of the energies $\varepsilon_i^v - \varepsilon_j^c$ which are always positive. Furthermore, a stronger attractive nuclear potential leads to deeper core states so that (11.14) also becomes more repulsive. This tendency was pointed out by Phillips and Kleinman and Antoncik and derived in a very general form as the "cancellation theorem" by Cohen and Heine [488]. Thus $\hat{V}^{\mathrm{PKA}}$ is much weaker than the original $V(\mathbf{r})$, but it is a more complicated non-local operator. In addition, the smooth pseudo-functions $\tilde{\psi}_i^v(\mathbf{r})$ are *not orthonormal* because the complete function ψ_i^v also contains the sum over core orbitals in Eq. (11.4). Thus the solution of the pseudopotential equation (11.12) is a generalized eigenvalue problem.[4] Furthermore, since the core states are still present in the definition, (11.14), this transformation does not lead to a "smooth" pseudopotential.

The full advantages of the pseudopotential transformation are realized by using *both* the formal properties of pseudopotential $\hat{V}^{\mathrm{PKA}}$ and the fact that the same scattering properties can be reproduced by different potentials. Thus the pseudopotential can be chosen to be both smoother and weaker than the original potential V by taking advantage of the non-uniqueness of pseudopotentials, as discussed in more detail in following sections.

Even though the potential operator is a more complex object than a simple local potential, the fact that it is weaker and smoother (i.e. it can be expanded in a small number of Fourier components) has great advantages, conceptually and computationally. In particular, it immediately resolves the apparent contradiction (see Ch. 12) that the valence bands $\varepsilon_{n\mathbf{k}}^v$ in many materials are nearly-free-electron-like, even though the wavefunctions $\psi_{n\mathbf{k}}^v$ must be very non-free-electron-like since they must be orthogonal to the cores. The resolution is that the bands are determined by the secular equation for the smooth, nearly-free-electron-like $\tilde{\psi}_{n\mathbf{k}}^v$ that involves the weak pseudopotential $\hat{V}^{\mathrm{PKA}}$ or $\hat{V}^{\mathrm{model}}$.

[4] "Norm-conserving" potentials described in Sec. 11.4 remove this complication; however, non-orthogonality is resurrected in "ultrasoft" pseudopotentials, which are formally similar to the operator construction described here (see Sec. 11.10).

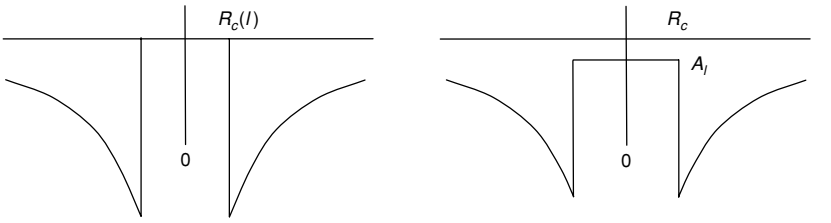


Figure 11.3. Left: "Empty core" model potential of Ashcroft [489] in which the potential is zero inside radius $R_c(l)$, which is different for each l. Right: Square well model potential with value A_l inside a cutoff radius R_c, proposed by Abarenkov and Heine [490] and fit to atomic data by Animalu and Heine [491,492] (see also by Harrison [468]). The fact that the potentials are weak, zero, or even positive inside a cutoff radius R_c is an illustration of the "cancellation theorem" [488].

11.3 Model ion potentials

Based upon the foundation of pseudopotentials in scattering theory, and the transformation of the OPW equations and the cancellation theorem, the theory of pseudopotentials has become a fertile field for generating new methods and insight for the electronic structure of molecules and solids. There are two approaches: (1) to define ionic pseudopotentials, which leads to the problem of interacting valence-only electrons, and (2) to define a *total pseudopotential* that includes effects of the other valence electrons. The former is the more general approach since the ionic pseudopotentials are more transferable with a single ion potential applicable for the given atom in different environments. The latter approach is very useful for describing the bands accurately if they are treated as adjustable empirical potentials; historically empirical pseudopotentials have played an important role in the understanding of electronic structures [467,469], and they reappear in Sec. 12.6 as a useful approach for understanding bands in a plane wave basis.

Here we concentrate upon ionic pseudopotentials and the form of model potentials that give the same scattering properties as the pseudopotential operators of Eqs. (11.13) and (11.14) or more general forms. Since a model potential replaces the potential of a nucleus and the core electrons, it is spherically symmetric and each angular momentum l, m can be treated separately, which leads to non-local l-dependent model pseudopotentials $V_l(r)$. The qualitative features of l-dependent pseudopotentials can be illustrated by the forms shown in Fig. 11.3. Outside the core region, the potential is $Z_{\rm ion}/r$, i.e. the combined Coulomb potential of the nucleus and core electrons. Inside the core region the potential is expected to be repulsive [488] to a degree that depends upon the angular momentum l, as is clear from the analysis of the repulsive potential in (11.14).

The dependence upon l means that, in general, a pseudopotential is a non-local operator that can be written in "semilocal" (SL) form

$$\hat{V}_{\rm SL} = \sum_{lm} |Y_{lm}\rangle V_l(r) \langle Y_{lm}|, \tag{11.15}$$

where $Y_{lm}(\theta, \phi) = P_l(\cos(\theta))e^{im\phi}$. This is termed semilocal (SL) because it is non-local in the angular variables but local in the radial variable: when operating on a function

$f(r, \theta', \phi')$, $\hat{V}_{\rm SL}$ has the effect

$$\left[\hat{V}_{\rm SL} f\right]_{r,\theta,\phi} = \sum_{lm} Y_{lm}(\theta, \phi) V_l(r) \int \mathrm{d}(\cos\theta')\mathrm{d}\phi' Y_{lm}(\theta', \phi') f(r, \theta', \phi'). \quad (11.16)$$

All the information is in the radial functions $V_l(r)$ or their Fourier transforms, which are defined in Sec. 12.4. An electronic structure involves calculation of the matrix elements of $\hat{V}_{\rm SL}$ between states ψ_i and ψ_j

$$\langle \psi_i | \hat{V}_{\rm SL} | \psi_j \rangle = \int \mathrm{d}r \ \psi_i(r, \theta, \phi) \left[\hat{V}_{\rm SL} \psi_j\right]_{r,\theta,\phi}. \quad (11.17)$$

(Compare this with Eq. (11.41) for a fully non-local separable form of the pseudopotential.)

There are two approaches to the definition of potentials:

- Empirical potentials fitted to atomic or solid state data. Simple forms are the "empty core" [489] and square well [490–492] models illustrated in Fig. 11.3. In the latter case, the parameters were fit to atomic data for each l and tabulated for many elements by Animalu and Heine [491, 492] (tables given also by Harrison [468]).[5]
- "*Ab initio*" potentials constructed to fit the valence properties calculated for the atom. The advent of "norm-conserving" pseudopotentials provided a straightforward way to make such potentials that are successfully transferrable to calculations on molecules and solids.

11.4 Norm-conserving pseudopotentials (NCPPs)

Pseudopotentials generated by calculations on atoms (or atomic-like states) are termed "*ab initio*" because they are *not fitted to experiment*. The concept of "norm-conservation" has a special place in the development of *ab initio* pseudopotentials; at one stroke it simplifies the application of the pseudopotentials and it makes them more accurate and transferable. The latter advantage is described below, but the former can be appreciated immediately. In contrast to the PKA approach (Sec. 11.2) (where the equations were formulated in terms of the smooth part of the valence function $\tilde{\psi}_i^v(\mathbf{r})$ to which another function must be added, as in Eq. (11.4)), norm-conserving pseudofunctions $\psi^{\rm PS}(\mathbf{r})$ are normalized and are solutions of a model potential chosen to reproduce the valence properties of an all-electron calculation. A schematic example is shown on the right-hand side of Fig. 11.2, which illustrates the difference from the un-normalized smooth part of the OPW. In the application of the pseudopotential to complex systems, such as molecules, clusters, solids, etc., *the valence pseudofunctions satisfy the usual orthonormality conditions* as in Eq. (7.9),

$$\langle \psi_i^{\sigma,\rm PS} | \psi_j^{\sigma',\rm PS} \rangle = \delta_{i,j}\delta_{\sigma,\sigma'}, \quad (11.18)$$

[5] Construction of such model potentials from atomic information presents a conceptual problem: the potential represents the effects of $\hat{V}^{\rm PKA}$, which depends upon the valence eigenvalue ε_i^v in the atom which is defined relative to a reference energy equal to zero at infinity; however, the goal is to apply the pseudopotentials to infinite solids, where this is not a well-defined reference energy, and to molecules, where the levels are shifted relative to the atom. How can one relate the eigenvalues of the two types of systems? This was a difficult issue in the original pseudopotentials that was resolved by the conditions for construction of "norm-conserving" pseudopotentials described in Sec. 11.4.

so that for the Kohn–Sham equations have the same form as in (7.11),

$$(H_{\rm KS}^{\sigma,{\rm PS}} - \varepsilon_i^\sigma)\psi_i^{\sigma,{\rm PS}}(\mathbf{r}) = 0, \tag{11.19}$$

with $H_{\rm KS}^{\sigma,{\rm PS}}$ given by (7.12) and (7.13), and the external potential given by the pseudopotential specified in the section following.

Norm-conservation condition

Quantum chemists and physicists have devised pseudopotentials called, respectively, "shape-consistent" [493,494] and "norm-conserving" [471].[6] The starting point for defining norm-conserving potentials is the list of requirements for a "good" *ab initio* pseudopotential given by Hamann, Schluter, and Chiang (HSC) [471]:

1. All-electron and pseudo valence eigenvalues agree for the chosen atomic reference configuration.
2. All-electron and pseudo valence wavefunctions agree beyond a chosen core radius R_c.
3. The logarithmic derivatives of the all-electron and pseudo wavefunctions agree at R_c.
4. The integrated charge inside R_c for each wavefunction agrees (norm-conservation).
5. The *first energy derivative* of the logarithmic derivatives of the all-electron and pseudo wavefunctions agrees at R_c, and therefore for all $r \geq R_c$.

From Points 1 and 2 it follows that the NCPP equals the atomic potential outside the "core region" of radius R_c; this is because the potential is uniquely determined (except for a constant that is fixed if the potential is zero at infinity) by the wavefunction and the energy ε, that need not be an eigenenergy. Point 3 follows since the wavefunction $\psi_l(r)$ and its radial derivative $\psi_l'(r)$ are continuous at R_c for any smooth potential. The dimensionless logarithmic derivative D is defined by

$$D_l(\varepsilon, r) \equiv r\psi_l'(\varepsilon, r)/\psi_l(\varepsilon, r) = r\frac{\rm d}{{\rm d}r}\ln\psi_l(\varepsilon, r), \tag{11.20}$$

also given in (J.5).

Inside R_c the pseudopotential and radial pseudo-orbital $\psi_l^{\rm PS}$ differ from their all-electron counterparts; however, Point 4 requires that the integrated charge,

$$Q_l = \int_0^{R_c} {\rm d}r r^2 |\psi_l(r)|^2 = \int_0^{R_c} {\rm d}r \phi_l(r)^2, \tag{11.21}$$

is the same for $\psi_l^{\rm PS}$ (or $\phi_l^{\rm PS}$) as for the all-electron radial orbital ψ_l (or ϕ_l) for a valence state. The conservation of Q_l insures that: (a) the total charge in the core region is correct, and (b) the normalized pseudo-orbital is equal[7] to the true orbital outside of R_c (in contrast to the smooth orbital of (11.6) which equals the true orbital outside R_c only if it is not normalized). Applied to a molecule or solid, these conditions ensure that the normalized

[6] Perhaps the earliest work was that of Topp and Hopfield [495].

[7] Equality can be strictly enforced only for local functionals, not for non-local cases as in Hartree–Fock and EXX potentials. For example, see [496].

pseudo-orbital is correct in the region outside R_c between the atoms where bonding occurs, and that the potential outside R_c is correct as well since the potential outside a spherically symmetric charge distribution depends only on the total charge inside the sphere.

Point 5 is a crucial step toward the goal of constructing a "good" pseudopotential: one that can be generated in a simple environment like a spherical atom, and then used in a more complex environment. In a molecule or solid, the wavefunctions and eigenvalues change and a pseudopotential that satisfies Point 5 will reproduce the changes in the eigenvalues to linear order in the change in the self-consistent potential. At first sight, however, it is not obvious how to satisfy the condition that the *first energy derivative* of the logarithmic derivatives $\mathrm{d}D_l(\varepsilon, r)/\mathrm{d}\varepsilon$ agree for the pseudo- and the all-electron wavefunctions evaluated at the cutoff radius R_c and energy ε_l chosen for the construction of the pseudopotential of angular momentum l.

The advance due to HSC [471] and others [493,494] was to show that Point 5 is implied by Point 4. This "norm-conservation condition" can be derived straightforwardly, e.g. following the derivation of Shirley et al. [497], which uses relations due to Luders [498] (see Exercises 11.8 and 11.9 for intermediate steps). The radial equation for a spherical atom or ion, (10.12), which can be written

$$-\frac{1}{2}\phi_l''(r) + \left[\frac{l(l+1)}{2r^2} + V_{\mathrm{eff}}(r) - \varepsilon\right]\phi_l(r) = 0, \tag{11.22}$$

where a prime means derivative with respect to r, can be transformed by defining the variable $x_l(\varepsilon, r)$

$$x_l(\varepsilon, r) \equiv \frac{\mathrm{d}}{\mathrm{d}r}\ln\phi_l(r) = \frac{1}{r}[D_l(\varepsilon, r) + 1]. \tag{11.23}$$

It is straightforward to show that (11.22) is equivalent to the non-linear first-order differential equation,

$$x_l'(\varepsilon, r) + [x_l(\varepsilon, r)]^2 = \frac{l(l+1)}{r^2} + 2[V(r) - \varepsilon]. \tag{11.24}$$

Differentiating this equation with respect to energy gives

$$\frac{\partial}{\partial\varepsilon}x_l'(\varepsilon, r) + 2x_l(\varepsilon, r)\frac{\partial}{\partial\varepsilon}x_l(\varepsilon, r) = -1. \tag{11.25}$$

Combining this with the relation valid for any function $f(r)$ and any l,

$$f'(r) + 2x_l(\varepsilon, r)f(r) = \frac{1}{\phi_l(r)^2}\frac{\partial}{\partial r}[\phi_l(r)^2 f(r)], \tag{11.26}$$

multiplying by $\phi_l(r)^2$ and integrating, one finds at radius R

$$\frac{\partial}{\partial\varepsilon}x_l(\varepsilon, R) = -\frac{1}{\phi_l(R)^2}\int_0^R \mathrm{d}r\phi_l(r)^2 = -\frac{1}{\phi_l(R)^2}Q_l(R), \tag{11.27}$$

or in terms of the dimensionless logarithmic derivative $D_l(\varepsilon, R)$

$$\frac{\partial}{\partial\varepsilon}D_l(\varepsilon, R) = -\frac{R}{\phi_l(R)^2}\int_0^R \mathrm{d}r\phi_l(r)^2 = -\frac{R}{\phi_l(R)^2}Q_l(R). \tag{11.28}$$

This shows immediately that if $\phi_l^{\rm PS}$ has the same magnitude as the all-electron function ϕ_l at R_c and obeys norm-conservation (Q_l the same), then the first energy derivative of the logarithmic derivative $x_l(\varepsilon, R)$ and $D_l(\varepsilon, R)$ is the same as for the all-electron wavefunction. This also means that the norm-conserving pseudopotential has the same scattering phase shifts as the all-electron atom to linear order in energy around the chosen energy ε_l, which follows from expression (J.6), which relates to $D_l(\varepsilon, R)$ and the phase shift $\eta_l(\varepsilon, R)$.[8]

11.5 Generation of l-dependent norm-conserving pseudopotentials

Generation of a pseudopotential starts with the usual all-electron atomic calculation as described in Ch. 10. Each state l, m is treated independently except that the total potential is calculated self-consistently for the given approximation for exchange and correlation and for the given configuration of the atom. The next step is to identify the valence states and generate the pseudopotentials $V_l(r)$ and pseudo-orbitals $\psi_l^{\rm PS}(r) = r\phi_l^{\rm PS}(r)$. The procedure varies with different approaches, but in each case one first finds a total "screened" pseudopotential acting on valence electrons in the atom. This is then "unscreened" by subtracting from the total potential the sum of Hartree and exchange–correlation potentials $V_{H\rm xc}^{\rm PS}(r) = V_{\rm Hartree}^{\rm PS}(r) + V_{\rm xc}^{\rm PS}(r)$

$$V_l(r) \equiv V_{l,\rm total}(r) - V_{H\rm xc}^{\rm PS}(r), \tag{11.29}$$

where $V_{H\rm xc}^{\rm PS}(r)$ is defined for the valence electrons in their pseudo-orbitals. Further aspects of "unscreening" are deferred to Sec. 11.6.

It is useful to separate the ionic pseudopotential into a local (l-independent) part of the potential plus non-local terms

$$V_l(r) = V_{\rm local}(r) + \delta V_l(r). \tag{11.30}$$

Since the eigenvalues and orbitals are required to be the same for the pseudo and the all-electron case for $r > R_c$, each potential $V_l(r)$ equals the local (l-independent) all-electron potential, and $V_l(r) \to -\frac{Z_{\rm ion}}{r}$ for $r \to \infty$. Thus $\delta V_l(r) = 0$ for $r > R_c$ and all the long-range effects of the Coulomb potential are included in the local potential $V_{\rm local}(r)$. Finally, the "semilocal" operator (11.15) can be written as

$$\hat{V}_{\rm SL} = V_{\rm local}(r) + \sum_{lm} |Y_{lm}\rangle \delta V_l(r) \langle Y_{lm}|. \tag{11.31}$$

Even if one requires norm-conservation, there is still freedom of choice in the form of $V_l(r)$ in constructing pseudopotentials. There is no one "best pseudopotential" for any given element – there may be many "best" choices, each optimized for some particular use of the pseudopotential. In general, there are two overall competing factors:

[8] This relation is very important and used in many contexts: in App. J it is seen to be the Friedel sum rule, which has important consequences for resistivity due to impurity scattering in metals. In Ch. 17 it is used to relate the band width to the value of the wavefunction at the boundary of a sphere.

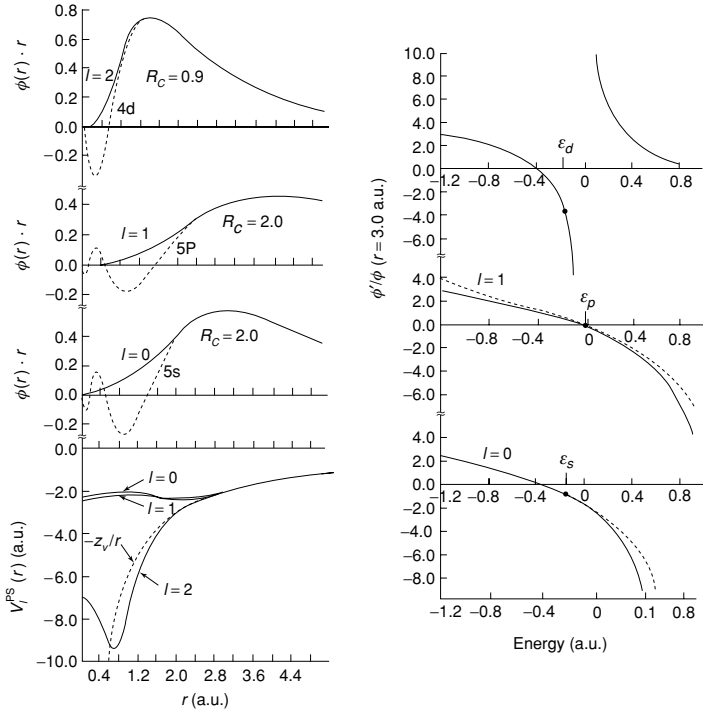


Figure 11.4. Example of norm-conserving pseudopotentials, pseudofunctions, and logarithmic derivative for the element Mo. Left bottom: $V_l(r)$ in Rydbergs for angular momentum $l = 0, 1, 2$ compared to Z_{ion}/r (dashed). Left top: All-electron valence radial functions $\phi_l(r) = r\psi_l(r)$ (dashed) and norm-conserving pseudofunctions. Right: Logarithmic derivative of the pseudopotential compared to the full atom calculation; the points indicate the energies, ε, where they are fitted. The derivative with respect to the energy is also correct due to the norm-conservation condition (11.27). From [471].

- Accuracy and transferability generally lead to the choice of a small cutoff radius R_c and "hard" potentials, since one wants to describe the wavefunction as well as possible in the region near the atom.
- Smoothness of the resulting pseudofunctions generally leads to the choice of a large cutoff radius R_c and "soft" potentials, since one wants to describe the wavefunction with as few basis functions as possible (e.g. plane waves).

Here we will try to present the general ideas in a form that is the basis of widely used methods, with references to some of many proposed forms that cannot be covered here.

An example of pseudopotentials [471] for Mo is shown in Fig. 11.4. A similar approach has been used by Bachelet, Hamann, and Schlüter (BHS) [499] to construct pseudopotentials for all elements from H to Po, in the form of an expansion in gaussians with tabulated coefficients. These potentials were derived starting from an assumed form of the potential and varying parameters until the wavefunction has the desired properties, an approach also

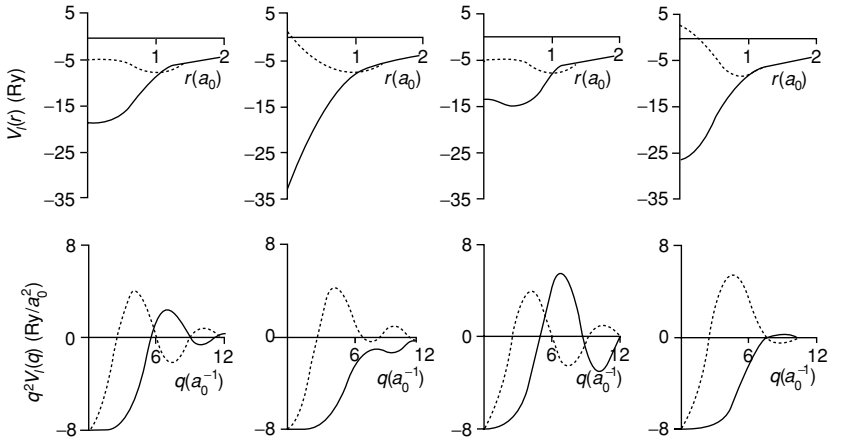


Figure 11.5. Comparison of pseudopotentials for carbon (dotted line for s and solid line for p) in real space and reciprocal space, illustrating the large variations in potentials that are all norm-conserving and have the same phase shifts at the chosen energies. In order from left to right generated using the procedures of: Troullier and Martins [502]; Kerker [501]; Hamann, Schlüter, and Chiang [471]; Vanderbilt [500]. From Troullier and Martins [502].

used by Vanderbilt [500]. A simpler procedure is that of Christiansen et al. [493] and Kerker [501], which defines a pseudo-wavefunction $\phi_l^{\rm PS}(r)$ with the desired properties for each l and numerically inverts the Schrödinger equation to find the potential $V_l(r)$ for which $\phi_l^{\rm PS}(r)$ is a solution with energy ε. The wavefunction outside the radius R_c is the same as the true function and at R_c it is matched to a parameterized analytic function. Since the energy ε is fixed (often it is the eigenvalue from the all-electron calculation, but this is not essential) it is straightforward to invert the Schrödinger equation for a nodeless function $\phi_l^{\rm PS}(r)$ for each l separately, yielding

$$V_{l,\rm total}(r) = \varepsilon - \frac{\hbar^2}{2m_e}\left[\frac{l(l+1)}{2r^2} - \frac{\frac{\mathrm{d}^2}{\mathrm{d}r^2}\phi_l^{\rm PS}(r)}{\phi_l^{\rm PS}(r)}\right]. \tag{11.32}$$

The analytic form chosen by Kerker is $\phi_l^{\rm PS}(r) = \mathrm{e}^{p(r)}$, $r < R_c$, where $p(r) =$ polynomial to fourth power with coefficients fixed by requiring continuous first and second derivatives at R_c and norm-conservation.

One of the important considerations for many uses is to make the wavefunction as smooth as possible, which allows it to be described by fewer basis functions, e.g. fewer Fourier components. For example, the BHS potentials [499] are a standard reference for comparison; however, they are generally harder and require more Fourier components in the description of the pseudofunction than other methods. Troullier and Martins [502] have extended the Kerker method to make it smoother by using a higher order polynomial and matching more derivatives of the wavefunction. A comparison of different pseudopotentials for carbon is given in Fig. 11.5 showing the forms both in real and reciprocal space. The one-dimensional radial transforms $V_l(q)$ (or "form factors") for each l are defined in Sec. 12.4; these are

the functions that enter directly in plane wave calculations and their extent in Fourier space determines the number of plane waves needed for convergence. A number of authors have proposed ways to make smoother potentials to reduce the size of calculations. One approach [503, 504] is to minimize the kinetic energy of the pseudofunctions explicitly for the chosen core radius. This can be quantified by examination of the Fourier transform and its behavior at large momentum q. Optimization of the potentials can be done in the atom, and the results carry over to molecules and solids, since the convergence as a function of the range $q_{\max}$ is the same in all the cases.

The forms used in chemistry literature [494] have generally been much more rapidly varying, often singular at the origin. In recent work, however, Hartree–Fock pseudopotentials that have no singularities [505, 506] have been generated for use in many-body quantum chemical calculations.

Relativistic effects

Effects of special relativity can be incorporated into pseudopotentials, since they originate deep in the interior of the atom near the nucleus, and the consequences for valence electrons can be easily carried into molecular or solid state calculations. This includes shifts due to scalar relativistic effects and spin–orbit interactions. The first step is generation of a pseudopotential from a relativistic all-electron calculation on the atom for both $j = l + 1/2$ and $j = l - 1/2$. From the two potentials we can define [413, 499]

$$V_l = \frac{l}{2l+1}\left[(l+1)V_{l+1/2} + l V_{l-1/2}\right], \tag{11.33}$$

$$\delta V_l^{\mathrm{so}} = \frac{2}{2l+1}\left[V_{l+1/2} - V_{l-1/2}\right]. \tag{11.34}$$

Scalar relativistic effects are included in the first term and the spin–orbit effects are included in a short–range non-local term [449, 450],

$$\delta \hat{V}_{\mathrm{SL}}^{\mathrm{so}} = \sum_{lm} |Y_{lm}\rangle \delta V_l^{\mathrm{so}}(r) \mathbf{L}\cdot\mathbf{S} \langle Y_{lm}|. \tag{11.35}$$

11.6 Unscreening and core corrections

In the construction of *ab initio* pseudopotentials there is a straightforward one-to-one relation of the valence pseudofunction and the *total* pseudopotential. It is then a necessary step to "unscreen" to derive the bare ion pseudopotential which is transferable to different environments. However, the process of "unscreening" is not so straightforward. If the effective exchange–correlation potential were a linear function of density (as is the Hartree potential V_{Hartree}) there would be no problem, and (11.29) could be written as

$$V_{l,\mathrm{total}} = V_l(r) + V_{\mathrm{Hartree}}([n^{\mathrm{PS}}], \mathbf{r}) + V_{\mathrm{xc}}([n^{\mathrm{PS}}], \mathbf{r}), \tag{11.36}$$

where the notation $[n^{\mathrm{PS}}]$ means the quantity is evaluated as a functional of the pseudonymity n^{PS}. This is true for the Hartree potential, but the fact that V_{xc} is a non-linear functional of n (and may also be non-local) leads to difficulties and ambiguities. (Informative discussions can be found in [507] and [496].)

Non-linear core corrections

So long as the exchange–correlation functional involves only the density or its gradients at each point, then the unscreening of the potential in the atom can be accomplished by defining the effective exchange–correlation potential in (11.29) as

$$\tilde{V}_{\rm xc}(\mathbf{r}) = V_{\rm xc}([n^{\rm PS}], \mathbf{r}) + \left[V_{\rm xc}([n^{\rm PS} + n^{\rm core}], \mathbf{r}) - V_{\rm xc}([n^{\rm PS}], \mathbf{r})\right]. \tag{11.37}$$

The term in square brackets is a core correction that significantly increases the transferability of the pseudopotential [507]. There are costs, however: the core charge density must be stored along with the pseudodensity and the implementation in a solid must use $\tilde{V}_{\rm xc}(\mathbf{r})$ defined in (11.37), and the rapidly varying core density would be a disadvantage in plane wave methods. The second obstacle can be overcome [507] using the freedom of choice inherent in pseudopotentials by defining a smoother "partial core density" $n^{\rm core}_{\rm partial}(r)$ that can be used in (11.37). The original form proposed by Louie, Froyen, and Cohen [507] is[9]

$$n^{\rm core}_{\rm partial}(r) = \begin{cases} \frac{A\sin(Br)}{r}, & r < r_0, \\ n^{\rm core}(r), & r > r_0, \end{cases} \tag{11.38}$$

where A and B are determined by the value and gradient of the core charge density at r_0, a radius chosen where $n^{\rm core}$ is typically 1 to 2 times $n^{\rm valence}$. The effect is particularly large for cases in which the core is extended (e.g. the 3d transition metals where the 3p "core" states strongly overlap the 3d "valence" states) and for magnetic systems where there may be a large difference between up and down valence densities even though the fractional difference in total density is much smaller. In such cases description of spin-polarized configurations can be accomplished with a spin-independent ionic pseudopotential, with no need for separate spin-up and spin-down ionic pseudopotentials.

Non-local E_{xc} functionals

There is a complication in "unscreening" in cases where the $E_{\rm xc}$ functional is intrinsically non-local, as in Hartree–Fock and exact exchange (EXX). In general it is not possible to make a potential that keeps the wavefunctions outside a core radius the same as in the original all-electron problem because the non-local effects extend to all radii. The issues are discussed thoroughly in [496].

11.7 Transferability and hardness

There are two meanings to the word "hardness." One meaning is a measure of the variation in real space which is quantified by the extent of the potential in Fourier space. In general, "hard" potentials describe the properties of the localized rigid ion cores and are more transferable from one material to another; attempts to make the potential "soft" (i.e. smooth) have tended to lead to poorer transferability. However, there is considerable effort to make

[9] The form in (11.38) has a discontinuity in the second derivative r_0, which causes difficulties in conjunction with GGA functionals. This problem is readily remedied by using a more flexible functional form.

accurate, transferable potentials that nevertheless do not extend far in Fourier space, e.g. the "optimized" pseudopotentials [503].

The second meaning is a measure of the ability of the valence pseudo-electrons to describe the response of the system to a change in the enviroment properly [508–510]. We have already seen that norm-conservation guarantees that the electron states of the atom have the correct first derivative with respect to change in energy. This meaning of "hardness" is a measure of the faithfulness of the response to a change in potential. Potentials can be tested versus spherical perturbations (change of charge, state, radial potential) using the usual spherical atom codes. Goedecker and Maschke [508] have given an insightful analysis in terms of the response of the charge density in the core region; this is relevant since the density is the central quantity in density functional theory and the integrated density is closely related to norm-conservation conditions. Also tests with non-spherical perturbations ascertain the performance with relevant perturbations, in particular, the polarizability in an electric field [510].

Tests in spherical boundary conditions

We have seen in Sec. 10.7 that some aspects of solids are well modeled by imposing different spherical boundary conditions on an atom or ion. A net consequence is that the valence wavefunctions tend to be more concentrated near the nucleus than in the atom. How well do pseudopotentials derived for an isolated atom describe such situations? The answer can be found directly using computer programs for atoms and pseudoatoms (Ch. 24); examples are given in the exercises. These are the types of tests that should be done *whenever generating a new pseudopotential*.

11.8 Separable pseudopotential operators and projectors

It was recognized by Kleinman and Bylander (KB) [472] that it is possible to construct a *separable* pseudopotential operator, i.e. $\delta V(\mathbf{r}, \mathbf{r}')$ written as a sum of products of the form $\Sigma_i f_i(\mathbf{r}) g_i(\mathbf{r}')$. KB showed that the effect of the semilocal $\delta V_l(r)$ in (11.30) can be replaced, to a good approximation, by a separable operator $\delta\hat{V}_{\rm NL}$ so that the total pseudopotential has the form

$$\hat{V}_{\rm NL} = V_{\rm local}(r) + \sum_{lm} \frac{|\psi^{\rm PS}_{lm}\delta V_l\rangle\langle\delta V_l\psi^{\rm PS}_{lm}|}{\langle\psi^{\rm PS}_{lm}|\delta V_l|\psi^{\rm PS}_{lm}\rangle}, \tag{11.39}$$

where the second term written explicitly in coordinates, $\delta\hat{V}_{\rm NL}(\mathbf{r}, \mathbf{r}')$, has the desired separable form. Unlike the semilocal form (11.15), it is fully non-local in angles θ, ϕ and radius r. When operating on the reference atomic states $\psi^{\rm PS}_{lm}$, $\delta\hat{V}_{\rm NL}(\mathbf{r}, \mathbf{r}')$ acts the same as $\delta V_l(r)$, and it can be an excellent approximation for the operation of the pseudopotential on the valence states in a molecule or solid.

The functions $\langle\delta V_l\psi^{\rm PS}_{lm}|$ are *projectors* that operate upon the wavefunction

$$\langle\delta V_l\psi^{\rm PS}_{lm}|\psi\rangle = \int {\rm d}\mathbf{r}\ \delta V_l(r)\psi^{\rm PS}_{lm}(\mathbf{r})\psi(\mathbf{r}). \tag{11.40}$$

Each projector is localized in space, since it is non-zero only inside the pseudopotential cutoff radius where $\delta V_l(r)$ is non-zero. This is independent of the extent of the functions $\psi^{\rm PS}_{lm} = \psi_{lm}(r)P_l(\cos(\theta))e^{im\phi}$, which have the extent of atomic valence orbitals or can even be non-bound states.

The advantage of the separable form is that matrix elements require only products of projection operations (11.40)

$$\langle\psi_i|\delta\hat{V}_{\rm NL}|\psi_j\rangle = \sum_{lm}\langle\psi_i|\psi^{\rm PS}_{lm}\delta V_l\rangle\frac{1}{\langle\psi^{\rm PS}_{lm}|\delta V_l|\psi^{\rm PS}_{lm}\rangle}\langle\delta V_l\psi^{\rm PS}_{lm}|\psi_j\rangle. \tag{11.41}$$

This can be contrasted with (11.17), which involves a radial integral for each pair of functions ψ_i and ψ_j. This leads to savings in computations that can be important for large calculations. However, it does lead to an additional step which may lead to increased errors. Although, the operation on the given atomic state is unchanged, the operations on other states at different energies may be modified, and care must be taken to ensure that there are no artificial "ghost states" introduced. (As discussed in Exercise 11.12, such ghost states at low energy are expected when $V_{\rm local}(r)$ is attractive and the non-local $\delta V_l(r)$ are repulsive. This choice should be avoided [511].)

It is straightforward to generalize to the case of spin–orbit coupling, using the states of the atom derived from the Dirac equation with total angular momentum $j = l \pm \frac{1}{2}$ [450, 472]. The non-local projections become

$$\hat{V}^{j=l\pm\frac{1}{2}}_{\rm NL} = V_{\rm local}(r) + \sum_{lm}\frac{|\psi^{\rm PS}_{l\pm\frac{1}{2},m}V_{l\pm\frac{1}{2}}\rangle\langle\delta V_{l\pm\frac{1}{2}}\psi^{\rm PS}_{l\pm\frac{1}{2},m}|}{\langle\psi^{\rm PS}_{l\pm\frac{1}{2},m}|\delta V_{l\pm\frac{1}{2}}|\psi^{\rm PS}_{l\pm\frac{1}{2},m}\rangle}. \tag{11.42}$$

The KB construction can be modified to generate the separable potential directly without going through the step of constructing the semilocal $V_l(r)$ [474]. Following the same procedure as for generating the norm-conserving pseudopotential, the first step is to define pseudofunctions $\psi^{\rm PS}_{lm}(\mathbf{r})$ and a local pseudopotential $V_{\rm local}(r)$ which are equal to the all-electron functions outside a cutoff radius $r > R_c$. For $r > R_c$, $\psi^{\rm PS}_{lm}(\mathbf{r})$ and $V_{\rm local}(r)$ are chosen in some smooth fashion as was done in Sec. 11.5. If we now define new functions

$$\chi^{\rm PS}_{lm}(\mathbf{r}) \equiv \left\{\varepsilon_l - \left[-\frac{1}{2}\nabla^2 + V_{\rm local}(r)\right]\right\}\psi^{\rm PS}_{lm}(\mathbf{r}), \tag{11.43}$$

it is straightforward to show that $\chi^{\rm PS}_{lm}(\mathbf{r}) = 0$ outside R_c and that the operator

$$\delta\hat{V}_{\rm NL} = \sum_{lm}\frac{|\chi^{\rm PS}_{lm}\rangle\langle\chi^{\rm PS}_{lm}|}{\langle\chi^{\rm PS}_{lm}|\psi^{\rm PS}_{lm}\rangle} \tag{11.44}$$

has the same properties as the KB operator (11.39), i.e. $\psi^{\rm PS}_{lm}$ is a solution of $\hat{H}\psi^{\rm PS}_{lm} = \varepsilon_l\psi^{\rm PS}_{lm}$ with $\hat{H} = -\frac{1}{2}\nabla^2 + V_{\rm local} + \delta\hat{V}_{\rm NL}$.

11.9 Extended norm conservation: beyond the linear regime

Two general approaches have been proposed to extend the range of energies over which the phase shifts of the original all-electron potential are described. Shirley and coworkers [497]

have given general expressions that must be satisfied for the phase shifts to be correct to arbitrary order in a power series expansion in $(\varepsilon - \varepsilon_0)^N$ around the chosen energy ε_0.

A second approach is easier to implement and is the basis for further generalizations that hold great promise for future work in electronic structure (see Sec. 17.8). The construction of the projectors can be done at any energy ε_s and the procedure can be generalized to satisfy the Schrödinger equation at more than one energy for a given l, m [473, 474]. (Below we omit superscript PS and subscript l, m for simplicity.) If pseudofunctions ψ_s are constructed from all-electron calculations at different energies ε_s, one can form the matrix $B_{s,s'} = \langle \psi_s | \chi_{s'} \rangle$, where the χ_s are defined by (11.43). In terms of the functions $\beta_s = \sum_{s'} B^{-1}_{s,s'} \chi_{s'}$, the generalized non-local potential operator can be written

$$\delta \hat{V}_{\mathrm{NL}} = \sum_{lm} \left[\sum_{s,s'} B_{s,s'} |\beta_s\rangle \langle \beta_{s'}| \right]_{lm} . \tag{11.45}$$

It is straightforward to show (Exercise 11.13) that each ψ_s is a solution of $\hat{H}\psi_s = \varepsilon_s \psi_s$. With this modification, the non-local separable pseudopotential can be generalized to agree with the all-electron calculation to arbitrary accuracy over a desired energy range.

The transformation (11.45) exacts a price; instead of the simple sum of products of projectors in (11.41), matrix elements of (11.45) involve a matrix product of operators. For the spherically symmetric pseudopotential, the matrix is $s \times s$ and is diagonal in l, m. (A similar idea is utilized in Sec. 17.8 to transform the equations for the general problem of electron states in a crystal.)

11.10 Ultrasoft pseudopotentials

One goal of pseudopotentials is to create pseudofunctions that are as "smooth" as possible, and yet are accurate. For example, in plane wave calculations the valence functions are expanded in Fourier components, and the cost of the calculation scales as a power of the number of Fourier components needed in the calculation (see Ch. 12). Thus one meaningful definition of maximizing "smoothness" is to minimize the range in Fourier space needed to describe the valence properties to a given accuracy. "Norm-conserving" pseudopotentials achieve the goal of accuracy, usually at some sacrifice of "smoothness."

A different approach known as "ultrasoft pseudopotentials" reaches the goal of accurate calculations by a transformation that re-expresses the problem in terms of a smooth function and an auxiliary function around each ion core that represents the rapidly varying part of the density. Although the equations are formally related to the OPW equations and the Phillips–Kleinman–Antoncik construction given in Sec. 11.2, ultrasoft pseudopotentials are a practical approach for solving equations beyond the applicability of those formulations. We will focus upon examples of states that present the greatest difficulties in the creation of accurate, smooth pseudofunctions: valence states at the beginning of an atomic shell, 1s, 2p, 3d, etc. For these states, the OPW transformation has no effect since there are no core states of the same angular momentum. Thus the wavefunctions are nodeless and extend into

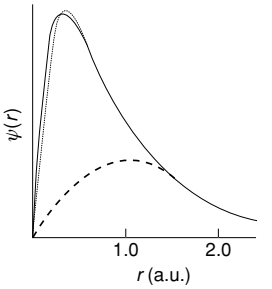


Figure 11.6. 2p radial wavefunction $\psi(r)$ for oxygen treated in the LDA, comparing the all-electron function (solid line), a pseudofunction generated using the Hamann–Schluter–Chiang approach([471] (dotted line), and the smooth part of the pseudofunction $\tilde{\psi}$ in the "ultrasoft" method (dashed line). From [474].

the core region. Accurate representation by norm-conserving pseudofunctions requires that they are at best only moderately smoother than the all-electron function (see Fig. 11.6).

The transformation proposed by Blöchl [473] and Vanderbilt [474] rewrites the non-local potential in (11.45) in a form involving a smooth function $\tilde{\phi} = r\tilde{\psi}$ that is *not norm conserving*. (We follow the notation of [474], omitting the labels PS, l, m, and σ for simplicity.) The difference in the norm equation (11.21), from that norm-conserving function $\phi = r\psi$ (either an all-electron function or a pseudofunction) is given by

$$\Delta Q_{s,s'} = \int_0^{R_c} \mathrm{d}r \, \Delta Q_{s,s'}(r), \tag{11.46}$$

where

$$\Delta Q_{s,s'}(r) = \phi_s^*(r)\phi_{s'}(r) - \tilde{\phi}_s^*(r)\tilde{\phi}_{s'}(r). \tag{11.47}$$

A new non-local potential that operates on the $\tilde{\psi}_{s'}$ can now be defined to be

$$\delta \hat{V}_{\mathrm{NL}}^{\mathrm{US}} = \sum_{s,s'} D_{s,s'} |\beta_s\rangle\langle\beta_{s'}|, \tag{11.48}$$

where

$$D_{s,s'} = B_{s,s'} + \varepsilon_{s'} \Delta Q_{s,s'}. \tag{11.49}$$

For each reference atomic states s, it is straightforward to show that the smooth functions $\tilde{\psi}_s$ are the solutions of the *generalized eigenvalue problem*

$$\left[\hat{H} - \varepsilon_s \hat{S}\right] \tilde{\psi}_s = 0, \tag{11.50}$$

with $\hat{H} = -\frac{1}{2}\nabla^2 + V_{\mathrm{local}} + \delta \hat{V}_{\mathrm{NL}}^{\mathrm{US}}$ and $\hat{S}$ an overlap operator,

$$\hat{S} = \hat{\mathbf{1}} + \sum_{s,s'} \Delta Q_{s,s'} |\beta_s\rangle\langle\beta_{s'}|, \tag{11.51}$$

which is different from unity only inside the core radius. The eigenvalues ε_s agree with the all-electron calculation at as many energies s as desired. The full density can be constructed

from the functions $\Delta Q_{s,s'}(r)$, which can be replaced by a smooth version of the all-electron density.

The advantage of relaxing the norm-conservation condition $\Delta Q_{s,s'} = 0$ is that each smooth pseudofunction $\tilde{\psi}_s$ can be formed independently, with only the constraint of matching the value of the functions $\tilde{\psi}_s(R_c) = \psi_s(R_c)$ at the radius R_c. Thus it becomes possible to choose R_c much larger than for a norm-conserving pseudopotential, while maintaining the desired accuracy by adding the auxiliary functions $\Delta Q_{s,s'}(r)$ and the overlap operator $\hat{S}$. An example of the un-normalized smooth function for the 2p state of oxygen is shown in Fig. 11.6, compared to a much more rapidly varying norm-conserving function.

In a calculation that uses an "ultrasoft pseudopotential" the solutions for the smooth functions $\tilde{\psi}_i(\mathbf{r})$ are orthonormalized according to

$$\langle \tilde{\psi}_i | \hat{S} | \tilde{\psi}_{i'} \rangle = \delta_{i,i'}, \tag{11.52}$$

and the valence density is defined to be

$$n_v(\mathbf{r}) = \sum_i^{\text{occ}} \tilde{\psi}_i^*(\mathbf{r}) \tilde{\psi}_{i'}(\mathbf{r}) + \sum_{s,s'} \rho_{s,s'} \Delta Q_{s,s'}(\mathbf{r}), \tag{11.53}$$

where

$$\rho_{s,s'} = \sum_i^{\text{occ}} \langle \tilde{\psi}_i | \beta_{s'} \rangle \langle \beta_s | \tilde{\psi}_i \rangle. \tag{11.54}$$

The solution is found by minimizing the total energy

$$\begin{aligned} E_{\text{total}} = &\sum_i^{\text{occ}} \langle \tilde{\psi}_n | -\frac{1}{2}\nabla^2 + V_{\text{local}}^{\text{ion}} + \sum_{s,s'} D_{s,s'}^{\text{ion}} \, |\beta_s\rangle\langle\beta_{s'}| \, |\tilde{\psi}_n\rangle \\ &+ E_{\text{Hartree}}[n_v] + E_{II} + E_{\text{xc}}[n_v], \end{aligned} \tag{11.55}$$

which is the analog of (7.5) and (9.3), except that now the normalization condition is given by (11.52).[10] If we define the "unscreened" bare ion pseudopotential by $V_{local}^{ion} \equiv V_{local} - V_{H\text{xc}}$, where $V_{H\text{xc}} = V_H + V_{\text{xc}}$, and similarly $D_{s,s'}^{ion} \equiv D_{s,s'} - D_{s,s'}^{H\text{xc}}$ with

$$D_{s,s'}^{Hxc} = \int \mathrm{d}\mathbf{r} V_{Hxc}(\mathbf{r}) \Delta Q_{s,s'}(r), \tag{11.56}$$

this leads to the generalized eigenvalue problem

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{local}} + \delta \hat{V}_{\text{NL}}^{\text{US}} - \varepsilon_i \hat{S} \right] \tilde{\psi}_i = 0, \tag{11.57}$$

where $\delta \hat{V}_{\text{NL}}^{\text{US}}$ is given by the sum over ions of (11.48). Fortunately, such a generalized eigenvalue problem is not a major complication with iterative methods (see App. M).

[10] Note that one can add a "non-linear core correction" in E_{xc} just as in other pseudopotential methods.

11.11 Projector augmented waves (PAWs): keeping the full wavefunction

The projector augmented wave (PAW) method [475, 476, 512] is a general approach to solution of the electronic structure problem that reformulates the OPW method, adapting it to modern techniques for calculation of total energy, forces, and stress. Like the "ultrasoft" pseudopotential method, it introduces projectors and auxiliary localized functions. The PAW approach also defines a functional for the total energy that involves auxiliary functions and it uses advances in algorithms for efficient solution of the generalized eigenvalue problem like (11.57). However, the difference is that the PAW approach keeps the full all-electron wavefunction in a form similar to the general OPW expression given earlier in (11.1); since the full wavefunction varies rapidly near the nucleus, all integrals are evaluated as a combination of integrals of smooth functions extending throughout space plus localized contributions evaluated by radial integration over muffin-tin spheres, as in the augmented plane wave (APW) approach of Ch. 16.

Here we only sketch the basic ideas of the definition of the PAW method for an atom, following [475]. Further developments for calculations for molecules and solids [475, 476, 512] are deferred to Sec. 13.2. Just as in the OPW formulation, one can define a smooth part of a valence wavefunction $\tilde{\psi}_i^v(\mathbf{r})$ (a plane wave as in (11.1) or an atomic orbital as in (11.4)), and a linear transformation $\psi^v = \mathcal{T}\tilde{\psi}^v$ that relates the set of all-electron valence functions $\psi_j^v(\mathbf{r})$ to the smooth functions $\tilde{\psi}_i^v(\mathbf{r})$. The transformation is assumed to be unity except with a sphere centered on the nucleus, $\mathcal{T} = \mathbf{1} + \mathcal{T}_0$. For simplicity, we omit the superscript v, assuming that the ψs are valence states, and the labels i, j. Adopting the Dirac notation, the expansion of each smooth function $|\tilde{\psi}\rangle$ in partial waves m within each sphere can be written (see Eqs. (J.1) and (16.5)),

$$|\tilde{\psi}\rangle = \sum_m c_m |\tilde{\psi}_m\rangle, \tag{11.58}$$

with the corresponding all-electron function,

$$|\psi\rangle = \mathcal{T}|\tilde{\psi}\rangle = \sum_m c_m |\psi_m\rangle. \tag{11.59}$$

Hence the full wavefunction in all space can be written

$$|\psi\rangle = |\tilde{\psi}\rangle + \sum_m c_m \left\{|\psi_m\rangle - |\tilde{\psi}_m\rangle\right\}, \tag{11.60}$$

which has the same form as Eqs. (11.4) and (11.8).

If the transformation $\mathcal{T}$ is required to be linear, then the coefficients must be given by a projection in each sphere

$$c_m = \langle \tilde{p}_m | \tilde{\psi}\rangle, \tag{11.61}$$

for some set of projection operators $\tilde{p}$. If the projection operators satisfy the biorthogonality condition,

$$\langle \tilde{p}_m | \tilde{\psi}_{m'}\rangle = \delta_{mm'}, \tag{11.62}$$

then the one-center expansion $\sum_m |\tilde{\psi}_m\rangle\langle \tilde{p}_m|\tilde{\psi}\rangle$ of the smooth function $\tilde{\psi}$ equals $\tilde{\psi}$ itself.

The resemblance of the projection operators to the separable form of pseudopotential operators (Sec. 11.8) is apparent. Just as for pseudopotentials, there are many possible choices for the projectors with examples given in [475] of smooth functions for $\tilde{p}(\mathbf{r})$ closely related to pseudopotential projection operators. The difference from pseudopotentials, however, is that the transformation $\mathcal{T}$ still involves the full all-electron wavefunction

$$\mathcal{T} = \mathbf{1} + \sum_m \left\{ |\psi_m\rangle - |\tilde{\psi}_m\rangle \right\} \langle \tilde{p}_m|. \tag{11.63}$$

Furthermore, the expressions apply equally well to core and valence states so that one can derive all-electron results by applying the expressions to all the electron states.

The general form of the PAW equations can be cast in terms of transformation (11.63). For any operator $\hat{A}$ in the original all-electron problem, one can introduce a transformed operator $\tilde{A}$ that operates on the smooth part of the wavefunctions

$$\tilde{A} = \mathcal{T}^\dagger \hat{A} \mathcal{T} = \hat{A} + \sum_{mm'} |\tilde{p}_m\rangle \left\{ \langle \psi_m|\hat{A}|\psi_{m'}\rangle - \langle \tilde{\psi}_m|\hat{A}|\tilde{\psi}_{m'}\rangle \right\} \langle \tilde{p}_{m'}|, \tag{11.64}$$

which is very similar to a pseudopotential operator as in (11.39). Furthermore, one can add to the right-hand side of (11.64) any operator of the form

$$\hat{B} - \sum_{mm'} |\tilde{p}_m\rangle \langle \tilde{\psi}_m|\hat{B}|\tilde{\psi}_{m'}\rangle \langle \tilde{p}_{m'}|, \tag{11.65}$$

with no change in the expectation values. For example, one can remove the nuclear Coulomb singularity in the equations for the smooth function, leaving a term that can be dealt with in the radial equations about each nucleus.

The expressions for physical quantities in the PAW approach follow from (11.63) and (11.64). For example, the density is given by[11]

$$n(\mathbf{r}) = \tilde{n}(\mathbf{r}) + n^1(\mathbf{r}) - \tilde{n}^1(\mathbf{r}), \tag{11.66}$$

which can be written in terms of eigenstates labeled i with occupations f_i as

$$\tilde{n}(\mathbf{r}) = \sum_i f_i |\tilde{\psi}_i(\mathbf{r})|^2, \tag{11.67}$$

$$n^1(\mathbf{r}) = \sum_i f_i \sum_{mm'} \langle \tilde{\psi}_i|\tilde{\psi}_m\rangle \psi_m^*(\mathbf{r}) \psi_{m'}(\mathbf{r}) \langle \tilde{\psi}_{m'}|\tilde{\psi}_i\rangle, \tag{11.68}$$

and

$$\tilde{n}^1(\mathbf{r}) = \sum_i f_i \sum_{mm'} \langle \tilde{\psi}_i|\tilde{\psi}_m\rangle \tilde{\psi}_m^*(\mathbf{r}) \tilde{\psi}_{m'}(\mathbf{r}) \langle \tilde{\psi}_{m'}|\tilde{\psi}_i\rangle. \tag{11.69}$$

The last two terms are localized around each atom and the integrals can be done in spherical coordinates with no problems from the string variations near the nucleus, as in augmented methods. Section 13.2 is devoted to the PAW method and expressions for other quantities in molecules and condensed matter.

[11] The equations are modified if the core functions are not strictly localized in the augmentation spheres [513].

11.12 Additional topics

Operators with non-local potentials

The non-local character of pseudopotentials leads to complications that the user should be aware of. One is that the usual relation of momentum and position matrix elements does not hold [514, 515]. The analysis at Eq. (19.31) shows that for non-local potentials the correct relation is

$$[H, \mathbf{r}] = i\frac{\hbar}{m_e}\mathbf{p} + [\delta V_{nl}, \mathbf{r}]\,, \tag{11.70}$$

where δV_{nl} denotes the non-local part of the potential. The commutator can be worked out using the angular projection operators in δV_{nl} [514, 515].

Reconstructing the full wavefunction

In a pseudopotential calculation, only the pseudowavefunction is determined directly. However, the full wavefunction is required to describe many important physical properties, e.g. the Knight shift and the chemical shift measured in nuclear resonance experiments [516, 517]. These provide extremely sensitive probes of the environment of a nucleus and the valence states, but the information depends critically upon the perturbations of the core states. Other experiments, such as core level photoemission and absorption, involve core states directly.

The OPW and PAW methods provide the core wavefunctions. Is it possible to reconstruct the core wavefunctions from a usual pseudopotential calculation? The answer is yes, within some approximations. The procedure is closely related to the PAW transformation (11.63). For each scheme of generating "*ab initio*" pseudopotentials, one can formulate an explicit way to reconstruct the full wavefunctions given the smooth pseudofunction calculated in the molecule or solid. Such reconstruction has been used, e.g. by Mauri and coworkers, to calculate nuclear chemical shifts [517, 518].

Pseudohamiltonians

A pseudohamiltonian is a more general object than a pseudopotential; in addition to changing the potential *the mass is varied to achieve the desired properties of the valence states.* Since the pseudohamiltonian is chosen to represent a spherical core, the *pseudo kinetic energy operator* is allowed only to have a mass that can be different for radial and tangential motion and whose magnitude can vary with radius [519]. Actual pseudohamiltonians derived thus far have assumed that the potential is local [519–521]. If such a form can be found it will be of great use in Monte Carlo calculations where the non-local operators are problematic [519, 520]; however, it has so far not proven possible to derive pseudohamiltonians of general applicability.

Beyond the single-particle approximation

It is also possible to define pseudopotentials that describe the effects of the cores *beyond the independent-electron approximation* [493, 522–524]. At first sight, it seems impossible to define a hamiltonian for valence electrons only, omitting the cores, when all electrons are identical. However, a proper theory can be constructed relying on the fact that all low-energy excitations can be mapped one-to-one onto a valence-only problem. In essence, the outer valence electrons can be viewed as *quasiparticles* that are renormalized by the presence of the core electrons. Further treatment is beyond the scope of the present work, but extensive discussion and actual pseudopotentials can be found in [522, 523].

SELECT FURTHER READING

Basic theory:

Ashcroft, N. W. and Mermin, N. D., *Solid State Physics*, W. B. Saunders Company, Philadelphia, 1976.

Jones, W. and March, N. H., *Theoretical Solid State Physics, Vol. I*, John Wiley and Sons, New York, 1976.

Kittel, C., *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1996.

Ziman, J. M., *Principles of the Theory of Solids*, Cambridge University Press, Cambridge, 1989.

History and early work:

Heine, V., in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1970, p. 1.

Empirical pseudopotential method:

Cohen, M. L. and Chelikowsky, J. R., *Electronic Structure and Optical Properties of Semiconductors*, 2nd ed., Springer-Verlag, Berlin, 1988.

Cohen, M. L. and Heine, V., in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1970, p. 37.

Recent developments:

Blöchl, P. E., "Generalized separable potentials for electronic-structure calculations," *Phys. Rev. B* 41:5414–5416, 1990.

Blöchl, P. E., "Projector augmented-wave method," *Phys. Rev. B* 50:17953–17979, 1994.

Hamann, D. R., Schlüter, M. and Chiang, C., "Norm-conserving pseudopotentials," *Phys. Rev. Lett.* 43:1494–1497, 1979.

Kleinman, L. and Bylander, D. M., "Efficacious form for model pseudopotentials," *Phys. Rev. Lett.* 48:1425–1428, 1982.

Kresse, G. and Joubert, D., "From ultrasoft pseudopotentials to the projector augmented-wave method," *Phys. Rev. B* 59:1758–1775, 1999.

Pickett, W. E., "Pseudopotential methods in condensed matter applications," *Computer Physics Reports* 9:115, 1989.

Singh, D. J., *Planewaves, Pseudopotentials, and the APW Method*, Kluwer Academic Publishers, Boston, 1994, and references therein.

Vanderbilt, D., "Soft self-consistent pseudopotentials in a generalized eigenvalue formalism," *Phys. Rev. B* 41:7892, 1990.

Exercises

11.1 Consider s-wave ($l = 0$) scattering in the example illustrated in Fig. 11.1. Using formula (J.4) for the radial wavefunction ψ, with the definition $\phi = r\psi$, and the graphical construction indicated in Fig. 11.1, show that the scattering length approaches a well-defined limit as $\kappa \to 0$, and find the relation to the phase shift $\eta_0(\varepsilon)$.

11.2 The pseudopotential concept can be illustrated by a square well in one dimension with width s and depth $-V_0$. (See also Exercises 11.6 and 11.14; the general solution for bands in one dimension in Exercise 4.22; and relations to the plane wave, APW, KKR, and MTO methods, respectively, in Exercises 12.6, 16.1, 16.7, and 16.13.)

A plane wave with energy $\varepsilon > 0$ travelling to the right has a reflection coefficient r and transmission coefficient t (see Exercise 4.22).

(a) By matching the wavefunction at the boundary, derive r and t as a function of V_0, s, and ε. Note that the phase shift δ is the shift of phase of the transmitted wave compared to the wave in the absence of the well.

(b) Show that the same transmission coefficient t can be found with different V_0' and/or s' at a chosen energy ε_0.

(c) Combined with the analysis in Exercise 4.22, show that a band in a one-dimensional crystal is reproduced approximately by the modified potential. The bands agree exactly at energy $\varepsilon_k = \varepsilon_0$ and have errors linear in $\varepsilon_k - \varepsilon_0 +$ higher order terms.

11.3 Following Eq. (11.9) it is stated that if $u_{li} = \psi_{li}^c$ in the OPW, then the smooth function $\tilde{\psi}_l^v(\mathbf{r})$ has no radial nodes. Show that this follows from definition of the OPW.

11.4 Verify expression (11.10) for the norm of an OPW. Show this means that different OPWs are not orthonormal and each has norm less than unity.

11.5 Derive the transformation from the OPW equation (11.11) to the pseudopotential equation (11.12) for the smooth part of the wavefunction.

11.6 Consider the one-dimensional square well defined in Exercise 11.2. There (and in Exercise 4.22) the scattering was considered in terms of left and right propagating waves ψ_l and ψ_r. However, pseudopotentials are defined for eigenstates of the symmetry. In one dimension the only spatial symmetry is inversion, so that all states can be classified as even or odd. Here we construct a pseudopotential; the analysis is also closely related to the KKR solution in Exercise 16.7.

(a) Using linear combinations of ψ_l and ψ_r, construct even and odd functions, and show they have the form

$$\begin{aligned} \psi^+ &= \mathrm{e}^{-ik|x|} + (t+r)\mathrm{e}^{ik|x|}, \\ \psi^- &= \left[\mathrm{e}^{-ik|x|} + (t-r)\mathrm{e}^{ik|x|}\right]\mathrm{sign}(x). \end{aligned} \tag{11.71}$$

(c) From the relation of t and r given in Exercise 4.22, show that the even and odd phase shifts are given by

$$\begin{aligned} \mathrm{e}^{2i\eta^+} &\equiv t + r = \mathrm{e}^{i(\delta+\theta)}, \\ \mathrm{e}^{2i\eta^-} &\equiv t - r = \mathrm{e}^{i(\delta-\theta)}, \end{aligned} \tag{11.72}$$

where $t = |t|\mathrm{e}^{i\delta}$ and $\theta \equiv \cos^{-1}(|t|)$.

(d) Repeat the analysis of Exercise 11.2 and show that the band of a one dimensional crystal at a given energy ε is reproduced by a pseudopotential if *both* phase shifts $\eta^+(\varepsilon)$ and $\eta^-(\varepsilon)$ are correct.

11.7 Find the analytic formulas for the Fourier transforms of a spherical square well potential $V(r) = v_0$, $r < R_0$, and a gaussian potential $V(r) = A_0 \exp -\alpha r^2$, using the expansion of a plane wave in spherical harmonics.

11.8 Show that the radial Schrödinger equation can be transformed to the non-linear first-order differential equation (11.24).

11.9 Show that (11.26) indeed holds for any function f and that this relation leads to (11.27) with the choice $f(r) = (\partial/\partial\varepsilon)x_l(\varepsilon, r)$. To do this use the fact that $\phi = 0$ at the origin so that the final answer depends only upon $f(R)$ and $\phi(R)$ at the outer radius.

11.10 Show that the third condition of norm conservation (agreement of logarithmic derivatives of the wavefunction) ensures that the potential is continuous at R_c.

11.11 Use a code from the on-line source given in Ch. 24 to generate a "high quality" (small R_c) pseudopotential for Si in the usual atomic ground state $3s^2 3p^2$. Check that the eigenvalues are the same as the all-electron calculation.
(a) Use the same pseudopotential to calculate the eigenvalues in various ionization states $+1$, $+2$, $+3$, $+4$. How do the eigenvalues agree with the all-electron results.
(b) Repeat for a poorer quality (larger R_c) pseudopotential. Is the agreement worse? Why or why not?
(c) Carry out another set of calculations for a "compressed atom," i.e. confined to a radius $\approx \frac{1}{2}$ the nearest neighbor distance. (This may require changes in the code.) Calculate the changes in eigenvalues using the all-electron code and using the same pseudopotential, i.e. one derived from the "compressed atom." How do they agree?
(d) Non-linear core correlation corrections can also be tested. In many generation codes, the corrections can simply be turned on or off. One can also calculate explicitly the exchange–correlation energy using the pseudo and the entire density. The largest effects are for spin polarized transition metals, e.g. Mn $3d^{5\uparrow}$ compared to $3d^{4\uparrow}\ 3d^{1\downarrow}$.

11.12 Show that unphysical "ghost states" can occur at low energies as eigenvalues of the hamiltonian with the non-local potential operator (11.39) if $V_{\text{local}}(r)$ is chosen to be large and negative (attractive) so that the non-local $\delta V_l(r)$ must be large and positive. Hint: Consider the limit of a very large negative $V_{\text{local}}(r)$ acting on a state that is orthogonal to $\phi_l(r)$.

11.13 Show that each ψ_s is a solution of $\hat{H}\psi_s = \varepsilon_s \psi_s$ if the "ultrasoft" potential is constructed using (11.45).

11.14 The square well in one dimension considered in Exercises 11.2 and 11.6 illustrates ideas of the OPW and pseudopotential methods and also shows close relations to other methods (see Exercise 11.2). In this example we consider a bound state with $\varepsilon < 0$, but similar ideas apply for $\varepsilon > 0$ (Exercise 11.2).
(a) A deep well has states analogous to core states with $\varepsilon_c \ll 0$. Consider a well with width $s = 2a_0$ and depth $-V_0 = -12Ha$. Solve for the two lowest "core" states using the approximation that they are bound states of an infinite well. Solve for the third "valence" state by matching the wavefunction.

(b) Construct a generalized OPW-like valence state using the definition $\psi^v(x) = \tilde{\psi}^v(x) + \sum_j B_j u_j(x)$, analogous to (11.4). Rather than using the expressions in Fourier space, it is easiest to use the definition $B_j = \langle u_j | \tilde{\psi}^v \rangle$. The overlap B_j is zero for one of the "core" states; give the reason and generalize the argument to apply to core states of an atom in three dimensions. Show that the "smooth state" $\tilde{\psi}^v$ is indeed smoother than the original ψ^v.
(c) Construct the PKA pseudopotential analogous to (11.13) and show that its operation on $\tilde{\psi}^v$ is effectively that of a weaker potential.
(d) Construct a model potential with the same width s but weaker potential V_0' that has the same logarithmic derivative at the "valence" energy ε. Is this potential norm-conserving?
(e) Construct a norm-conserving potential, which can be done by first finding a nodeless norm-conserving wavefunction and inverting it as in (11.32). If the form of the wavefunction is analytic, e.g. a polynomial inside the well, all steps can be done analytically.
(f) Write a computer code to integrate the one-dimensional Schrödinger equation and evaluate the logarithmic derivative as a function of energy near ε and compare the results for the original problem with the pseudopotentials from parts (d) and (e).
(g) Transform the potential to a separable form as in Sec. 11.8. There is only one projector since only one state is considered. Show that for a symmetric well in one dimension the general form involves only two projectors for even and odd functions.
(h) Generate an "ultrasoft" potential and the resulting generalized eigenvalue problem analogous to (11.57). Discuss the relation to the OPW method and PKA form of the potential.
(i) Generate a PAW function and show the relation to the OPW and APW methods (part (b) above and Exercise 16.1).

PART IV

THE THREE BASIC METHODS DETERMINATION OF ELECTRONIC STRUCTURE

There are nine and sixty ways of constructing tribal lays,
And every single one of them is right!
Rudyard Kipling, *In the Neolithic Age*

Overview of Chapters 12–17

There are three basic approaches to the calculation of independent-particle electronic states in materials. There are no fundamental disagreements: all agree when applied carefully and taken to convergence. Indeed, each of the approaches leads to instructive, complementary ways to understand electronic structure and each can be developed into a general framework for accurate calculations.

- Each method has its advantages: each is most appropriate for a range of problems and can provide particularly insightful information in its realm of application.
- Each method has its pitfalls: the user beware. It is all too easy to make glaring errors or over-interpret results if the user does not understand the basics of the methods.

The three types of methods and their characteristic pedagogical values are:

1. **Plane wave and grid methods** provide general approaches for solution of differential equations, including the Schrödinger and Poisson equations. At first sight, plane waves and grids are very different, but in fact each is an effective way of representing smooth functions. Furthermore, grids are involved in modern efficient plane wave calculations that use fast Fourier transforms.

   Chapter 12 is devoted to the basic concepts and methods of electronic structure. Plane waves are presented first because of their simplicity and because Fourier transforms provide a simple derivation of the Bloch theorem. Since plane waves are eigenfunctions of the Schrödinger equation with constant potential, they are the natural basis for description of bands in the nearly-free-electron approximation which provides important insight into band structures of many materials including sp-bonded metals, semiconductors, *etc.* Pseudopotentials are intertwined with plane wave methods because they allow calculations to be done with a feasible number of plane waves. The basic ideas can be understood in terms of empirical pseudopotentials which provide a compact description

of bands in terms of a few Fourier components of the potentials. Real-space grids provide an alternative way to solve the equations, which is especially useful for finite systems.

Chapter 13 is devoted to self-consistent "*ab initio*" methods that utilize plane waves (and/or grids) to solve the Kohn–Sham equations. Because of the simplicity of plane waves, they are often the basis of choice for development of new methods, such as Car-Parrinello quantum molecular dynamics simulations (Ch. 18), efficient iterative methods (App. M), and many other innovations. Plane waves and grids are appropriate for smooth functions, *namely* pseudopotentials or related operators. "Norm-conserving" potentials provide accurate solutions with pseudofunctions that are orthonormal solutions of ordinary differential equations.

Two approaches have brought the OPW approach into the framework of total energy functionals: ultrasoft pseudopotentials and projector augmented wave (PAW) formulation. With "ultrasoft" pseudopotentials the problem is cast in terms of localized spherical functions and smooth wavefunctions that obey a generalized eigenvalue equation with an OPW-type hamiltonian. The projector augmented wave (PAW) formulation completes the transformation by expressing the wavefunctions as a sum of smooth functions plus core functions, just as in the OPW approach. Unlike pseudopotentials, the PAW method keeps the entire set of all-electron core functions and the smooth parts of the valence functions. Matrix elements involving core functions are treated using muffin-tin spheres as in augmented methods (Ch. 16). Nevertheless, the ultrasoft and PAW methods maintain the advantage of pseudopotentials that forces can be calculated easily.

2. **Localized atomic(-like) orbitals (LCAO)** provide a basis that captures the essence of the atomic-like features of solids and molecules. They provide a satisfying, localized description of electronic structure widely used in chemistry, in recently developed "order-N" methods (Ch. 23), and in constructing useful models.

   Chapter 14 defines the orbitals and presents basic theory. In particular, local orbitals provide an illuminating derivation (indeed the original derivation of Bloch in 1928) of the Bloch theorem. The semiempirical tight-binding method, associated with Slater and Koster, is particularly simple and instructive since one needs only the matrix elements of the overlap and hamiltonian. Tables of tight-binding matrix elements can be used to determine electronic states, total energies, and forces with very fast, simple calculations.

   Chapter 15 is devoted to methods for full calculations done with localized bases such as gaussians, Slater-type orbitals, and numerical radial atomic-like orbitals. Calculations can vary from quick (and often dirty) to highly refined with many basis orbitals per atom. Even in the latter case, the calculations can be much smaller than with plane waves or grids. However, compared to general bases like plane waves and grids, it is harder to reach convergence and greater care is needed in constructing basis functions of sufficient quality.

3. **Atomic sphere methods** are the most general methods for precise solution of the Kohn–Sham equations. The basic idea is to divide the electronic structure problem, providing efficient representation of atomic-like features that are rapidly varying near each nucleus and smoothly varying functions between the atoms.

Chapter 16 is devoted to the original methods in which smooth functions are "augmented" near each nucleus by solving the Schrödinger equation in the sphere at each energy and matching to the outer wavefunction. The resulting APW and KKR methods are very powerful, but suffer from the fact that they require solution of non-linear equations. The Green's function KKR method is particularly elegant, providing local information as well as global information such as the Fermi surface. The non-linearity does not present any problem in a Green's function approach; however, it is difficult to extend the KKR approach beyond the muffin-tin potential approximation. Muffin-tin orbitals (MTOs) are localized, augmented functions that can lead to physically meaningful descriptions of electronic states in terms of a *minimal basis*, including the concept of "canonical bands," described in terms of structure constants and a very few "potential parameters."

Chapter 17 deals with the advance that has made augmented methods much more tractable and useful: the "L" methods that make use of linearization of the equations around reference energies. This allows any of the augmented methods to be written in the familiar form of a secular equation linear in energy involving a hamiltonian and overlap matrix. The simplification has led to further advances, e.g. the development of full-potential methods, so that LAPW provides the most precise solutions of the Kohn–Sham equations available today. The LMTO approach describes electronic states in terms of a reduced linear hamiltonian with basis functions that are localized and designed to provide understanding of the electronic states. LMTO involves only a small basis and can be cast in the form of an "*ab initio*" orthogonal tight-binding hamiltonian with all matrix elements derived from the fundamental Kohn–Sham hamiltonian. It is also possible to go beyond linearization and a methodology is provided by the "NMTO" generalization to order N.

12

Plane waves and grids: basics

Summary

Plane waves and grids provide general methodologies for solution of differential equations including the Schrödinger and Poisson equations: in many ways they are very different and in other ways they are two sides of the same coin. Plane waves are especially appropriate for periodic crystals where they provide intuitive understanding as well as simple algorithms for practical calculations. Methods based upon grids in real space are most appropriate for finite systems and are prevalent in many fields of science and engineering. We introduce them together because modern electronic structure algorithms use both plane waves and grids with fast Fourier transforms.

This chapter is organized first to give the general equations in a plane wave basis and a transparent derivation of the Bloch theorem, complementary to the one given in Ch. 4. The remaining sections are devoted to relevant concepts and useful steps, such as nearly-free-electron approximation and empirical pseudopotentials, that reveal the characteristic properties of electronic bands in materials. This lays the ground work for the full solution of the density functional equations using *ab initio* non-local pseudopotentials given in Ch. 13.

12.1 The independent-particle Schrödinger equation in a plane wave basis

The eigenstates of any independent particle Schrödinger-like equation in which each electron moves in an effective potential $V_{eff}(\mathbf{r})$,[1] such as the Kohn–Sham equations, satisfy the eigenvalue equation

$$\hat{H}_{eff}(\mathbf{r})\psi_i(\mathbf{r}) = \left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{eff}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}). \tag{12.1}$$

In a solid (or any state of condensed matter) it is convenient to require the states to be normalized and obey periodic boundary conditions in a large volume Ω that is allowed to go to infinity. (Any other choice of boundary conditions will give the same result in the large

[1] The derivations in this section also hold if the potential is a non-local operator acting only on valence electrons (as for a non-local pseudopotential) or is energy dependent (as in the APW method). See Exercise 12.8.

Ω limit [90].) Using the fact that any periodic function can be expanded in the complete set of Fourier components, an eigenfunction can be written

$$\psi_i(\mathbf{r}) = \sum_{\mathbf{q}} c_{i,\mathbf{q}} \times \frac{1}{\sqrt{\Omega}} \exp(\mathrm{i}\mathbf{q}\cdot\mathbf{r}) \equiv \sum_{\mathbf{q}} c_{i,\mathbf{q}} \times |\mathbf{q}\rangle, \tag{12.2}$$

where $c_{i,\mathbf{q}}$ are the expansion coefficients of the wavefunction in the basis of orthonormal plane waves $|\mathbf{q}\rangle$ satisfying

$$\langle \mathbf{q}'|\mathbf{q}\rangle \equiv \frac{1}{\Omega}\int_{\Omega} \mathrm{d}\mathbf{r}\ \exp(-\mathrm{i}\mathbf{q}'\cdot\mathbf{r})\exp(\mathrm{i}\mathbf{q}\cdot\mathbf{r}) = \delta_{\mathbf{q},\mathbf{q}'}. \tag{12.3}$$

Inserting (12.2) into (12.1), multiplying from the left by $\langle \mathbf{q}'|$ and integrating as in (12.3) leads to the Schrödinger equation in Fourier space

$$\sum_{\mathbf{q}} \langle \mathbf{q}'|\hat{H}_{eff}|\mathbf{q}\rangle c_{i,\mathbf{q}} = \varepsilon_i \sum_{\mathbf{q}} \langle \mathbf{q}'|\mathbf{q}\rangle c_{i,\mathbf{q}} = \varepsilon_i c_{i,\mathbf{q}'}. \tag{12.4}$$

The matrix element of the kinetic energy operator is simply

$$\langle \mathbf{q}'| -\frac{\hbar^2}{2m_e}\nabla^2|\mathbf{q}\rangle = \frac{\hbar^2}{2m_e}|q|^2\delta_{\mathbf{q},\mathbf{q}'} \rightarrow \frac{1}{2}|q|^2\delta_{\mathbf{q},\mathbf{q}'}, \tag{12.5}$$

where the last expression is in Hartree atomic units. For a crystal, the potential $V_{eff}(\mathbf{r})$ is periodic and can be expressed as a sum of Fourier components (see Eqs. (4.7) to (4.11))

$$V_{eff}(\mathbf{r}) = \sum_{m} V_{eff}(\mathbf{G}_m)\exp(\mathrm{i}\mathbf{G}_m\cdot\mathbf{r}), \tag{12.6}$$

where $\mathbf{G}_m$ are the reciprocal lattice vectors, and

$$V_{eff}(\mathbf{G}) = \frac{1}{\Omega_{\mathrm{cell}}}\int_{\Omega_{\mathrm{cell}}} V_{eff}(\mathbf{r})\exp(-\mathrm{i}\mathbf{G}\cdot\mathbf{r})\mathrm{d}\mathbf{r}, \tag{12.7}$$

with Ω_{cell} the volume of the primitive cell. Thus the matrix elements of the potential

$$\langle \mathbf{q}'|V_{eff}|\mathbf{q}\rangle = \sum_{m} V_{eff}(\mathbf{G}_m)\delta_{\mathbf{q}'-\mathbf{q},\mathbf{G}_m}, \tag{12.8}$$

are non-zero only if $\mathbf{q}$ and $\mathbf{q}'$ differ by some reciprocal lattice vector $\mathbf{G}_m$.

Finally, if we *define* $\mathbf{q} = \mathbf{k} + \mathbf{G}_m$ and $\mathbf{q}' = \mathbf{k} + \mathbf{G}_{m'}$ (which differ by a reciprocal lattice vector $\mathbf{G}_{m''} = \mathbf{G}_m - \mathbf{G}_{m'}$), then the Schrödinger equation for any given $\mathbf{k}$ can be written as the matrix equation

$$\sum_{m'} H_{m,m'}(\mathbf{k})c_{i,m'}(\mathbf{k}) = \varepsilon_i(\mathbf{k})c_{i,m}(\mathbf{k}), \tag{12.9}$$

where[2]

$$H_{m,m'}(\mathbf{k}) = \langle \mathbf{k}+\mathbf{G}_m|\hat{H}_{eff}|\mathbf{k}+\mathbf{G}_{m'}\rangle = \frac{\hbar^2}{2m_e}|\mathbf{k}+\mathbf{G}_m|^2\delta_{m,m'} + V_{eff}(\mathbf{G}_m - \mathbf{G}_{m'}). \tag{12.10}$$

[2] The effective potential $V_{eff}(\mathbf{G}_m - \mathbf{G}_{m'})$ must be generalized for non-local potentials to depend on all the variables $V_{eff}(\mathbf{K}_m, \mathbf{K}_{m'})$, where $\mathbf{K}_m = \mathbf{k} + \mathbf{G}_m$ (see Sec. 12.4).

Here we have labeled the eigenvalues and eigenfunctions $i = 1, 2, \ldots$, for the discrete set of solutions of the matrix equations for a given $\mathbf{k}$. Equations (12.9) and (12.10) are the basic Schrödinger equations in a periodic crystal, leading to the formal properties of bands derived in the next section as well as to the practical calculations that are the subject of the remainder of this chapter.

12.2 The Bloch theorem and electron bands

The fundamental properties of bands and the Bloch theorem have been derived from the translation symmetry in Sec. 4.3; this section provides an alternative, simpler derivation[3] in terms of the Fourier analysis of the previous section.

1. **The Bloch theorem.** Each eigenfunction of the Schrödinger equation, (12.9), for a given $\mathbf{k}$ is given by (12.2), with the sum over $\mathbf{q}$ restricted to $\mathbf{q} = \mathbf{k} + \mathbf{G}_m$, which can be written
$$\psi_{i,\mathbf{k}}(\mathbf{r}) = \sum_m c_{i,m}(\mathbf{k}) \times \frac{1}{\sqrt{\Omega}} \exp(\mathrm{i}(\mathbf{k} + \mathbf{G}_m) \cdot \mathbf{r}) = \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r}) \frac{1}{\sqrt{N_{\mathrm{cell}}}} u_{i,\mathbf{k}}(\mathbf{r}), \quad (12.11)$$
where $\Omega = N_{\mathrm{cell}}\Omega_{\mathrm{cell}}$ and
$$u_{i,\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega_{\mathrm{cell}}}} \sum_m c_{i,m}(\mathbf{k}) \exp(\mathrm{i}\mathbf{G}_m \cdot \mathbf{r}), \quad (12.12)$$
which has the periodicity of the crystal. This is the Bloch theorem also stated in (4.33): any eigenvector is a product of $\exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r})$ and a periodic function. Since we require $\psi_{i,\mathbf{k}}(\mathbf{r})$ to be orthonormal over the volume Ω, then $u_{i,\mathbf{k}}(\mathbf{r})$ are orthonormal in one primitive cell, i.e.
$$\frac{1}{\Omega_{\mathrm{cell}}} \int_{\mathrm{cell}} \mathrm{d}\mathbf{r} u^*_{i,\mathbf{k}}(\mathbf{r}) u_{i',\mathbf{k}}(\mathbf{r}) = \sum_m c^*_{i,m}(\mathbf{k}) c_{i',m}(\mathbf{k}) = \delta_{i,i'}, \quad (12.13)$$
where the final equation means the $c_{i,m}(\mathbf{k})$ are orthonormal vectors in the discrete index m of the reciprocal lattice vectors.
2. **Bands of eigenvalues.** Since the Schrödinger equation, (12.9), is defined for each $\mathbf{k}$ separately: each state can be labeled by the wavevector $\mathbf{k}$ and the eigenvalues and eigenvectors for each $\mathbf{k}$ are independent unless they differ by a reciprocal lattice vector. In the limit of large volume Ω, the $\mathbf{k}$ points become a dense continuum and the eigenvalues $\varepsilon_i(\mathbf{k})$ become continuous *bands*. At each $\mathbf{k}$ there are a discrete set of eigenstates labeled $i = 1, 2, \ldots$, that may be found by diagonalizing the hamiltonian, (12.10), in the basis of discrete Fourier components $\mathbf{k} + \mathbf{G}_m$, $m = 1, 2, \ldots$.
3. **Conservation of crystal momentum.** Since any state can be labeled by a well-defined $\mathbf{k}$ it follows that $\mathbf{k}$ is *conserved* in a way analogous to ordinary momentum in free space; however, in this case $\mathbf{k}$ is *conserved modulo addition of any reciprocal lattice vector* $\mathbf{G}$. In fact, it follows from inspection of the Schrödinger equation, (12.9), with the hamiltonian, (12.10), that the solutions are periodic in $\mathbf{k}$, so that all unique solutions are given by $\mathbf{k}$ in one primitive cell of the reciprocal lattice.

[3] This derivation follows the "second proof" of the Bloch theorem given by Ashcroft and Mermin [84].

4. **The role of the Brillouin zone.** Since all possible eigenstates are specified by the wavevector **k** within any one primitive cell of the periodic lattice in reciprocal space, the question arises: is there a "best choice" for the cell? The answer is "yes." The first Brillouin zone (BZ) is the uniquely defined cell that is the most compact possible cell, and it is the cell of choice in which to represent excitations. It is unique among all primitive cells because its boundaries are the bisecting planes of the **G** vectors where Bragg scattering occurs (see Sec. 4.2). Inside the Brillouin zone there are no such boundaries: the bands must be continuous and analytic inside the zone. The boundaries are of special interest since every boundary point is a **k** vector for which Bragg scattering can occur; this leads to special features, such as zero group velocities due to Bragg scattering at the BZ boundary. The construction of the BZ is illustrated in Figs. 4.1, 4.2, 4.3, and 4.4, and widely used notations for points in the BZ of several crystals are given in Fig. 4.10.
5. **Integrals in k space** For many properties such as the counting of electrons in bands, total energies, etc., it is essential to integrate over **k** throughout the BZ. As pointed out in Sec. 4.3, an intrinsic property of a crystal expressed "per unit cell" is an average over **k**, i.e. a sum over the function evaluated at points **k** divided by the number of values N_k, which in the limit is an integral. For a function $f_i(\mathbf{k})$, where i denotes the discrete band index, the average value is

$$\bar{f}_i = \frac{1}{N_k}\sum_{\mathbf{k}} f_i(\mathbf{k}) \rightarrow \frac{\Omega_{\text{cell}}}{(2\pi)^d}\int_{\text{BZ}} \mathrm{d}\mathbf{k}\ f_i(\mathbf{k}), \tag{12.14}$$

where Ω_{cell} is the volume of a primitive cell in real space and $(2\pi)^d/\Omega_{\text{cell}}$ is the volume of the BZ. Specific algorithms for integration over the BZ are described in Sec. 4.6.

12.3 Nearly-free-electron approximation

The nearly-free-electron approximation (NFEA) is the starting point for understanding bands in crystals. Not only is it a way to illustrate the properties of bands in periodic crystals, but the NFEA quantitatively describes bands for many materials. In the homogeneous gas, described in Ch. 5, the bands are simply the parabola $\varepsilon(\mathbf{q}) = (\hbar^2/2m_e)|\mathbf{q}|^2$. The first step in the NFEA is to plot the free-electron bands in the BZ of the given crystal. The bands are still the simple parabola $\varepsilon(\mathbf{q}) = (\hbar^2/2m_e)|\mathbf{q}|^2$, but they are plotted as a function of **k** where $\mathbf{q} = \mathbf{k} + \mathbf{G}_m$, with **k** restricted to the BZ. Thus for each Bravais lattice, the free-electron bands have a characteristic form for lines in the Brillouin zone, with the energy axis scaled by $\Omega^{-2/3}$, where Ω is the volume of the primitive cell. By this simple trick we can plot the bands that result from the Schrödinger equation, (12.9), for a vanishing potential.

An example of a three-dimensional fcc crystal is shown in Fig. 12.1. The bands are degenerate at high symmetry points like the zone center, since several **G** vectors have the same modulus. Introduction of a weak potential on each atom provides a simple way of understanding NFEA bands, which are modified near the zone boundaries. An excellent example is Al, for which bands are shown in Fig. 16.6, compared to the free-electron parabolic dispersion. The bands are very close to free-electron-like, yet the Fermi surface is highly modified by the lattice effects because it involves bands very near zone boundary

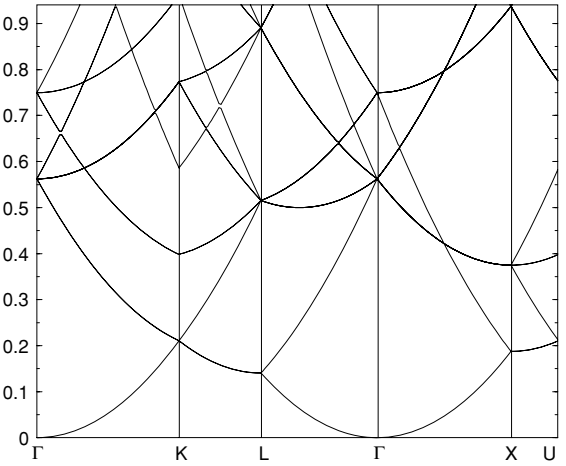


Figure 12.1. Free-electron bands plotted in the BZ of a fcc crystal. The BZ is shown in Fig. 4.10, which defines the labels. Compare this with the actual bands of Al in Fig. 16.6 that were calculated using the KKR method. Al is an ideal case where the bands are well explained by a weak pseudopotential [467–469, 529].

points where degeneracies are lifted and there are first-order effects on the bands. The bands have been calculated using many methods: the KKR method is effective since one is expanding around the analytic free-electron Green's function (outside the core) [527]; the OPW [528] and pseudopotential methods [481] make use of the fact that for weak effective scattering only a few plane waves are needed. Computer programs available on-line (Ch. 24) can be used to generate the bands and understand them in terms of the NFEA using only a few plane waves. See Exercises 12.4, 12.7, 12.11, and 12.12.

The fcc NFEA bands provide an excellent illustration of the physics of band structures. For sp-bonded metals like Na and Al, the NFEA bands are very close to the actual bands (calculated and experimental). The success of the NFEA directly demonstrates the fact that the bands can in some sense be considered "nearly-free" even though the states must actually be very atomic-like with structure near the nucleus so that they are properly orthogonal to the core states. The great beauty of the pseudopotential, APW, and KKR methods is that they provide a very simple explanation in terms of the weak scattering properties of the atom even though the potential is strong.

12.4 Form factors and structure factors

An important concept in the Fourier analysis of crystals is the division into "structure factors" and "form factors." For generality, let the crystal be composed of different species of atoms each labeled $\kappa = 1, n_{\text{species}}$, and for each κ there are n^κ identical atoms at positions $\tau_{\kappa,j}$, $j = 1, n^\kappa$ in the unit cell. Any property of the crystal, e.g. the potential, can be written,

$$V(\mathbf{r}) = \sum_{\kappa=1}^{n_{\text{species}}} \sum_{j=1}^{n^\kappa} \sum_{\mathbf{T}} V^\kappa(\mathbf{r} - \tau_{\kappa,j} - \mathbf{T}), \tag{12.15}$$

where $\mathbf{T}$ denotes the set of translation vectors. It is straightforward (Exercise 12.2) to show that the Fourier transform of (12.15) can be written as

$$V(\mathbf{G}) \equiv \frac{1}{\Omega_{\text{cell}}} \int_{\Omega_{\text{cell}}} V(\mathbf{r}) \exp(\mathrm{i}\mathbf{G}\cdot\mathbf{r})\mathrm{d}\mathbf{r} = \sum_{\kappa=1}^{n_{\text{species}}} \frac{\Omega^\kappa}{\Omega_{\text{cell}}} S^\kappa(\mathbf{G}) V^\kappa(\mathbf{G}), \tag{12.16}$$

where the **structure factor** for each species κ is

$$S^\kappa(\mathbf{G}) = \sum_{j=1}^{n^\kappa} \exp(\mathrm{i}\mathbf{G}\cdot\tau_{\kappa,j}) \tag{12.17}$$

and the **form factor** is[4]

$$V^\kappa(\mathbf{G}) = \frac{1}{\Omega^\kappa} \int_{\text{all space}} V^\kappa(\mathbf{r}) \exp(\mathrm{i}\mathbf{G}\cdot\mathbf{r})\mathrm{d}\mathbf{r}. \tag{12.18}$$

The factors in (12.16)–(12.17) have been chosen so that $V^\kappa(|\mathbf{G}|)$ is defined in terms of a "typical volume" Ω^κ for each species κ, so that $V^\kappa(|\mathbf{G}|)$ is independent of the crystal. In addition, the structure factor is defined so that $S^\kappa(\mathbf{G}=0) = n^\kappa$. These are arbitrary – but convenient – choices; other authors may use different conventions.

Equation (12.16) is particularly useful in cases where the potential is a sum of spherical potentials in real space,

$$V^\kappa(\mathbf{r} - \tau_{\kappa,j} - \mathbf{T}) = V^\kappa(|\mathbf{r} - \tau_{\kappa,j} - \mathbf{T}|). \tag{12.19}$$

This always applies for nuclear potentials and bare ionic pseudopotentials. Often it is also a reasonable approximation for the total crystal potential as the sum of spherical potentials around each nucleus.[5] Using the well-known expansion of plane waves in spherical harmonics, (J.1), Eq. (12.18) can be written as [104,413,470]

$$V^\kappa(\mathbf{G}) = V^\kappa(|\mathbf{G}|) = \frac{4\pi}{\Omega^\kappa} \int_0^\infty \mathrm{d}r r^2 \; j_0(|\mathbf{G}|r) V^\kappa(r). \tag{12.20}$$

For a nuclear potential, $V^\kappa(\mathbf{G})$ is simply

$$\begin{aligned} V^\kappa_{\text{nucleus}}(|\mathbf{G}|) &= \frac{4\pi}{\Omega^\kappa} \frac{-Z^\kappa_{\text{nucleus}} e^2}{|\mathbf{G}|^2}, \quad \mathbf{G} \neq 0, \\ &= 0, \quad \mathbf{G} = 0, \end{aligned} \tag{12.21}$$

where the divergent $\mathbf{G} = 0$ term is treated separately, as discussed in Sec. 3.2 and App. F. For a bare pseudopotential, the potential form factor (12.20) is the transform of the pseudopotential $V_l(\mathbf{r})$, given in Ch. 11. Again the $\mathbf{G} = 0$ term must be treated carefully. One procedure is to calculate the potential and total energy of point ions of charge Z^κ in a compensating background that represents the $\mathbf{G} = 0$ Fourier component of the electron density.

[4] Note the difference from (4.11), between (12.18) and where for the latter the integral is over the cell instead of all space; Exercise 12.3 shows the equivalence of the expressions.

[5] Many studies have verified that the total potential $V(\mathbf{r})$ is close to the sum of neutral atom potentials. This is especially true for examples like transition metals where the environment of each atom is nearly spherical. See Ch. 16.

In that case, there is an additional contribution that arises from the fact that the ion is not a point charge [530],

$$\alpha^\kappa = \int 4\pi r^2 \mathrm{d}r \left[V^\kappa_{\mathrm{local}}(r) - \left(-\frac{Z^\kappa}{r} \right) \right]. \tag{12.22}$$

Each ion contributes a constant term in the total energy (see Eq. (13.1) below) equal to $(N_e/\Omega)\alpha^\kappa$, where N_e/Ω is the average electron density.

The generalization of (12.16) to non-local potentials $V^\kappa_{\mathrm{NL}}(\mathbf{r}, \mathbf{r}')$ is straightforward. For each $\mathbf{k}$ and basis vectors $\mathbf{G}_m$ and $\mathbf{G}_{m'}$, it is convenient to define $\mathbf{K}_m = \mathbf{k} + \mathbf{G}_m$ and $\mathbf{K}_{m'} = \mathbf{k} + \mathbf{G}_{m'}$. The structure factor $S(\mathbf{G})$ still depends only upon $\mathbf{G} = \mathbf{K}_m - \mathbf{K}_{m'} = \mathbf{G}_m - \mathbf{G}_{m'}$, but the matrix elements of the semilocal form factor are more complicated since the matrix elements depends upon two arguments. Using the fact that the spherical symmetry of the non-local operator guarantees that it can be written as a function of the magnitudes $|\mathbf{K}_m|$, $|\mathbf{K}_{m'}|$ and the angle θ between $\mathbf{K}_m$ and $\mathbf{K}_{m'}$, the matrix elements of the semilocal form factor (11.15), are (Exercise 12.9)

$$\delta V^\kappa_{NL}(\mathbf{K}_m, \mathbf{K}_{m'}) = \frac{4\pi}{\Omega^\kappa} \sum_l (2l+1) P_l(\cos(\theta)) \int_0^\infty \mathrm{d}r r^2 \; j_l(|\mathbf{K}_m| r) j_l(|\mathbf{K}_{m'}| r) \delta V^\kappa_l(r). \tag{12.23}$$

This formula has the disadvantage that it must be evaluated for each $|\mathbf{K}_m|$, $|\mathbf{K}_{m'}|$, and θ, i.e. for a three-dimensional object. In order to treat this in a computationally efficient manner, one can discretize this function on a grid and interpolate during an actual calculation.

The separable Kleinman–Bylander form, (11.39), is simpler because it is a sum of products of Fourier transforms. Each Fourier transform is a one-dimensional function of $|\mathbf{K}_m|$ (and the same function of $|\mathbf{K}_{m'}|$) which is much more convenient. The form in k space is analogous to that in real space [413, 472]. (Here we denote the azimuthal quantum number as m_l to avoid confusion with the index m for basis functions $\mathbf{G}_m$.)

$$\delta V^\kappa_{NL}(\mathbf{K}_m, \mathbf{K}_{m'}) = \sum_{lm_l} \frac{Y^*_{lm_l}(\hat{\mathbf{K}}_m) T^*_l(|\mathbf{K}_m|) \times T_l(|\mathbf{K}_{m'}|) Y_{lm_l}(\hat{\mathbf{K}}_{m'})}{\langle \psi^{PS}_{lm} | \delta V_l | \psi^{PS}_{lm} \rangle}, \tag{12.24}$$

where $T_l(q)$ is the Fourier transform of the radial function $\psi^{PS}_l(r)\delta V_l(r)$. The simplicity of this form has led to its widespread use in calculations. Furthermore it is straightforward to extend to "ultrasoft" potentials that involve additional projectors (see Sec. 11.10).

12.5 Approximate atomic-like potentials

A first step in including the effects of the nuclei is to assume that the potential is a sum of atomic-like potentials. This gives all the qualitative features of the bands and often given semi-quantitative results. One procedure is simply to use the potential directly from an atomic calculation; another is to assume the potential has some simple analytic form. For example, if we approximate the electrons as nearly-free-electron-like then the total potential due to the nuclei and electrons to first order in perturbation theory is given by

$$V_{total}(\mathbf{G}) \approx V_{\mathrm{screened}}(\mathbf{G}) \equiv V_{\mathrm{bare}}(\mathbf{G})/\epsilon(\mathbf{G}), \tag{12.25}$$

where V_{bare} is a bare nuclear or ionic potential and $\epsilon(\mathbf{G})$ is the screening function. In the NFE limit, the screening is evaluated for the homogeneous gas, so it is isotropic $\epsilon(|\mathbf{G}|)$ and a reasonable approximation is the Thomas–Fermi screening, where ϵ can be written

$$\epsilon(|\mathbf{G}|) = \frac{|\mathbf{G}|^2}{|\mathbf{G}|^2 + k_0^2}, \tag{12.26}$$

using Eqs. (5.20) and Eq. (5.21), where k_0 is dependent only upon the electron density (i.e. r_s). Furthermore, since the screening is linear in this approximation, the total potential is a sum of spherical screened nuclear or ionic potential which are neutral and atomic-like.

This approach was instrumental in the early work on *ab initio* pseudopotentials, e.g. the Heine–Abarenkov potentials [467, 468, 490] that are derived from atomic data and have been very successfully used in solids with an approximate screening function such as Eq. (12.26). A simple, instructive example is hydrogen at high pressure, i.e. high density or small $r_s \approx 1$. This corresponds to about 10 GPa, pressures that can be found in the interiors of the giant planets. At such densities, hydrogen is predicted to form a monatomic crystal with nearly-free-electron bands. Since the "bare" potential is just $\propto 1/|\mathbf{G}|^2$, it is easy to work out the screened potential in the Thomas–Fermi approximation. Exercise 12.13 calculates the appropriate form factors, estimates band structure in perturbation theory, carries out calculations using available programs (or by writing one's own), and compares with fully self-consistent calculations.

This approximation is sufficient to illustrate two points. First, the total potential near each nucleus is very well approximated by a spherical atomic-like form. This is widely used in augmented methods such as APW, KKR, and LMTO that treat the region around the nucleus using spherical coordinates (Chs. 16 and 17). Second, the approximation demonstrates the problems with the straightforward application of plane waves. Except for the lowest Z elements, materials with core electrons require huge numbers of plane waves (see Exercise 12.10). This is why pseudopotentials (Ch. 11) are so intimately related to the success of plane wave methods.

12.6 Empirical pseudopotential method (EPM)

Even though the general ideas of pseudopotentials have been known for many years [58–60], and model potentials close to those used in recent work were already applied to solids as early as the 1930s [59, 60], the modern use of pseudopotentials started with the work of Phillips and Kleinman [481], and Antonchik [479,480]. Those authors realized that the band structure of sp-bonded metals and semiconductors could be quantitatively described by a few numbers: the values of the spherical atomic-like potentials at a few lowest reciprocal lattice vectors. By fitting to experimental data, a few parameters could be used to describe a tremendous amount of data related to the band structure, effective masses and band gaps, optical properties, etc. The "empirical pseudopotential" method has been described in detail by Heine and Cohen [467, 469], who showed the connections to the underlying theory. Applications to metals are covered thoroughly by Harrison [468], and a very complete

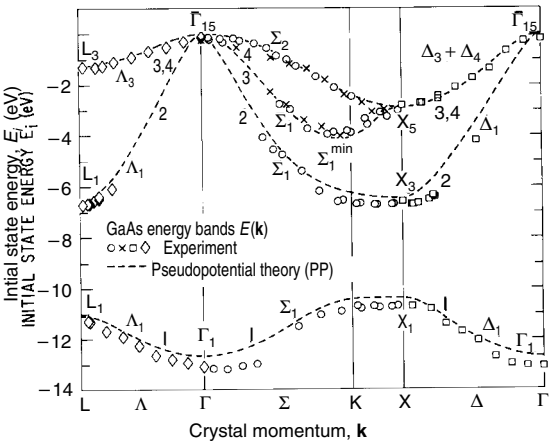


Figure 12.2. Experimental energy bands of GaAs measured by photoemission in [214] compared to empirical pseudopotential calculations [215]. The pseudopotential was fitted earlier to independent optical data, so this is a test of the transferability of information within an independent-particle theory. From [214].

exposition of the method and results for semiconductors has been given by Cohen and Chelikowsky [470].

The EPM method has played an important role in understanding electronic structure, especially for the sp-bonded metals and semiconductors. As an example, Fig. 12.2 shows the bands of GaAs measured [214] by photoemission spectroscopy are compared with EPM bands calculated [215] many years before. The agreement with the photoemission data is nearly perfect for this non-local pseudopotential that was adjusted to fit the band gaps, effective masses, and optical spectra [470]. Comparison of Fig. 12.2 with Fig. 2.25 shows the agreement with inverse photoemission and recent many-body calculations, and the fact that the adjusted EPM provides a better description of the bands than do LDA calculations. The pseudocharge density has been calculated for many materials [470]: as illustrated in Fig. 12.3, the results show the basic features of the chemical bonding and the nature of individual states. Thus the EPM plays two important roles:

- On a fundamental level, the EPM provides stimulus for the development of independent-particle band methods for solids because of its success in describing many different experiments within a single independent-particle theory.
- On a practical level, the EPM approach continues to be important because it allows bands of many important materials to be described using a few parameters, namely the first few Fourier components of the potential, Eq. (12.20).

The method is more than just a fitting procedure if one makes the approximation that the total potential is a sum of spherical potentials that *have analytic form* and are *transferable between different structures*. Although this is an approximation, it has been tested in many cases and, at least, provides semiquantitative results. With the assumption of transferability, the EPM method can readily be applied to calculations for many structures and for properties

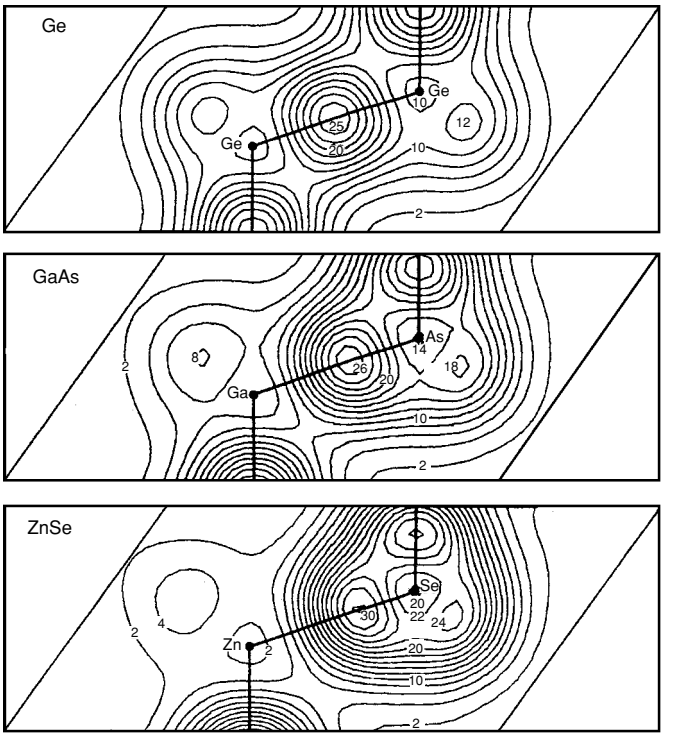


Figure 12.3. Theoretical calculations of the valence charge density of semiconductors showing the formation of the covalent bond and the progression to more ionic character in the series Ge, GaAs, and ZnSe. The results have the same basic features as full self-consistent calculations such as that for Si in Fig. 2.2. From [105].

like electron–phonon interactions (see, e.g. [531]), where the distorted lattice is simply viewed as a different structure.

The simplicity of the EPM makes possible calculations not feasible using *ab initio* pseudopotentials. It is a great advantage to have an analytic representation since it can be used for any structure. For example, EPM calculations for films [532] and "quantum dots" formed from thousands of atoms [533, 534] have been carried out using the iterative methods discussed in App. M. For example, [533, 534] report calculations of the electronic structure of pyramidal quantum dots containing up to 250,000 atoms, using spin–orbit-coupled, nonlocal, empirical pseudopotentials and with results that differ from those found using the effective-mass approximation.

A computer code for EPM (and tight-binding) calculations is available on-line as described in Ch. 24 and in schematic form in App. N. The code is modular, separating aspects that are common to all band methods from those that are specific to one method. The code includes examples of local empirical potentials for Si, Ga, and As [532] and example results including those given in Figs. 14.6 and 14.7. Options are given for a user to create new potentials. See Exercises 12.11, 12.12, and 12.13 for examples of problems illustrating EPM calculations.

12.7 Calculation of electron density: introduction of grids

One of the most important operations is the calculation of the density of electrons n. The general form for a crystal treated in independent-particle theory, e.g. Eqs. (3.42) or (7.2), can be written as

$$n(\mathbf{r}) = \frac{1}{N_k} \sum_{\mathbf{k},i} f(\varepsilon_{i,\mathbf{k}}) n_{i,\mathbf{k}}(\mathbf{r}), \text{ with } n_{i\mathbf{k}}(\mathbf{r}) = |\psi_{i,\mathbf{k}}(\mathbf{r})|^2, \tag{12.27}$$

which is an average over $\mathbf{k}$ points (see Eq. (12.14)), with i denoting the bands at each $\mathbf{k}$ point (including the spin index σ) and $f(\varepsilon_{i,\mathbf{k}})$ denoting the Fermi function. For a plane wave basis, expression (12.11) for the Bloch functions leads to

$$n_{i,\mathbf{k}}(\mathbf{r}) = \frac{1}{\Omega} \sum_{m,m'} c^*_{i,m}(\mathbf{k}) c_{i,m'}(\mathbf{k}) \exp(\mathrm{i}(\mathbf{G}_{m'} - \mathbf{G}_m) \cdot \mathbf{r}) \tag{12.28}$$

and

$$n_{i,\mathbf{k}}(\mathbf{G}) = \frac{1}{\Omega} \sum_{m} c^*_{i,m}(\mathbf{k}) c_{i,m''}(\mathbf{k}), \tag{12.29}$$

where m'' denotes the $\mathbf{G}$ vector for which $\mathbf{G}_{m''} \equiv \mathbf{G}_m + \mathbf{G}$.

The symmetry operations R_n of the crystal can be used as in Secs. 4.5 and 4.6 to find the density in terms only of the $\mathbf{k}$ points in the IBZ,

$$n(\mathbf{r}) = \frac{1}{N_k} \sum_{i,\mathbf{k}} n_{i,\mathbf{k}}(\mathbf{r}) = \frac{1}{N_{\text{group}}} \sum_{R_n} \sum_{\mathbf{k}}^{\text{IBZ}} w_{\mathbf{k}} \sum_i f(\varepsilon_{i,\mathbf{k}}) n_{i,\mathbf{k}}(R_n \mathbf{r} + \mathbf{t}_n), \tag{12.30}$$

and

$$n(\mathbf{G}) = \frac{1}{N_{\text{group}}} \sum_{R_n} \exp(\mathrm{i} R_n \mathbf{G} \cdot \mathbf{t}_n) \sum_{\mathbf{k}}^{\text{IBZ}} w_{\mathbf{k}} \sum_i f(\varepsilon_{i,\mathbf{k}}) n_{i,\mathbf{k}}(R_n \mathbf{G}). \tag{12.31}$$

The phase factor due to the translation $\exp(\mathrm{i} R_n \mathbf{G} \cdot \mathbf{t}_n)$ follows from (12.28).

Despite the simplicity of (12.29), it is not the most efficient way to calculate the density $n(\mathbf{r})$ or $n(\mathbf{G})$. The problem is that finding all the Fourier components using (12.29) involves a double sum, i.e. a convolution in Fourier space that requires N_G^2 operations, where N_G is the number of $\mathbf{G}$ vectors needed to describe the density. For large systems this becomes very expensive. On the other hand, if the Bloch states are known on a grid of N_R points in real space, the density can be found simply as a square, in N_R operations. The trick is to use a fast Fourier transform (FFT) that allows one to transform from one space to the other in $N \log N$ operations, where $N = N_R = N_G$. The flow chart, Fig. 12.4, illustrates the algorithm, and the general features for all such operations are described in Sec. M.11. A great advantage is that $n(\mathbf{r})$ is needed to find $\epsilon_{xc}(\mathbf{r})$ and $V_{xc}(\mathbf{r})$. The inverse transform

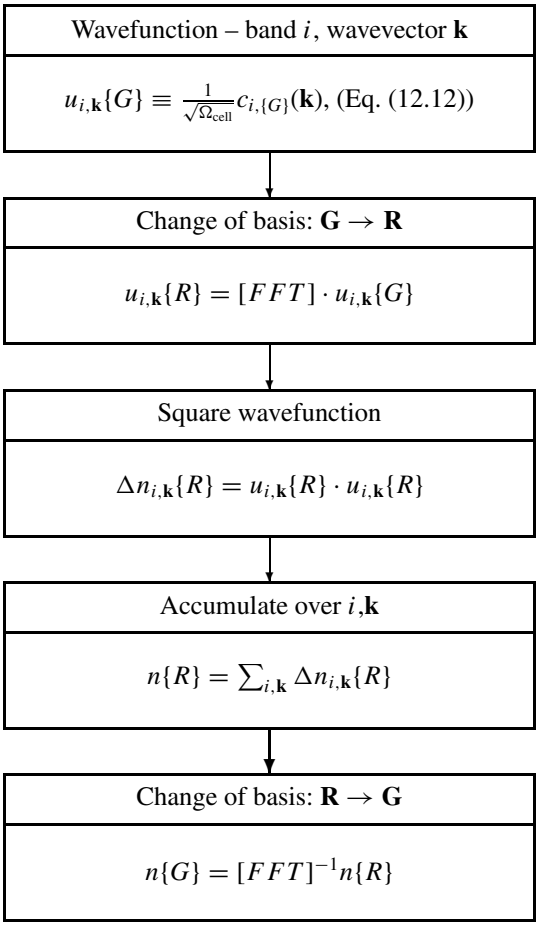


Figure 12.4. Calculation of the density using Fourier transforms and grids. The notation $\{G\}$ and $\{R\}$ denotes the sets of N $\mathbf{G}$ vectors and N grid points $\mathbf{R}$. Since the fast Fourier transform (FFT) scales as $N \ln N$, the algorithm is faster than the double sum needed to calculate $n\{G\}$ that scales as N^2. In addition, the result is given in both real and reciprocal space, needed for calculation of the *exchange–correlation* and Hartree terms. The algorithm is essentially the same as used in iterative methods, App. M.

can be used to find $n(\mathbf{G})$ which can be used for solving the Poisson equation in Fourier space.

It is relevant to note that the density n requires Fourier components that extend twice as far in each direction as those needed for the wavefunction ψ because $n \propto |\psi|^2$. Also the FFT requires a regular grid in the form of a parallelepiped, whereas the wavefunction cutoff is generally a sphere with $(1/2)|\mathbf{k} + \mathbf{G}|^2 < E_{\text{cutoff}}$. Thus the number of points in the FFT grid for density $N = N_R = N_G$ is roughly an order of magnitude larger than the number N_G^{wf} of $\mathbf{G}$ vectors in the basis for the wavefunctions. *Nevertheless, the FFT approach is much more efficient for large systems* since the number of operations scales as $N \log N$.

12.8 Real-space methods

Since the Kohn–Sham equations are a set of coupled second-order differential equations, it is natural to ask: why not use methods widely employed in many areas of computational physics, finite element, finite difference, multi-grid, wavelets, etc? In fact, such methods are used for problems like quantum dots in semiconductors and are under development for other areas of electronic structure. There is a recent review by Beck [525] and the methods will be only briefly mentioned here. As we have already pointed out in the calculation of the density, many operations are easier in real space. For example, if $\psi_i(\mathbf{r})$ is explicitly represented on a grid, then $n(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2$ with no need for an FFT as required in the plane wave method. The Hartree potential can be found using FFTs or real-space multi-grid algorithms which have been highly optimized. For solution of both the Poisson and Schrödinger equations, real-space methods are particularly advantageous for finite systems, where the wavefunctions vanish outside a boundary and the Coulomb potentials, in general, do not obey periodic boundary conditions.

Other advantages can be appreciated only in terms of the iterative methods described in App. M. All such methods require the operation $\hat{H}\psi$ instead of diagonalization of a matrix. The action of a local potential on the wavefunction is simply a point-by-point multiplication in real space, $V(\mathbf{r})\psi(\mathbf{r})$. Non-local pseudopotentials can also be handled since the non-locality extends only over the small core region; the procedure is in some ways simpler than for plane waves since the wavefunctions are already in real space. Thus the solution of the Kohn–Sham equations in real-space methods has the same form as for plane waves except that no Fourier transform is needed.

Finite difference

In a finite difference (FD) method the kinetic energy laplacian operator is evaluated from values of the function at a set of grid points. For example, the FD method of Chelikowsky et al. [526, 535] uses higher order expansions for the kinetic energy laplacian operator, separable in the x, y, z orthogonal components. For a uniform orthogonal three-dimensional (3D) grid with points (x_i, y_j, z_k), the mth order approximation is

$$\left[\frac{\partial^2\psi}{\partial x^2}\right]_{x_i,y_j,z_k} = \sum_{-m}^{m} C_m \psi(x_i + mh, y_j, z_k) + O(h^{2m+2}), \tag{12.32}$$

where h is the grid spacing and m is a positive integer. As illustrated on the left-hand side of Fig. 12.5, the laplacian at the central point is computed in terms of values of the function on the "cross" of points along the axes; the size of the dots for the 25 points represents the decreasing magnitude of C_m. This approximation is accurate to $O(h^{2m+2})$ assuming ψ can be approximated by a polynomial in h.[6] Algorithms are available to compute the coefficients C_m for any grid to arbitrary order in h [536]. Expansion coefficients for $m = 1, 6$ for a uniform grid are given in Table I of [535].

[6] The method can readily be extended to non-orthogonal systems and non-uniform grids, but at the price of having to compute many different sets of C.

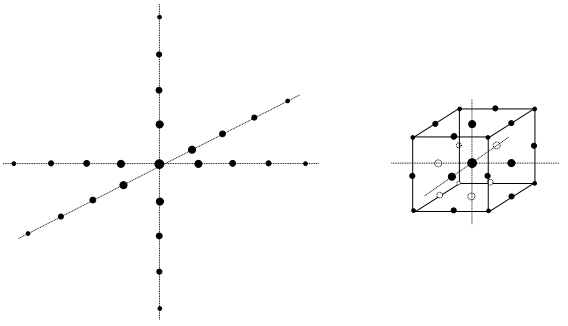


Figure 12.5. Two examples of stencils for finite difference calculation of the laplacian. The size of the points represents the weights schematically. Left: Orthogonal "cross" with 25 points [535]. Right: More compact cube of 27 points that has been used with the "Mehrstellen" (Numerov) method [209].

A different approach uses the "Mehrstellen" operator, which is an extension of the Numerov method (Sec. L.1) to higher dimensions(see [537], p. 164 as cited in [209]). As illustrated on the right-hand side of Fig. 12.5, the 27 points are more compact in space than the 25-point cross. This is an advantage, especially for finite systems where the more extended "cross" leads to larger boundary effects.

There are a number of working algorithms applied to many problems [209,526,535,538–540]. Applications are left to the following chapter since they involve full self-consistent calculations; however, the basic ideas of the laplacian operators belong here. Calculations using finite difference algorithms have been applied to many problems, including clusters and other finite systems [359,526,535,541,542]. Examples are shown in Ch. 20. A multigrid method [209, 538] based upon the Mehrstellen form for the laplacian has been applied to many periodic and non-periodic problems, such as the C–BN nanotube junction shown in Fig. 2.21. Real-space methods can also be combined [539, 540] with adaptive grids to increase resolution where needed.

Finite elements and multi-resolution

Finite elements are widely used in many fields [543]; they form a localized basis in which variational calculations can be done, unlike the finite difference method which simply approximates the laplacian. A finite element basis is usually chosen to be a set of functions, each of which is strictly localized, that overlap so that together they can form an approximation to a smooth function. Examples are triangular "hat" functions and polynomial spline functions. The former describe a piecewise linear function and the latter a smoother approximation. Matrix elements of the operators are integrals $\langle m|\hat{O}|m'\rangle$ just as for any other basis. Examples are piecewise cubic functions of Pask et al. [544, 545]; a B-spline basis closely related to traditional finite element bases, Hernandez et al. [546]; and piecewise third-order polynomials, Tsuchida and Tsukada [547]. See [525] for a more complete review.

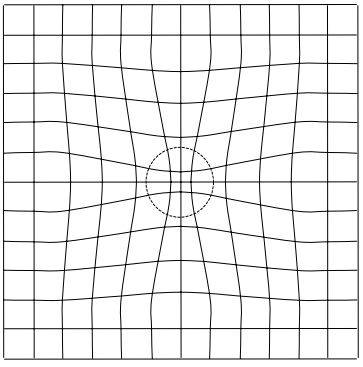


Figure 12.6. Schematic figure of adaptive coordinates defined by a smooth transformation that maps a regular grid onto a non-uniform set of points. The transformation can be thought of as working in curved space. The example shows a transformation to provide greater resolution near an atom.

Multiresolution denotes the ability to describe all regions with desired accuracy: those where there are strong variations and high resolution is needed, and others where less resolution is required. There are two general types of approaches. First, there are the well-developed finite difference and finite element methods with non-uniform grids or using non-uniform placement of localized basis functions with different widths. These work well where the structure is known in advance. Recently there has been great interest in two developments, multi-grid methods [548,549] and wavelet-type bases [550,551]. Multi-grid methods can be used with any algorithm for solution of the differential equations, and it works by cycling up and down between levels of resolution to use the speed of coarse functions while adding corrections due to fine functions. There are full-functioning codes for electronic structure that use the multigrid approach [209, 538, 552], with applications to problems such as the C–BN nanotube junction illustrated in Fig. 2.21. Wavelets involve localized bases that are optimized in important ways, e.g. the Daubichies wavelets are orthonormal at the same level and between levels. These approaches are described in reviews [525, 553] which are oriented toward electronic structure and cannot be covered in more detail here.

Adaptive curvilinear coordinates

An attractive idea suggested by Gygi [554] and Hamann [555] is to warp the grid using a smooth transformation as illustrated in Fig. 12.6. The transformation can be defined in terms of a smooth set of basis functions that map the regular points $\mathbf{r}_i$ specifying the grid to the points $\mathbf{r}'_i(\mathbf{r}_i)$. For example, the transformation can be specified in terms of plane waves. The method can be adaptive in the sense that one can make an algorithm to determine where more resolution is needed and adjust the adaption. In any case, the resulting equations expressed on the regular grid $\mathbf{r}_i$ have the form of operators in curved space. An alternative approach uses a local set of fixed transformations [539,540] around each atom that overlap

to form the complete transformation; these are easy to visualize and are equivalent to a form of global transformation [556].

SELECT FURTHER READING

Basic aspects of plane wave methods:

Ashcroft, N. W. and Mermin, N. D. *Solid State Physics*, Saunders, W.B. Company, Philadelphia, 1976.
Kittel, C. *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1996.
Ziman, J. M. *Principles of the Theory of Solids*, Cambridge University Press, Cambridge, 1989.
Marder, M. *Condensed Matter Physics*, John Wiley and Sons, New York, 2000.

The empirical pseudopotential method:

Heine, V. in *Solid State Physics*, edited by Ehrenreich, H. Seitz, F. and Turnbull, D. Academic, New York, 1970, p. 1.
Cohen, M. L. and Heine, V. in *Solid State Physics*, edited by Ehrenreich, H. Seitz, F. and Turnbull, D. Academic, New York, 1970, p. 37.
Cohen, M. L. and Chelikowsky, J. R. *Electronic Structure and Optical Properties of Semiconductors*, 2nd ed., Springer-Verlag, Berlin, 1988.

Grid methods:

Beck, T. L. "Real-space mesh techniques in density-functional theory," *Rev. Mod. Phys.* 72:1041–1080, 200.

Finite difference method:

Chelikowsky, J. R. Troullier, N. and Saad, Y. "Finite-difference-pseudopotential method: Electronic structure calculations without a basis," *Phys. Rev. Lett.* 72:1240–1243, 1994.

Multigrid method:

Briggs, E. L. Sullivan, D. J. and Bernholc, J. "Real-space multigrid-based approach to large-scale electronic structure calculations," *Phys. Rev. B* 54:14362–14375, 1996.

Exercises

12.1 See many excellent problems (and solutions) on the nearly-free-electron approximation in the book by Mihaly and Martin [248].

12.2 Show that the Fourier transform of (12.15) leads to the expression in terms of form and structure factors given in (12.16).

12.3 Show the equivalence of expressions (4.11) and (12.18) which express the final Fourier component in two ways, one an integral over the cell and the other as a structure factor times an integral for one unit only but over all space.

12.4 Plot the bands for a nearly-free-electron system in one dimension if the lattice constant is a.
(a) First plot the bands using analytic expressions for the energy in the free-electron limit.
(b) Then qualitatively sketch the changes if there is a small lattice potential.

(c) Use the empirical pseudopotential program (Sec. 12.6) or write your own to calculate the bands for a pure sine wave potential $V(x) = V_0 \sin(2\pi x/a)$. This is the Matthieu potential for which there are solutions; check your results with known results.

12.5 Consider a one-dimensional crystal with potential $V(x) = V_0 \cos(2\pi x/a)$ as in Exercise 12.4. In this exercise make the simplifying approximation that a state with wavevector k is the solution of the 2×2 hamiltonian

$$\begin{vmatrix} \frac{k^2}{2} - \varepsilon(k) & V_0 \\ V_0 & \frac{(k-G)^2}{2} - \varepsilon(k) \end{vmatrix} = 0, \tag{12.33}$$

where $G = 2\pi/a$. Give the analytic expressions for the bands $\varepsilon(k)$ and the periodic part of the Bloch functions $u_k(x)$. If there are two electrons per cell, give the expression for the density $n(x)$ as an integral over k. Evaluate the density using a grid of "special" k points (Sec. 4.6). Note that more points are required for an accurate answer if V_0 is small. Plot the lowest two bands and the electron density for the case where $V_0 = \frac{1}{4}(\pi/a)^2$ in atomic units. (See Exercise 21.12 Wannier functions and Exercise 22.10 for polarization using a variation of this model.)

12.6 Consider a one-dimensional crystal with a square well potential which in the cell at the origin has the form $V(x) = V_0$ for $-s/2 < x < s/2$ and $V = 0$ otherwise. The potential is repeated periodically in one dimension with $V(x + Na) = V(x)$, with cell length $a > s$. (See also Exercises 11.2, 11.6, 11.14; the general solution for bands in one dimension in Exercise 4.22; and relations to the APW, KKR, and MTO methods, respectively, in Exercises 16.1, 16.7, and 16.13.)
(a) First find the Fourier transform of the potential $V(G)$.
(b) Next construct a computer code or use one like that in App. N to solve for the bands. As an explicit example, choose $a = 4$, $s = 2$, and $V_0 = 0.2$ in atomic units and choose a sufficient number of plane waves so that the solution is accurate.
(c) Compare the results with the solutions in Exercise 16.1 in which the bands are found by matching the wavefunctions at the boundary, i.e. a simple example of the APW method. Of course, the result must be the same as derived by other methods: compare and contrast the plane approach with the general solution for any potential in one dimension given in Exercise 4.22.

12.7 Find the bands for Al using a simple empirical pseudopotential. One source is the paper by Segall [527] that shows bands similar to those in Fig. 16.6 calculated with $V(1\,1\,1) = 0.0115Ha$ and $V(111) = 0.0215Ha$ and mass $m^* = 1.03m$. (The last can be included as a scaling factor.) Use the NFEA to calculate energies at the X point analytically. Use the empirical pseudopotential program (Sec. 12.6) to generate full bands.

12.8 Show that the derivations in Sec. 12.1 also hold for non-local potentials as given in Eq. (12.24).

12.9 Derive the semilocal and separable forms of the pseudopotential in Eqs. (12.23) and (12.24). Hint: Use the definitions of the potential operators in real space in Ch. 11 and the expansion of a plane wave in spherical harmonics, Eq. (J.1).

12.10 Pseudopotentials are used because calculations with the full nuclear Coulomb potential are very expensive for heavy atoms of nuclear charge Z. Derive the power law with which the number of plane waves needed scales with Z. Do this by using perturbation theory for very high Fourier components, where the matrix element is given by $V(G)$ and the energy denominator

is approximately given by the kinetic energy. Argue that screening is not effective for high Fourier components.

12.11 Use the empirical pseudopotential program (App. N) to find the bands and charge densities of Si in the diamond structure at the lattice constant 5.431 Å. The bands should be insulating and the bonds should be visible in the charge density.
(a) Verify that the minimum along the Δ direction (see Fig. 4.10) is qualitatively the same as in experiment, which is given in many texts, e.g. [86], and is given by the more accurate tight-binding bands shown in Fig. 14.6.
(b) Now compress the system until it is metallic (this can only be done in theory; in reality it transforms). Can you tell when the system becomes a metal just from the density? In principle, if you had the exact functional, what aspect of the density would be the signature of the insulator–metal transition?
(c) Do a similar calculation replacing the Si atoms with Al, still in the diamond structure with lattice constant 5.431 Å. (Of course this is a theoretical structure.) There are three Al electrons/atom, i.e. six electrons per cell and it turns out to be a metal. Show that it must be metallic *without doing the calculation*. Does the density plot look a lot like Si? Can you find any feature that shows it is a metal?

12.12 Use the empirical pseudopotential program (App. N) to find the bands for GaAs.
(a) Verify that it has a direct gap at Γ.
(b) Displace the atoms in the unit cell a small amount along the (111) direction. Check the spitting of the top of the valence band at Γ. Is the spitting what you expect?
(c) Repeat with the displacement in the (100) direction.

12.13 This exercise is to work out the form factor for the screened H potential in the Thomas–Fermi approximation and calculate the bands for fcc H at very high density, $r_s = 1.0$.
(a) Estimate the deviation of the bands from the free electron parabola by calculating the gaps at the X and L points of the BZ in lowest non-zero-order perturbation theory.
(b) Carry out calculations using the empirical pseudopotential program (App. N) and compare with the results from perturbation theory.
(c) Compare with the simple expression for the band width in Exercise 10.13 and with fully self-consistent band structure results as described in Exercise 13.5.

13

Plane waves and grids: full calculations

Summary

The subject of this chapter is the role of plane waves and grids in modern electronic structure calculations, which builds upon the basic formulation of Ch. 12. Plane waves have played an important role from the early OPW calculations to widely used methods involving norm-conserving pseudopotentials. Plane waves continue to be the basis of choice for many new developments, such as quantum molecular dynamics simulations (Ch. 18), owing to the simplicity of operations. Efficient iterative methods (App. M) have made it feasible to apply plane waves to large systems, and recently developed approaches such as "ultrasoft" pseudopotentials and projector augmented waves (PAWs Ch. 11) have made it feasible to apply plane waves to difficult cases such as materials containing transition metals. Real-space grids are an intrinsic part of efficient plane wave calculations and there is a growing development of real-space methods, including multigrids, finite elements, wavelets, *etc.*

Basic Schrödinger-like equations for eigenstates expanded in a plane wave basis can be found in Sec. 12.1 and related equations for real-space grids in Sec. 12.8. These methods are appropriate in cases where the potentials and wavefunctions are smooth. Thus application of these methods to real materials means that they must be combined with a transformation to remove the core states, such as OPWs, pseudopotentials, or PAWs (Ch. 11). Many aspects of pseudopotential calculations have been given in Sec. 12.6. The additional steps that are required for a full self-consistent "*ab initio*" calculation are:

- If the calculation is "*ab initio*", i.e. there are no parameters, then the pseudopotential must be derived from theoretical calculations, usually on an atom, as described in Ch. 11. Such pseudopotentials are "bare potentials" and the total potential is determined. This is one of two essential steps that take the calculation beyond the empirical pseudopotential method of Sec. 12.6.
- The total effective potential in the Kohn–Sham Schrödinger-like equations is a sum of the "bare" ion pseudopotentials and the effective potentials from the valence electrons, the Hartree, and the exchange–correlation potential. This requires that the equations be solved self-consistently as described in general in Ch. 9.

- The primary results in a Kohn–Sham density functional theory are the total energy and related quantities such as forces and stresses, which are sufficient for ground state properties. In addition, there are eigenvalues and eigenvectors that are only approximately related to true excitation energies, as discussed in Chs. 7 and 20.

Grid methods are in fact closely related to plane waves and are discussed in Secs. 12.7 and 12.8. Since the only fundamental difference is the way the kinetic energy is treated, all the methods for smooth functions described here can be translated into real-space grid methods. However, such methods are not well developed at present; we will not reiterate the details and merely refer the reader back to Sec. 12.8 for the general approaches.

13.1 "*Ab initio*" pseudopotential method

Expressions for total energy, force, and stress in Fourier space

The starting point for derivation of the full Kohn–Sham theory is the total energy for which general expressions have been given in Chs. 8 and 9; the subject of this section is the derivation of explicit expressions in reciprocal space. For example, the variational expression for energy (Eqs. (7.5) or (9.7)) in terms of the output wavefunctions and density can be written [104, 413, 530, 561]

$$E_{\text{total}}[V_{\text{eff}}] = \frac{1}{N_k}\sum_{\mathbf{k},i} w_{k,i}\left\{\sum_{m,m'} c^*_{i,m}(\mathbf{k})\left[\frac{\hbar^2}{2m_e}|\mathbf{K}_m|^2\delta_{m,m'} + V_{\text{ext}}(\mathbf{K}_m,\mathbf{K}_{m'})\right]c_{i,m'}(\mathbf{k})\right\}$$
$$+\sum_{\mathbf{G}}\epsilon_{\text{xc}}(\mathbf{G})n(\mathbf{G}) + \frac{1}{2}4\pi e^2\sum_{\mathbf{G}\neq 0}\frac{n(\mathbf{G})^2}{G^2} + \gamma_{\text{Ewald}} + \left(\sum_{\kappa}\alpha_\kappa\right)\frac{N_e}{\Omega}. \quad (13.1)$$

Since E_{total} is the total energy per cell, the average over $\mathbf{k}$ and sum over bands is the same as for the density in (12.27). Similarly, the sums can be reduced to the IBZ just as in (4.44). The potential terms involve $\mathbf{K}_m \equiv \mathbf{k} + \mathbf{G}_m$; the xc term is the total exchange–correlation energy; and the final three terms are considered below. Alternatively, one can use expression Eq. 9.7 for the energy, in which the eigenvalues replace the term in square brackets in Eq. 13.1. As discussed in Ch. 9, the form in (13.1) is manifestly a functional of V_{eff}, which determines each term (except the final two terms that depend only upon the structure and number of electrons).

Correct treatment of the Coulomb terms is accomplished by *consistently* separating out the $\mathbf{G} = 0$ components in the potential and the total energy. The Hartree term in 13.1 is the Coulomb interaction of the electrons with themselves *excluding the divergent term due to the average electron density*. Similarly, *the* $\mathbf{G} = 0$ *Fourier component of the local potential is defined to be zero* in (13.1). Both these terms are included in the Ewald term γ_{Ewald}, which is the energy of point ions in a compensating background (see App. F, Eq. (F.5)), i.e. this term includes the ion–ion terms as well as the interactions of the average electron density with the ions and with itself. *Only by combining the terms together is the expression well defined.* The final term in (13.1) is a contribution due to the non-Coulombic part of the local pseudopotential (see Eq. (12.22)) where $\frac{N_e}{\Omega}$ is the average electron density.

Following the analysis of Sec. 9.2, one can define a functional[1]

$$\tilde{E}_{\rm total} = \frac{1}{N_k}\sum_{\mathbf{k},i} w_{k,i}\varepsilon_i + \sum_{\mathbf{G}} [\epsilon_{\rm xc}(\mathbf{G}) - V_{\rm xc}(\mathbf{G})]\, n(\mathbf{G}) + \left[\gamma_{\rm Ewald} - \frac{1}{2}4\pi e^2 \sum_{\mathbf{G}\neq 0} \frac{n(\mathbf{G})^2}{G^2}\right] + \left(\sum_{\kappa}\alpha_\kappa\right)\frac{N_e}{\Omega}, \tag{13.2}$$

where all terms involve the *input density* $n \equiv n^{\rm in}$. This expression is not variational but instead is a saddle point as a function of $n^{\rm in}$ around the consistent solution $n^{\rm out} = n^{\rm in}$. It is very useful because it often converges faster to the final consistent energy so that it is useful at every step of a self-consistent calculation. Furthermore, it is the basis for useful approximations, e.g. stopping after one step and never evaluating any output quantity other than the eigenvalues [144, 415, 417–419].

The force on any atom $\tau_{\kappa,j}$ can be found straightforwardly from the "force theorem" or "Hellmann–Feynman theorem" given in Sec. 3.3. For this purpose, expression (13.1) is the most useful and the explicit expression for Eq. (3.20) in Fourier components can be written

$$\mathbf{F}_j^\kappa = -\frac{\partial E}{\partial \tau_{\kappa,j}} = -\frac{\partial \gamma_{\rm Ewald}}{\partial \tau_{\kappa,j}} - i\sum_m \mathbf{G}_m e^{i\mathbf{G}_m\cdot\tau_{\kappa,j})} V^\kappa_{\rm local}(\mathbf{G}_m) n(\mathbf{G}_m) \frac{-i}{N_k}\sum_{\mathbf{k},i} w_{k,i}\sum_{m,m'} c^*_{i,m}(\mathbf{k})\left[\mathbf{K}_{m,m'} e^{i(\mathbf{K}_{m,m'}\cdot\tau_{\kappa,j})}\delta V^\kappa_{\rm NL}(\mathbf{K}_m,\mathbf{K}_{m'})\right] c_{i,m'}(\mathbf{k}), \tag{13.3}$$

where the Ewald contribution is given in (F.10). Here the external pseudopotential has been separated into the local part, which contains the long-range terms, and the short-range non-local operator $\delta V^\kappa_{\rm ext}(\mathbf{K}_m, \mathbf{K}_{m'})$, with $\mathbf{K}_{m,m'} \equiv \mathbf{K}_m - \mathbf{K}_{m'}$. The expression for stress in Fourier components is given in Sec. G.3.

Solution of the Kohn–Sham equations

The Kohn–Sham equation is given by (12.9) and (12.10) with the local and non-local parts of the pseudopotential specified by the formulas of Sec. 12.4. Consistent with the definitions above, the local part of the potential in the Kohn–Sham equation can be written straightforwardly as the Fourier transform of the external local potential (12.16), Hartree, and xc potentials in (7.13),

$$V^\sigma_{\rm KS,local}(\mathbf{G}) = V_{\rm local}(\mathbf{G}) + V_{\rm Hartree}(\mathbf{G}) + V^\sigma_{\rm xc}(\mathbf{G}), \tag{13.4}$$

where all $\mathbf{G} = 0$ Fourier components are omitted. The $\mathbf{G} = 0$ term represents the average potential which is only a shift in the zero of energy that has no consequence for the bands, since the zero of energy is arbitrary in an infinite crystal [184, 290, 562]. The full potential is Eq. (13.4) plus the non-local potential Eqs. (12.23) or (12.24).

[1] The electron Coulomb term on the second line cancels the double counting in the eigenvalues. The terms are arranged in two neutral groupings: the difference of the ion and the electron terms in the square bracket and the sum of eigenvalues that are the solution of the Kohn–Sham equation with a neutral potential.

The equations are solved by the self-consistent cycle shown in Fig. 9.1, where the solution of the equations for a fixed potential is the same as for a non-self-consistent EPM calculation. The new steps that must be added are:

- Calculation of the output density $n^{\text{out}}(\mathbf{G})$
- Generation of a new input density $n^{\text{in}}(\mathbf{G})$, which leads to the new effective potential
- After self-consistency is reached, calculation of the total energy (Eqs. (13.1), (13.2), or related variational formulas using the expressions of Sec. 9.2), forces, stress, etc.

Approach to self-consistency

The plane waves framework affords a simple case in which to discuss the approach to self-consistency, bringing out issues addressed in Sec. 9.3. The simplest approach – that works very well in many cases – is linear mixing

$$V_{i+1}^{\sigma,\text{in}}(\mathbf{G}) = \alpha V_i^{\sigma,\text{out}}(\mathbf{G}) + (1-\alpha) V_i^{\sigma,\text{in}}(\mathbf{G}). \tag{13.5}$$

Choice of α by trial-and-error is often sufficient since the same value will apply to many similar systems.

In order to go further and analyze the convergence, one can treat the region near convergence where the error in the output density or potential is proportional to the error in the input potential δV^{in}. Using the definition of the dielectric function, the error in the output potential is given by[2]

$$\delta V^{\text{out}}(\mathbf{G}) = \sum_{\mathbf{G}'} \epsilon(\mathbf{G}, \mathbf{G}') \delta V^{\text{in}}(\mathbf{G}'). \tag{13.6}$$

(Note that this *does not apply to the* $\mathbf{G} = 0$ *component*, which is fixed at zero.) It follows that the error in the output density $\delta n^{\text{out}}(\mathbf{G}) = \delta V^{\text{out}}(\mathbf{G})(G^2/4\pi e^2)$ is also governed by the dielectric function, and the kernel χ in Eq. (9.21) is related by $\chi(\mathbf{G}, \mathbf{G}') = \epsilon(\mathbf{G}, \mathbf{G}')G'^2/G^2$. In general the dielectric function approaches unity for large $\mathbf{G}$ or $\mathbf{G}'$, however, it may be much larger than unity for small wavevectors. For example, for Si, $\epsilon \approx 12$ for small wavevectors, so that *the error in the output potential (or density) is 12 times larger than the error in the input!* For a metal, the problem is worse since ϵ diverges.

How can the iterations reach the solution? There are two answers. First, for crystals with small unit cells, this is *not a problem* because all the $\mathbf{G} \neq 0$ components of the potential are for large values of $|\mathbf{G}|$, and the $\mathbf{G} \equiv 0$ is taken care of in combination with the Ewald term (see App. F). It is only if one has small non-zero components that problems arise. This happens for large unit cells and is called the "charge sloshing problem." It is worst for cases like metal surfaces where the charge can "slosh" to and from the surface with essentially no cost in energy. In such cases the change is in the right direction but one must take only small steps in that direction. If a linear mixing formula is used, then the mixing of the output must be less than $1/\epsilon(G_{\text{min}})$ for convergence.

[2] Note the similarity to the Thomas–Fermi expression, (12.25). The reason that we have ϵ here instead of ϵ^{-1} is that here we are considering the response to an internal field.

The relation to the dielectric function also suggests an improved way to reach convergence. It follows from (13.6) that the exact potential can be reached after one step (see also Eqs. (9.21) and (9.23)) by solving the equation

$$\delta V^{\text{out}}(\mathbf{G}) \equiv V^{\text{out}}(\mathbf{G}) - V_{\text{KS}}(\mathbf{G}) = \sum_{\mathbf{G}'} \epsilon(\mathbf{G}, \mathbf{G}') \left[V^{\text{in}}(\mathbf{G}') - V_{\text{KS}}(\mathbf{G}') \right] \tag{13.7}$$

for the converged Kohn–Sham potential $V_{\text{KS}}(\mathbf{G})$. The input and output potentials are known from the calculation; however, the problem is that it is very difficult to find $\epsilon(\mathbf{G}, \mathbf{G}')$ – a more difficult problem than solving the equations. Nevertheless, approximate forms for $\epsilon(\mathbf{G}, \mathbf{G}')$ such as the diagonal Thomas–Fermi form, Eq. (12.26), can be used to improve the convergence [428]. One can also take advantage of the fact that a linear mixing leads to exponential convergence (or divergence) near the solution; by fitting three points to an exponential, the solution for an infinite number of steps can be predicted [563].

From a numerical point of view the dielectric matrix (or the second derivatives defined in Ch. 9) are nothing more than the Jacobian. Since it is in general not known, approximations (such as approximate dielectric functions) are really "preconditioners" as discussed in the chapter on iterative methods and in App. M. This leads to the practical approach now widely used: for high Fourier components the dielectric function is near unity and nearly diagonal, so one can use (13.5) with $\alpha \approx 0.5$ to 1; for low Fourier components one can use general numerical approaches to build up the Jacobian iteratively as the calculations proceed, e.g. the Broyden-type methods described in Sec. 9.3.

13.2 Projector augmented waves (PAWs)

The projector augmented wave (PAW) method [475] described in Sec. 11.11 is analogous to pseudopotentials in that it introduces projectors acting on smooth valence functions $\tilde{\psi}^v$ that are the primary objects in the calculation. It also introduces auxiliary localized functions like the "ultrasoft" pseudopotential method. However, the localized functions actually keep all the information on the core states like the OPW and APW (see Ch. 16) methods. Thus many aspects of the calculations are identical to pseudopotential calculations; e.g. all the operations on smooth functions with FFTs, generation of the smooth density, *etc.*, are the same. However, since the localized functions are rapidly varying, augmentation regions around each nucleus (like the muffin-tin spheres in Ch. 16) are introduced and integrals within each sphere are done in spherical coordinates.

The expressions given in Sec. 11.11 apply here also. The linear transformation to the all-electron valence functions $\psi^v = \mathcal{T}\tilde{\psi}^v$ is assumed to be a sum of non-overlapping atom-centered contributions $\mathcal{T} = \mathbf{1} + \sum_{\mathbf{R}} \mathcal{T}_{\mathbf{R}}$, each localized to a sphere denoted Ω_{vecr}. If the smooth wavefunction is expanded in spherical harmonics inside each sphere, omitting the labels v and i as in (11.58),

$$|\tilde{\psi}\rangle = \sum_m c_m |\tilde{\psi}_m\rangle, \tag{13.8}$$

with the corresponding all-electron function,

$$|\psi\rangle = \mathcal{T}|\tilde{\psi}\rangle = \sum_m c_m |\psi_m\rangle. \tag{13.9}$$

Hence the full wavefunction in all space can be written

$$|\psi\rangle = |\tilde{\psi}\rangle + \sum_{\mathbf{R}m} c_{\mathbf{R}m} \left\{ |\psi_{\mathbf{R}m}\rangle - |\tilde{\psi}_{\mathbf{R}m}\rangle \right\}. \tag{13.10}$$

The biorthogonal projectors $\langle \tilde{p}_{\mathbf{R}m}|$ in each sphere are the same as defined in (11.62) since the spheres are non-overlapping.

Thus the expressions carry over with the generalization to many spheres, for example, the density given by Eqs. (11.66)–(11.69). Here it is particularly relevant to give the form for the total energy, from which follow the basic Kohn–Sham equations and expressions for forces, *etc.* [475, 476]. Like the density, the energy can be written as a sum of three terms,

$$E_{\text{total}} = \tilde{E}_{\text{total}} + E^1_{\text{total}} + \tilde{E}^1_{\text{total}}, \tag{13.11}$$

where $\tilde{E}$ denotes the energy due to the smooth functions evaluated in Fourier space or a grid that extends throughout space, $\tilde{E}^1$ denotes the same terms evaluated only in the spheres on radial grids, and E^1 the energy in the spheres with the full functions. The classical Coulomb terms are straightforward in the sense that they are given directly by the density; however, they can be rearranged in different ways to improve convergence of the Coulomb sums. In the PAW approach, an additional density is added to both auxiliary densities in $\tilde{n}(\mathbf{r})$ and $\tilde{n}^1(\mathbf{r})$ so that the multi-pole moments of the terms $n^1(\mathbf{r}) - \tilde{n}^1(\mathbf{r})$ in (11.66) vanish. Thus the electrostatic potential due to these terms vanishes outside the augmentation spheres around each atom, just as is accomplished in full-potential LAPW methods [564]. A discussion of different techniques for the additional density terms [475, 476] is given in [476]. The expression for E_{xc} also divides into three terms with each involving the total density evaluated in the different regions [476]. It is not hard to derive the expressions for the Kohn–Sham equations by functional derivatives of the total energy and a detailed account can be found in [475].

It is advantageous that expressions for the total energy are closely related in the ultrasoft and the PAW formulations, differing only in the choice of auxiliary functions and technical aspects. Thus the expressions for forces and stress are also essentially the same. In particular, the large intra-atomic terms do not enter the derivatives and forces can be derived by derivatives of the structure constants [565]. Stress can also be derived as referred to in [476]. Computer programs for atomic calculations needed to construct the PAW basis functions, as well as PAW calculations on solids using plane waves, are available on-line (see Ch. 24) and are described in papers by Holzwarth and coworkers [559, 560].

13.3 Simple crystals: structures, bands, . . .

Plane waves are the method of choice for Kohn–Sham calculations in important classes of problems. The ideal examples are crystals with small primitive cells and atoms accurately represented by pseudopotentials. If the cell in real space is small, then a relatively small set of plane waves $\mathbf{k} + \mathbf{G}_m$ is an effective basis that takes advantage of the periodicity. There are no shape approximations and they are applicable to open structures with no extra effort. Operations are simple and it is straightforward to calculate total energy, electron density, forces, stresses, *etc.* Norm-conserving pseudopotentials (Sec. 11.9) provide a direct way

to extend the approach to *any* problem if enough electrons are included in the valence shells and enough plane waves are used. Ultrasoft pseudopotentials and the PAW method (Secs. 11.10, 11.11, and 13.2) permit accurate calculations with a smaller set of plane waves.

Because plane waves are so pervasive, many applications are given in other chapters. Examples include the calculations of total energies and phase transition pressures for Si in Figs. 2.4 and 2.5; the first stress calculations illustrated by Fig. 2.6; bands in C_{60} solids in Fig. 2.19; Wannier functions and resulting bands shown in Figs. 21.4 and 21.6; and "frozen phonons" illustrated in Fig. 2.8.

Among the most extensive and important types of calculations are total energies calculated by the *ab initio* pseudopotential method to determine the equilibrium structures of a crystal under pressure, pioneered in the work of Yin and Cohen [103, 561]. Figure 2.5 compares LDA calculations with experiment for transition pressures from the tetrahedrally coordinated diamond or zinc-blende structures to higher coordinated metallic or NaCl structures. As emphasized in Sec. 2.2, the results are in remarkable agreement with experiment, which is also found for a large range of materials, including semiconductors, sp-bonded metals, the chalcogenides, etc. The *ab initio* pseudopotential method using plane waves has played a key role in the theoretical development of electronic structure owing to the ease with which different structures can be studied, including low-symmetry ones. It should be emphasized that other full-potential methods can be applied to this problem, and in general, these theoretical methods agree well.

Comparison of methods

It is essential to establish the level of accuracy of the different methods and to demonstrate that they agree when done carefully. For this, one need use only one functional, namely the local density approximation. Comparison with experimental results is another matter, as these results are often improved with GGA and other functionals. Table 13.1 shows the calculated lattice constants a and bulk moduli B for selected crystals, as well as the magnetic moment of bcc Fe. The agreement between methods is excellent except for CaF_2, where there is large core polarization not taken into account in the present implementation of PAW that assumes rigid cores. The comparison with experiment for Fe is greatly improved with GGA functionals [566]. Similar tests [476] on molecules and solids, ranging from the first row elements to transition metals, show similar accuracy; this confirms the applicability and accuracy of the different methods when treated carefully.

Stability of various phases

As stressed in Sec. 2.2 the stability of phases as a function of volume or pressure is perhaps the most fundamental thermodynamic property of matter. Examples were given of pressure-induced phase transformations of open-structure materials, such as diamond and silicon, to other structures including more close-packed metals. Figure 13.1 shows two examples of calculations of energy versus volume that illustrate different points. On the left are energies of nitrogen in the molecular α-phase and various possible high-pressure non-molecular

Table 13.1. Calculated properties of selected crystals using the local density approximation and various methods that involve plane waves: norm-conserving pseudopotentials (NCPPs), projector augmented waves (PAWs), "ultrasoft" pseudopotentials (USPPs) and linearized APWs (Ch. 17).

	C		Si		CaF_2		bcc Fe		
Method	*a*	*B*	*a*	*B*	*a*	*B*	*a*	*B*	*m*
NCPP[a]	3.54	460	5.39	98	5.21	90	2.75[c]	226[c]	
PAW[a]	3.54	460	5.38	98	5.34	100			
PAW[b]	3.54	460	5.40	95	5.34	101	2.75	247	2.00
USPP[b]	3.54	461	5.40	95	5.34	101	2.72	237	2.08
LAPW[a]	3.54	470	5.41	98	5.33	110	2.72[d]	245[d]	2.04[d]
EXP[a]	3.56	443	5.43	99	5.45	85–90	2.87[d]	172[d]	2.12[d]

Since the results depend upon many details of the calculations, the values shown are mainly from two references that carried out careful comparisons: [a] Holzwarth, et al. [512] and [b] Kresse and Joubert [476]. Other values for Fe are from [c] Cho and Scheffler [568] and [d] Stixrude, et al. [566]. References for experimental values are cited in [512] and [566].

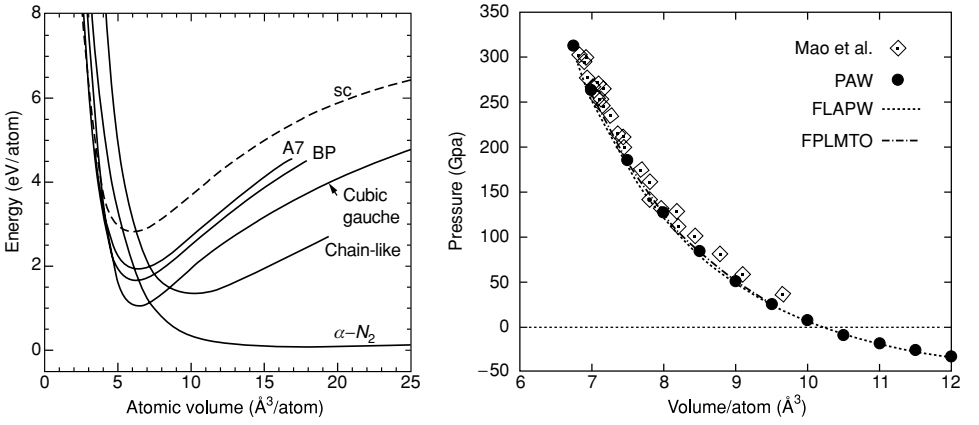


Figure 13.1. Left: Total energy of nitrogen versus volume in molecular structures and the "cubic gauche" non-molecular structure predicted at high pressure [121]. Right: Energy versus volume for hcp Fe showing agreement of calculations performed using ultrasoft pseudopotentials, the PAW method, and full-potential LMTO and LAPW calculations [164].

structures calculated [121] using the LDA and norm-conserving pseudopotentials. An outstanding aspect of this work is the creativity of the authors in finding a structure called "cubic gauche" (CG) that was not known before for any material, but which was predicted to be significantly more stable than any other structure suggested previously. This is an example of the theoretical prediction of a new polymeric phase of N: there is no experimental confirmation of this phase as yet, but it may be consistent with recent high-pressure experiments [567].

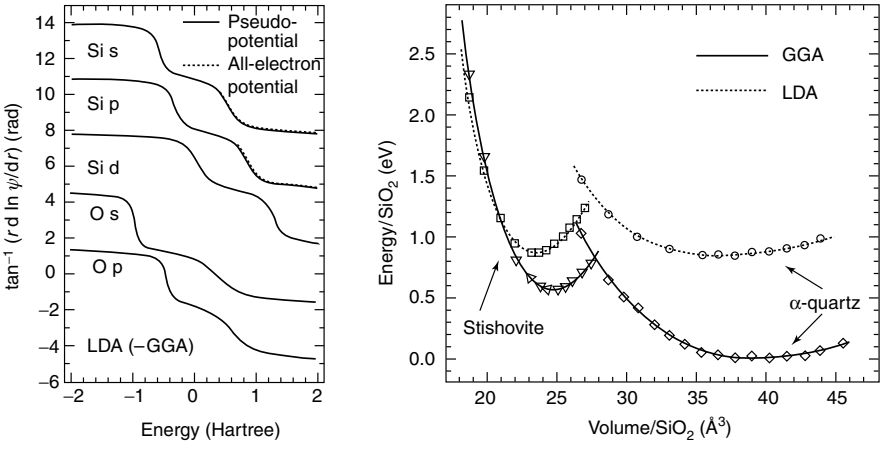


Figure 13.2. Right: Total energy of SiO_2 in α-quartz and in the known high-pressure stishovite phase calculated using the LDA and PW91 GGA [569] (Sec. 8.2). With the LDA, stishovite is found to be the most stable form with the lowest energy, whereas the GGA leads to the collect result that α-quartz has lower energy and stishovite is stabilized under pressure. The quality of the pseudopotential is shown by comparion of the pseudopotential phase shifts $\tan^{-1}$ with those from the all-electron calculation (offset by multiples of 2π for clarity). From Hamann [380].

On the right-hand side of Fig. 13.1, the equation of state of hcp Fe at high pressure is shown. This is the pressure range relevant for understanding the behavior of Fe in the core of the Earth. The curves show the theoretical agreement of calculations made using ultrasoft pseudopotentials, the PAW method, and full-potential LMTO and LAPW calculations [164]. This in agreement with the results in Tab. 13.1 and shows that, with care, the approaches can be used under extreme conditions. An important point for our purpose, is that the same PAW methods have been used in full-thermal simulations of liquid Fe under Earth-core conditions, as described in Sec. 18.6.

Another example of the stability of crystal phases as a function of pressure is demonstrated by SiO_2, which has various low-pressure polymorphs, including α-quartz structure and silica glass, that consists of corner-sharing SiO_4 tetrahedra. A high-pressure phase, stishovite, has tetragonal rutile structure with octahedrally coordinated SiO effectively in three-fold coordination. Figure 13.2 shows the energy versus volume for the two structures and for two functionals: the LDA and the PW91 GGA [569] (Sec. 8.2) calculated by Hamann [380]. As is found in other cases, the LDA favors the dense phase: in this case indicating that α-quartz and silica (window) glass would be unstable at ordinary pressure! The GGA favors more open structures and leads to proper ordering of the structures, with the energy difference in at least qualitative agreement with experiment.

The calculations for SiO_2 also illustrate other points. The left-hand side of Fig. 13.2 shows the agreement of the all-election and pseudopotential atomic phase shifts over the energy range of the valence electrons (see also Sec. 11.4). The calculations were done using adaptive coordinates [555, 570], illustrated in Fig. 12.6, in order to accommodate hard pseudopotentials needed for accurate calculations involving oxygen. Finally, careful

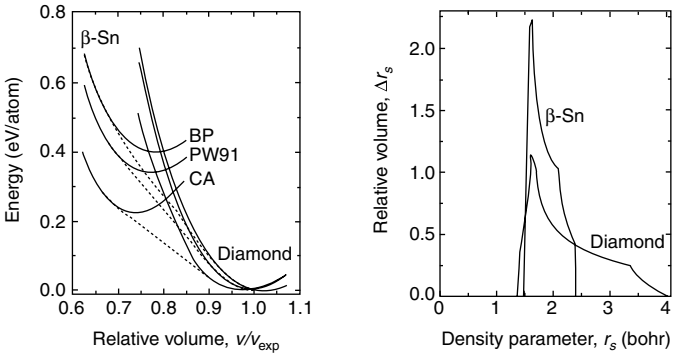


Figure 13.3. Left: Calculated energies for Si comparing different functionals [127]. The LDA is denoted as "CA" and other functionals are described in Ch. 8: Perdew–Wang 91 (PW91) and Becke–Parr (BP). The energies are arbitrarily set equal for the diamond structure. In fact, the GGA functionals lead to lower total energies compared to LDA. Transition pressures are increased for the GGA functionals compared to the LDA because lowering of the total energy is more pronounced for the open diamond structure (which has larger gradients) than for the more close-packed β-Sn structure. Right: Density of points in the unit cell at which the valence electron density has a value $n = 1/(\frac{4\pi}{3}r_s^3)$ plotted versus r_s (see text for description). From [127].

calculation with the all-electron LAPW method (Ch. 17) found energy differences between the two phases that agree to within ≈0.02 eV, only barely visible on the scale of Fig. 13.2.

Comparison of functionals: phase transition in Si

The most studied element is Si for which there have been calculations on an extensive set of structures [123] including diamond structures, various compact distorted tetrahedral structures, and many metallic structures. This illustrates the great advantage of the plane wave pseudopotential method that can treat many structures in an unbiased way. Silicon is also a test case where changes in transition pressure for different functionals have been calculated [127]. Figure 13.3 shows a comparison of LDA (denoted "CA") and typical GGA functionals (see Ch. 8). The transition pressure changes from 8.0 GPa for the LDA to 12.2 GPa for the Perdew–Wang 91 (PW91) and 14.6 GPa for the Becke–Parr (BP) functionals, which can be compared with the reported experimental pressure of ≈12.5 GPa [120].

The right-hand side of Fig. 13.3 is an illuminating plot that reveals several aspects of Si in diamond and high-pressure β-Sn structures. The plot shows the density of points in the unit cell at which the valence electron density (the core density is not included) has the value $n = 1/(\frac{4\pi}{3}r_s^3)$. The horizontal range shows the maximum and minimum values of the local r_s parameter and the vertical axis shows the normalized number of points at each value of r_s. Since the density is periodic in the cell, this plot has the same critical point structure as the density of states discussed in Sec. 4.7 and illustrated in Fig. 14.3. Clearly, the high-pressure phase has only a small range of r_s so that it is nearly a homogeneous gas. The diamond structure has a larger range and thus larger gradients and larger effects of GGA. On the left-hand side of Fig. 13.3 the energies are arbitrarily set equal for the diamond

structure (since only relative energies matter for the transition) but, in fact, the GGAs *lower all the energies, with the diamond structure lowered most* compared to the LDA.

Electronic bands

Among the enumerable calculations of bands, it is appropriate to consider only a few examples that illustrate characteristic aspects of pseudopotentials and plane waves (or grids), careful comparisons to other methods, and salient points about the bands in density functional theory. As a first point, it should be emphasized that all calculational techniques, when carried out carefully, yield essentially the same results for the bands.

Many calculations have shown that the dispersion in both the occupied and unoccupied bands of many materials are well predicted by LDA and GGA functionals. However, the great failure of these functionals is that the band gap between the occupied and unoccupied bands is too small. An extreme example is Ge which is predicted to be a metal in the LDA, a result first found using the all-electron relativistic LMTO approach [571]. These effects are illustrated for Ge and GaAs in Figs. 2.25, 17.9, and 17.8, all of which show that the valence bands are in excellent agreement with experiment but the conduction bands are shifted downward. These examples show that accurate results can be found using carefully constructed pseudopotentials with plane waves or with gaussians. However, they also show that care must be taken in the generation and use of pseudopotentials to capture all the relevant effects.

To show the power of plane wave calculations in problems that are very non-plane-wave-like, Fig. 13.4 shows the electronic bands of the perovskite structure (Fig. 4.8) ferroelectric

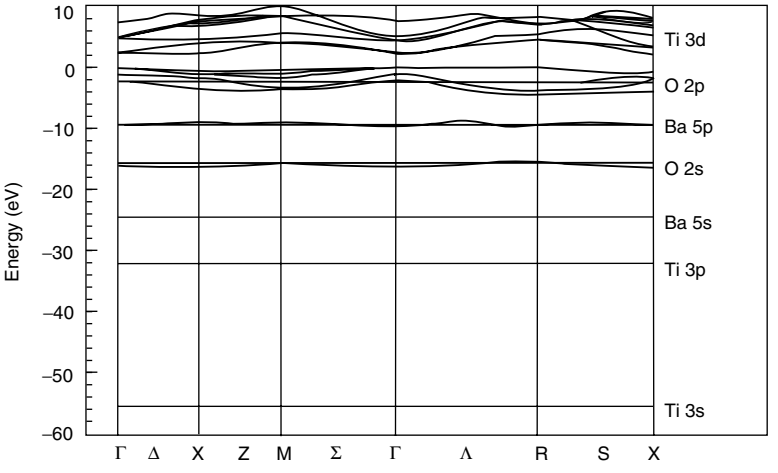


Figure 13.4. Example of $BaTiO_3$ bands calculated [572] using plane waves and the "band-by-band" minimization method (Sec. M.8). This is an ionic system with narrow bands each derived primarily from one atomic-like state that is identified in the figure. Calculation of the Wannier functions for each of the separated groups of bands has been used to analyze the origin of the anomalous effective charges (Tab. 22.1) and understanding of the ferroelectric moment.

material $BaTiO_3$, as calculated [572] using plane waves and the "band-by-band" minimization method (Sec. M.8). This is an example of a very ionic material with narrow bands each derived primarily from one atomic-like state that is identified in the figure. Yet there are crucial effects due to the fact that the system is not completely ionic. Because there is incomplete charge transfer, i.e. the bands have mixed atomic character, there are large redistributions of charge as the atoms move. This effect of hybridization (or covalency), especially involving the Ti 3d and O 2p states, leads to the extremely large anomalous effective charges shown in Tab. 22.1 and to the quantitative theory of ferroelectric polarization [572, 573]. As a practical matter, it essential, for accurate results, that the "semi-core states" are treated as valence states (in this case Ti 3p and 3s have similar radial extent to the 3d state). Thus the Ti pseudopotential only eliminates the Ti 1s, 2s, and 2p core states.

A major challenge is to go beyond the simple LDA and GGA functionals. As discussed in Secs. 8.7 and 8.8, orbital-dependent functionals include crucial effects such as the "band gap discontinuity;" however, these functionals are much more difficult to use. Because of their simplicity, pseudopotentials and plane waves are the methods of choice for exploratory calculations to test methods and new ideas. One example is the calculation of bands using the non-local exact exchange (EXX) functional. There is a marked improvement in the band gaps, as shown in Fig. 2.26, without affecting detrimentally the good agreement for total energies [223]. In particular, Ge is predicted to be a semiconductor with a gap in very good agreement with experiment. In finite systems – atoms, molecules and clusters – grids are particularly appropriate for calculations of excitations [526] and examples of time-dependent density functional theory calculations are illustrated in Ch. 20.

13.4 Supercells: surfaces, interfaces, phonons, defects

Plane waves also turn out to be the method of choice for a very different set of problems: surfaces, interfaces, phonons, defects, *etc.* Despite the fact that these are all cases in which there is no periodicity in at least one direction, plane waves are nevertheless an effective basis. If one adopts the simple brute-force method of forming a "supercell," then the problem is made artificially periodic and all the usual plane wave methods apply. Despite the obvious disadvantage that many plane waves are required, the combination of efficient iterative methods (App. M) and powerful computers makes this an extremely effective approach. From a more fundamental point of view, the variation in calculated properties with the size of the "supercell" is an example of finite-size-scaling, which is a powerful method to extrapolate to the large-system limit. This is the opposite of the approach in Ch. 23 to treat localized properties; each has advantages and disadvantages.

Schematic examples of supercells in one dimension are shown in Fig. 13.5. The figure shows an example of perfect crystal disturbance (a phonon displacement), a superlattice formed by different atoms, and a slab with vacuum on either side. In fact, the periodic cell made by repeating the cell shown is possibly a real physical problem of interest in its own right. In addition, the limit of infinite horizontal cell size can be used in various ways to represent the physical limit of an isolated plane of displaced atoms, an isolated interface, and a surface. Although each case may be better treated by some other method, the fact

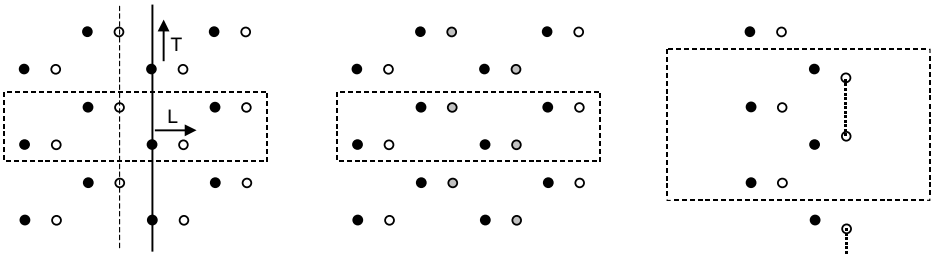


Figure 13.5. Examples of the use of "supercells" for the case of zinc-blende crystals with the long axis in the [1 0 0] direction and the vertical axis in the [1 1 0] direction. The unit cell is shown by the dashed boundary. Left: Perfect crystal with one plane of atoms displaced; as described in Sec. 19.2, calculation of forces on other planes, e.g. the one shown by the dashed line, allows calculation of longitudinal (L) or transverse (T) phonon dispersion curves in the [1 0 0] direction. Middle: interface between two crystals, e.g. GaAs–AlAs, which allows calculation of band offsets, interface states, *etc.* Right: Two surfaces of a slab created by removing atoms, leaving a vacuum spacer. Complicated reconstructions often lead to increased periodicity in the surface and larger unit cells as indicated.

that plane waves can treat all cases and can be used to describe the limits accurately is an enormous advantage.

"Frozen" phonons and elastic distortions

An example of the use of supercells in calculating phonon properties is given in Sec. 19.2. Any particular cell can be used to calculate all the vibrational properties at the discrete wavevectors corresponding to the reciprocal lattice vectors of the supercell. This is very useful for many problems, and a great advantage is that non-linear, anharmonic effects can be treated with the same method (see, e.g. Fig. 2.8). Furthermore, it is possible [574–576] to derive full dispersion curves from the direct force calculations on each of the atoms in a supercell, like those shown on the left-hand side of Fig. 13.5. The requirement is that the cell extended to twice the range of the forces, which can be accomplished. An example of phonon dispersion curves in GaAs in the [100] direction is shown in Fig. 19.2. The inverse dielectric constant ϵ^{-1} and the effective charges Z_I^* can also be calculated from the change in the potentials due to an induced dipole layer in the supercell [574] or by finite wavevector analysis [577].

Interfaces

The presence of interfaces between bulk materials is one on the major problems of materials science [182]. Interfaces determine the strength of real materials, electrical properties, *etc.* Perhaps the most perfect interfaces are those grown with atomic precision in semiconductors. Four-fold coordination can be maintained across the interface with essentially no defects. The change is just that due to chemical identity of the atoms and relaxations of the structure. Therefore this is a case where detailed comparison can be made between theory

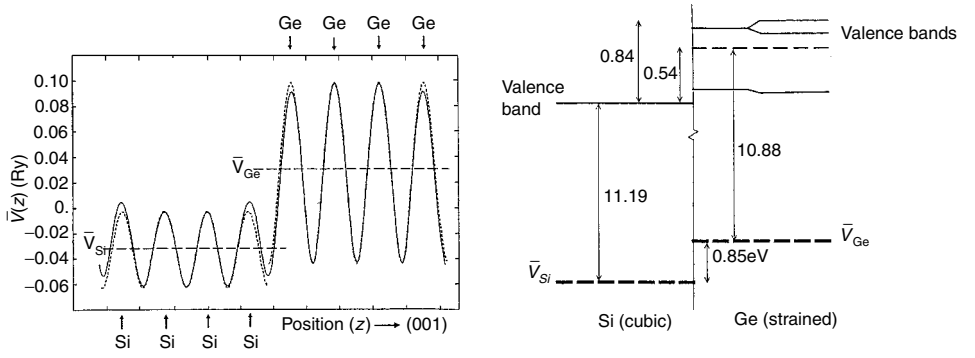


Figure 13.6. Average potential and band alignment at a strained layer interface Ge grown on a Si substrate. The left panel shows the average potential offset determined by the interface calculation. The right panel shows the alignment of the bands relative to the average potential fixed by a bulk calculation performed using the same potential. The bands shown are the top of the valence band (omitting spin–orbit splitting), which is three-fold degenerate in cubic Si but split in the strained Ge. Growth on a Ge substrate leads to strained Si, and intermediate cases and alloys are expected to be described by interpolation.

and experiment. Extensive calculations have been done for many combinations of semiconductors, e.g. using plane waves [183, 578, 579] and using the LMTO method [580, 581]

A representative example [578] of Si/Ge is shown in Fig. 13.6. In this case, there is a uniaxial strain induced by a lattice matching condition in addition to the change in chemical identity. The parallel lattice constant is fixed by the substrate, creating a uniaxial strain in the thin layer. The figure illustrates the results for strained Ge grown on a Si substrate. The left panel shows the planar averaged potential,[3] which quickly reaches the periodic bulk potential, with an offset determined by the calculations. The relative alignment of the Si and Ge bands (the band "offset") is fixed by referring the bulk of the bands of each material to the average potentials indicated. The dilation leads to a shift of the average potential and shifts of all the bands, and the uniaxial part of the strain leads to splitting of the top of the valence bands, as shown on the right-hand side of the figure. Similarly, growth on a Ge substrate leads to strained Si. Intermediate cases and alloys can be treated by interpolation.

For polar interfaces, simple electron counting [176, 582] shows that there can be filled bands only if the stoichiometry is satisfied, which may require mixing of atoms in the interface layer(s). Furthermore, the band offset can be changed by dipoles due to arrangement of the atoms as has been demonstrated for ZnSe/Ge/GaAs (1 0 0) interfaces [583]. This is in fact the same physics as the change in work function of a metal due to surface effects, and illustrates the fact that the absolute value of the potential in a crystal is *not an intrinsic property*; instead, it is fixed by the boundary conditions and/or external fields.

[3] The potential shown in Fig. 13.6 is the local pseudopotential; if done correctly, the final results are invariant to choice of pseudopotential or all-electron potential since the alignment is due to an intrinsic part plus terms that depend only on the interface dipole (App. F).

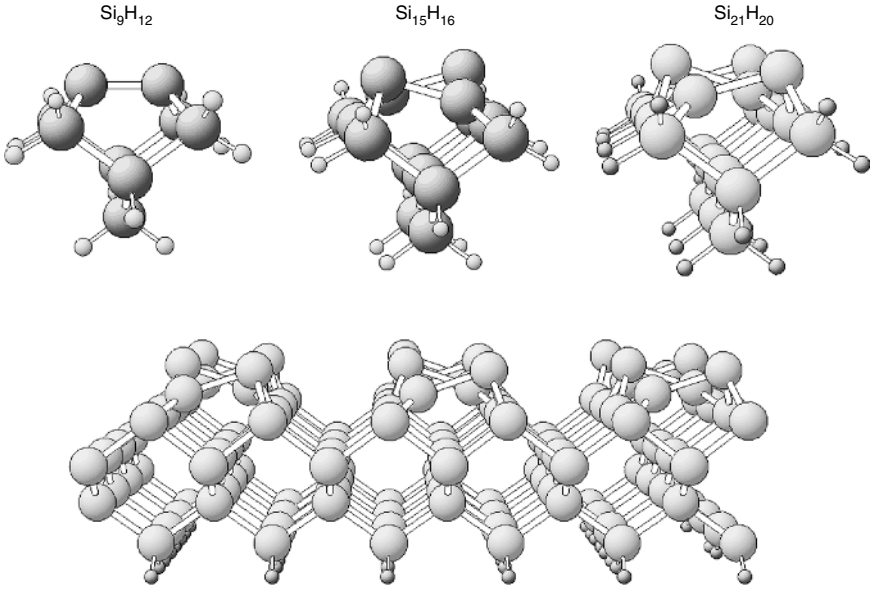


Figure 13.7. Models of the Si(1 0 0) surface including clusters of different size and the slab geometry shown at the bottom, treated by plane wave calculations [585]. The issue is whether or not the Si dimers buckle as shown in the figure; this has not been firmly established experimentally. In each case, the atoms at the top are allowed to relax to model the surface and search for dimerization. Other "surface" atoms are held in bulk positions and are passivated with hydrogen to tie off dangling bonds. Small clusters are often used because traditional quantum-chemical many-body methods are feasible only for small systems. However, the calculation of [585] leads to the conclusion that large clusters or a slab is required for this delicate problem. From [585].

Surfaces

Examples of surface structures determind by plane wave total energy calculations were given in Sec. 2.8 for ZnSe, which illustrating the many issues that occur in polar semiconductors. In particular, surface stoichiometry must be considered as a variable, which means that the surfaces are governed by the grand potential [179, 180], Eq. (2.7). The results of [180], shown in Fig. 2.15, predict a sequence of different reconstructions and changes in stoichiometry as a function of the chemical potential for either of the elements.

Elemental solids are simpler but nevertheless can exhibit an array of surface reconstructions. Calculated absolute surface energies for various surfaces of C, Si, and Ge are reported in [584] which gives many earlier references. A well-known example of surface reconstruction is the (1 0 0) surface of Si and Ge. As indicated in Fig. 13.7, each surface atom has broken bonds and the surface reconstructs with atoms dimerizing to form a new bond so that each surface atom has three bonds. This leaves one electron per surface atom: for a symmetric unbuckled dimer, the two extra electrons can make a π bond, whereas if the dimer buckles so that the two atoms have inequivalent positions, the two electrons can form a "lone pair" in the lower energy state. However, it is not firmly settled whether or

not the buckled dimer is the lowest energy state. Density functional theory (DFT) calculations using LDA and various GGAs have been done by many groups in attempts to resolve the structure. The effects are sufficiently subtle that the results can depend upon whether the calculations are done using finite clusters or slabs. Particularly extensive studies have been reported in [585] and [586], which give earlier references. Figure 13.7 shows geometries that represent different approaches varying from small clusters (hydrogen terminated everywhere except in the desired surface region) to a slab. All were treated with plane waves [585], illustrating the ability of plane waves to treat the different geometries in an unbiased way. The conclusion is that electronic interactions between adjacent Si dimers in a row are essential in stabilizing the buckled ground state, so that large clusters must be used to model the Si (1 0 0) surface adequately. Both density functional theory [585] and quantum Monte Carlo calculations [586] find the lowest energy state to be the buckled dimer.

The calculations also predict the electronic structure that can be measured experimentally. For example, surface bands for Ge [587] are shown in Fig. 15.2, comparing GW quasiparticle and LDA bands. That work uses gaussians, but essentially the same results are found with plane waves.

13.5 Clusters and molecules

Finite systems such as clusters and molecules can be studied conveniently using either plane waves or grids. Grids are an obvious choice for localized functions since only that part of the grid where the functions are non-zero need be considered. In the case of plane waves, it is essential to construct a supercell in each dimension in which the system is localized. Since the supercell must be large enough so that any spurious interactions with the images are negligible, this means that many plane waves must often be used; nevertheless, this may still be an effective way of solving the problem.

Calculations that employ grids are featured in Ch. 20 on excitations, where finite difference methods [526,535] have been used effectively for total energy minimization and time-dependent DFT studies of finite systems from atoms to clusters of hundreds of atoms [359]. Multigrid methods have been developed, with results illustrated in Fig. 2.21 for a boron–carbon nanotube junction.

As an example of plane wave calculations, Fig. 13.8 from [206] shows the bands for a small-diameter carbon nanotube. Nanotubes are described in more detail in Sec. 14.7 where they are considered as an excellent example where tight-binding is the natural description. Yet plane waves provide essential results and insights. The example in Fig. 13.8 is for a tube that would be an insulator if it were simply "rolled graphite" and, indeed, that is the result of the simplest tight-binding models. However, for small tubes there can be large changes. The band labeled (a) was discovered in plane wave calculations [206] to be pushed down to cross the Fermi energy and create a metallic state. This is due to curvature and resulting mixing of states, which is shown on the right-hand side of the figure by the charge density for this band. The fact that the density is higher on the outside than the inside shows that the wavefunction is not simply derived from the π state of graphite, which would yield equal

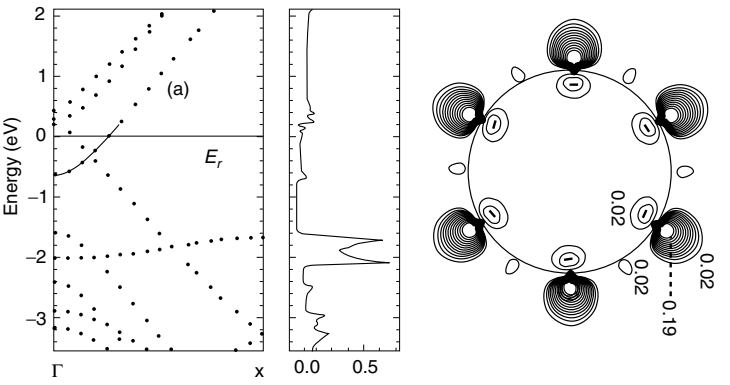


Figure 13.8. Electronic states of small nanotubes are significantly modified from simple "rolled graphite." Left: Bands for a (6,0) "zig-zag" tube (see Sec. 14.7) calculated with plane waves [206]. The fact that graphene-like bands are strongly mixed, causes one band to be lowered below the Fermi level to make the tube metallic. Right: density of the lowered band, which is primarily on the outside, not symmetric as would be the case if it were radial graphene-like p states.

density inside and out. The same effect is captured using other methods [588] and improved non-orthogonal tight-binding, as discussed in Sec. 14.7.

SELECT FURTHER READING

General references are given in Ch. 12. For recent developments with plane waves:

Denteneer, P. J. H. and van Haaringen, W., "The pseudopotential-density functional method in momentum space: details and test cases," *J. Phys. C* 18:4127, 1985. [581]

Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A. and Joannopoulos, J. D., "Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients," *Rev. Mod. Phys.* 64:1045–1097, 1992.

Pickett, W. E., "Pseudopotential methods in condensed matter applications," *Computer Physics Reports* 9:115, 1989.

Singh, D. J., *Planewaves, Pseudopotentials, and the APW Method*, Kluwer Academic Publishers, Boston, 1994, and references therein.

Srivastava, G. P. and Weaire, D., "The theory of the cohesive energy of solids," *Advances in Physics* 36:463–517, 1987.

PAW method:

Blöchl, P. E., "Projector augmented-wave method," *Phys. Rev. B* 50:17953–17979, 1994.

Holzwarth, N. A. W., Tackett, A. R. and Matthews, G. E., "A projector augmented wave (PAW) code for electronic structure calculations, part I: atompaw for generating atom-centered functions," *Comp. Phys. Commun.* 135:329–347, 2001.

Holzwarth, N. A. W., Tackett, A. R. and Matthews, G. E., "A projector augmented wave (PAW) code for electronic structure calculations, part II: pwpaw for periodic solids in a plane wave basis," *Comp. Phys. Commun.* 135:348–376, 2001.

Kresse, G. and Joubert, D., "From ultrasoft pseudopotentials to the projector augmented-wave method," *Phys. Rev. B* 59:1758–1775, 1999.

Grid methods:

Beck, T. L., "Real-space mesh techniques in density-functional theory," *Rev. Mod. Phys.* 72:1041–1080, 200.

Also see other references cited in Ch. 12.

Exercises

13.1 The on-line site in Ch. 24 has links to planewave codes and many examples and tutorials for calculations on real materials. These codes can be used in the calculations for metallic H in Ex. 13.5

13.2 Show that the Eq. (9.7) leads to the expression Eq. (13.2) written in Fourier components. In particular, show that the groupings of terms lead to two well-defined neutral groupings: the difference of the ion and the electron terms in the square bracket and the sum of eigenvalues that are the solution of a the Kohn–Sham equation with a neutral potential.

13.3 Derive the result that the α parameter in the linear mixing scheme (13.5) must be less than $1/\epsilon(G_{\min})$ for convergence. Show that this is a specific form of the general equations in Sec. 9.3 and is closely related to Exercise 9.8. In this case it is assumed that $\epsilon_{\max}$ occurs for $G = G_{\min}$. Discuss the validity of this assumption. Justify it in the difficult extreme case of a metal surface as discussed in Sec. 13.1.

13.4 Show that the density distribution of periodic quantities in real space, e.g. the distribution of local values of $r_s \propto n(\mathbf{r})^{-1/3}$ shown in Fig. 13.3 have the same types of critical points as the distributions, such as the density of states, which is a density in $\mathbf{k}$-space. Hint: The basic arguments hold for any crystal represented in either space.

13.5 This exercise is to calculate the band structure of metallic H at high density ($r_s = 1$ is a good choice) in the fcc structure and to compare with (1) the free-electron bands expected for that density and (2) bands calculated with the Thomas–Fermi approximation for the potential (Exercise 12.13). Use the Coulomb potential for the proton and investigate the number of plane waves required for convergence. (There is no need to use a pseudopotential at high density, since a feasible number of planes is sufficient.) Comparison with the results of Exercise 12.13 can be done either by comparing the gaps at the X and L points of the BZ in lowest non-zero-order perturbation theory, or by carrying out the full band calculation with the Thomas–Fermi potential.

14

Localized orbitals: tight-binding

Summary

Localized functions afford a satisfying description of electronic structure and bonding in an intuitive localized picture. They are widely used in chemistry and have been revived in recent years in physics for efficiency in large simulations, especially "order-N" methods (Ch. 23). The semi-empirical tight-binding method is particularly simple and instructive since the basis is not explicitly specified and one needs only the matrix elements of the overlap and the hamiltonian. This chapter starts with a definition of the problem of electronic structure in terms of localized orbitals, and considers various illustrative examples in the tight-binding approach. Many of the concepts and forms carry over to full calculations with localized functions that are the subject of the following chapter, Ch. 15.

The hallmark of the approaches considered in this chapter and the next is that the wavefunction is expanded in a linear combination of fixed energy-independent orbitals, each associated with a specific atom in the molecule or crystal. For example, the linear combination of atomic orbitals (LCAO) formulation denotes a basis of atomic or modified atomic-like functions. Such a basis provides a natural, physically motivated description of electronic states in materials; in fact, possibly the first theory of electrons in a crystal was the tight-binding[1] method developed by Bloch [36] in 1928. The history of this approach is summarized nicely by Slater and Koster [589], who point out that the seminal work of Bloch considered only the simplest s-symmetry function and the first to consider a basis of different atomic orbitals were Jones, Mott, and Skinner [594] in 1934.

We will highlight three ways in which the local orbital, or tight-binding, formulation plays an important role in electronic structure:

- Of all the methods, perhaps tight-binding provides the simplest understanding of the fundamental features of electronic bands. In particular, this provided the original derivation

[1] Here "tight-binding" means "highly localized atomic states," whereas it has taken different meaning (Sec. 14.4) in more recent years.

of the Bloch theorem [36], which will also suffice for us to derive the theorem yet again in Sec. 14.1.

- Empirical tight-binding methods can provide accurate, useful descriptions of electronic bands and total energies. In this approach, one assumes a form for the hamiltonian and overlap matrix elements without actually specifying anything about the orbitals except their symmetry. The values of the matrix elements may be derived approximately or may be fitted to experiment or other theory. This is generically called "tight-binding" and is widely used as a fitting procedure or as a quick (and sometimes dirty) electronic structure method. As such it is the method of choice for development of many ideas and new methods, e.g. "order-N" techniques in Ch. 23. This is the subject of later sections in this chapter.
- Finally, local orbitals can be used as a basis to carry out a full self-consistent solution of independent-particle equations. Analytic forms, especially gaussians, are extensively used in chemistry, where standard basis sets have been developed. Alternatively, one can use atomic-like orbitals with all integrals calculated numerically. Localized orbital methods are powerful, general tools, and are the subject of Ch. 15.

14.1 Localized atom-centered orbitals

A local orbital basis is a set of orbitals $\chi_\alpha(\mathbf{r} - \mathbf{R}_I)$, each associated with an atom at position $\mathbf{R}_I$. In order to simplify notation, we will let m denote both α and site I, so that $m = 1, \ldots, N_{\text{basis}}$ labels all the states in the basis, which can also be written $\chi_m(\mathbf{r} - \mathbf{R}_m)$.[2] In a crystal, the atoms in a unit cell are at positions $\tau_{\kappa,j}$, where $\tau_{\kappa,j}$ is the position of $j = 1, \ldots, n^\kappa$ atoms of type κ. The composite index $\{\kappa, j, \alpha\} \rightarrow m$ allows the entire basis to be specified by $\chi_m(\mathbf{r} - (\tau_m + \mathbf{T}))$, where $\mathbf{T}$ is a translation vector. The matrix elements of the hamiltonian of a state m in the cell at the origin and state m' in the cell labeled by translation vector $\mathbf{T}$ is

$$H_{m,m'}(\mathbf{T}) \;=\; \int \mathrm{d}\mathbf{r}\; \chi_m^*(\mathbf{r} - \tau_m)\hat{H}\,\chi_{m'}[\mathbf{r} - (\tau_{m'} + \mathbf{T})], \tag{14.1}$$

which applies to any orbitals m and m' in cells separated by the translation $\mathbf{T}$, since the crystal is translationally invariant. Similarly, the overlap matrix is given by

$$S_{m,m'}(\mathbf{T}) \;=\; \int \mathrm{d}\mathbf{r}\; \chi_m^*(\mathbf{r} - \tau_m)\chi_{m'}[\mathbf{r} - (\tau_{m'} + \mathbf{T})]. \tag{14.2}$$

The Bloch theorem for the eigenstates can be derived by defining a basis state with wavevector $\mathbf{k}$,

$$\chi_{m\mathbf{k}}(\mathbf{r}) = A_{m\mathbf{k}} \sum_{\mathbf{T}} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}}\chi_m[\mathbf{r} - (\tau_m + \mathbf{T})], \tag{14.3}$$

where $\mathrm{A}_{m\mathbf{k}}$ is a normalization factor (Exercise 14.3). The analysis proceeds much like the

[2] Here the subscript m is a generic label for a basis function as in other chapters; when used in combination with l, m denotes the azimuthal quantum number, e.g. in Sec. 14.2.

derivation of the Bloch theorem in a plane wave basis in Secs. 12.1 and 12.2, except that here the wavevector $\mathbf{k}$ is restricted to the Brillouin zone. This is sufficient since the phase factor $e^{i\mathbf{k}\cdot\mathbf{T}}$ in (14.3) is unchanged if a reciprocal lattice vector is added. Using the translation invariance of the hamiltonian, it is straightforward (Exercise 14.2) to show that matrix elements of the hamiltonian with basis functions $\chi_{m\mathbf{k}}$ and $\chi_{m'\mathbf{k}'}$ are non-zero only for $\mathbf{k} = \mathbf{k}'$, with

$$H_{m,m'}(\mathbf{k}) \;=\; \int \mathrm{d}\mathbf{r}\; \chi^*_{m\mathbf{k}}(\mathbf{r})\hat{H}\chi_{m'\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{T}} e^{i\mathbf{k}\cdot\mathbf{T}} H_{m,m'}(\mathbf{T}), \tag{14.4}$$

and

$$S_{m,m'}(\mathbf{k}) \;=\; \int \mathrm{d}\mathbf{r}\; \chi^*_{m\mathbf{k}}(\mathbf{r})\chi_{m'\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{T}} e^{i\mathbf{k}\cdot\mathbf{T}} S_{m,m'}(\mathbf{T}). \tag{14.5}$$

Since the hamiltonian conserves $\mathbf{k}$, an eigenfunction of the Schrödinger equation in a basis always can be written in the form

$$\psi_{i\mathbf{k}}(\mathbf{r}) \;=\; \sum_m\; c_m(\mathbf{k})\; \chi_{m\mathbf{k}}(\mathbf{r}), \tag{14.6}$$

and the secular equation for wavevector $\mathbf{k}$ is

$$\sum_{m'} \left[H_{m,m'}(\mathbf{k}) - \varepsilon_i(\mathbf{k}) S_{m,m'}(\mathbf{k}) \right] c_{i,m'}(\mathbf{k}) = 0. \tag{14.7}$$

This has the same form as Eq. (12.9) except that in (14.7) the orbitals are not assumed to be orthonormal. The only fundamental sense in which local orbitals are different from any other basis is that the locality of $\chi_m(\mathbf{r} - (\tau_m + \mathbf{T}))$ is expected to cause $H_{m,m'}(\mathbf{T})$ and $S_{m,m'}(\mathbf{T})$ to decrease and become negligible for large distances $|\tau_m - (\tau_{m'} + \mathbf{T})|$.

14.2 Matrix elements with atomic orbitals

Much can be gained from consideration of the symmetries of the basis orbitals and the crystal or molecule. This is the basis for tight-binding approaches (Sec. 14.4) and continues to be essential in full calculations (Ch. 15). An appropriate choice for basis functions is a set of atomic-like functions centered on the atom sites. On each site κ, j the basis functions can be written as radial functions multiplied by spherical harmonics,

$$\chi_\alpha(\mathbf{r}) \rightarrow \chi_{nlm}(\mathbf{r}) = \chi_{nl}(r) Y_{lm}(\hat{\mathbf{r}}), \tag{14.8}$$

where n indicates different functions with the same angular momentum. Real basis functions can also be defined using the real angular functions $S^+_{lm} = \frac{1}{\sqrt{2}}(Y_{lm} + Y^*_{lm})$ and $S^-_{lm} = \frac{1}{\sqrt{2}i}(Y_{lm} - Y^*_{lm})$ defined in Eq. (K.11). These are useful for visualization and in actual calculations, but the Y_{lm} are most convenient for symmetry analysis. Examples of real s, p, and d orbitals are given in Fig. 14.1.

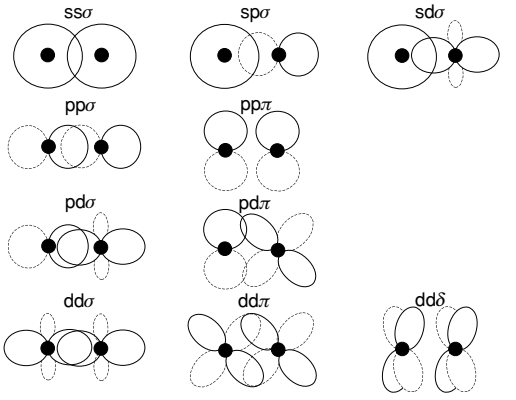


Figure 14.1. Schematic figures of local orbitals indicating all possible overlap and two-center hamiltonian matrix elements for s, p, and d orbitals, which are classified by the angular momentum about the axis with the notation σ ($m = 0$), π ($m = 1$), and δ ($m = 2$). The orbitals shown are the real combinations of the angular momentum eigenstates. Positive and negative lobes are denoted by solid and dashed lines, respectively. Note that the sign of the p orbitals must be fixed by convention; here and in Tab. 14.1 the positive p_x lobe is along the positive x-axis, *etc.*

The matrix elements, Eqs. (14.1) and (14.2), can be divided into one-, two-, and three-center terms. The simplest is the overlap matrix S in (14.2), which involves only one center if the two orbitals are on the same site ($\mathbf{T} = 0$ and $\tau_m = \tau_{m'}$) and two centers otherwise. The hamiltonian matrix elements in (14.1) consist of kinetic and potential terms with

$$\hat{H} = -\frac{1}{2}\nabla^2 + \sum_{\mathbf{T}\kappa j} V^{\kappa}[|\mathbf{r} - (\tau_{\kappa j} + \mathbf{T})|], \tag{14.9}$$

where the first term is the usual kinetic energy and the second is the potential decomposed into a sum of spherical terms centered on each site κ, j in the unit cell.[3] The kinetic part of the hamiltonian matrix element always involves one or two centers. However, the potential terms may depend upon the positions of other atoms; they can be divided into the following.

- One-center, where both orbitals and the potential are centered on the same site. These terms have the same symmetry as an atom in free space.
- Two-center, where the orbitals are centered on different sites and the potential is on one of the two. These terms have the same symmetry as other two-center terms.
- Three-center, where the orbitals and the potential are all centered on different sites. These terms can also be classified into various symmetries based upon the fact that three sites define a triangle.

[3] This decomposition can always be done formally. Often $V^{\kappa}(|\mathbf{r}|)$ can be approximated as spherical atomic-like potentials associated with atom of type κ.

- A special class of two-center terms with both orbitals on the same site and the potential centered on a different site. These terms add to the one-center terms above, but depend upon the crystal symmetry.

Two-center matrix elements

Two-center matrix elements play a special role in calculations with local orbitals and are considered in more detail here. The analysis applies to all overlap terms and to any hamiltonian matrix elements that involve only orbitals and potentials on two sites. For these integrals the problem is the same as for a diatomic molecule in free space with cylindrical symmetry. The orbitals can be classified in terms of the azimuthal angular momentum about the line between the centers, i.e. the value of m with the axis chosen along the line, and the only non-zero matrix elements are between orbitals with the same $m = m'$. If $K_{lm,l'm'}$ denotes an overlap or two-center hamiltonian matrix element for states lm and $l'm'$, then in the standard form with orbitals quantized about the axis between the pair of atoms, the matrix elements are diagonal in mm' and can be written $K_{lm,l'm'} = K_{ll'm}\delta_{m,m'}$. The quantities $K_{ll'm}$ are independent matrix elements that are irreducible, i.e. they cannot be further reduced by symmetry. By convention the states are labeled with l or l' denoted by s, p, d, ..., and $m = 0, \pm 1, \pm 2, \ldots$, denoted by σ, π, $\delta, \ldots$, leading to the notation $K_{\mathrm{ss}\sigma}$, $K_{\mathrm{sp}\sigma}$, $K_{\mathrm{pp}\pi}, \ldots$.

Figure 14.1 illustrates the orbitals for the non-zero σ, π, and δ matrix elements for s, p, and d orbitals. The orbitals shown are actually the real basis functions $S^{\pm}_{lm}$ defined in Eq. (K.11) as combinations of the $\pm m$ angular momentum eigenstates. These are oriented along the axes defined by the line between the neighbors and two perpendicular axes. All states except the s state have positive and negative lobes, denoted by solid and dashed lines, respectively. Note that states with odd l are odd under inversion. Their sign must be fixed by convention (typically one chooses the positive lobe along the positive axis). (The direction of the displacement vector is defined to lie between the site denoted by the first index and that denoted by the second index.) For example, in Fig. 14.1, the $K_{\mathrm{sp}\sigma}$ matrix element in the top center has the negative lobe of the p function oriented toward the s function. Interchange of the indices leads to $K_{\mathrm{ps}\sigma} = -K_{\mathrm{sp}\sigma}$ and, more generally, to $K_{ll'm} = (-1)^{l+l'} K_{l'lm}$ (Exercise 14.4).

An actual set of basis functions is constructed with the quantization axis fixed in space, so that the functions must be transformed to utilize the standard irreducible form of the matrix elements. Examples of two-center matrix elements of s and $\mathrm{p}_i = \{\mathrm{p}_x, \mathrm{p}_y, \mathrm{p}_z\}$ orbitals for atoms separated by displacement vector $\mathbf{R}$ are shown in Fig. 14.2. Each of the orbitals on the left-hand side can be expressed as a linear combination of orbitals that have the standard form oriented along the rotated axes, as shown on the right. An s orbital is invariant and a p orbital is transformed to a linear combination of p orbitals. The only non-zero matrix elements are the σ and π matrix elements, as shown. The top row of the figure illustrates the transformation of the p_x orbital needed to write the matrix element $K_{\mathrm{s},\mathrm{p}_x}$ in terms of $K_{\mathrm{sp}\sigma}$ and the bottom row illustrates the relation of $K_{\mathrm{p}_x,\mathrm{p}_z}$ to $K_{\mathrm{pp}\sigma}$ and $K_{\mathrm{pp}\pi}$. Specific relations for all s and p matrix elements are given in Tab. 14.1. Matrix elements

Table 14.1. Table of two-center matrix elements for either the overlap or the hamiltonian, with real orbitals s and p_x, p_y, p_z. The vector $\mathbf{R}$ between sites, as shown in Fig. 14.2, is defined to have direction components $\hat{\mathbf{R}} \equiv \{x, y, z\}$, and the matrix element is expressed in terms of the σ and π integrals $K_{ss\sigma}$, $K_{sp\sigma}$, and $K_{pp\pi}$. Examples of matrix elements are shown; all other s and p matrix elements are related by symmetry. Expressions for d orbitals are given in many places including [344], [589], and [590], and arbitrary angular momenta can be treated using the procedures in Sec. N.5.

Element	Expression
$K_{s,s}$	$K_{ss\sigma}$
K_{s,p_x}	$x^2 K_{sp\sigma}$
K_{p_x,p_x}	$x^2 K_{pp\sigma} + (1-x^2) K_{pp\pi}$
K_{p_x,p_z}	$xz(K_{pp\sigma} - K_{pp\pi})$

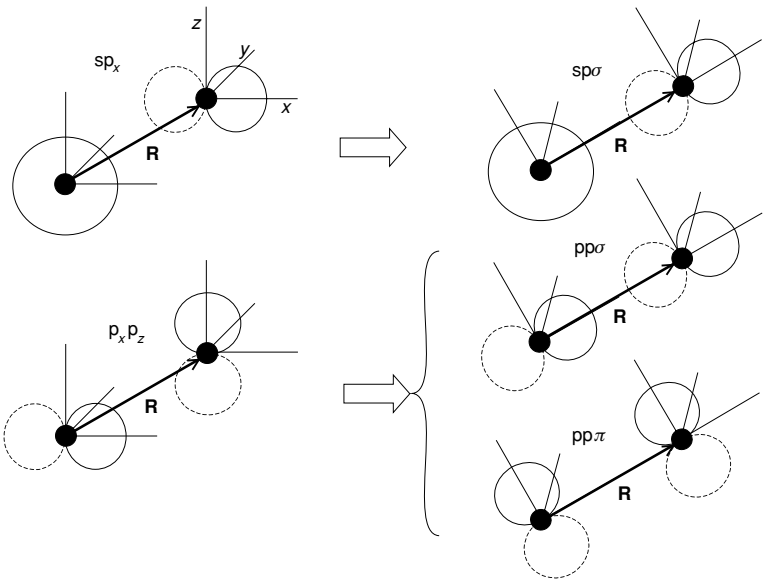


Figure 14.2. Schematic figures for two-center matrix elements of s and $p_i = \{p_x, p_y, p_z\}$ orbitals for atoms separated by displacement vector $\mathbf{R}$. Matrix elements are related to σ and π integrals by the transformation to a combination of orbitals that are aligned along $\mathbf{R}$ and perpendicular to $\mathbf{R}$. The top figure illustrates the transformation to write a real matrix element K_{s,p_x} in terms of $K_{sp\sigma}$: the s orbital is unchanged and the p_x orbital is written as a sum of the σ orbital, which is shown, and the π orbitals, which are not shown because there is no spπ matrix element. The lower figure illustrates the transformation needed to write K_{p_x,p_z} in terms of $K_{pp\sigma}$ and $K_{pp\pi}$. The coefficients of the transformation for all s and p matrix elements are given in Tab. 14.1. Matrix elements for arbitrary angular momenta can be found using the rotation matrix method described in Sec. N.5.

for arbitrary angular momenta can be found using the rotation matrix method described in Sec. N.5.

For the lowest angular momenta it is instructive to write out the expressions explicitly. If the basis orbitals are defined in a system of fixed axes $\hat{x}$, $\hat{y}$, $\hat{z}$, then they must be transformed to new axes for each pair of neighbors. If the vector between the neighbors is $\mathbf{R}$, the new axes are $\hat{x}'$, $\hat{y}'$, $\hat{z}'$ with $\hat{z}'$ parallel to $\mathbf{R}$. The transformation is illustrated in Fig. 14.2 for the examples of s, p_x and p_z, p_x matrix elements; the explicit coefficients for the transformation are given in Tab. 14.1, which is sufficient to generate all sp matrix elements.

Three-center matrix elements

The hamiltonian matrix elements, in general, depend upon the presence of other atoms, resulting in three-center or multi-center matrix elements. Such terms are discussed in Ch. 15 since they arise naturally in the integrals required in a local orbital basis. In this chapter we consider only the "empirical tight-binding" or "semiempirical tight-binding" approaches that involve only the matrix elements $H_{m,m'}(\mathbf{T})$ and $S_{m,m'}(\mathbf{T})$ expressed in a parameterized form, without an explicit representation for the basis orbitals.

The only rigorous result that one can apply immediately to the nature of matrix elements (14.1) and (14.2) is that they must obey crystal symmetry. This is often very helpful in reducing the number of parameters to a small number for a high-symmetry crystal as is done in the tables of Papaconstantopoulos [591] for many crystalline metals. In this form, the tight-binding method is very useful for interpolation of results of more expensive methods [591].

14.3 Slater–Koster two-center approximation

Slater and Koster (SK) [589] developed the widely used [344, 590] approach that bears their name. They proposed that the hamiltonian matrix elements be approximated with the two-center form and fitted to theoretical calculations (or empirical data) as a simplified way of describing and extending calculations of electronic bands. Within this approach, all matrix elements have the same symmetry as for two atoms in free space given in Fig. 14.2 and Tab. 14.1. This is a great simplification that leads to an extremely useful approach to understanding electrons in materials. Of all the methods for treating electrons, the SK approach provides the simplest, most illuminating picture of electronic states. In addition, more accurate treatments involving localized orbitals (Ch. 15) are often best understood in terms of matrix elements having SK form plus additional terms that modify the conclusions in quantitative ways.

Slater and Koster gave extensive tables for matrix elements, including the s and p matrix elements given in Tab. 14.1. In addition, they presented expressions for d states and analytic formulas for bands in several crystal structures. Examples of the latter are given below in Sec. 14.4 to illustrate useful information that can be derived. However, the primary use of the SK approach in electronic structure has become the description of complicated systems,

including the bands, total energies, and forces for relaxation of structures and molecular dynamics. These different applications have very different requirements that often lead to different choices of SK parameters.

For the bands, the parameters are usually designed to fit selected eigenvalues for a particular crystal structure and lattice constant. For example, the extensive tables derived by Papaconstantopoulos [591] are very useful for interpolation of results of more expensive methods. It has been pointed out by Stiles [595] that for a fixed ionic configuration, effects of multi-center integrals can be included in two-center terms that can be generated by an automatic procedure. This makes it possible to describe any band structure accurately with a sufficient number of matrix elements in SK form. However, the two-center matrix elements are not transferable to different structures.

On the other hand, any calculation of total energies, forces, *etc.*, requires that the parameters be known *as a function of the positions of the atoms*. Thus the choices are usually compromises that attempt to fit a large range of data. Such models are fit to structural data and, in general, are only qualitatively correct for the bands. Since the total energy depends only upon the occupied states, the conduction bands may be poorly described in these models. Of particular note, Harrison [344, 590] has introduced a table that provides parameters for any element or compound. The forms are chosen for simplicity, generality, and ability to describe many properties in a way that is instructive and useful, albeit approximate. The basis is assumed to be orthonormal, i.e. $S_{mm'} = \delta_{mm'}$. The diagonal hamiltonian matrix elements are given in a table for each atom. Any hamiltonian matrix element for orbitals on neighboring atoms separated by a distance R is given by a factor times $1/R^2$ for s and p orbitals and $1/R^{l+l'}$ for $l > 1$. The form for s and p orbitals comes from scaling arguments on the homogeneous gas [590] and the form for higher angular momenta is taken from muffin-tin orbital theory (Sec. 16.7).

Many other SK parameterizations have been proposed, each tailored to particular elements and compounds. Examples are given in Secs. 14.4–14.8, chosen to illustrate various aspects of electronic structure calculations in the present and other chapters. Care must be used in applying the different parameterizations to the appropriate problems.

14.4 Tight-binding bands: illustrative examples

This section is concerned with electronic bands calculated using tight-binding with the SK two-center form for the hamiltonian. First, we consider simple cases that can be worked out analytically, with further examples in exercises. This is followed by applications that illustrate the power of the approach for relevant problems, such as the electronic structure of nanotubes.

s-bands in line, square, and cubic Bravais lattices

The simplest possible example of bands is for s-symmetry states on each site I in a Bravais lattice so that there is only one band. As a further simplification, we consider the case of

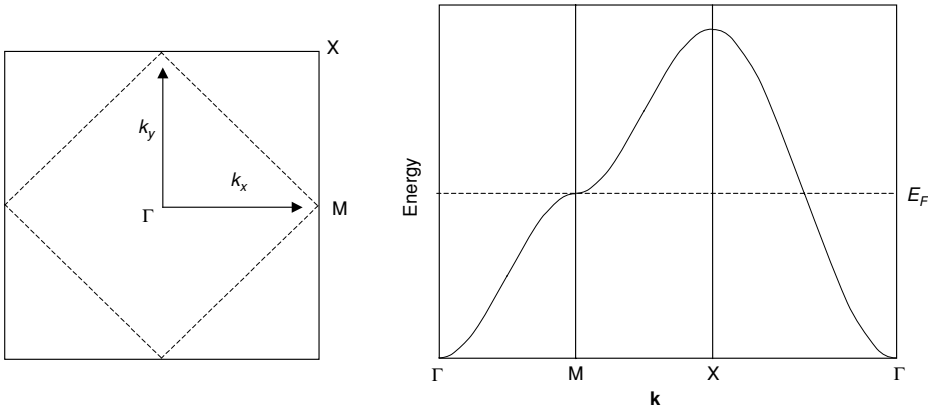


Figure 14.3. Tight-binding bands in the square lattice with only an s state on each site and nearest neighbor interactions. The left figure shows the BZ and the dashed line shows the Fermi surface in the case of a half-filled band. The right figure shows the bands with **k** along the lines between the high-symmetry points.

orthogonal basis states and non-zero hamiltonian matrix elements $\langle I|\hat{H}|I'\rangle \equiv t$ only if I and I' are nearest neighbors. The on-site term can be chosen to be zero, $\langle 0|\hat{H}|0\rangle = 0$. There are three cases (line, square, and cubic lattices) that can be treated together. For the cubic lattice with spacing a the general expressions (14.4) and (14.7) reduce to

$$\varepsilon(\mathbf{k}) = H(\mathbf{k}) = 2t\left[\cos(k_x a) + \cos(k_y a) + \cos(k_z a)\right]. \tag{14.10}$$

The bands for the square lattice in the x, y-plane are given by this expression, omitting the k_z term; for a line in the x-direction, only the k_x term applies.

This simple example leads to useful insights. In particular, the bands are symmetric about $\varepsilon(\mathbf{k}) = 0$ in the sense that every state at $+\varepsilon$ has a corresponding state at $-\varepsilon$. This can be seen by plotting the bands in two ways: first in the usual Brillouin zone centered on $\mathbf{k} = 0$, and second in a cell of the reciprocal lattice centered on $\mathbf{k} = (\pi/a, \pi/a, \pi/a)$. Since $\cos(k_x a - \pi) = -\cos(k_x a)$, *etc.*, it follows that the bands have exactly the same shape except that the sign of the energy is changed,

$$\varepsilon(\mathbf{k}) = -\varepsilon[\mathbf{k} - (\pi/a, \pi/a, \pi/a)]. \tag{14.11}$$

The same arguments apply to the line and square: the line has a simple cosine band and the bands for a square lattice are illustrated in Fig. 14.3. The densities of states (DOS) for one, two, and three dimensions are shown in Fig. 14.4. The shapes can be found analytically in this case, which is the subject of Exercise 14.5.

There are several remarkable consequences in the case of the square. The energy $\varepsilon(\mathbf{k}) = 0$ at a zone face $\mathbf{k} = (\pi/a, 0)$, which is easily verified using (14.10) and omitting the k_z term. This is a saddle point since the slope vanishes and the bands curve upward and downward in

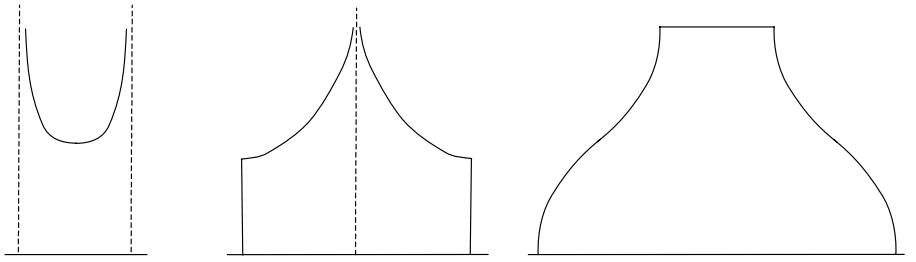


Figure 14.4. Schematic densities of states (DOS) for an s-band in a one-dimensional (1-D) line, a two-dimensional (2-D) square, and a three-dimensional (3-D) simple cubic lattice with nearest neighbor interactions t. The bands are symmetric about the center and the width of each segment is $4t$, i.e. the width of the 1-D DOS is $4t$, that of the 2-D DOS is $8t$ divided into two parts, and the 3-D band has width $12t$ divided into three parts of equal width. The special property of the square lattice in this leads to a logarithmic singularity at the band center. See also Exercise 14.5 for further information.

different directions as shown in Fig. 14.3. This leads to a density of states with a logarithmic divergence at $\varepsilon = 0$ (Exercise 14.6). Furthermore, for a half-filled band (one electron per cell), the Fermi surface is at energy $\varepsilon(\mathbf{k}) = 0$. This leads to the result shown in Fig. 14.4 that the Fermi surface is a square (Exercise 14.6) rotated by $\pi/4$ with half the volume of the Brillouin zone, and the density of states diverges at $\varepsilon = E_F$ as shown in Fig. 14.3. If there are second-neighbor interactions, the symmetry of the bands in $\pm\varepsilon$ is broken and the Fermi surface is no longer square.

Non-orthogonal orbitals

Solution of the tight-binding equations in terms of non-orthogonal orbitals can be done simply in terms of the overlap matrix S using Eq. (14.7). The matrix elements of S can be parameterized in the same way as the hamiltonian, with the added benefit that the two-center form is rigorous and each orbital is to be normalized so that $S_{mm} = 1$. The effect can be illustrated by line, square, or simple cubic lattices, with nearest neighbor overlap defined to be s. Then the solution for the bands, Eq. (14.10), is generalized to

$$\varepsilon(\mathbf{k}) = \frac{H(\mathbf{k})}{S(\mathbf{k})} = \frac{2t\left[\cos(k_x a) + \cos(k_y a) + \cos(k_z a)\right]}{1 + 2s\left[\cos(k_x a) + \cos(k_y a) + \cos(k_z a)\right]}. \tag{14.12}$$

The effect of non-zero s is discussed in Exercises 14.15 and 14.16. In this case, the symmetry about $\varepsilon = 0$ is broken, so that the conclusions on bands and the Fermi surface no longer apply. In fact s has an effect like longer range hamiltonian matrix elements, indeed showing strictly infinite range but rapid exponential decay.

Non-orthogonal orbitals play an essential role in realistic tight-binding models. As discussed more completely in Sec. 14.9, it is never rigourously consistent to cut off the hamiltonian matrix elements while assuming orthogonal orbitals. This is a manifestation of the

well-known properties of Wannier functions (Ch. 21) and the fact that Wannier functions are very environment dependent. On the other hand, non-orthogonal functions can be much more useful because they are more transferable between different environments. This is illustrated by examples in Secs. 14.9 and 21.4.

Two atoms per cell

The simplest possible example of bands for an ionic crystal with two types of atoms is the generalization of the model to two s-symmetry states on two sites in the unit cell. Consider the same structures as above (a line, square, or cube) but with alternation of the two types of atoms on the sites. In three dimensions this leads to the NaCl structure shown in Fig. 4.7. This figure also illustrates the case of the square (one plane of the NaCl structure) and the line (one line of that structure). The bands can be illustrated by the one-dimensional case, in which case the two bands are the eigenvalues of the secular equation

$$\begin{vmatrix} \frac{\Delta}{2} - \varepsilon(k) & 2t\cos(ka) \\ 2t\cos(ka) & -\frac{\Delta}{2} - \varepsilon(k) \end{vmatrix} = 0, \tag{14.13}$$

where the on-site matrix elements are chosen as $\pm\Delta/2$ in order to have average value zero and t is the matrix element between nearest neighbors. It is simple to derive the analytic bands $\varepsilon(k) = \pm\sqrt{(2t\cos(ka))^2 + (\Delta/2)^2}$, which are symmetric in $\pm\varepsilon$ and have a minimum gap Δ between the bands. Exercise 14.17 considers the extension to non-orthogonal orbitals.

14.5 Square lattice and CuO_2 planes

The problem of an s band in a square lattice has a particularly noteworthy application in the case of the cuprate high-temperature superconductors.[4] Figure 4.5 shows the square lattice structure of CuO_2 planes that is the common feature of these materials, e.g. each of the planes in the bilayer in $YBa_2Cu_3O_7$ shown in Fig. 17.3. Extensive calculations, exemplified by the bands presented in Fig. 17.4, have shown that the primary electronic states at the Fermi energy are a single band formed from Cu d and O p states. The band has the same symmetry as $d_{x^2-y^2}$ states centered on each Cu (where x and y are in directions toward the neighboring Cu atoms. This can be understood in terms of the Cu and O states shown in Fig. 14.5. The three states per cell form a bonding, a non-bonding, and an anti-bonding combination, with the anti-bonding band crossing the Fermi level. In fact, the single anti-bonding band has the same symmetry as a Cu $d_{x^2-y^2}$ band with an effective hamiltonian matrix element (Exercise 14.14) so that the problem is equivalent to a model with one $d_{x^2-y^2}$

[4] This is a well-known case [216] where the simple LDA and GGA functionals predict a metal at half-filling whereas the real solution is an antiferromagnetic insulator. Nevertheless, the metallic state created by doping appears to be formed from the band as described here.

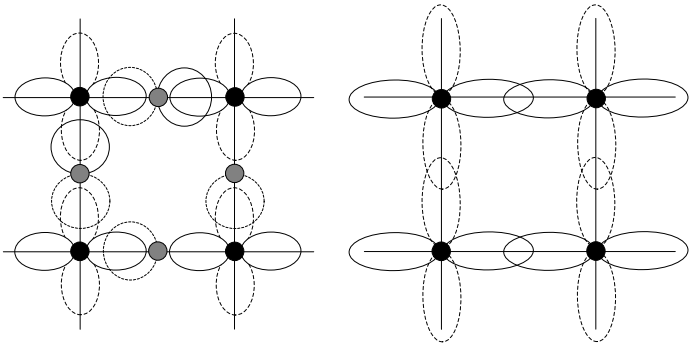


Figure 14.5. Tight-binding models representing electronic states in the square lattice structure of a CuO plane. On the left is shown the three-band model representing one Cu $d_{x^2-y^2}$ and two O p states per cell. As discussed in the text, the most relevant states are the anti-bonding combination of Cu and O states that are equivalent to the model shown on the right with modified orbitals that have $d_{x^2-y^2}$ symmetry around each Cu site. The effective orbitals are more extended, as shown on the right; actual calculated orbitals shown in Fig. 17.11 are more realistic and show extended shape with $d_{x^2-y^2}$ symmetry. Finally, the band is isomorphic to a single s band because, by symmetry, all hamiltonian matrix elements have the same symmetry; e.g. the nearest neighbor elements all have the same sign, as is evident from the right-hand figure.

state per Cu, as shown on the right-hand side of Fig. 14.5. This highly schematic figure is supported by detailed calculations of the one-band orbital shown in Fig. 17.11. The orbital has $d_{x^2-y^2}$ and is extended in the direction of the Cu–O bonds. Indeed, it is extended in the directions along the Cu–O bonds, with large amplitude on the O sites. If the orbitals are required to be orthonormal, like Wannier functions, then each orbital must also extend to the neighbouring Cu sites.

Finally, the problem is isomorphic to a single s band; this occurs because nearest neighbor $d_{x^2-y^2}$ states always have lobes of the same sign ($++$ or $--$) so that the matrix elements are equal for all four neighbors, exactly as for s symmetry states. Thus the simplest model for the bands is a single s band, with dispersion shown in Fig. 14.4, and a square Fermi surface at half-filling. In fact, there are second-neighbor interactions which modify the bands and the calculated Fermi surface [456, 458]. The single band resulting from the orbital in Fig. 17.11 is shown in Fig. 17.12. It accurately describes the actual band, and its dispersion is significantly different from the nearest neighbor model due to longer range matrix elements in a realistic model.

14.6 Examples of bands: semiconductors and transition metals

In this section two simple examples of tight-binding bands are given using the Slater–Koster approximation: Si and Ni. The simplest case is the bands of Si with a minimum basis of one s state and 3 p states per atom. A number of features can be derived analytically (Exercise 14.23). In particular, the states at $\mathbf{k} = 0$ are pure s or pure p; the lowest state is the bonding combination of the s states and the top of the valence band is the three-fold

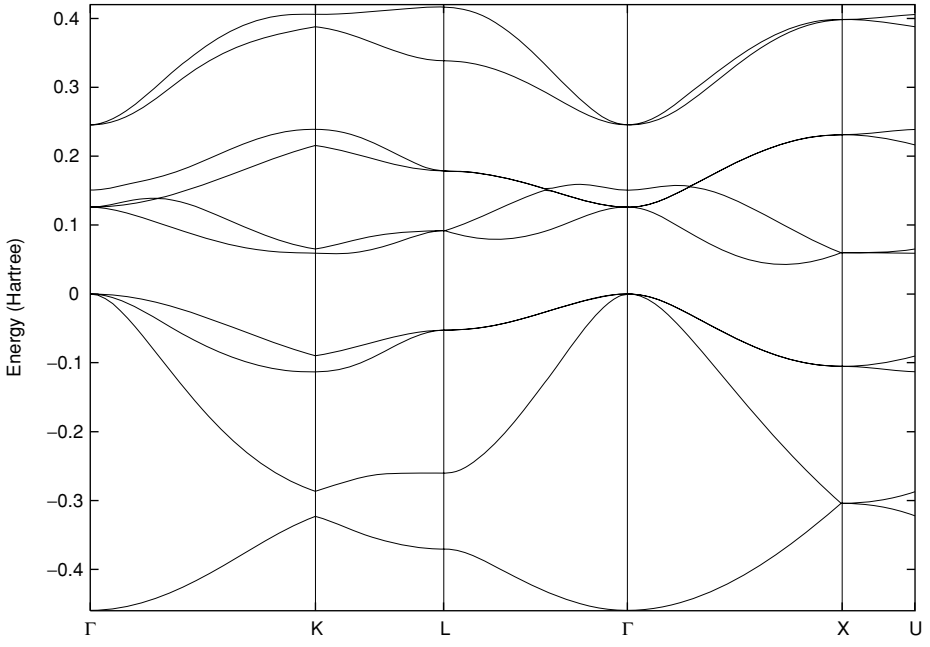


Figure 14.6. Band structure of Si calculated with the Slater–Koster parameters determined by Vogl and coworkers [596] with an added s* orbital, which is the simplest addition that leads to a reasonable description of the lowest conduction bands. Provided by N. Romero; programs described schematically in App. N and available on-line (Ch. 24).

degenerate bonding combination of p_x, p_y, and p_z states. The eigenvalues are readily derived in terms of the matrix elements of the hamiltonian, the on-site energies E_s and E_p, and the matrix elements $H_{ss\sigma}$, $H_{sp\sigma}$, $H_{pp\sigma}$, and $H_{pp\pi}$.

The bands of Si shown in Fig. 14.6 were calculated using SK parameters [596] that were designed to describe energies of electronic states accurately, especially those near the band gap that are most relevant for determination of the electronic properties. By including a second s symmetry state (called s*) this model provides the simplest quantitative description of the lowest conduction band. In fact, full calculations show that the actual effect is due to the admixture of Si 3d states in the conduction bands. For example, a non-orthogonal tight-binding model, constructed as described in Sec. 14.9, including d states has been shown [597] to reproduce well both the LDA bands and total energies.

The bands derived from the parameters in Harrison's "universal" table [344, 590] are considered in Exercise 14.24. They have the same qualitative form for the valence bands as the more accurate bands shown in Fig. 14.6; however, the conduction bands are qualitatively incorrect. This illustrates that care must be taken in using an approach like tight-binding; the detailed shape of the conduction band is not given well by a tight-binding approach designed to describe properties like the total energy that depend only upon the filled bands.

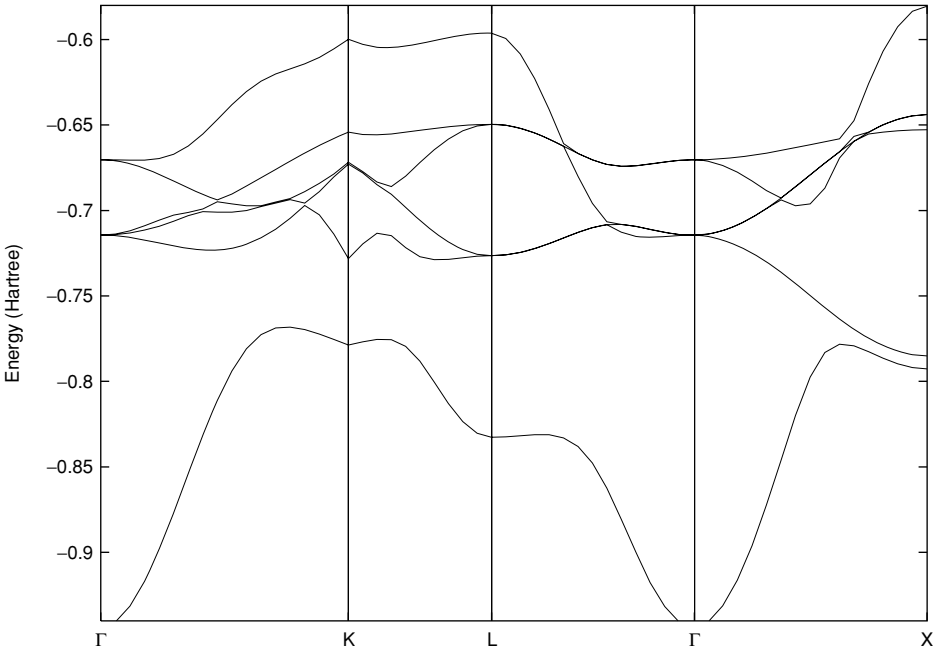


Figure 14.7. Bands of Ni calculated with parameters from the table of Harrison [344, 590]. This is a case where the main features of the narrow d bands and wide s band are well described; compare with full calculations in Figs. 16.4 and 16.5 and the canonical LMTO fcc bands in Fig. 16.12. The high-energy parts of the band are not correct because the s band is not sufficient. See also caption of Fig. 14.6.

The bands of transition metals have features that are dominated by localized d states. Tight-binding is a very natural approach for these states, as is also emphasized in the tight-binding LMTO method (Secs. 16.7 and 17.5). Figure 14.7 shows the bands of Ni calculated using Harrison's parameters [344, 590]. The bands have the right features, compared to full calculations, for narrow d bands, s–d hybridization, and the s band. The highest energy part of the bands plotted is inadequate; other states become more relevant and the bands continue to higher energies with no gaps.

14.7 Electronic states of nanotubes

Nanotubes are ideal for illustrating the use of tight-binding to reveal the most vital information about the electronic structure in a simple, illuminating way. Carbon nanotubes were discovered in 1991 by Iijima [196] and recognized to be nanoscale versions of "microtubes," long tubular graphitic structures that had been grown using iron particles for catalysis [205]. In a perfect single-wall tube, each carbon atom is equivalent and is at a junction of three hexagons. The various ways a sheet of graphene (i.e. a single honeycomb-structure

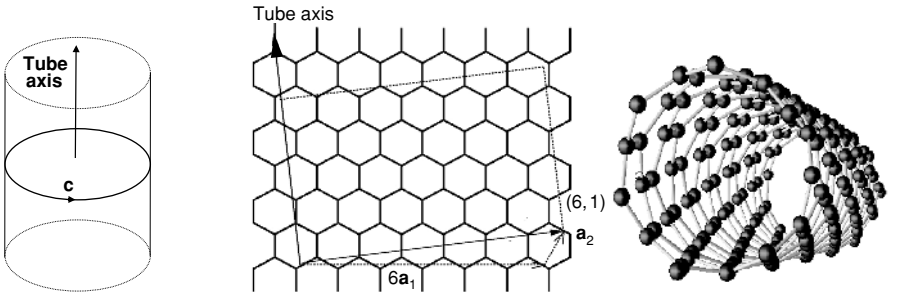


Figure 14.8. Rolling of a graphene sheet to form a nanotube. On the left is shown a tube with circumference indicated by the circle **c**. The middle and right figures show a graphene translation vector $\mathbf{c} = n\mathbf{a}_1 + m\mathbf{a}_2$ with $n = 6$ and $m = 1$ and the chiral (6, 1) tube formed by rolling to join the site related by **c**. The basis vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ shown are the same as in Fig. 4.5 and in [204, 598]; in this notation the example shown is a (6,1) nanotube. Provided by J. Kim; programs available on-line (Ch. 24).

plane of carbon) can be rolled into a tube leads to an enormous variety of semiconductors and metals [203–205]. These can be understood in terms of a simple tight-binding model for the electrons, based upon modifying the bands of graphene in ways that are readily understood.

The structures of nanotubes are defined in terms of a graphene layer as shown in Fig. 14.8. The vector indicated connects atoms in the layer that are equivalent in the tube, i.e. the tube is defined by rolling the plane of graphene to bring those points together. The tube axis is perpendicular to the vector. The convention is to label the vector with multiples of graphene translation vectors, $\mathbf{a}_1$ and $\mathbf{a}_2$, defined as in Fig. 4.5. The example shown is for a (6, 1) tube which denotes $(6 \times \mathbf{a}_1, 1 \times \mathbf{a}_2)$ and which defines the chiral tube shown on the right. Special examples of "zig-zag" $(n, 0)$ and "armchair" (n, n) tubes are shown in Fig. 14.9. These are not chiral, but have very different properties due to the underlying atomic structure. See Exercise 14.20.

The first step is to make a simple model for the bands of graphene, which has the planar honeycomb structure shown in Fig. 4.5. The Brillouin zone is shown in Fig. 14.9 (the same as for the three-dimensional hexagonal zone in Fig. 4.10 with $k_z = 0$), where K denotes the corner and M the edge center. Full calculations like those shown in Fig. 2.29 demonstrate the well-known fact that the bands of graphite at the Fermi energy are π bands, composed of electronic states that are odd in reflection in the plane. For a single, flat graphene sheet, symmetry forbids coupling of π bands to σ bands that are well below and well above the Fermi energy. The π bands are well represented as linear combinations of p_z orbitals of the C atoms, where z is perpendicular to the plane. Since graphene has two atoms per cell, the p_z states form two bands. If there is a nearest neighbor hamiltonian matrix element t, the bands are given by [204]

$$\left|\hat{H}(\mathbf{k}) - \varepsilon(\mathbf{k})\right| = \begin{vmatrix} -\varepsilon(\mathbf{k}) & H_{12}(\mathbf{k}) \\ H_{12}^*(\mathbf{k}) & -\varepsilon(\mathbf{k}) \end{vmatrix} = 0, \tag{14.14}$$

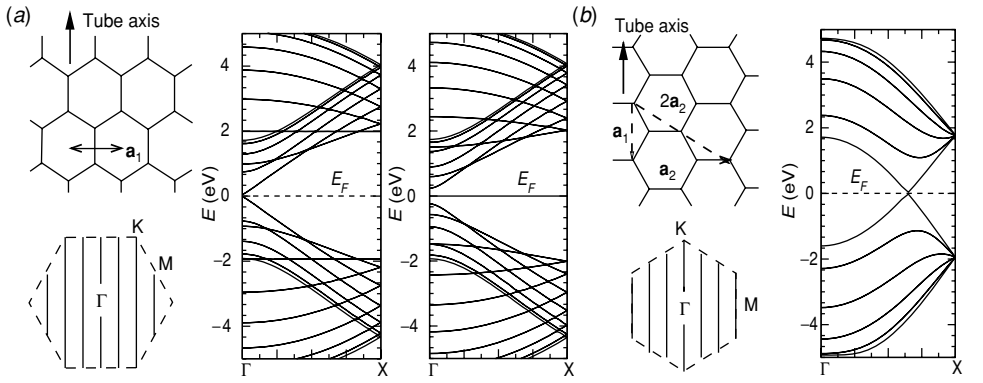


Figure 14.9. Structures, Brillouin zones, and bands of the "zig-zag" (*a*) and "armchair" (*b*) nanotubes. The bands are calculated using the orthogonal tight-binding model of Xu et al. [162]. The zig-zag tubes shown are denoted (13, 0) and (12, 0); the latter is insulating and the former has a small gap that is due to curvature. However, for smaller tubes, the curvature can induce large effects, including band overlap leading to metallic tubes as shown in Fig. 13.8 and discussed in the text. The armchair (n, n) tubes are always metallic since the lines of allowed states with k along the tube always include Γ and one of the K points. (*b*) Illustrates the bands for a (3, 3) tube. In each case the bands of the tube are plotted in the one-dimensional BZ denoted $\Gamma \rightarrow$ X. Provided by J. Kim; programs available on-line (Ch. 24).

where (with the lattice oriented as in Figs. 4.5 and 14.9(a))

$$H_{12}(\mathbf{k}) = t\left[\mathrm{e}^{\mathrm{i}k_y a/\sqrt{3}} + 2\mathrm{e}^{-\mathrm{i}k_y a/2\sqrt{3}}\cos\left(k_x\frac{a}{2}\right)\right], \tag{14.15}$$

where a is the lattice constant. This is readily solved to yield the bands [204]

$$\varepsilon(\mathbf{k}) = \pm|H_{12}(\mathbf{k})| = \pm t\left[1 + 4\cos\left(k_y\frac{\sqrt{3}a}{2}\right)\cos\left(k_x\frac{a}{2}\right) + 4\cos^2\left(k_x\frac{a}{2}\right)\right]^{1/2}. \tag{14.16}$$

The most remarkable feature of the graphene bands is that they touch at the corners of the hexagonal Brillouin zone, e.g. the point denoted K ($k_x = 4\pi/3a, k_y = 0$) shown in the BZ in Fig. 14.9. This and other aspects are brought out in Exercise 14.19. Note also that the bands are symmetric in $\pm\varepsilon$. Since there is one π electron per atom, the band is half-filled and the bands touch with finite slope at the Fermi energy, i.e. a Fermi surface consisting of points. It is this unusual feature that gives rise to the grand array of possibilities for the electronic structure of nanotubes.

The first approximation is to assume that the bands are unchanged from graphene and the only effect is that certain states are allowed by the boundary condition. The condition on allowed functions is that they must be single valued in the circumferential direction but have Bloch boundary conditions along the tube axis. This leads to allowed $\mathbf{k}$ vectors, that are shown as lines in Fig. 14.9. The resulting bands have been analyzed in general [203–205] with the simple conclusion that there is a gap between filled and

empty states unless the allowed k lines pass through the point K. If they do include K, e.g. the armchair tube, then the interesting situation of one-dimensional metallic bands arises.

A convenient approach to finding the actual bands is through tight-binding models that provide a beautiful description of the electronic properties of graphene. Examples are shown in Fig. 14.9 for the insulating even-order zig-zag, the almost metallic odd-order zig-zag (see the following paragraph), and metallic armchair tubes. The calculations are done with a simple orthogonal tight-binding model [162] that describes graphene well; however, it cannot be expected to describe all possible effects in nanotubes, as shown below.

The next approximation is to include expected effects due to curvature [203, 204]. The simplest depends only upon symmetry: curvature makes bonds along the axis inequivalent to those at an angle to the axis. Therefore, the k point where the bands touch moves away from the point K opening a small gap, as shown in Fig. 14.9 for the (13, 0) zig-zag tube. The gap is expected to increase for small tubes with larger curvature. On the other hand, there is no effect upon the band crossing at point K, which is along the tube axis for the armchair tube, so that it remains metallic in all cases.

Finally, one can ask: have all the effects been included? The answer is, "no." As shown in Fig. 13.8, calculations [206] on small tubes have shown that the bands can be qualitatively changed from the graphene-like states considered thus far. The reason is that in small tubes there is large mixing of the graphene-like states, including a particularly strong admixture of π states with a σ anti-bonding state that pushes a band below the Fermi level. This leads to the prediction [206] that small diameter nanotubes can be metallic due to band overlap, even in cases where analogy to graphene would expect an insulator. This effect is also found in high-quality local orbital calculations [588].

What is required for a tight-binding model to capture such effects? There is no unique answer because it is difficult to construct simple models that can describe many different (unforeseen) geometries. Nevertheless, there is considerable success using ideas outlined in Sec. 14.9. Notably, the tight-binding model of [599] has been fitted to LDA calculations of many properties (eigenvalues, total energies, phonons, *etc.*) of carbon in various coordinations and geometries. The non-orthogonal basis improves transferability to different structures. Indeed, results using this model lead to bands in good agreement with both the plane wave calculations [206] for small tubes, as shown in Fig. 13.8, and with the graphene-like bands for larger tubes.

Boron nitride nanotubes

Nanotubes of boron nitride have been proposed theoretically [207] and later made experimentally [208]. Structures for the tubes are allowed if they maintain the B–N equal stoichiometry, and the tubes always have a large gap due to the difference between the B and N atoms. Thus the electronic properties are very different from carbon nanotubes and BN tubes hold the potential to act like one-dimensional semiconductors in the III–V family.

Like other III–V materials they exhibit piezoelectric and pyroelectric effects, but in this case the one dimensionality leads to extreme anisotropy and novel electric polarization and piezoelectric effects [202, 600].

In addition, heterostructure tubes can potentially be created that could be one-dimensional analogs of semiconductor quantum structures. For example, Fig. 2.21 shows the electron density for the highest occupied state in metal–semiconductor junctions of BN–C nanotubes, calculated [202] using real-space methods [209] described in Chs. 12 and 13. The system can be treated as a supercell much like the semiconductor structures in Fig. 13.6; however, there are new aspects introduced by the geometry, by the fact that a single defect can have qualitative effects in one dimension, *etc.* The basic features of the electronic states can also be understood using tight-binding. The requirements are that the model must contain information on C and BN, the B–NC and N–C interactions at the interface, and the relative energies of the bands in the C and BN tubes. The weakness of tight-binding is that these parameters must be obtained from some other calculation; the strength is that once the parameters are obtained and shown to be reasonably transferable, tight-binding methods make possible very fast, illuminating calculations on complicated structures.

14.8 Total energy, force, and stress in tight-binding

The total energy in any self-consistent method, such as the Kohn–Sham approach, can be written as in Eqs. (9.7) and (9.9) expressed as a sum of eigenvalues (Eq. (9.6)) plus the interaction of the nuclei and a correction needed to avoid double counting the interactions

$$E_{\text{total}} = \sum_i \varepsilon_i f(\varepsilon_i) + F[n]. \tag{14.17}$$

Here $f(\varepsilon_i)$ is the Fermi function and i labels eigenstates including the spin index; in a crystal the sum is over all bands and $\mathbf{k}$ in the BZ. In terms of E_{pot} in (9.3), $F[n]$ is given by

$$\begin{aligned} F[n] &\equiv E_{\text{pot}}[n] - \int \mathrm{d}\mathbf{r}\; V_{\text{KS}}(\mathbf{r}) n(\mathbf{r}) \\ &= E_{II} - E_{\text{Hartree}}[n] + \int \mathrm{d}\mathbf{r}\; [\epsilon_{\text{xc}}(\mathbf{r}) - V_{\text{xc}}(\mathbf{r})] n(\mathbf{r}), \end{aligned} \tag{14.18}$$

where V_{KS} is the Kohn–Sham potential taken to be spin independent for simplicity.

In the tight-binding method, the parameterized hamiltonian matrix elements lead to the eigenvalues ε_i. How can the second term be included in such an approach? How can it be approximated as a function of the positions of the nuclei, even though the full theory defines $F[n]$ as a complicated functional of the density? An elegant analysis of the problem has been given by Foulkes and Haydock [417] based upon the expression for the energy, Eq. (9.9). They used the variational properties of that functional and the choice of the density as a sum of neutral spherical atom densities, which is a good approximation [415, 416, 601]. It immediately follows that the difference of the Hartree and ion–ion terms in (14.18) is a

sum of short-range, pair-wise interactions between atoms. The exchange–correlation term in (14.18) is also short range and it can be approximated as a pair potential. Thus we are led to the approximation that F can be expressed as a sum of terms $f(|\mathbf{R}_I - \mathbf{R}_J|)$ that depend only on distances to near neighbors, so that the total energy in the tight-binding approach can be expressed as

$$E_{\text{total}} = \sum_{i=1}^{N} \sum_{m,m'} c_{i,m}^{*} H_{m,m'} c_{i,m'} + \sum_{I<J} f(|\mathbf{R}_I - \mathbf{R}_J|), \tag{14.19}$$

where the eigenvectors are given by

$$\psi_i = \sum_m c_{i,m} \chi_m. \tag{14.20}$$

Finally, defining the density matrix,

$$\rho_{m,m'} = \sum_{i=1}^{N} c_{i,m}^{*} c_{i,m'}, \tag{14.21}$$

the energy can be written

$$E_{\text{total}} = \sum_{m,m'} \rho_{m,m'} H_{m,m'} + \sum_{I<J} f(|\mathbf{R}_I - \mathbf{R}_J|) = \text{Tr}\{\hat{\rho}\hat{H}\} + \sum_{I<J} f(|\mathbf{R}_I - \mathbf{R}_J|). \tag{14.22}$$

The added functions F or f can be found by fitting to additional information related to the total energy, e.g. the elastic constants or phonon frequencies. There can be no unique form of F, however, because of the fundamental ambiguity in separating the two terms in (14.17). An arbitrary function of the nuclear positions can be added to the hamiltonian matrix elements, shifting all eigenvalues rigidly with no change in the physics. The function F must be chosen consistent with the choice in the matrix elements. One choice of F is a sum of pair potentials, as is done in many models such as that of Harrison [344, 590] and models mentioned below. A different approach to is define eigenvalues that include a shift due to repulsive affects [602],

$$\varepsilon_i' \equiv \varepsilon_i + F/N_e. \tag{14.23}$$

Then the total energy is simply

$$E_{\text{total}} = \sum_{i=1}^{N} \varepsilon_i' = \text{Tr}\{\hat{\rho}\hat{H}'\}, \tag{14.24}$$

and the challenge is to parameterize the tight-binding matrix elements to describe both the eigenvalues and the total energy. Considerable success has been demonstrated with this form (see Sec. 14.9). In any case, the results depend upon the availability of calculated and/or experimental energies and the adequacy of the forms chosen to represent different structures.

Forces can be found by taking derivatives of the energy using the force theorem just as in other methods. In tight-binding, the matrix elements are considered as functions of the positions of the nuclei and the expression follows from the condition that the energy is variational w.r.t. the density matrix $\hat{\rho}$ (Exercise 14.21). Taking the derivative of (14.22) with respect to the position of atom I leads to

$$\mathbf{F}_I = -\mathrm{Tr}\left\{\hat{\rho}\frac{\partial \hat{H}}{\partial \mathbf{R}_I}\right\} - \sum_{J\neq I}\frac{\partial f(|\mathbf{R}_I - \mathbf{R}_J|)}{\partial \mathbf{R}_I}, \tag{14.25}$$

where the last term is absent if equation (14.24) is used. Pressure and stress are also straightforward since the stress tensor, Eq. (G.4), can be written as the sum of terms with the form

$$\sigma_{\alpha\beta} = -\frac{1}{\Omega}\frac{\partial E_{\mathrm{total}}}{\partial u_{\alpha\beta}} = -\mathrm{Tr}\left\{\hat{\rho}\frac{\partial \hat{H}}{u_{\alpha\beta}}\right\} - \sum_{J\neq I}\frac{\partial f(|\mathbf{R}_I - \mathbf{R}_J|)}{\partial u_{\alpha\beta}}. \tag{14.26}$$

The first term involves the derivative of the matrix elements with distance and the final term is a sum of two-body contributions as treated in Sec. G.2.

Examples of tight-binding models for total energies

Perhaps the most useful and widely used tight-binding formulations are for carbon and silicon, for which there are several very successful parameterizations. For carbon, these are extremely useful for the amazing variety of structures found, including graphite, diamond, buckyballs, nanotubes, and amorphous and liquid carbon. For example, the form of Xu et al. [162] was constructed to fit the total energies of C in low-coordination structures, the chain, graphitic and diamond. The potential has been used for many simulations and agrees well with other calculations, e.g. the bands of nanotubes shown in Fig. 14.9 and simulations of liquid carbon [162] leading to the radial density distribution in Fig. 18.2. Other potentials are also successful and widely used, such as [603] and forms given in Sec. 14.9.

A number of parameterizations, e.g. those in [604–607], have been developed for Si and applied to problems involving defects, diffusion, and many other interesting properties. For example, Fig. 23.8 shows the results of an O(N) molecular dynamics calculation [608,609] of the structure of complex {311} defects done using the model of Kwon et al. [605], with checks on smaller cells with plane wave density functional theory calculations.

14.9 Transferability: non-orthogonality and environment dependence

There is a basic difficulty in generating tight-binding models that can describe very different structures, in particular, those with different numbers of near neighbors such as open- and close-packed structures [590]. In models that have only two-center matrix elements, the

values of the matrix elements must take into account effects of three-center terms. These effects change drastically between structures such as diamond (four nearest neighbors that lead to $4 \times 3 = 12$ three-center terms) and a close-packed structure (12 nearest-neighbors that lead to $12 \times 11 = 132$ three-center terms). There are two primary approaches toward making tight-binding models that are transferable between such different structures. One is to define *environment-dependent tight-binding matrix elements*, the values of which depend upon the presence of other neighbors. The other approach involves non-orthogonal tight-binding, which is more transferable than orthogonal forms. The reasons are brought out in Sec. 21.4, where it is clear that a small basis of orthogonal functions *must* be long-ranged and environment dependent in order to be accurate; on the other hand, non-orthogonal functions can accurately describe bands even if they are short-range atomic-like functions that are almost environment independent. Indeed, such functions are also the bread and butter of the local orbital methods of Ch. 15.

The different models can be exemplified by the extensive body of work of Cohen, Mehl, and Papaconstantopoulos [602], developed in many subsequent papers,[5] which utilizes non-orthogonal tight-binding with environment-dependent matrix elements. This approach employs shifted eigenvalues ε_i', defined in (14.23), with the diagonal on-site matrix elements dependent upon a sum of densities representing the neighboring atoms. The explicit form suggested [602] for the state at atom I with angular momentum l and spin σ is

$$H_{Il\sigma} = \sum_{n=0}^{3} b_{Il\sigma}^{(n)} \rho_{I\sigma}^{2n/3}, \tag{14.27}$$

where the $b_{Il\sigma}^{(n)}$ are parameters and $\rho_{I\sigma}$ depends upon the surrounding atoms

$$\rho_{I\sigma} = \sum_{J \neq I} \exp(-\lambda_{IJ\sigma}^2 R_{IJ})\ f\left(\frac{R_{IJ} - R_0}{R_c}\right). \tag{14.28}$$

Here the exponential factor represents a density assigned to a neighboring atom, with the scales set by the parameters $\lambda_{IJ\sigma}$, and f is a cutoff factor taken to be the Fermi function. The spin dependence is needed for magnetic systems. The intersite matrix elements of the hamiltonian and overlap are each parameterized and have the same same functional form that can be written

$$K_\gamma(R) = \left(\sum_{n=0}^{3} c_\gamma^{(n)} R^n\right) \exp(-g_\gamma^2 R)\ f\left(\frac{R - R_0}{R_c}\right), \tag{14.29}$$

where $K_\gamma(R)$ denotes either hamiltonian and overlap matrix elements, and the subscript γ denotes ssσ, spσ, spπ, *etc*.

[5] See links at sites given in Ch. 24.

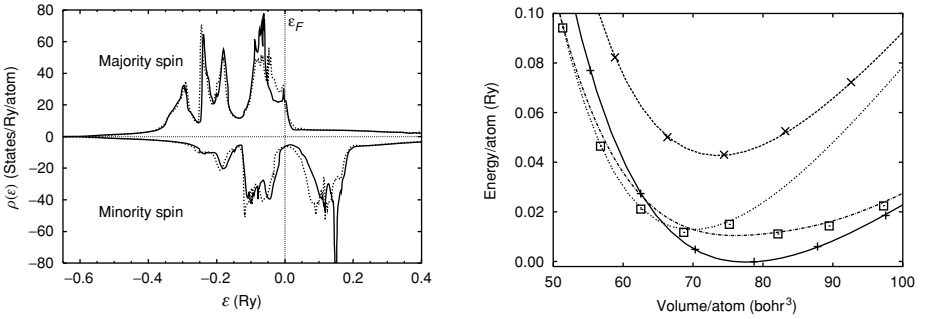


Figure 14.10. Left: Density of states of ferromagnetic Fe in the bcc structure with $a = 5.40a_0$, comparing LAPW (solid) and fitted tight-binding (dashed curves). Right: Total energy versus volume for Fe in various structures, comparing points from LAPW calculations (+, ferromagnetic bcc; square, ferromagnetic fcc; ×, paramagnetic bcc) and lines from the tight-binding calculations. The left figure and the most of the total energies illustrate the quality of the fit. The total energy of the ferromagnetic fcc state was not fitted and the results are a test of transferability. For example, the tight-binding calculations reproduce the delicate collapse of the moment at $\approx 70a_0^3$ where the paramagnetic and ferromagnetic energies cross. From [610].

This form has been applied to a large number of systems, including carbon [599], silicon [597,599], and many transition metals. As an example, Fig. 14.10 shows the results for the density of states of ferromagnetic Fe in the bcc structure and the total energies in fcc and bcc structures, both paramagnetic and ferromagnetic. The tight-binding parameters are the same for all structures and are fitted with a total of 106 parameters[6] to the LAPW results at several volumes for paramagnetic and ferromagnetic bcc and paramagnetic fcc Fe. The total energies and moments for ferromagnetic fcc Fe were not fitted but reproduce correctly the energies, including collapse of the moment to form a paramagnetic state at $\approx 70a_0^3$. It is evident that the fits are extremely good and useful for analysis, given that the tight-binding calculations are orders of magnitude faster than the LAPW ones.

SELECT FURTHER READING

Classic paper:

Slater, J. C. and Koster, G. F., "Simplified LCAO method for the periodic potential problem," *Phys. Rev.* 94:1498–1524, 1954.

Books on electronic structure:

Harrison, W. A., *Electronic Structure and the Properties of Solids*, Dover, New York, 1989.
Harrison, W. A., *Elementary Electronic Structure*, World Publishing, Singapore, 1999.

Book on tight-binding:

Papaconstantopoulos, D. A., *Handbook of Electronic Structure of Elemental Solids*, Plenum, New York, 1986.

[6] There are 40 parameters each for intersite hamiltonian and overlaps, and 13 on-site terms (s, p, d_{e_g}, $d_{t_{2g}}$).

Collection on applications in material science:

Tight-Binding Approach to Computational Materials Science, edited by P. E. A. Turchi, A. Gonis, and L. Colombo, Materials Research Society, Warrendale PA, 1998. [603]

Review article:

Goringe, C. M., Bowler, D. R. and Hernandez, E., "Tight-binding modelling of materials," *Rep. Prog. Phys.* 60:1447–1512, 1997. [604]

Exercises

14.1 See many excellent problems (and solutions) on tight-binding bands, densities of states, and the meaning of the bands in the book by Mihaly and Martin [248].

14.2 Using translation invariance of the matrix elements, show that matrix elements of the hamiltonian with basis functions $\chi_{m\mathbf{k}}$ and $\chi_{m'\mathbf{k}'}$ are non-zero only for $\mathbf{k} = \mathbf{k}'$, i.e. the Bloch theorem, and derive the form given in (14.4).

14.3 Derive the factor $A_{m\mathbf{k}}$ in (14.3) required for the Bloch basis states $\chi_{m\mathbf{k}}(\mathbf{r})$ to be normalized. Show that $A_{m\mathbf{k}} = 1$ if the functions $\chi_m(\mathbf{r} - (\tau_m + \mathbf{T}))$ are orthonormal and that in general $A_{m\mathbf{k}} = (S_{m,m}(\mathbf{k} = 0))^{-\frac{1}{2}}$, where $S(\mathbf{k})$ is defined in (14.5). This relation is used in Exercise 21.2 in examples of Wannier functions.

14.4 Show that, in general, one has the relation $K_{ll'm} = (-1)^{l+l'} K_{l'lm}$ under interchange on the indices of the K matrix. This follows from a consistent definition of the orbitals.

14.5 Show that for an s band in a line, square lattice, and simple cubic lattice with only nearest neighbor hamiltonian matrix elements, the respective densities of states (DOS) have the forms shown schematically in Fig. 14.4. First determine the form for the DOS for the one-dimensional line analytically. Then use this result along with the fact that the bands, (14.10), are simply a sum of cosines for orthogonal directions to derive the form of the DOS for the square and simple cubic lattice. Show that the bands are divided into segments of width $4t$ as stated, and show that in three dimensions the DOS is exactly symmetric and flat in the central range.

14.6 Consider an s band in a square lattice with nearest neighbor matrix element t and one electron per cell. Show that the Fermi surface is a square as shown in Fig. 14.3 and there is a divergence in the density of states at the Fermi energy as shown in Fig. 14.4.

14.7 Derive the expression for the tight-binding s band $\varepsilon(\mathbf{k})$ in a simple cubic crystal. Assume the states are orthonormal and have hamiltonian matrix elements t_1, t_2, and t_3 for the first three neighbors. The bands for $t_2 = t_3 = 0$ are an approximation for the s-like conduction bands in CsCl which has simple cubic structure. Compare with a calculated band structure in the literature, or using a code like that in App. N, using the fact that the states at $\mathbf{k} = 0$ can be classified into purely s and p symmetry and are derived mainly from Cs states.

14.8 Derive the expression for an s band $\varepsilon(\mathbf{k})$ in a fcc crystal with nearest neighbor hamiltonian matrix element t assuming the states are orthonormal. This should be a qualitative approximation

for the lowest conduction band in a fcc metal like Al or an ionic insulator like NaCl which has the fcc structure. Compare with the nearly-free-electron bands in Fig. 12.1, Fig. 16.6, calculated band structures in the literature, or using a code like that in App. N. (Note that there is a relation to the expressions derived in Exercise 14.7 for second neighbors in a simple cubic lattice. Explain the relation in detail.)

14.9 Derive the expression for an s band in a hcp crystal with nearest neighbor hamiltonian matrix element t assuming the states are orthonormal. Assume the c/a ratio is the ideal value. Explicitly evaluate the bands in the direction along the c axis perpendicular to the hexagonal planes. Show that the lower and upper bands touch at the zone boundary, i.e. there is no gap at the zone boundary. Explain why this happens even though there are two atoms per primitive cell.

14.10 Derive expressions for p bands respectively in simple cubic and fcc crystals with nearest neighbor hamiltonian matrix element $t_{pp\sigma}$ and $t_{pp\pi}$. Compare with calculated bands in the literature for the Cl p state in CsCl and NaCl, respectively, to find reasonable values of $t_{pp\pi}$ and $t_{pp\sigma}$.

14.11 There is a close relation of p bands to the equations for phonons as expressed in Exercises 19.3–19.5. As an example, derive the explicit relation of tight-binding equations for p bands in Exercise 14.10 and the phonon dispersion curves in Exercise 19.6 for a nearest neighbor central potential model.

14.12 Consider the one-dimensional tight-binding model with two atoms per cell labeled A and B. If the basis is one s state on each atom, the model can denoted pictorially by $[-\varepsilon_A - t_1 - \varepsilon_B - t_2 - \cdots]$, where $\varepsilon_A, \varepsilon_B$ are the on-site energies and t_1, t_2 the hopping matrix elements. By varying the parameters, this model describes a symmetric ionic crystal ($\varepsilon_A \neq \varepsilon_B$, $t_1 = t_2$), a molecular elemental crystal ($\varepsilon_A = \varepsilon_B$, $t_1 \neq t_2$), and any ionic/molecular combination. Derive the bands as a function of the parameters and show that there is a gap between the two bands for all cases except the one-atom/cell limit where $\varepsilon_A = \varepsilon_B, t_1 = t_2$. See Exercises 21.10, 21.11, 22.8, and 22.9 for examples of Wannier functions, polarization, and effective charges using this model.

14.13 Consider a model like that in Exercise 14.12 except that the state on the B atom has p symmetry. For on-site energies $\varepsilon_A > \varepsilon_B$, this is a one-dimensional model for an ionic crystal like NaCl. From the symmetry of the crystal, show that two bands are formed from the s and p_x states decoupled from bands formed by the orthogonal p_y and p_z states. Assume the states are orthonormal and there are only nearest neighbor hamiltonian matrix elements of magnitude t. In terms of $\Delta = \varepsilon_A - \varepsilon_B$ and t, give analytic expressions for s–p_x bands $\varepsilon_i(k)$. Describe any simplifications in the expressions at $k = 0$ and the BZ boundary. Plot the bands in the Brillouin zone for the case $\Delta = 4t$ and show there is qualitative agreement with published bands of NaCl in the (100) direction. What values of Δ and t provide a reasonable description of NaCl bands? Suggest changes that would better describe the bands.

14.14 Show that the model of Cu and O states shown on the left-hand side of Fig. 14.5 leads to effective model nearest neighbor interactions between Cu states as shown on the right-hand side of the figure. Hint: Construct a 3×3 matrix and diagonalize to find the highest band that corresponds to the band that crosses the Fermi energy in Fig. 17.11.

14.15 Show that the expression for bands with non-orthogonal basis orbitals, Eq. (14.12), is correct. The bands are no longer symmetric about $\varepsilon = 0$. Why is this? What is the physical interpretation? See the following problem for more general properties.

14.16 This problem is to analyze the general consequences of the overlap term in a non-orthogonal basis. Show that the effect of the overlap can be transformed to an orthogonal form, with the result that the hamiltonian matrix elements have infinite range, decaying exponentially. This is the correct result as shown by the decay of orthonormal Wannier functions in Ch. 21. Thus show that rigorously it can never be fully consistent to assume that the hamiltonian matrix elements are finite range and yet the orbitals are orthogonal. The same conclusion is found in Exercise 21.2.

14.17 Find the explicit expression for the generalization of Eq. (14.13) to non-orthogonal orbitals with a nearest neighbor overlap s. Is the minimum gap increased, decreased, or left unchanged by inclusion of non-zero s.

14.18 Give a simple argument why "cosine" appeared many times in this chapter, whereas "sine" did not appear at all.

14.19 This problem relates to the structure and bands of a plane of graphene. Show that the Brillouin zone has the shape and orientation shown in the two cases in Fig. 14.9; also show that one of the K points is given by $k_x = 4\pi/3a, k_y = 0$ and find the coordinates of all six K points. Show that the bands indeed touch at all six K points.

14.20 Show that rolling of a graphene sheet to form $(n, 0)$ and (n, n) tubes leads, respectively, to the structures and BZs for the "zig-zag" and "armchair" tubes that are shown in Fig. 14.9. For the armchair tubes show that the allowed states always include the states at the K point in graphene, so that simple mapping of graphene bands always leads to the prediction of metallic bands. For the zig-zag tubes, give the conditions for which the allowed states include the graphene K point.

14.21 Show that the expressions for the force and stress theorems in tight-binding form, Eqs. (14.25) and (14.26), follow immediately from the condition that energy is minimum w.r.t. the coefficients in the wavefunctions.

14.22 Consider a heteropolar diatomic molecule with a total of two electrons. The hamiltonian is approximated by a orthogonal tight-binding model with one state per atom and hamiltonian matrix elements $H_{11} = E_1$, $H_{22} = E_2$, and $H_{12} = H_{21} = t(x)$, where x is the distance between atoms. Find the analytic expression for the ground state energy E.
(a) Calculate the force on atoms 1 and 2 directly from the derivative of the analytic expression for the energy, and also from the force theorem.
(b) Do the same for a generalized force $\mathrm{d}E/\mathrm{d}\Delta$, where $\Delta = E_1 - E_2$.

14.23 Find expressions for the valence and conduction band eigenvalues in a diamond-structure crystal at $\mathbf{k} = 0$ in terms of the matrix elements of the hamiltonian, the on-site energies E_s and E_p, and the matrix elements $H_{\mathrm{ss}\sigma}$, $H_{\mathrm{sp}\sigma}$, $H_{\mathrm{pp}\sigma}$, and $H_{\mathrm{pp}\pi}$. Do this in two steps. First, show that the eigenstates at $\mathbf{k} = 0$ are pure s or pure p. Next, use this fact to find expressions for the four eigenvalues for bonding and anti-bonding s and p states. Assuming four electrons per atom, identify the valence and conduction states and the gap between filled and empty states at

$\mathbf{k} = 0$. Find numerical values for Si using Harrison's "universal" table [344, 590] and compare with Fig. 14.6.

14.24 Project: Calculate the bands of Si using the parameters in Harrison's "universal" table [344, 590]. Construct a simple tight-binding code (or use one available on-line at the site in Ch. 24) to calculate the bands and compare these to those shown in Fig. 14.6. The valence bands should be similar but the conduction bands are quite different.

15

Localized orbitals: full calculations

Summary

As emphasized in the previous chapter, localized functions provide an intuitive description of electronic structure and bonding. This chapter is devoted to quantitative methods in which the wavefunction is expanded as a linear combination of localized atomic(-like) orbitals, such as gaussians, Slater-type orbitals, and numerical radial atomic-like orbitals. Such calculations can be very efficient; they can also be very accurate, as shown by the highly developed codes used in chemistry; and they provide the basis for creation of new methods, such as "order-N" (Ch. 23) and Green's function approaches. There is a cost, however: full self-consistent DFT calculations require specification of the basis, and the price paid for efficiency is loss of generality (in contrast to the "one basis fits all" philosophy of plane wave methods). Since details depend upon the basis, we can only describe general principles with limited examples.

It is instructive to note that there are important connections to localized muffin-tin orbitals (MTOs) (Ch. 16), the linear LMTO method (Ch. 17). This has led to an "*ab initio* tight-binding" method (Sec. 17.6) in which a minimal basis of orthogonal localized orbitals is derived from the Kohn–Sham hamiltonian.

15.1 Solution of Kohn–Sham equations in localized bases

The subject of this chapter is the class of general methods for electronic structure calculations in terms of the localized atom-centered orbitals defined in Sec. 14.1. The orbitals may literally be atomic orbitals, leading to the LCAO method or various atomic-like orbitals. These are powerful methods widely used in chemistry (see, e.g. [247, 261, 611–613]) and of increasing importance in condensed matter (see, e.g. [613–615, 617]). Unlike the tight-binding methods of the previous chapter, these methods are fully "*ab initio*," i.e. they involve no parameters and solve the full Kohn–Sham or Hartree–Fock equations in a basis of orbitals. Unlike plane waves, however, the orbitals must be chosen for the given system to be accurate and efficient, and there is a problem of "overcompleteness" if one attempts to go to convergence. Nevertheless, there is great experience in constructing appropriate orbitals, so that localized orbitals are often the basis of choice, providing crucial understanding and calculational procedures that can be both fast and accurate with careful choice of orbitals.

In constructing desirable localized basis functions, there are two (often competing) considerations: reduction of the number of basis functions and ease of computation of the needed integrals. The former consideration means each function must be well tailored to the problem, which has led to many choices; only a few examples can be considered here. These competing requirements have led to the two general classes of orbitals discussed in Secs. 15.2 and 15.4, that involve, respectively, analytic basis functions and numerical orbitals.

The goal of having a small basis leads to some overall conclusions that can be seen from general principles. The most common approach is to use atom-centered orbitals that are the products of radial functions and spherical harmonics defined in Eq. (14.8). The primary degrees of freedom are captured by a small set of l, m and radial functions the shape of which must be optimized. It is often advantageous to choose a small set of radial functions that are optimal for a given environment. However, we shall concentrate upon more general, flexible methods with a basis of several radial orbitals for each l, m channel.

Basis functions: naming conventions and examples

The common notation in the field is that multiple radial functions for the same l, m are denoted "multiple-zeta", i.e. single-ζ or "SZ," double-ζ or "DZ," triple-ζ or "TZ" for 1, 2, or 3 radial functions, *etc.* The nomenclature arises from the use of ζ to denote the range of the basis functions. There are some general guidelines for the choice of optimal radial basis functions. For example, it is well known that in a molecule or solid, the localized orbitals typically are best described by atomic-like orbitals with shorter range and larger amplitude at the nucleus than in the atom [247, 613]. This is a direct consequence of the fact that the fundamental driving force for the binding of molecules or solids is the lowering of total energy because the electrons can be closer to the nuclei without paying as much cost in kinetic energy, compared to electrons in isolated atoms. Furthermore, the long-range exponential tails of the atomic orbitals are irrelevant or incorrect in regions that overlap other atoms. Thus basis orbitals tend not to be as extended as atomic functions. Different radial functions can be generated in many ways. One of the most elegant uses the ideas of the energy derivative of the wavefunction $\dot{\psi}$ derived in Sec. 17.1. Using the same principles as invoked in the LMTO approach and in norm-conserving pseudopotentials (Sec. 11.9), the change in the wavefunction in different environments is described to linear order by a combination of ψ and $\dot{\psi}$. Thus, $\psi, \dot{\psi}, \ddot{\psi}$, *etc*., form a possible set of localized orthonormal radial functions. On the other hand, it is often essential to include longer range functions, e.g. to describe the decay of wavefunctions in the vacuum around molecules or at surfaces.

Since, the environment of an atom in a molecule or solid is not spherical, in general the basis requires higher angular momenta than the minimal basis in the atom. The first such functions are called "polarization functions," which have angular momentum l^+ one unit larger than the maximum occupied state in the atom. It is pertinent to note that it is *not* appropriate to use the atomic state of angular momentum l^+. Such a state tends to be very diffuse and not relevant to the actual change in the function in the molecule or solid. A much better choice [247, 613, 617] is the actual change in the wavefunction of angular

momentum l upon application of a weak electric field; this is a real "polarization function" that is localized and captures the essence of the lowest-order effect of the non-spherical environment. Inclusion of polarization functions in the basis is denoted by "P," e.g. "TZP" for triple-ζ with polarization functions.

The solution of the Kohn–Sham equations has exactly the same form as for the tight-binding equations of Ch. 14 except that the matrix elements must be computed explicitly and the potential must be derived self-consistently. Thus, as in all Kohn–Sham or Hartree–Fock methods, the key problem is to calculate the integrals for the matrix elements of the hamiltonian, the solution of the Poisson equation, and the generation of the potential in the self-consistency cycle. The ease with which one can do these operations is greatly affected by the choice of the basis functions, which has led to a number of methods. Furthermore, it is one of the major reasons for the development of standard sets of basis functions, as given in references such as [247, 261, 611, 612].

15.2 Analytic basis functions: gaussians

By far the most useful and used basis functions for electronic structure calculations of molecules are gaussians multiplied by polynomials, apparently first adopted by Boys [618] and expounded upon in many texts such as [247, 261, 611, 612]. The great virtue is that *all* matrix elements can be computed analytically, greatly simplifying and speeding up calculations.[1]

Gaussians have the property, illustrated in Fig. 15.1, that the product of any two gaussians is a gaussian

$$\mathrm{e}^{-\alpha|\mathbf{r}-\mathbf{R}_A|^2}\mathrm{e}^{-\beta|\mathbf{r}-\mathbf{R}_B|^2} = K_{AB}\mathrm{e}^{-\gamma|\mathbf{r}-\mathbf{R}_C|^2}, \tag{15.2}$$

where (Exercise 15.2)

$$\gamma = \alpha + \beta, \tag{15.3}$$

$$\mathbf{R}_C = \frac{\alpha\mathbf{R}_A + \beta\mathbf{R}_B}{\alpha + \beta}, \tag{15.4}$$

and

$$K_{AB} = \left[\frac{2\alpha\beta}{\pi(\alpha+\beta)}\right]^{3/4} \mathrm{e}^{-\frac{\alpha\beta}{\gamma}|\mathbf{R}_A-\mathbf{R}_B|^2}. \tag{15.5}$$

[1] The general principle that determines the usefulness of the analytic basis functions is the existence of an "expansion theorem" for the orbital centered on one site in terms of the basis functions on neighboring sites

$$\chi_\alpha(\mathbf{r}-\mathbf{R}) = \sum_{\alpha'} B_{\alpha\alpha'}(\mathbf{R},\mathbf{R}')\,\chi_{\alpha'}(\mathbf{r}-\mathbf{R}'), \tag{15.1}$$

which greatly facilitates evaluation of the integrals. Examples of functions that possess this property are polynomials multiplied by gaussians ($r^\beta\ \mathrm{e}^{-\alpha r^2}$), Slater-type orbitals ($r^\beta\ \mathrm{e}^{-\alpha r}$), and spherical Bessel, Neumann, and Hankel functions. The advantages of the expansion theorem are emphasized in Chs. 16 and 17, where the expansion formulas for spherical Bessel, Neumann, and Hankel functions are crucial to the formulation of the KKR and (L)MTO methods.

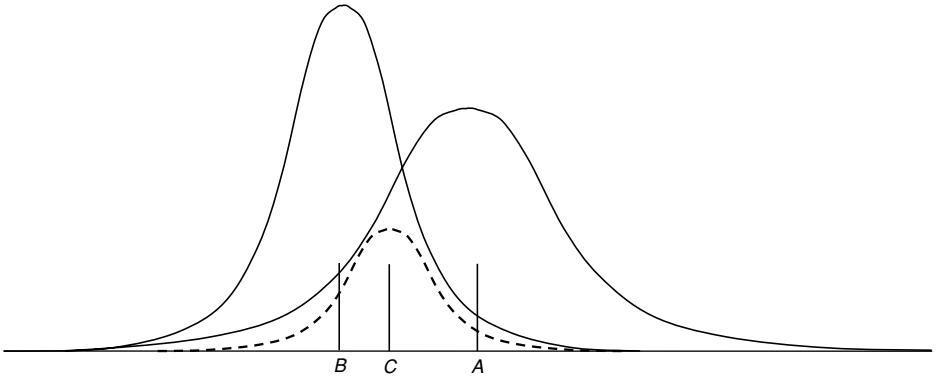


Figure 15.1. The overlap of two gaussians is another gaussian, with center, width, and weight given by (15.2)–(15.5). From this basic result, matrix elements of kinetic energy and polynomials can be constructed by simple procedures (see text).

Analytic relations can be found for gaussians multiplied by any polynomial of the radius by differentiating the above formulas using the fact that $(\mathrm{d}/\mathrm{d}x)\mathrm{e}^{x^2} = 2x\mathrm{e}^{x^2}$ and so forth up to any power. Similarly, it is straightforward to evaluate the laplacian applied to any gaussian multiplied by a polynomial. Thus for a basis set consisting of gaussians times polynomials (and spherical harmonics) centered at any site, *all multi-center integrals can be evaluated analytically*.

The expressions for the overlap and kinetic energy matrix elements can be easily derived (Exercise 15.3). The charge density $|\psi(\mathbf{r})|^2$, where $\psi(\mathbf{r})$ is a sum of such basis functions, is also readily expressed as a sum of gaussians. Potential matrix elements depend upon the form of the potential. Two cases are of particular interest: if the potential is a sum of gaussians, the matrix elements are simply a sum of analytic three-center integrals; in addition, matrix elements of the Coulomb interaction with the nuclei and between the electrons can be computed analytically in terms of "Boys functions." Since these are the only integrals needed in Hartree–Fock calculations, gaussians have long enjoyed their status as the basis of choice. (For density functional theory some of the advantage is lost since the exchange–correlation potential is a non-linear functional of the density that is not directly expressible as gaussians even if the density is a sum of gaussians.)

Detailed expressions for the total energy, *etc.*, are not given here, since they can be found from the expressions in Sec. 15.5. However, here is one major point. The Hartree–Fock equations can be written directly in terms of four center integrals, since the Coulomb matrix elements involve four orbitals. This is an effective approach for small systems; however, it scales as N^4_{orbital}. For large systems, especially for the Kohn–Sham equations, it is more effective to generate the total potential due to all occupied orbitals and to evaluate the matrix elements using grids [262].

The downside of gaussians is that they are eigenstates of a harmonic oscillator, which has little in common with potentials in a material made of atoms. For this reason there is great use of standardized "slater type orbitals" (STOs), which are sums of gaussians

with fixed coefficients [247,261,611,612]. A STO retains all the nice features of gaussians while at the same time having a form closer to an atomic-type orbital. In essence, a STO is a radial orbital that is expanded in a convenient basis; however, instead of allowing all the coefficients of the gaussians to vary, standard optimized sets have been generated that can be used to compare calculations with identical bases and to achieve different levels of accuracy with different bases.

15.3 Gaussian methods: ground state and excitation energies

Electronic structure calculations using gaussians and STOs are far too numerous to attempt to summarize here and they are covered in great detail elsewhere, for example, in references such as [247,261,611,612]. The methods are so successful that there are many commercially available codes adapted to molecular systems. In addition, gaussian bases can be efficient for the periodic systems that are more relevant to the subject matter treated here. In particular, gaussian methods capable of Hartree–Fock calculations for crystals have a special role in current developments of electronic structure. Many recent calculations have used the CRYSTAL code [614, 615] for which further information is available at the website given in Ch. 24.

Of course, gaussian bases can be used to calculate ground state properties including energies, forces, atomic geometries, and other properties using either Hartree–Fock or density functional theory methods. There are many examples [247, 261, 611, 612] and a recent review [619] brings out the point that gaussian methods are very efficient and can be applied to complex systems. For example, linear-scaling "order-N" methods have been developed and applied to problems such as the structures of large RNA molecules [619] (see caption of Fig. 23.10). Calculations of the energies of the Ge crystal and its surfaces are illustrated by the work shown in Figs. 2.25 and 15.2.

Perhaps the most salient advantage of gaussian bases is that Coulomb integrals can be computed analytically. It is for this reason that gaussians have been the workhorse of Hartree–Fock calculations and it for this reason that gaussians can play a special role in any problem involving Coulomb integrals. This includes Hartree–Fock, "exact exchange" (EXX; Sec. 8.7), "hybrid density functional theory" (Sec. 8.8), and the many-body "GW" calculations, all of which involve computation of exchange integrals. These methods are particularly relevant for the key problem in density functional theory calculations: accurate, robust prediction of the band gap and other electronic excitations. Therefore, gaussians are in many ways the basis of choice for this important area of electronic structure.

A relevant example is given in Fig. 2.25 which shows the bands for Ge calculated with a gaussian basis, both using the standard local density approximation (LDA) and the many-body "GW" quasiparticle approach. Both calculations agree well with plane wave pseudopotential and LMTO results, e.g. as shown by comparison of the LDA bands in Fig. 2.25 with those shown in Fig. 17.9. Figure 2.25 illustrates the (in)famous zero-band gap for Ge in the LDA that is greatly improved in GW results. The flexibility of the gaussian approach is illustrated by application of the same methods to more complex systems, e.g. to electronic bands for the dimerized Ge (1 0 0) surface [587] shown in Fig. 15.2, which

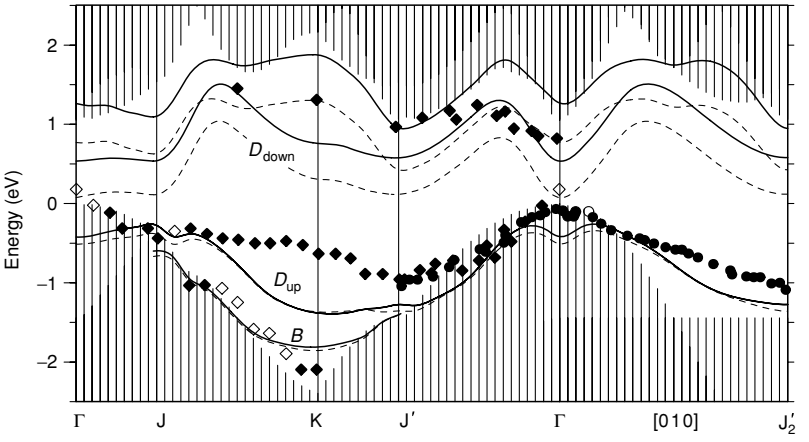


Figure 15.2. Bands for buckled dimer in Ge [587] for a structure that is similar to the Si surface shown in Fig. 13.7. The bulk states are denoted by shaded areas, and surface states in the gap by lines. Bands denoted D and B are surface states associated with the surface "dangling" bonds and with the "back" bond between the first and second layers. The dashed lines are for the LDA (showing the zero-gap problem) and the solid lines are calculated in the GW approximation. Dots indicate experimental results referenced in [587]. From [587].

compares the LDA and GW quasiparticle bands with experimental results. Note that the occupied LDA valence bands agree well with GW calculations, including surface states, but the LDA conduction bands are too low, exhibiting once again the zero-gap problem.

Among the most promising approaches for improved excitation energies within density functional theory are EXX and hybrid functional approaches (Secs. 8.7 and 8.8). For example, calculations of bands have been done using the CRYSTAL code for materials such as La_2CuO_4 [620] and UO_2 [621], where the correct insulating antiferromagnetic state is found unlike the usual LDAs or GGAs that predict a metallic non-magnetic state. A simple example is Si for which the band structure [622] resulting from the hybrid "B3LYP" functional[2] is shown in Fig. 15.3. The figure shows good agreement with the points calculated from GW [219] and quantum Monte Carlo (QMC) [623] methods, which in turn are close to observed energies. For example, the lowest gap at the X point is found to be 1.57 eV compared with 1.51 (QMC), 1.43 (GW), and 1.25 eV from empirical potential results [105] fitted to experiment. This can be compared with the Hartree–Fock overestimate of 5.3 eV and LDA underestimate of 0.63 eV. Along with other work, e.g. the EXX calculations [223] summarized in Fig. 2.26, these results show the promise of methods with improved treatment of exchange.[3]

[2] The B3LYP functional derived by Becke [404] was adjusted to fit data on molecules; it is given by Eq. (8.37) with $a_0 = 0.2$, $a_x = 0.72$, and $a_c = 0.81$, with the LYP correlation functional.

[3] Care needs to be taken in comparing methods with one another and with experiment due to basis set limitations, which is a practical consideration in local orbital calculations in solids. The higher bands of Si in Fig. 15.3 are qualitatively incorrect due to the limited basis set. In addition, features such as the energy of the lowest bands at X are known to be very sensitive to the basis and require d states as discussed in Sec. 14.6.

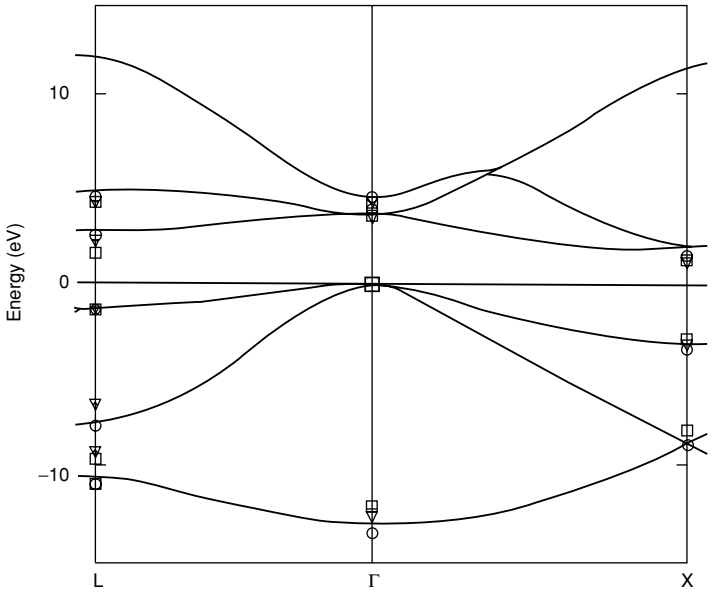


Figure 15.3. Bands of Si calculated [622] using a gaussian basis (CRYSTAL code [614, 615]) and the "B3LYP" hybrid functional [404] (see text). As expected for a hybrid functional the results are intermediate between Hartree–Fock and LDA, and the results are in general agreement with GW [219] (squares), QMC [623] (circles), and experimental (triangles) energies. The curves can also be compared to the fitted tight-binding bands in Fig. 14.6. The values for lowest gap at the X point given in the text illustrate the improvement over LDA or Hartree–Fock. Adapted from [622].

15.4 Numerical orbitals

Efficient algorithms using numerical orbitals can be constructed for either all-electron [624, 625] or pseudopotential methods [616, 617]. Construction of the orbitals is very similar in either case. However, evaluation of the matrix elements in a local orbital calculation may be different. Since pseudopotentials and pseudofunctions are smooth, it is possible to carry out many integrals directly on a grid. For all-electron states, on the other hand, it is essential to have a method that accurately integrates the region around each nucleus where the wavefunctions vary rapidly.

Construction of orbitals

Localized orbitals can be constructed from atomic-like programs with spherically symmetric potentials. It is possible to use the atomic orbitals themselves, but their long-range tails are not desirable. Since they are not really appropriate in a solid or molecule, the tails actually decrease the accuracy of the calculations and at the same time introduce troublesome long-range terms. It is more desirable to define shorter range, more "compressed" orbitals that are better suited to the final application. The effects are relevant primarily for the valence states since the core states are localized and little affected by boundary conditions.

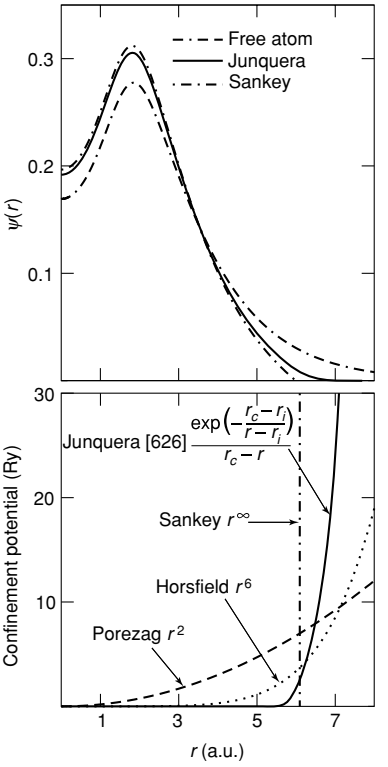


Figure 15.4. Confining potentials and pseudo wavefunctions for numerical basis functions of Mg. The orbital is the solution with an atomic potential modified by the addition of a confining potential as shown at the bottom for various choices. From Junquera [626], where the citations to other potentials are given.

Many different procedures have been proposed for generation of compressed, short-range orbitals. Each involves modification of the atomic potential so that it is strongly repulsive at large distance. This makes the orbitals confined, which is more realistic and, in addition, has advantages for the calculation since fewer matrix elements need be included. However, such confined orbitals may not be sufficient, especially at surfaces, since they cut off the charge density in the vacuum in an unphysical way. Examples of types of potentials for generating orbitals that have been used for calculations on solids include those shown in Fig. 15.4. The orbitals [620] constructed by the hard-wall confining potential have the advantage that they are strictly localized, but the disadvantage that the second derivative has a divergence that makes a finite contribution to kinetic matrix elements. The other confining potentials are constructed to have chosen degrees of confinement versus smoothness.

Integrals involving the orbitals

Many of the needed integrals depend only upon the positions of the atoms. Efficiency is not a premium for these terms because they can be calculated in advance and used later. This

includes the overlap, kinetic energy matrix element, and non-local pseudopotential terms in Eqs. (14.2) and (14.1). The first two are two-center terms that are functions only of the distance R between the centers for the case where the angular momentum axis is along the line joining the centers. Rotation to treat the general case is given in Sec. 14.2. An effective procedure is to calculate the values at many discrete values of R and interpolate to find the matrix elements during a calculation. Thus we can view these matrix elements as known two-body functions,

$$S_{m,m'}(R) \;=\; \int \mathrm{d}\mathbf{r}\; \chi^*_m(\mathbf{r})\chi_{m'}(\mathbf{r}-\mathbf{R}), \tag{15.6}$$

and

$$T_{m,m'}(R) \;=\; \int \mathrm{d}\mathbf{r}\; \chi^*_m(\mathbf{r})\frac{1}{2}\nabla^2\chi_{m'}(\mathbf{r}-\mathbf{R}), \tag{15.7}$$

where m denotes the atom type and orbital on that atom.

Potential terms are not so easy to express since they involve three centers (wavefunctions on two centers and the potential on a third). However, non-local pseudopotential terms have special features that can be used to advantage. First, non-local terms are fixed for each atom and never change during a calculation. Second, if a separable form (the Kleinman–Bylander form of Sec. 11.8, the "ultrasoft" pseudopotential of Sec. 11.10, or the PAW in Sec. 11.11) is used, then all three-center terms factorize into sums of products of two-center terms. These can be tabulated in advance as a function of distance.

Thus we are left with the problem of treating the matrix elements of the local potentials. This includes the full Kohn–Sham potential or any local parts of the potential. In a pseudopotential method, this can be treated exactly as is done with plane waves – sums on a grid. All operations on the grid, such as finding the density and exchange–correlation functions, can be done exactly as for plane waves. The only differences are that the wavefunctions in the local orbital basis must be transferred to the grid and the integrals are done by summing on the grid points to find the matrix elements

$$V^{\mathrm{local}}_{m,m'}(\mathbf{T}) \;=\; \int \mathrm{d}\mathbf{r}\; \chi^*_m(\mathbf{r}-\tau_m)V^{\mathrm{local}}(\mathbf{r})\chi_{m'}(\mathbf{r}-(\tau_{m'}+\mathbf{T})). \tag{15.8}$$

If the wavefunctions are smooth, for example, with pseudopotentials, the integral in (15.8) can be carried out on a regular grid [617]. In an all-electron method, the integration must be done carefully around each nucleus since the wavefunctions vary rapidly. One approach [627] is to use the "muffin-tin" partitioning of space illustrated in Fig. 16.1, in which case the integration can be done on radial grids around each nucleus and uniform grids in the interstitial regions very much like the procedures in augmented methods. Another general approach is to break up the integral into overlapping domains using functions $\alpha_i(\mathbf{r})$ that together cover all space. If we define the normalized weight functions $w_i(\mathbf{r}) = \alpha_i(\mathbf{r})/\sum_j \alpha_j(\mathbf{r})$, any integral can be written as

$$\int \mathrm{d}\mathbf{r}\, f(\mathbf{r}) = \sum_i \int \mathrm{d}\mathbf{r}\; w_i(\mathbf{r}) f(\mathbf{r}). \tag{15.9}$$

Each integral on the right-hand side can be done on a different grid; in particular, radial grids can be used around each atom to deal with the rapid variations near the nucleus [624, 625, 628].

15.5 Localized orbitals: total energy, force, and stress

Any of the expressions for the total energy in Sec. 9.2 can be used in a local orbital basis. The expressions can be written most compactly in terms of the density matrix $\rho_{mm'}$. If the eigenvectors are $\psi_i(\mathbf{r}) = \sum_m c_{im}\chi_m(\mathbf{r} - \mathbf{R}_m)$, then $\rho_{mm'}$ is given by

$$\rho_{mm'} = \sum_i f(\varepsilon_i) c^*_{im} c_{im'}. \tag{15.10}$$

In the present case, $\chi_m(\mathbf{r} - \mathbf{R}_m)$ are the localized basis functions, where m denotes both orbital α and site I, and $\mathbf{R}_m$ denotes the position of atom $\mathbf{R}_I$ on which orbital m is centered.

All quantities in the total energy expressions can be cast in terms of $\rho_{mm'}$. In particular, the sum of eigenvalues in (9.4) and (9.6) is given by

$$E_s = \sum_{i=1}^{N} \varepsilon_i f(\varepsilon_i) = \sum_{mm'} \rho_{mm'} [H_{KS}]_{mm'}, \tag{15.11}$$

where $[H_{KS}]_{mm'}$ denotes matrix elements of the Kohn–Sham effective hamiltonian, and the spatial electron density is given by

$$n(\mathbf{r}) = \sum_i f(\varepsilon_i) |\psi_i(\mathbf{r})|^2 = \sum_{mm'} \rho_{mm'} \chi^*_m(\mathbf{r} - \mathbf{R}_m) \chi_{m'}(\mathbf{r} - \mathbf{R}_{m'}). \tag{15.12}$$

These are sufficient to determine the energy in any of the various expressions in the Kohn–Sham approach.

It is useful to be more specific because certain forms are advantageous in a localized basis [617, 629]. In particular, calculation of forces, stress, and Coulomb energies can be facilitated by the choice of the energy functional. Calculation of Coulomb terms (including the local pseudopotential term) can be done conveniently by grouping terms due to each nucleus (or ion) I with a compensating localized, spherical charge $n_I(\mathbf{r})$, as is described in Sec. F.4. Any convenient localized electron density can be used; if the orbitals are constructed from an atomic-like calculation, an obvious choice is the density resulting from these orbitals. The needed expressions for the Coulomb terms are given in Secs. F.3 and F.4 in terms of the electron density written as

$$n(\mathbf{r}) \equiv \sum_I n_I(\mathbf{r}) + \delta n(\mathbf{r}), \tag{15.13}$$

which is equivalent to Eq. (F.17).

Inserting the expression for the Coulomb and local potential terms, Eqs. (F.18) and (F.19), into that for total energy (F.15), a convenient form[4] for the total energy is [617, 629]

$$E_{tot} = \sum_{mm'} \rho_{mm'} \left[T_{mm'} + V^{NL}_{mm'} \right] + \sum_{I<J} U^{\mathrm{NA}}_{IJ}(|\mathbf{R}_I - \mathbf{R}_J|) + \sum_I U^{\mathrm{NA}}_I + \int \mathrm{d}\mathbf{r} V^{\mathrm{NA}}(\mathbf{r}) \delta n(\mathbf{r}) + \frac{1}{2} \int \mathrm{d}\mathbf{r} \delta V_{\mathrm{Hartree}}(\mathbf{r}) \delta n(\mathbf{r}) + \int \mathrm{d}\mathbf{r} \epsilon_{\mathrm{xc}}(\mathbf{r}) n(\mathbf{r}). \tag{15.14}$$

Here U^{NA}_I denotes the potential energy of the neutral atom, $U^{\mathrm{NA}}_{IJ}(|\mathbf{R}_I - \mathbf{R}_J|)$ is the classical interaction of two neutral atom densities, and the two terms involving $\delta n(\mathbf{r})$ are the first- and second-order changes in the potential energy due to the changes in the density in the solid. As discussed in Sec. F.4, the local part of the pseudopotential is included in U^{NA}_I and $V^{\mathrm{NA}}(\mathbf{r})$. Since the exchange–correlation energy is a non-linear functional of ρ, it cannot be divided in the same way and is left as the usual expression. Similarly, it is more convenient to express the first term that involves the density matrix directly as shown.

Force and stress

Expression (15.14) is particularly appropriate for calculation of derivatives. The solution of the self-consistent Kohn–Sham equations leads to a minimization of the total energy with respect to the coefficients c_{im} in the wavefunction. Therefore the derivative of E_{tot} with respect to $\rho_{mm'}$ vanishes. Thus the derivatives of the first three terms in (15.14) can be considered as functions of the distances between the atoms, since $T_{mm'}$ and $V^{NL}_{mm'}$ are functions only of distances, as shown in Sec. 15.4. It is straightforward [617, 629] to express their contribution to the force on atom I in terms of derivatives that involve its position relative to other atoms $\mathbf{R}_I - \mathbf{R}_J$. The contribution of such two-body terms to the stress can be derived from the analysis of Sec. G.2. The fourth term is a constant that has no effect.

The fifth term involving V^{NA} contributes a term of exactly the same form

$$-\int \mathrm{d}\mathbf{r} \frac{\partial V^{\mathrm{NA}}}{\partial \mathbf{R}_I} \delta n(\mathbf{r}), \tag{15.15}$$

since the "NA" terms move rigidly with the atom. For the stress this term can be included by scaling the density.

The last three terms all involve n or δn. They contribute to stress as in other formulas involving the density; however, their contributions to the force would vanish if the density had no explicit dependence on the atom positions. Indeed, this is the case for plane waves where the basis is not related to atom positions. However, the local orbitals are displaced with the atom; if the basis is not complete (Exercise 15.4), there are "Pulay corrections" due to the fact that the density changes to first order. These terms can be calculated using

[4] The term δE_{II} in (F.15) is omitted here. It should be added to cancel an unphysical term if the extent of the smeared ion cores is allowed to be so large that they overlap.

the fact that they arise from the change in $n(\mathbf{r})$ *for fixed* $\rho_{\alpha\beta}$. For any functional $F[n]$, the derivative is

$$-\frac{\partial F[n]}{\partial \mathbf{R}_I} = -\int \mathrm{d}\mathbf{r} \frac{\delta F[n]}{\delta n(\mathbf{r})} \frac{\partial n(\mathbf{r})}{\partial \mathbf{R}_I}, \tag{15.16}$$

where (here $m \to \alpha$, I to clarify the role of the position R_I)

$$\frac{\partial n(\mathbf{r})}{\partial \mathbf{R}_I} = \sum_{\alpha,\beta J} \left[\rho_{\alpha,\beta J} \frac{\partial \chi_\alpha^*(\mathbf{r} - \mathbf{R}_I)}{\partial \mathbf{R}_I} \chi_\beta(\mathbf{r} - \mathbf{R}_J) + \text{c.c.} \right]. \tag{15.17}$$

Since the functions χ_α are localized, the sum can be restricted to only include atoms J within some range of I.

15.6 Applications of numerical local orbitals

The primary application of numerical local orbitals is for density functional calculations, using either all-electron [624, 625] or pseudopotential methods [616, 617]. They can be applied to arbitrary molecules and crystals, e.g. the example of the ferroelectric distortion in $BaTiO_3$ illustrated in Sec. 2.5. Local orbitals are particularly efficient for complicated systems with many atoms per cell or with vacuum regions, where plane waves become expensive to use. For example, linear scaling calculations based upon numerical orbitals have been developed and applied to problems such as quantum molecular dynamics, structural relaxation and electronic states of large DNA molecules, as illustrated in Fig. 23.10. Another example is C_{60} bound to Si surfaces, with comparison of theoretical and experimental STM images shown in Fig. 2.20. This is a very large problem requiring detailed atomic-scale description of the Si bulk and surface, the C_{60} molecules, and the (unknown) binding mechanisms of the C_{60} to the surface. Further investigation has been done [630] using the SIESTA code, with Fig. 15.5 showing results of the reconstruction of the surface to bind the molecule in two different ways. To simulate the surface at terraces of the 7×7 reconstructed surface, a 2×2 adatom surface reconstruction was used. The structure of several possible adsorption configurations was optimized using the forces from the force theorem, finding good candidates for the two different adsorption states observed experimentally. While the C_{60} molecule remains nearly spherical, the silicon substrate is quite soft, especially the adatoms, which move substantially to form extra C–Si bonds at the expense of breaking Si–Si bonds. Structural relaxation has a large effect on the adsorption energies, which strongly depend on the adsorption configuration, and depend much less upon the charge transfer.

15.7 Green's function and recursion methods

One of the primary uses of local orbitals is Green's function-type methods that take advantage of the locality. For example, self-consistent calculations of localized defects in semiconductors were calculated using Green's functions to treat the infinite medium around the defect [631, 632]. Similarly, adatoms on metal surfaces have been treated using Green's function gaussian orbital codes [633, 634]. Although Green's function methods are widely

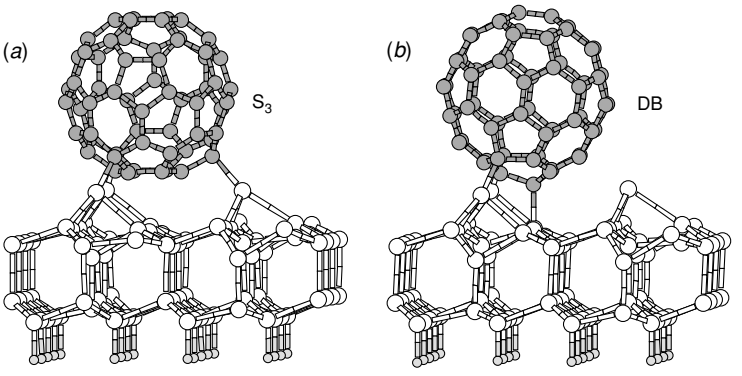


Figure 15.5. Atomic scale bonding of a C_{60} molecule to a Si (111) surface [630] calculated using the SIESTA code [617], which employs numerical local orbitals. The surface was modelled as a 2×2 adatom reconstruction and several possible adsorption configurations were optimized, leading to the two geometries labeled S_3 and DB which are proposed as candidates for the two different adsorption states observed experimentally. The C_{60} molecule remains nearly rigid, but the Si substrate deforms, especially the adatoms, which move substantially, breaking Si–Si bonds to form C–Si bonds. From [630].

used with simpler tight-binding hamiltonians (Ch. 23), these approaches have not been extensively used in fully self-consistent density functional theory calculations because of difficulties in calculation of the hamiltonian. Most self-consistent work has involved "supercell" methods (Sec. 13.4) and quantum molecular dynamics (Ch. 18) to treat such problems with periodic boundary conditions.

With the re-emergence of interest in local non-periodic methods and the advent of linear scaling methods, there is now renewed interest in Green's function approaches. Indeed, many approaches described in Ch. 23 are variations of Green's function methods that utilize localized functions.

15.8 Mixed basis

Mixed basis methods utilize a combination of localized and delocalized bases, e.g. the appealing choice of gaussians and plane waves [635]. This gives the possibility of two widely used methods, plane waves and gaussians, and any linear combination. The hallmark of a mixed basis approach is that both bases are used in the same region of space and the equations are expressed in terms of the usual overlap and hamiltonian matrix elements. The motivations have much in common with ultrasoft pseudopotential and projector augmented wave methods, which also include additive localized functions that augment the smooth functions near the nuclei. However, those methods transform the problem so that one needs to solve equations that involve only the smooth plane waves and the localized functions do not appear as explicit basis functions. This is a great advantage that allows much of the additional work to be done once and for all on the atomic reference state, simplifying the final calculation.

A different use of the mixed basis idea is to utilize plane waves and Gaussians in *different spatial directions* [636]. This can be used to advantage, for example, for surfaces that are periodic in two directions but not in the third. Thus the basis functions become

$$\chi_{\mathbf{k},m,n}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}(\mathbf{k}+\mathbf{G}_m)\cdot\mathbf{r}}\mathrm{e}^{-\alpha|z-z_n|^2}, \tag{15.18}$$

which denotes a Fourier component $\mathbf{k} + \mathbf{G}_m$ in the x, y plane of the surface and multiplied by a gaussian centered at position z_n. A surface or interface can thus be represented by "layer" wavefunctions that are extended in the plane and centered on atomic layers.

SELECT FURTHER READING

Methods in computational chemistry:

McWeeny, R. D. and Sutcliffe, B. T. *Methods of Molecular Quantum Mechanics, second edition*, Academic Press, New York, 1976.

Szabo, A. and Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory (Unabridged reprinting of 1989 version)*, Dover, Mineola, New York, 1996.

Jensen, F. *An Introduction to Computational Chemistry*, John Wiley and Sons, New York, 1998.

Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, Wiley, New York, 2002.

Books, papers, and other material on localized orbital methods:

Eschrig, H. *Optimized LCAO Methods*, Springer, Berlin, 1987.

Orlando, R. Dovesi, R. Roetti, C. and Saunders, V. R. "*Ab initio* Hartree-Fock calculations for periodic compounds: application to semiconductors," *J. Phys. Condens. Matter* 2:7769, 1990.

Saunders, V. R. Dovesi, R. Roetti, C. Causa, M. Harrison, N. M. Orlando, R. and Zicovich-Wilson, C. M. CRYSTAL 98 User's Manual (University of Torino, Torino). See http://www.theochem.unito.it/, 2003.

Delley, B. "From molecules to solids with the DMol3 approach," *J. Chem. Phys.* 113:7756–7764, 2000.

Soler, J. M. Artacho, E. Gale, J. Garcia, A. Junquera, J. Ordejon, P. and Sanchez-Portal, D. "The SIESTA method for *ab intio* order-N materials simulations," *J. Phys. : Condens. Matter* 14:2745–2779, 2002.

Exercises

15.1 The on-line site in Ch. 24 has links to local orbital codes and many examples and tutorials for Hartree–Fock and Kohn–Sham calculations on atoms, molecules, and crystals.

15.2 Show that the product of two gaussians is a gaussian as in Eq. (15.2), and derive the expressions for the coefficients in the product gaussian Eqs. (15.3)–15.5.

15.3 Find the analytic formula for the kinetic energy matrix element between gaussian basis functions with spreads α and β and separated by displacement $\mathbf{R}$.
(a) First consider only simple gaussians with s symmetry and not multiplied by powers of the radius.

(b) Then show that the formulas can be generalized to any l, m and power r^p by taking appropriate derivatives of expressions derived in (a). You do not need to work out all the detailed formulas, which can be found in texts.

15.4 Derive Eq. (15.17) using a chain rule and show that the right-hand side vanishes if the basis is complete. Hint: Use the completeness relation.

15.5 Construct a simple computer program for a gaussian s band in one dimension. This entails calculating the overlap and hamiltonian matrix elements that are analytic if we assume the potential is also a sum of gaussians centered on each atom. Vary the band shapes from nearly-nearest-neighbor tight-binding-like given in (14.12) to nearly-free-electron-like.

15.6 Use the results for the eigenvectors from Exercise 15.5 to construct Wannier functions. Construct the atom-centered "maximally projected" form defined in Sec. 21.2 with the phase (sign) chosen to maximize the function on the central atom.
(a) Show that the function has positive and negative values (a plot is best) and it is longer range than the gaussian basis function.
(b) With a careful fit to the long-range behavior (the log of the absolute value) of the Wannier function, show it decays exponentially as a function of distance as claimed in Sec. 21.2.

16

Augmented functions: APW, KKR, MTO

Summary

Augmentation provides a method of constructing a basis that is in some ways the "best of both worlds:" the smoothly varying parts of the wavefunctions between the atoms represented by plane waves or other smoothly varying functions, and the rapidly varying parts near the nuclei represented as radial functions times spherical harmonics inside a sphere around each nucleus. The solution of the equations becomes a problem of matching the functions at the sphere boundary. The original approach is the augmented plane wave (APW) method of Slater, which leads to equations similar to the pseudopotential and OPW equations, but with matrix elements of a more complicated, energy-dependent potential operator. The disadvantage of augmentation is that the matching conditions lead to non-linear equations, which has led to the now widely used linearized methods described in Ch. 17. The KKR method is a multiple-scattering Green's function approach that yields directly local quantities. The muffin-tin orbital (MTO) approach reformulates the KKR method, leading to physically meaningful descriptions of the electronic bands in terms of a small basis of localized, augmented functions.

16.1 Augmented plane waves (APWs) and "muffin tins"

The augmented plane wave (APW) method, introduced by Slater [54] in 1937, expands the eigenstates of an independent-particle Schrödinger equation in terms of basis functions, each of which is represented differently in the two characteristic regions illustrated in Fig. 16.1. In the region around each atom the potential is similar to the potential of the atom and the solution for the wavefunction is represented in a form appropriate to the central region of an atom. In the interstitial region between atoms the potential is smooth and the wavefunction is represented in a form appropriate to smooth variations coupling the atomic-like regions.

If the total effective potential is approximated as spherically symmetric $V_{\rm eff}(\mathbf{r}) \rightarrow V_{\rm eff}(r)$ within each sphere, and constant $V_{\rm eff}(\mathbf{r}) \rightarrow V_0$ in the interstitial region, it is termed a "muffin-tin potential." This approximation is very appropriate for many problems and allows for dramatic simplifications, since the wavefunctions can be represented in terms of the eigenstates

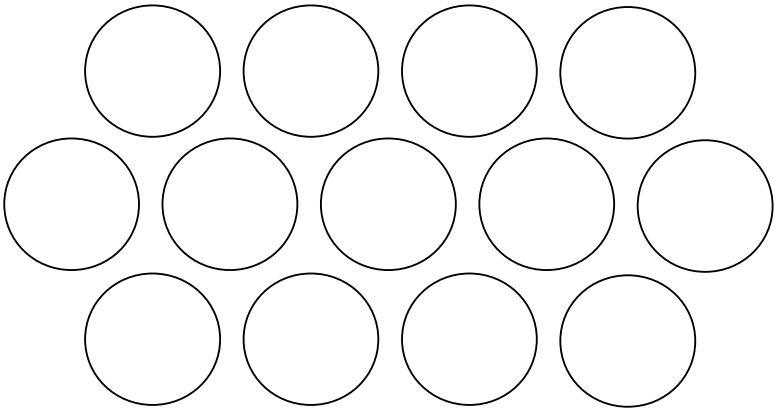


Figure 16.1. The "muffin-tin" division of space into intra-atomic spheres of radius S, and interstitial regions. This is the basis for representing wavefunctions differently in the different regions used in all augmented formulations. The muffin-tin potential approximates the potential in the two regions, but the division into spheres and interstitial regions is more general and can be applied to any potential. (The picturesque name derives from the fact that the figure looks like a pan for cooking muffins.)

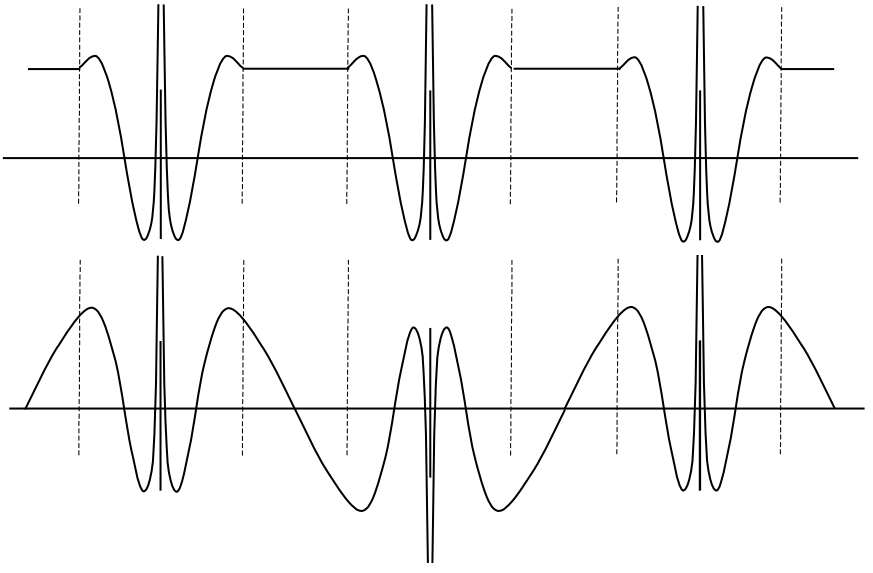


Figure 16.2. Schematic representation of the APW basis functions for $\mathbf{k} = 0$ (top) and the zone boundary (real part shown in bottom panel). The sphere boundaries are represented by the vertical dashed lines. In the interstitial region, each APW is a single plane wave. Inside each sphere the APW is a solution of the radial equation, with the boundary condition that it match the plane wave in value. A single APW has a discontinuous derivative at the boundaries. The solution for the eigenstate minimizes the final discontinuity in the derivative, leading to Bloch states like those illustrated in Fig. 4.11.

in each region, i.e. spherical harmonics around each atom and plane waves in between. The entire problem is recast into a matching or boundary condition problem. It is most instructive to first discuss the APW approach for such idealized muffin-tin potentials; however, we emphasize that the Schrödinger equation with a general potential (and the full Poisson equation) can be solved using the APW basis. Generalization of the equations to "full potential" problems is discussed in Ch. 17 after linearization is introduced.

In many ways, the APW approach brings together the "best of both worlds" with the ability to treat highly localized atomic-like states (e.g. core states) using atomic-like spherical functions and delocalized states using delocalized plane waves – the Bloch wavefunctions for the crystal illustrated in Fig. 4.11 are solved separately in the two regions in terms of the APW basis functions, defined in Eq. (16.2) and illustrated in Fig. 16.2). The disadvantage is the difficulty of matching the functions and solving the resulting non-linear equations in this basis. We will first describe the original non-linear methods which illustrate many of the main ideas and the relations to other methods, OPW, pseudopotential, and KKR, etc., through their common features of describing the valence states in terms of the scattering properties of the atoms, which in turn are determined by the phase shifts. A separate chapter (Ch. 17) is devoted to linearized methods because of their conceptual and practical importance in transforming the augmented methods into more useful forms.

Just as in the plane wave (see Eq. (12.11)) or OPW (see Eq. (11.1)) method , each Bloch function $\psi_{i,\mathbf{k}}(\mathbf{r})$ is expanded in a set of basis functions labeled by the reciprocal lattice vectors $\mathbf{G}_m$, $m = 1, 2, \ldots,$

$$\psi_{i,\mathbf{k}}(\mathbf{r}) = \sum_m c_{i,m}(\mathbf{k}) \chi^{\mathrm{APW}}_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}). \tag{16.1}$$

However, in the APW method each basis function $\chi_{\mathbf{k}+\mathbf{G}}(\mathbf{r})$ is represented as a single plane wave only in the interstitial region between atoms, and within a sphere of radius S around each atom the function is represented in spherical harmonics:

$$\chi^{\mathrm{APW}}_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}) = \begin{cases} \exp(\mathrm{i}(\mathbf{k}+\mathbf{G}_m)\cdot\mathbf{r}), & r > S, \\ \sum_{Ls} C_{Ls}(\mathbf{k}+\mathbf{G}_m)\psi_{Ls}(\varepsilon, \mathbf{r}), & r < S, \end{cases} \tag{16.2}$$

where the compact notation for the wavefunction,[1]

$$\psi_{Ls}(\varepsilon, \mathbf{r}) = i^l Y_L(\hat{r})\psi_{ls}(\varepsilon, r), \tag{16.3}$$

is introduced to simplify the notation. The angular momentum is indicated by upper case $L \equiv l, m_l$, $Y_L(\hat{r}) \equiv Y_{l,m_l}(\theta, \phi)$ are spherical harmonics, with r and $\hat{r}$ referred to an origin τ_s for each atom s in the unit cell. The function $\psi_{ls}(\varepsilon, r)$ is a solution of the radial Schrödinger

[1] The factor of i^l is introduced to simplify the coefficients that result when matching plane waves that have the factor i^l; see expansion (16.5).

equation regular at the origin at energy ε, i.e. it satisfies (10.12), written here keeping factors of $\hbar$ and m_e[2]

$$\left[\frac{\hbar^2}{2m_e}\left(-\frac{\mathrm{d}^2}{\mathrm{d}r^2}+\frac{l(l+1)}{r^2}\right)+V_s(r)-\varepsilon\right] r\psi_{ls}(r)=0. \tag{16.4}$$

It is important here that ε is a variable and need not be an eigenvalue. Schematic forms of two $\chi^{\mathrm{APW}}_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r})$ are shown along a line through the atoms in Fig. 16.2.

The coefficients $C_{Ls}(\mathbf{k}+\mathbf{G}_m)$ are obtained by requiring the waves to match at the surface of the muffin-tin spheres, i.e. that the phase shifts match as described in scattering theory, Sec. J.1. Using the expansion given in (J.1),

$$\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{r}}=4\pi\sum_L i^l\; j_l(qr)\; Y_L^*(\hat{\mathbf{q}})\; Y_L(\hat{\mathbf{r}}), \tag{16.5}$$

where $j_l(qr)$ are spherical Bessel functions, it follows that χ^{APW} is continuous at the sphere boundary if

$$C_{Ls}(\mathbf{K}_m)=4\pi\;\mathrm{e}^{\mathrm{i}\mathbf{K}_m\cdot\tau_s}\;j_l(|\mathbf{K}_m|S_s)\;\frac{Y_L^*(\hat{\mathbf{K}}_m)}{\psi_{ls}(\varepsilon,S_s)}, \tag{16.6}$$

where $\mathbf{K}_m=\mathbf{k}+\mathbf{G}_m$. An APW is, by construction, discontinuous in slope on the muffin-tin boundary (see Fig. 16.2), a fact that must be taken into account when applying the kinetic energy operator.

Within the APW basis, the secular equation can be written

$$\sum_m\left\{\langle m'|H-\varepsilon_{i,\mathbf{k}}|m\rangle+\langle m'|H^S|m\rangle\right\}c_{i,m}(\mathbf{k})=0, \tag{16.7}$$

where

$$\langle m'|H-\varepsilon_n(\mathbf{k})|m\rangle=\int_{\mathrm{cell}}\mathrm{d}^3r\;\chi^*_{\mathbf{k}+\mathbf{G}_{m'}}(\mathbf{r})\,[H-\varepsilon_n(\mathbf{k})]\,\chi_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}), \tag{16.8}$$

and the discontinuity is incorporated into the integral over the sphere surface(s) using Green's identity (Exercise 16.2)

$$\begin{aligned}\langle m'|H^S|m\rangle&=\int_S\mathrm{d}S\;\chi^*_{\mathbf{k}+\mathbf{G}_{m'}}(\mathbf{r})\left[\frac{\partial}{\partial r}\chi_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}^-)-\frac{\partial}{\partial r}\chi_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}^+)\right]\\&=\int_S\mathrm{d}S\chi^*_{\mathbf{k}+\mathbf{G}_{m'}}(\mathbf{r})\left[\frac{\partial}{\partial r}\ln\chi_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}^-)-\frac{\partial}{\partial r}\ln\chi_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}^+)\right]\chi_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}^-),\end{aligned} \tag{16.9}$$

where $+$ $(-)$ indicates just outside (inside) the sphere.

One way to proceed is to solve the secular equation (16.8) in terms of matrix elements of the hamiltonian and the overlap matrix, just as for any non-orthogonal orbital. However, one can take advantage of the fact that the basis functions are not fixed but instead are chosen to satisfy the Schrödinger equation inside each muffin-tin sphere at energy ε. Thus the integral

[2] The factor $\hbar^2/2m_e=\frac{1}{2}$ in Hartree atomic units, where $\hbar=m_e=e=1$ is used in this text. In the present chapter and the next, $\hbar$ and m_e are explicitly indicated where needed to avoid confusion with expressions in the literature, since many authors use "Rydberg atomic units," where $\hbar=2m_e=e^2/2=1$.

(16.8) is zero inside each sphere and needs to be evaluated only in the interstitial region where the hamiltonian is just the kinetic energy operator and χ is a plane wave. All the information about the way each atom affects the bands is incorporated into the boundary terms, i.e. boundary conditions upon the plane waves. This is, of course, not surprising since the wavefunction both inside and outside the spheres must each obey the Schrödinger equation in their respective regions subject to the boundary conditions.[3]

Following this approach, it is straightforward to cast the APW equation (16.7) in a form identical to that in plane wave Fourier methods,

$$\sum_{\mathbf{G}} \left\{ \left[\frac{\hbar^2}{2m_e} (\mathbf{k}+\mathbf{G})^2 - \varepsilon_{i,\mathbf{k}} \right] \delta_{\mathbf{G}'\mathbf{G}} + V^{\mathrm{APW}}_{\mathbf{G}',\mathbf{G}}(\varepsilon_k, \mathbf{k}) \right\} c_{i,\mathbf{G}}(\mathbf{k}) = 0, \tag{16.10}$$

where the first term is the usual kinetic energy for a plane wave extended throughout the cell including the sphere, the energy is relative to the constant in the muffin-tin potential, and all effects due to the potential in the sphere are collected into an APW "potential" $\hat{V}^{\mathrm{APW}}$, which is an operator that is both non-local and energy dependent. The matrix elements of $\hat{V}^{\mathrm{APW}}$ for a sphere at $\tau = 0$ in the unit cell are [639, 134]

$$\begin{aligned} V^{\mathrm{APW}}_{\mathbf{G}',\mathbf{G}}(\varepsilon_k, \mathbf{k}) = & -\frac{4\pi S^2}{\Omega_{\mathrm{cell}}} \left(\frac{\hbar^2}{2m_e} |\mathbf{k}+\mathbf{G}|^2 - \varepsilon_k \right) \frac{j_1\left(|\mathbf{G}-\mathbf{G}'|S\right)}{|\mathbf{G}-\mathbf{G}'|} \\ & + \frac{\hbar^2}{2m_e} \frac{4\pi S}{\Omega_{\mathrm{cell}}} \sum_l \left\{ (2l+1) P_l(\cos(\theta_{\mathbf{G}\mathbf{G}'}) j_l\left(|\mathbf{k}+\mathbf{G}'|S\right) j_l\left(|\mathbf{k}+\mathbf{G}|S\right) \right\} \\ & \times \Delta D_{l,\mathbf{G}}(\varepsilon_k), \end{aligned} \tag{16.11}$$

with $\theta_{\mathbf{G}\mathbf{G}'}$ the angle between vectors $\mathbf{k}+\mathbf{G}$ and $\mathbf{k}+\mathbf{G}'$, and S the sphere radius. For a crystal with more than one sphere centered at positions τ_s, it is simple to show that the potential is a sum of terms with phase factors $\exp(\mathrm{i}(\mathbf{G}-\mathbf{G}')\cdot\tau_s)$ just as in the plane wave method (Eq. 12.16 and Exercise 12.2). The first term in the APW potential operator (16.11) subtracts the kinetic energy for that part of the plane wave inside the spheres (see Loucks [639] p. 32–33), and the last term[4] includes all the effects of the atoms in terms of the difference of the dimensionless logarithmic derivative $\Delta D_{l,\mathbf{G}}(\varepsilon)$ from that of an empty sphere

$$\Delta D_{l,\mathbf{G}}(\varepsilon) = \left[r \frac{\mathrm{d}}{\mathrm{d}r} \ln \psi_l(\varepsilon, r) - r \frac{\mathrm{d}}{\mathrm{d}r} \ln j_l\left(|\mathbf{k}+\mathbf{G}|r\right) \right]_{r=S}, \tag{16.12}$$

which follows from (16.9) for the boundary "kink" term, with the function just inside the sphere given by $\psi_l(\varepsilon, r)$ and the function just outside by the partial wave component of the unscattered plane wave j_l. (The normalization is not needed for the logarithmic derivative.) It is interesting that the "potential" operator involves $\frac{\hbar^2}{2m_e}$; this is because it really is a "matching operator."

[3] This is exactly the same condition as used in Ch. 11 where the pseudowavefunctions were shown to equal the true wavefunctions in the outer part of the atom, so long as the eigenvalue was the same and the wavefunction satisfied the boundary conditions at the sphere boundary. Furthermore, the specification of the boundary condition in terms of the logarithmic derivative in (16.12) is the same as in (11.20) or (J.5); here the evaluation is done at the muffin-tin boundary and with the assumption that the total potential has the spherical muffin-tin form.

[4] Another form slightly different and more convenient for computation is given by Loucks [639] p. 37).

16.2 Solving APW equations: examples

The APW equations are more difficult than the usual independent-particle equations that are linear in energy ε, such as (12.9) for plane waves or (14.7) for localized orbitals, where all the eigenvalues and eigenvectors can be determined at once from a single diagonalization. Instead, the APW equations must be solved separately for each eigenstate as follows:

- Solution of the matrix equation (16.10) has exactly the same form as the usual linear equations, except that the potential operator depends upon the logarithmic derivatives in (16.12) that are functions of the energy $\varepsilon = \varepsilon_{i,k}$, which are not known in advance and are different for each band.
- In order to find the logarithmic derivatives, the radial equations (16.4) for $r\psi_l(\varepsilon, r) \equiv \phi_l(\varepsilon, r)$ must be solved for each band energy $\varepsilon_{i,k}$, individually. However, $\varepsilon_{i,k}$ are established only in conjunction with solution of the plane wave equations (16.10). In general, this requires "root tracing," i.e. searching for the zeros of the determinant on the APW matrix given in (16.10). This may be done by fixing ε and varying k or vice versa.
- There can be simplification in some cases, e.g. highly localized states, such as core states that are completely contained in the sphere, are fully specified by $\psi_l(\varepsilon, r)$ and there is no $\mathbf{k}$ dependence and it can often be considered to be the same as in an atom. This is termed the "frozen-core approximation."

Illustrative examples

Two limiting cases illustrate the power and generality of the APW method: the nearly-free-electron case and the opposite limit in which the spherical potential has a strong resonance. It is important that, despite the artificial division of space, the free-electron case is solved trivially. If the potential inside the muffin-tin sphere is set to be the same constant value as in the interstitial, exact solutions inside the spheres are spherical Bessel functions j_l, in which case the *difference* in logarithmic derivatives vanishes $\Delta D_l(\varepsilon_k) = 0$. It follows immediately that the eigenvalues of (16.10) are just those for free electrons $\varepsilon_k = \frac{1}{2}|\mathbf{k} + \mathbf{G}|^2$. If the phase shift $\Delta D_l(\varepsilon_k) \neq 0$ but is small, then the dispersion ε_k will be only slightly modified. This is the "nearly-free-electron case" and the APW method clarifies an important point: this *does not necessarily mean the potential is small or that the wavefunctions are close to a single plane wave. The difference in phase shift can be small even for strong potentials.* Explicit cases, where actual phase shifts are close to free-electron values, are for Na ([443], p. 242) and many examples in [640]. Logarithmic derivatives for Cu are presented in Fig. 16.3, which shows that the phase shifts for $l = 0, 1, 3$ are very close to the free-electron values (see Exercise 16.5). Thus the APW approach explains the nearly-free-electron character of bands in many materials that are weak scatterers just as well as the OPW or pseudopotential methods.

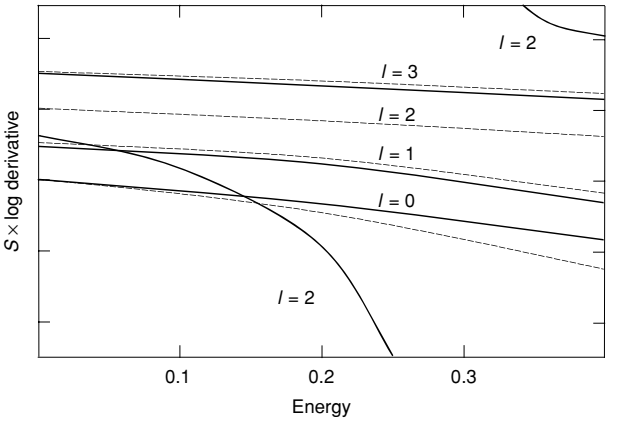


Figure 16.3. Logarithmic derivatives of the radial wavefunctions in Cu, defined in (16.12) as a function of energy in Hartrees for different angular momenta l, evaluated at a sphere radius $S = 2.415a_0$ appropriate for metallic Cu. For comparison are shown the free-electron curves (Exercise 16.5). The $l = 2$ d channels show strong resonance, leading to narrow bands, whereas the $l = 0, 1, 3$ channels reveal the nearly-free-electron behavior. These are essentially the same as for the Chodorow potential [645] used by Burdick [213] to calculate the bands shown in Fig. 2.24. From Kubler and Eyert [134], who attribute the figure to Slater and Mattheiss.

The opposite limit is a resonance at energy ε_0 where the phase shift becomes large. In an isolated atom, the logarithmic derivative $D(\varepsilon)$ evaluated at large radius diverges at $\varepsilon = \varepsilon_0$, signifying a bound state at that energy. (This is one of the standard methods to find bound states in actual atomic programs.) In a crystal, the fact that the phase shift at radius S changes rapidly with energy means that the Bloch boundary conditions for different $\mathbf{k}$ can be satisfied with only small changes in energy, i.e. a band $\varepsilon(\mathbf{k})$ with only a small dispersion. In the case of Cu, the $l = 2$ logarithmic derivative disperses rapidly corresponding to the d bands in Cu, which are much narrower than the s–p bands. In general, the bands start as parabolic and each resonance introduces a new band, which has the physical interpretation of each atomic state broadening into a band in the crystal.

Calculations of bands using the APW method

The power of the APW approach was first fully realized after the advent of electronic computers, in particular, for the first accurate calculations of bands of transition and rare earth metals. A well-known early example is the band structure of Cu calculated by Burdick [213] in 1963 and reproduced in Fig. 2.24 where the bands are compared with measurements from angle-resolved photoemission experiments by Thiry et al. [212] in 1979. The logarithmic derivatives in the APW equations were calculated using the Chodorow potential [645], derived in 1939 for the Cu atom and are essentially the same as shown in Fig. 16.3. The impressive agreement between measured and calculated energies is due to two factors: (1) the potential was modelled as a sum of atomic potentials fitted to the atom, and (2) Cu has a filled (closed-shell) d band and a wide s–p band, a case in which independent-particle

methods are expected to work well.[5] In general, one should exercise caution in identifying Kohn–Sham eigenvalues as excitation energies and one should not expect such agreement.

In Fig. 2.24, the bands are shown along high-symmetry lines of the Brillouin zone (BZ) depicted in Fig. 4.10, since Cu has the fcc crystal structure. The lowest band is minimum at the Γ point ($\mathbf{k} = 0$) in the BZ, with energy defined to be $\varepsilon = 0$; the label Γ_1 designates the same symmetry as an s wave function ($l = 0$). The nearly parabolic ε_k curve is modified because it mixes with narrower bands (examples of a resonance) that start at Γ with labels $\Gamma_{25'}$ and Γ_{12}, which are labels for d ($l = 2$) states in cubic symmetry: $\Gamma_{25'}$ is three-fold degenerate and transforms under rotations like xy, yz and zx, whereas Γ_{12} is two-fold degenerate and transforms like $x^2 - y^2$ and $2z^2 - x^2 - y^2$. The bands labeled Δ_5 and Λ_3 are two-fold degenerate d states along the lines shown. At higher energy around the Fermi level, the bands are also approximately parabolic; this is the feature that explains why Cu is a good electrical conductor. The states $X_{4'}$ at $\varepsilon \approx 0.80$ Ry and $\Lambda_{2'}$ at $\varepsilon \approx 0.61$ Ry have labels that designate p symmetries ($l = 1$) which is expected for a free-electron band, and the eigenvalues are quite close to free-electron energies at the density of one electron per Cu atom as discussed in Exercise 16.5.

Bands for the entire series of 3d transition metals are presented in Fig. 16.4, which shows narrow d bands crossing the wider s–p bands. The 3d bands are much broader than the 4f bands of the lanthanides, but narrow enough to indicate that many atomic-like properties carry over to the solid. At the end of the series is the noble metal Cu in which the d bands are filled, but only slightly below the Fermi energy, which leads to its closed-shell, non-magnetic behavior, and its yellow color. The transition metals have partially occupied d bands. This leads to their magnetic behavior and correlations among the d electrons. Nevertheless, independent-particle methods are in many ways adequate to describe the basic properties of these metals [132], e.g. the prediction of magnetism from the Stoner criterion, as illustrated in Fig. 2.7 [107, 134, 132]. The total energies are remarkably well described by the local density approximation, as illustrated in Fig. 2.3.

It should be emphasized, however, that density functional theory methods often lead to qualitatively wrong predictions for more strongly correlated electron systems, notably the rare earths and the transition metal oxides [216]. In a nutshell, methods that start from the homogeneous gas (LDAs and GGAs) tend to predict solutions that are too much like the gas – non-magnetic and metallic – whereas methods that involve Hartree–Fock exchange (Hartree–Fock itself and exact exchange (EXX)) tend to predict solutions that are too much like Hartree–Fock - too magnetic and insulating. Methods such as self-interaction correction (SIC) and "LDA+U" can apparently describe aspects of strongly correlated electron systems, but at the cost that the functionals are not universal.

The bands ε_k for the non-d states are close to free-electron bands; however, the wave functions are far from single plane waves. Although they are close to a single plane wave between the atomic spheres, inside each sphere the wave functions have all the oscillations

[5] The bands are also expected to be predicted qualitatively by density functional theory in LDA or GGA approximations, except that the energies of the filled d states will be too shallow and the s–p bands too broad in accordance with experience on many systems.

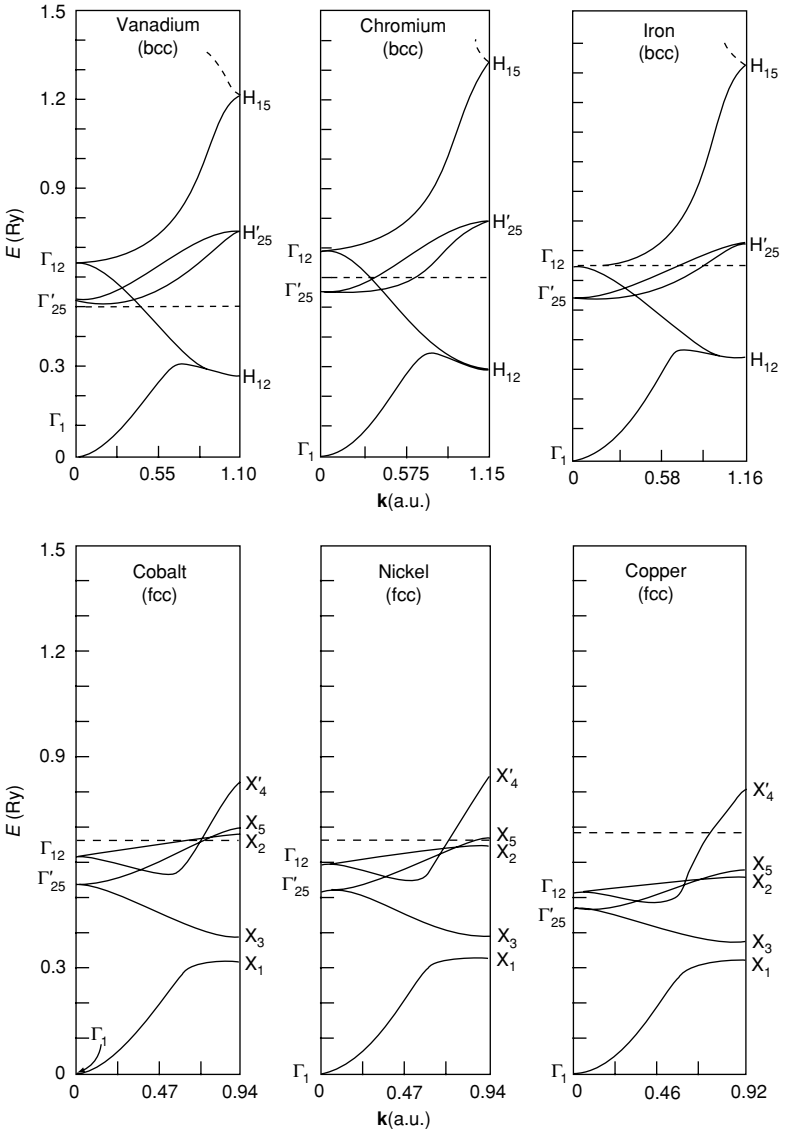


Figure 16.4. Bands of 3d metals showing the narrow d bands crossing the wide s band, and the progression of band filling across transition series. Calculations were done by Mattheiss [646] using the APW method.

characteristic of atomic 4s and 4p wave functions. The point is that a single plane wave joins smoothly onto the solution of the Schrödinger equation inside the sphere, illustrated in Fig. 4.11, so that the energy is essentially the same as that of the plane wave. In addition, the s–p states hybridize with the d states, which acts like a resonance in the scattering of the s electrons from the atoms. This is the basis of the"s–d model" [647], which describes the

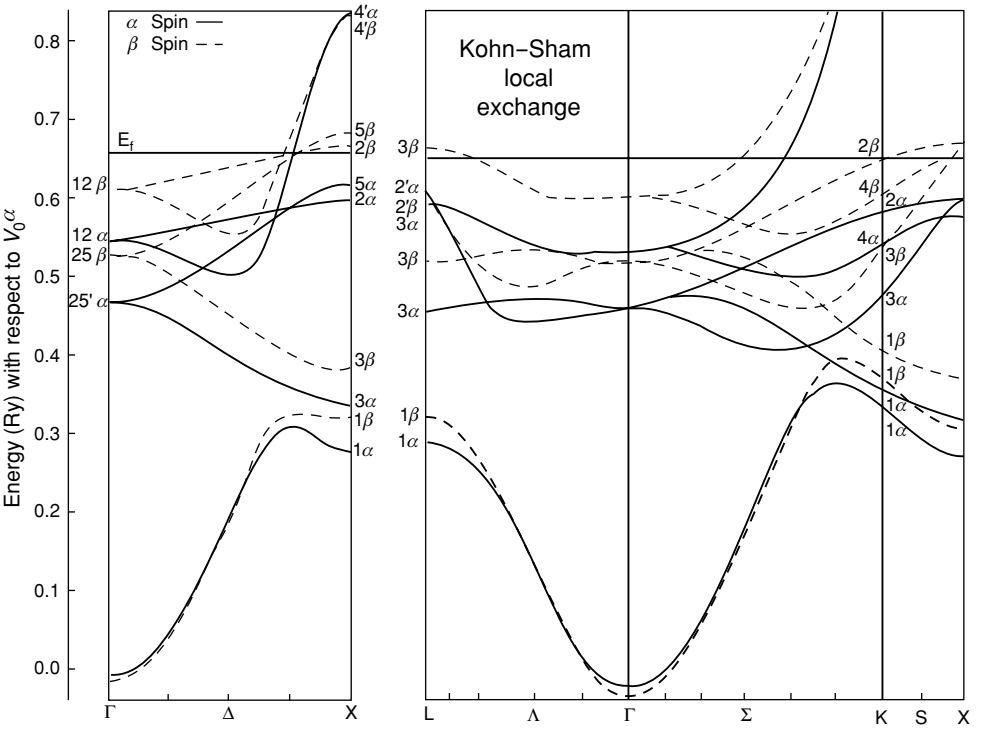


Figure 16.5. Spin polarized bands of ferromagnetic Ni in the fcc structure [648]. Solid lines indicate the majority-spin and dashed lines the minority-spin bands. The numbers indicate symmetry labels at high-symmetry points. This figure represents the results using the Kohn–Sham local spin density approximation. (The Slater local exchange gives poorer agreement with experiment, especially around the L point, as shown in [648].)

s–p bands near the energies of the d states and the resulting dispersion in the narrow d bands (Exercise 16.6).

As an example of spin-dependent bands in a ferromagnet, Fig. 16.5 shows the bands for Ni in the fcc structure calculated by Connolly [648]. The solid lines show the majority-spin and the dashed lines the minority-spin bands calculated with the local spin density formalism of Kohn and Sham. Connolly also found the bands using the Slater local exchange and concluded that it gives much poorer agreement with experiment. The larger exchange causes significant changes in the bands, particularly around the L point.

Practical aspects

How large is the secular equation? The number of plane waves is determined by the fact that they must accurately represent the variations in the wavefunctions in the interstitial region. These are determined by the fact that they must match the solutions inside the spheres which is incorporated into the logarithmic derivatives in (16.10). Thus, in general, we expect the number of plane waves (i.e. the number of APWs) to be comparable to the number of plane

waves needed in a norm-conserving pseudopotential calculation, so long as the atoms have relatively smooth orbitals and large interstitial regions (like Si). However, unlike norm-conserving pseudopotentials, the number of basis functions does not increase as the states become more localized around the atoms (e.g. 3d states in transition metals or 4f in rare earths) since the augmentation takes care of the regions around the atoms. Indeed, roughly 40 APW basis functions are needed for each atom in the unit cell for transition metals [134]. In this respect, APW methods are more closely related to "ultrasoft pseudopotential" and PAW methods (Sec. 11), which can also describe localized states, since they add localized functions and use plane waves only for the smooth part. In addition, spherical harmonics with high angular momenta are required for accurate description of the bands (see also Sec. 17.3).

The APW method is not restricted to muffin-tin potentials and it can be extended to general potentials [110]. The APW basis can be defined as before (using an effective muffin-tin potential determined by averaging the full potential) but now one must calculate matrix elements of the full potential. The solution has the same general form as (16.10) and (16.11), but it becomes more complicated because the matrix elements are no longer diagonal in angular momentum in the sphere nor in momentum in the interstitial region. The extension becomes much more feasible using linearized methods and further discussion is given in Sec. 17.9.

16.3 The KKR or multiple-scattering theory (MST) method

In the words of J. Ziman [637]:

> From mathematical point of view, the most refined method of calculating energy band structures is the subtle procedure invented independently by Korringa [649] and Kohn and Rostoker [650]. This method is indeed so fundamental that it is to be found in all its essentials in a study by Rayleigh [651] [in 1892] of the propagation of sound waves through an assembly of spheres.

The KKR method, also called "multiple-scattering theory" (MST) or Green's function method, finds the stationary values of the inverse transition matrix T rather than the hamiltonian. This is the method used in the pioneering work of Moruzzi and coworkers [106, 107], highlighted in Sec. 2.2, that first established the efficacy of density functional theory for calculation of properties of close-packed metals. In addition, KKR is the method of choice for most calculations on liquids, disordered systems, and impurities in various metallic and non-metallic hosts. Probably the most important features of the KKR or Green's function formulation are: (1) it separates the two aspects of the problem: the structure (positions of the atoms) from the scattering (chemical identity of the atoms); and (2) Green's functions provide a natural approach to a localized description of electronic properties that can be adapted to alloys and other disordered systems.

Here we consider only muffin-tin potentials, where each site can be viewed as a spherical scatterer; and electrons propagate between sites with the free propagator or Green's function. This greatly simplifies the equations and was the basis of all KKR calculations until recently when full potential methods have been introduced [652, 653]. The problem of overlapping

potentials is subtle and the reader is referred to papers in the collection in [642] and references given there.

A Green's function G describes propagation of a particle from one event to another [654], e.g. $G(\varepsilon, \mathbf{r}, \mathbf{r}')$ that describes propagation of an independent particle from point $\mathbf{r}$ to $\mathbf{r}'$ at energy ε. In terms of a reference Green's function G_0 (for example, the free-particle propagator given in (16.19) below) and scattering matrix elements t, representing single scattering events from any of the atoms in the system, the full Green's function can be written in schematic form as

$$\begin{aligned} G &= G_0 + G_0 t G_0 + G_0 t G_0 t G_0 + \cdots \\ &= G_0 + G_0 t G \Rightarrow \\ G &= (G_0^{-1} - t)^{-1}. \end{aligned} \tag{16.13}$$

Similarly, one can sum the series to write G as

$$G = G_0 + G_0 T G_0, \tag{16.14}$$

where T is the full multiple scattering matrix for the entire system

$$\begin{aligned} T &= t + t G_0 t + t G_0 t G_0 t + \cdots \\ &= t + t G_0 (t + t G_0 t + \cdots) \\ &= t + t G_0 T \Rightarrow \\ T &= (t^{-1} - G_0)^{-1}. \end{aligned} \tag{16.15}$$

The stationary states of the system are given by the poles of G or T as functions of ε and hence are obtained from the zeros of the determinant[6]

$$\det(t^{-1} - G_0) = 0. \tag{16.16}$$

For independent-particle electronic structure problems with hamiltonian $\hat{H} = -(\hbar^2/2m_e)\nabla^2 + V_{\text{eff}}(\mathbf{r})$,[7] a convenient starting point is to take G_0 to be the free Green's function. It is useful to first give the well-known solution for the Helmholtz equation $(\nabla^2 + \kappa^2) g(\mathbf{r} - \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}')$, for which the real solution is

$$g(x) = -\frac{1}{4\pi} \frac{\cos(\kappa x)}{x}, \tag{16.17}$$

where $x = |\mathbf{r} - \mathbf{r}'|$. Thus the Green's function for the Schrödinger equation with $V = 0$ satisfies

$$\left[-\frac{\hbar^2}{2m_e} \nabla_r^2 - (\varepsilon - V_0) \right] G_0(\varepsilon, \mathbf{r} - \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'), \tag{16.18}$$

[6] As indicated by the form of (16.14), one must take care to avoid spurious poles appearing in the final solution at the positions of the poles of G_0.

[7] Here $\hbar^2/2m_e$ is explicitly indicated to avoid confusion with references that assume $m_e = 1/2$.

which has the solution

$$G_0(\varepsilon, x) = \frac{2m_e}{\hbar^2}\frac{1}{4\pi}\frac{\cos(\kappa x)}{x}, \tag{16.19}$$

where $(\hbar^2/2m_e)\kappa^2 = \varepsilon - V_0$ and V_0 is the "muffin-tin zero" reference energy. For positive energies ε, $G_0(\varepsilon, x)$ is a slowly decaying oscillatory function; for negative ε, it decays exponentially.

Within the muffin-tin spherical approximation, the scattering amplitude $t(\varepsilon)$ of an electron from each sphere conserves angular momentum $L \equiv \{l, m\}$ referred to the center of that sphere. Because the scattering is unitary and independent of m [641, 655], $t_l(\varepsilon)$ and can be written in terms of the phase shift $\eta_l(\varepsilon)$ (see Sec. J.8)

$$t_l(\varepsilon) = \frac{i}{2\kappa}(\mathrm{e}^{i2\eta_l(\varepsilon)} - 1) = -\frac{1}{\kappa}\mathrm{e}^{\mathrm{i}\eta_l(\varepsilon)}\sin(\eta_l(\varepsilon)). \tag{16.20}$$

The scattering amplitude plays a key role in many phenomena in physics, such as Friedel oscillations around an impurity, resistivity due to impurities, *etc.* (see Sec. J.1). The scattering is represented pictorially in Fig. J.1. In the present case, the great advantage is that the electronic bands and Green's functions can be described by a few phase shifts $\eta_l(\varepsilon)$, typically $l \leq 3$.

The full solution for the multiple-scattering problem for the muffin-tin potential is given by (16.16), where G_0 depends only upon the structure and the energy ε, and t incorporates all the effects of the potential inside each sphere. The expressions needed are for the Green's function $G_0(\varepsilon, |\mathbf{r} - \mathbf{r}'|)$ when $\mathbf{r}$ and $\mathbf{r}'$ are in the same and different spheres. For different spheres, this requires that a spherical wave of angular momenta L about one sphere be expressed in terms of waves centered at another site, which involves a sum over L' at that sphere. The needed formulas are given in [11, 134, 641], which can be understood using the addition formula for plane waves

$$\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{R}_1)} = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{R}_2)}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{R}_2-\mathbf{R}_1)}, \tag{16.21}$$

together with the identity (16.5). The result for sites $\mathbf{R} \neq \mathbf{R}'$ is given by (see Exercise 16.3)

$$G_0(\varepsilon, |\mathbf{r} - \mathbf{r}'|) = \sum_{L,L'} i^l j_l(\kappa r) Y_L(\hat{\mathbf{r}}) B_{LL'} (-i)^{l'} j_{l'}(\kappa r') Y^*_{L'}(\hat{\mathbf{r}}'), \tag{16.22}$$

where $\mathbf{r}$ and $\mathbf{r}'$ are referred to the centers of their respective spheres, and $B_{LL'}$ denotes the "KKR structure constants"

$$B_{LL'}(\varepsilon, \mathbf{R} - \mathbf{R}') = -4\pi\kappa \sum_{L''} i^{l''} C^L_{L'L''} n_{l''}(\kappa|\mathbf{R} - \mathbf{R}'|) Y_{L''}(\widehat{\mathbf{R} - \mathbf{R}'}), \tag{16.23}$$

where the Cs are directly related to the Gaunt coefficients $c^{l''}(l\ m, l'\ m')$ defined in (K.14),

$$C^L_{L'L''} \equiv \int \mathrm{d}\Omega Y^*_L(\Omega) Y_{L'}(\Omega) Y_{L''}(\Omega) = \sqrt{\frac{2l''+1}{4\pi}} c^{l''}(l\ m, l'\ m'). \tag{16.24}$$

For the general case where the scattering amplitude $t_l(\varepsilon, \mathbf{R})$ is site-dependent, the resulting equation for the Green's function (16.13) is

$$\left[G_{LL'}(\varepsilon, \mathbf{R}, \mathbf{R}')\right]^{-1} = \left[\left[B_{L,L'}(\varepsilon, \mathbf{R} - \mathbf{R}')\right]^{-1} - t_l(\varepsilon, \mathbf{R})\delta_{\mathbf{R},\mathbf{R}'}\delta_{L,L'}\right], \tag{16.25}$$

and condition (16.16) for stationary states becomes

$$\det\left[t_l^{-1}(\varepsilon, \mathbf{R})\delta_{\mathbf{R}\mathbf{R}'}\delta_{LL'} - B_{LL'}(\varepsilon, \mathbf{R} - \mathbf{R}')\right] = 0, \tag{16.26}$$

where $\mathbf{R}$, $\mathbf{R}'$ denote centers of the spheres, and $B_{LL'}(\varepsilon, 0) \equiv 0$ for the same site. As it stands, this is a matrix equation in all the sites and angular momenta – a formal expression valid for crystals, molecules, and disordered solids.

KKR band structure equations

The original equation of Koringa [649] results if we consider a crystal where the scattering is the same at every site, centered on the translations vectors $\mathbf{T}$. Then the determinant equation can be solved separately for each wavevector $\mathbf{k}$. The structure constants can be defined as

$$B_{LL'}(\varepsilon, \mathbf{k}) = \sum_{\mathbf{T}\neq 0} B_{LL'}(\varepsilon, \mathbf{T})\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}}, \tag{16.27}$$

and the bands $\varepsilon = \varepsilon_{\mathbf{k}}$ are the solution of

$$\det\left[t_l^{-1}(\varepsilon)\delta_{LL'} - B_{LL'}(\varepsilon, \mathbf{k})\right] = 0. \tag{16.28}$$

The well-known form for the KKR equations is found by using expression (16.20) for t (only the real part is needed) in terms of the phase shifts,

$$\sum_{L'} \left[B_{LL'}(\varepsilon_{\mathbf{k}}, \mathbf{k}) \;+\; \kappa\;\cot(\eta_l(\varepsilon_{\mathbf{k}}))\;\delta_{LL'}\right]\; a_{L'}(\mathbf{k}) \;=\; 0. \tag{16.29}$$

This can be generalized straightforwardly to more than one atom per cell, $\alpha = 1, \ldots, N$, leading to one band per atom and angular momentum

$$\sum_{L'}\sum_{\beta=1}^{N}\left[B_{LL'}(\tau_\alpha - \tau_\beta, \varepsilon_{\mathbf{k}}, \mathbf{k}) \;+\; \kappa\;\cot\eta_{l\beta}(\varepsilon_{\mathbf{k}})\;\delta_{LL'}\;\delta_{\alpha\beta}\right]\; a_{L'\beta}(\mathbf{k}) = 0. \tag{16.30}$$

The dispersion relation $\varepsilon_{\mathbf{k}}$ can be found from the roots of the determinant of the matrix in square brackets. Often it is most effective to fix the energy and scan the wavevector $\mathbf{k}$ to find the roots, since the phase shifts depend only on energy and the structure constants depend only on $\mathbf{k}$ at a given energy. For example, the Fermi surface can be mapped out conveniently in this way.

The eigenvectors of (16.29) or (16.30) determine the wavefunction, since the eigenvectors of the Green's function are the same as those of the hamiltonian. Inside each sphere the solution is simply a linear combination of the augmentation functions (apart from a normalization factor)

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{L'}\sum_{\beta=1}^{N} a_{L'\beta}(\mathbf{k})\;\psi_{L'\beta}(\mathbf{r}), \tag{16.31}$$

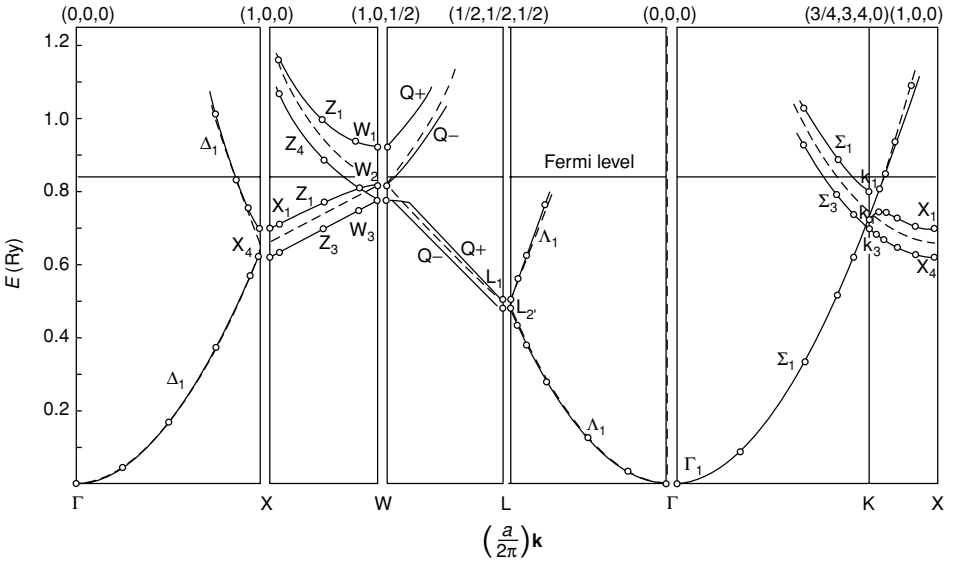


Figure 16.6. Band structure of Al (solid line) calculated by the KKR method [527] compared to free-electron bands (dashed lines). The results can also be easily understood in terms of a weak effective scattering in the plane wave method and the nearly-free-electron approximation, Ch. 12. From [527].

where $\psi_{L'\beta}$ are known since they were used to find the phase shifts and t matrix. Outside the spheres the wavefunction can be found from the Green's function equation (see Exercise 16.8)

$$\psi_{\mathbf{k}}(\mathbf{r}) = -\int \mathrm{d}\mathbf{r}' G_0(\varepsilon^n_{\mathbf{k}}, |\mathbf{r} - \mathbf{r}'|) V(\mathbf{r}') \psi_{\mathbf{k}}(\mathbf{r}'), \tag{16.32}$$

which can be evaluated with $\mathbf{r}'$ restricted to the interstitial region with boundary conditions on each sphere or with an integral over all space. The integral can be done in different ways, since the free Green's function G_0 given in 16.19 is long-ranged for energy in the continuum, but is exponential for energies below the continuum (see also Sec. 16.7).

The size of the secular equation depends on the maximum angular momentum needed. In the case of Cu and the elementary transition metals it is sufficient to take $l_{\max} = 3$ and therefore the rank of the secular equation is 16. Note that the basis is much smaller than in the APW (or plane wave) method; the number of functions is determined by the principle angular momenta of the atomic states needed and not by an accuracy criterion for representation of the wavefunctions.

The KKR method has provided some of the most influential and insightful examples of electronic structure calculations. For example, Fig. 16.6 shows the bands of Al [527]. This is an ideal case for the muffin-tin approximation and illustrates the simple physics that emerges from the KKR approach. The results are similar to previous OPW [528] and pseudopotential calculations [481], all showing the free-electron character of the valence bands. The KKR method conveniently integrates over all plane waves in the analytic Green's

function, whereas the plane wave methods make use of the fact that for weak effective scattering only a few plane waves are needed.

KKR is the method used for the first quantitative calculations of the total energy, equilibrium lattice constants, and bulk moduli given in Fig. 2.3; the density of states and Stoner interaction that led to Fig. 2.7; and hosts of other properties, as documented in [107, 132, 642] and many other sources. As a band method, however, it suffers from the same non-linearity difficulties as the APW method and it is very difficult to extend to a full potential [653]. Therefore, we focus upon Green's function approaches where KKR shines.

KKR Green's function equations

The power of the KKR approach is most apparent in its formulation as a Green's function method that determines electronic properties directly from $G_{LL'}(\varepsilon)$ in (16.13). The explicit form in real space for the muffin-tin potential is given by $G_{LL'}(\varepsilon, \mathbf{R}, \mathbf{R}')$ in (16.25). In a crystal with one atom per cell (the expressions are easily generalized to many atoms per cell) the Green's function is a function only of the relative separation $G_{LL'}(\varepsilon, \mathbf{R} - \mathbf{R}')$. It is most convenient to work with the Fourier transform $G_{LL'}(\varepsilon, \mathbf{k})$ which can be evaluated at each $\mathbf{k}$ separately, as follows from the Bloch theorem. Furthermore, in a crystal, the fact that $G_{LL'}(\varepsilon, \mathbf{k})$ is only a small matrix, of dimension determined by $l_{\max}$, is a great advantage: the inversion of such small matrices is of negligible computational cost and the method can be very efficient depending upon the effort required to set up the matrices.

The Green's function provides a spectral representation and many physical properties can be calculated as integrals over energy. The basic relations given in Sec. D.4 apply to any representation of a Green's function. In particular, the imaginary part of $G(\varepsilon, \mathbf{R})$ provides a *local density of states*, whereas $G(\varepsilon, \mathbf{k})$ provides a "*Bloch spectral representation*" i.e. energy and wavevector resolved spectra. For example, the density of states per unit energy ε in the L channel is given by the diagonal part of G with $L = L'$,

$$n_L(\varepsilon, \mathbf{k}) = -\frac{1}{\pi} \mathrm{Im}\, G_{LL}(\varepsilon + \mathrm{i}\delta, \mathbf{k}), \tag{16.33}$$

where δ is a positive infinitesimal. The total density of states at wavevector $\mathbf{k}$ is given by $n(\varepsilon, \mathbf{k}) = \sum_L n_L(\varepsilon, \mathbf{k})$, which is a sum of delta functions of unity weight at the band energies $\varepsilon = \varepsilon_i(\mathbf{k})$.

This Green's function approach provides a convenient of way of calculating the band structure. For example, the Fermi surface can be calculated directly as the locus of states with $\varepsilon_F = \varepsilon_i(\mathbf{k})$ by calculating only the Green's function at ε_F, without calculating the entire band structure. But how does one know ε_F and the potential V_{eff} from which the phase shifts are derived? The Fermi energy can be fixed by a fast procedure for counting the total number of states up to a given energy which is given by a formula due to Lloyd [656] that effectively evaluates the integral in (D.27). The potential is fixed by the density, which is considered next.

The density in real space $n(\mathbf{r})$ can be calculated from the projected density at each site $\mathbf{R}$ due to angular momentum component L. The local density of states is

$$n_L(\varepsilon, \mathbf{R}) = -\frac{1}{\pi} \mathrm{Im}\, G_{LL}(\varepsilon + \mathrm{i}\delta, \mathbf{R}), \tag{16.34}$$

and the total density in the sphere at site $\mathbf{R}$ is

$$n_{L,\mathbf{R}} = -\frac{1}{\pi}\int_{\infty}^{E_F} \mathrm{d}\varepsilon\ \mathrm{Im}\, G_{LL}(\varepsilon + \mathrm{i}\delta, 0). \tag{16.35}$$

Another quantity that is easily derived is the sum of eigenvalues of occupied states, which is given by

$$\sum_i \varepsilon_i = -\frac{1}{\pi}\sum_L \int_{\infty}^{E_F} \mathrm{d}\varepsilon\ \varepsilon\ \mathrm{Im}\, G_{LL}(\varepsilon + \mathrm{i}\delta, 0). \tag{16.36}$$

The last equation represents a way of summing the eigenvalues: each eigenvalue leads to a pole in $G(\varepsilon)$ which gives a contribution of ε to the integral in (16.36). This provides all the information needed to determine the total energy.

The integrals for total quantities can also be evaluated as contour integrals as shown in Fig. D.1 and given in Eqs. (D.27)–(D.28), which can be evaluated by a discrete sum over points on the contour in the complex plane. Thus one evaluates the Green's function for chosen complex energies z, so that there is no disadvantage due to the non-linear nature of the secular equations. Furthermore, wherever the contour is far from any pole, the Green's function $G_{LL}(z, \mathbf{R})$ decays exponentially as a function of distance $|\mathbf{R}|$, so that it can be evaluated using only a cluster of atoms. However, in a metal, the contour necessarily approaches the poles at the Fermi energy, and $G(z, \mathbf{R})$ must exhibit long-range oscillatory behavior in real space (Friedel or Ruderman–Kittel oscillations) due to the sharp cutoff in Fourier space at the Fermi surface.

16.4 Alloys and the coherent potential approximation (CPA)

Alloys represent important classes of materials ranging from metallic alloys, where mechanical and magnetic properties can be controlled, to semiconductors where delicate electronic properties are tuned by composition. There are two general types of theoretical approaches: direct calculations on selected supercells and methods that average over disorder. The former approach allows direct studies of effects of short-range order and can be very powerful using clusters and supercell methods (see, e.g. [657] and refences given there.) We will concentrate upon the coherent potential approximation (CPA), which provides an intuitive, yet accurate, approach when combined with Green's function methods. Such methods are widely applied in crystalline metallic alloys. The formulation that underlies present-day work is due to Soven [658] and Velicky, et al. [659], and earlier work of Lax [660] and Beeby [661].

The general idea of the CPA approach is to formulate an effective (or coherent) potential which, when placed on every site of the alloy lattice, will mimic the electronic properties of the actual alloy. As distinguished from a "virtual crystal approximation" in which the alloy is replaced by an average crystal potential, the coherent potential is derived from averaging the scattering properties of the different atoms embedded in an effective potential as illustrated in Fig. 16.7. Requiring the weighted site average to be the same as the effective potential results in a complex, energy-dependent CPA potential. This is readily treated in

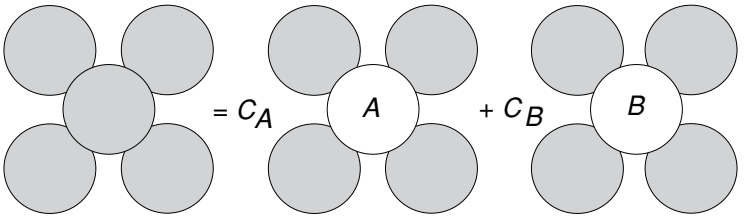


Figure 16.7. Schematic illustration of the averaging over sites in the CPA. The shaded spheres represent an effective average environment and the equation indicates that the average is required to equal the weighted average over sites A and B with concentrations C_A and C_B, each in the same average environment. This leads to the complex CPA potential most readily represented by a Green's function.

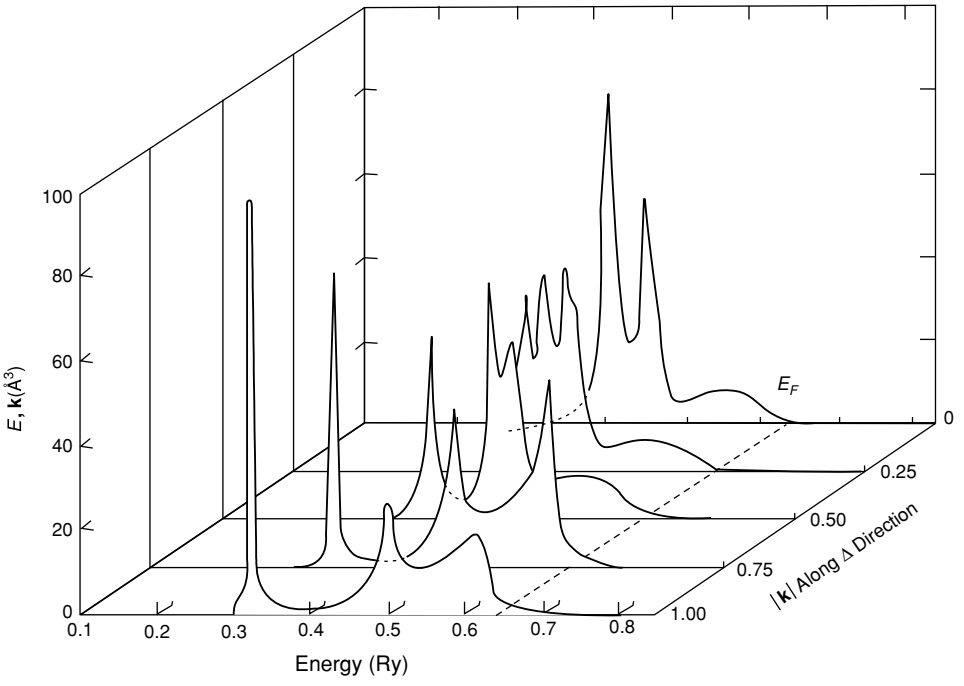


Figure 16.8. Bloch spectral function for a $Cu_{0.77}Ni_{0.23}$ alloy calculated using the KKR–CPA muffin-tin approximation for potential. The figure shows peaks that disperse revealing the underlying crystal-like bands and broadening that is due to the disorder treated in the coherent potential approximation (CPA). From [663].

terms of Green's functions in which a complex energy is naturally introduced. An early formulation of the KKR–CPA method, with application to Cu–Ni alloys, is that of Stocks, Temmerman, and Gyorffy [662].

As an example of the "band structure" of alloys calculated in the CPA approximation, Fig. 16.8 shows the Bloch spectral function at five **k** points along the Δ direction in a

$Cu_{0.77}Ni_{0.23}$ alloy [663]. The peaks (that would be delta functions in a perfect crystal) indicate effective bands that are broadened by scattering due to disorder. The energy- and **k**-dependent broadening is directly related to scattering rates and lengths, and therefore to transport properties such as resistivity [664].

The KKR–CPA equations can also yield total quantities such as energy, pressure, and magnetization in random substitutional alloys [666]. As an example, a recent calculation of the total energy and magnetic moments in disordered Fe_xCu_{1-x} alloys [665] finds an abrupt first-order transition from a non-magnetic Cu-like phase to a magnetic phase with a change in volume. Alloys can also be treated in a response function approach in which the differences are treated in perturbation theory (see, e.g. [236]).

16.5 Muffin-tin orbitals (MTOs)

Muffin-tin orbitals form a basis of localized augmented orbitals introduced by Andersen [667] in 1971 and subsequently extended into an entire methodology. The goal of the MTO approach is not merely to devise another band structure method but to provide a satisfying interpretation of the electronic structure of materials in terms of a *minimal basis* of orbitals. Like local orbital methods, the electronic states are described in a small number of meaningful orbitals; however, unlike those approaches the minimal basis can be accurate because the MTOs are generated from the Kohn–Sham hamiltonian itself.

This section is devoted to the MTO approach, which sets the stage for the linearized LMTO extension [643, 644] (Ch. 17) that exhibits the real power of the approach. The (L)MTO approach has led to many new concepts and methods, for example, "canonical bands" [643, 644], a new approach to the first-principles tight-binding method [668], and many other features. The (L)MTO methodology has been developed in a way most appropriate for close-packed solids, and the descriptions in the literature are often difficult to penetrate because the basic theory is interwoven with approximations and motivational aspects. The goal of the presentation here and in Sec. 17.5 is to bring out the simplicity of the (L)MTO approach, the ways in which the concepts enhance our understanding, difficulties in its use in structures that are open or have low symmetry, and the power of the method in actual calculations when used appropriately.

An MTO can be understood in terms of a single atomic sphere with a flat potential in all space outside the sphere, which is the subject of Sec. J.1 and is closely related to the KKR method. The MTO is defined to be a localized basis function continuous in value and derivative at the sphere boundary. Direct application of the KKR formalism would be to construct an orbital as the energy dependent $\psi_l(\varepsilon, r)$ inside the sphere as in (J.3), and matching the wavefunction outside the sphere, leading to the form $\propto j_l(\kappa r) - \tan(\eta_l(\varepsilon))\, n_l(\kappa r)$ outside the sphere, where j_l and η_l are spherical Bessel and Neumann functions. For negative energies, the Neumann functions are replaced by Hankel functions $h_l^{(1)} = j_l + i\eta_l$, which have the asymptotic form $i^{-l}e^{-|\kappa|r}/|\kappa|r$, and the Bessel functions are unbounded. Such orbitals are *not* suitable as basis functions since, at negative energies, they are normalizable only at ε corresponding to eigenvalues where the coefficient of the Bessel function vanishes.

The insight of Andersen [667] was to reformulate the problem defining a new set of functions that depend separately on κ and ε,

$$\chi_L^{\mathrm{MTO}}(\varepsilon,\kappa,\mathbf{r}) = i^l Y_L(\hat{r}) \begin{cases} \psi_l(\varepsilon,r) + \kappa\cot(\eta_l(\varepsilon))j_l(\kappa r), & r < S, \\ \kappa n_l(\kappa r), & r > S, \end{cases} \tag{16.37}$$

where $Y_L(\hat{r}) \equiv Y_l^m(\hat{r})$ and the factor i^l is a convenient definition (This is the same as adopted in [643, 134, 132] and it leads to bound state functions that are real, as shown in Sec. J.1). The definition in (16.37) leads to a very simple envelope function outside the sphere with the property that each MTO basis function is well defined, both inside the sphere (since $j_l(\kappa r)$ is regular at the origin) and outside the sphere (since $n_l(\kappa r)$ is regular at ∞). Furthermore, the states are normalizable for all negative energies for any κ. Of course, the χ^{MTO} cannot be eigenstates of a single-muffin-tin potential, but they are basis functions with desirable features for the many-site problem.

The form of (16.37) contains the seed of an idea that flows through the development of the MTO and LMTO methods: the wavefunction inside the sphere has been modified in a way that takes into account the presence of neighboring atoms to some approximation. The Bessel function $j_l(\kappa r)$ added for $r < S$ is a step toward incorporating into the wavefunction effects due to the neighbors so that a minimal basis of MTO functions χ^{MTO} can accurately describe the system.

The equations for many atoms can be derived using an expansion theorem of the form of (15.1), which expresses the tail of an MTO extending into another sphere in terms of functions centered on that sphere. Fortunately, there is a well-known expansion analogous to (16.22),

$$n_L(\kappa,\mathbf{r}-\mathbf{R}) = 4\pi \sum_{L'L''} C^L_{L'L''} n^*_{L''}(\kappa,\mathbf{R}-\mathbf{R}') j_{L'}(\kappa,\mathbf{r}-\mathbf{R}'), \tag{16.38}$$

where the $C^L_{L'L''}$ are defined by (16.24). At this point, the MTO basis can be used for calculation of bands by requiring that the total wave function be a solution both inside and outside the spheres, i.e. that the energy and κ be related by $(\hbar^2/2m_e)\kappa^2 = \varepsilon - V_0$. This amounts to a transformation of the KKR method and would lead to non-linear equations equivalent to (16.28) or (16.29).

However, the MTO approach can also be used in a different way. By treating the $\chi_L^{\mathrm{MTO}}(\varepsilon,\kappa,\mathbf{r})$ defined in (16.37) as functions of ε and κ separately, a judicious fixed choice of κ can be used to define a basis that greatly simplifies the problem and yet is accurate for many problems. This has the advantage that one can define structure constants $S_{L'L}(\mathbf{R})$ or $S_{L'L}(\mathbf{k})$ that depend only upon the structure (and the fixed value of κ); in contrast, the KKR "structure constants" are not really constant but are functions of energy ε. This leads to a second hallmark of the (L)MTO approach: developing the method in such a way to take advantage of the fact that an error in the wavefunction leads to higher-order errors in the energy and certain other properties, so that a minimal basis and energy-independent structure constants suffice for accurate calculations.

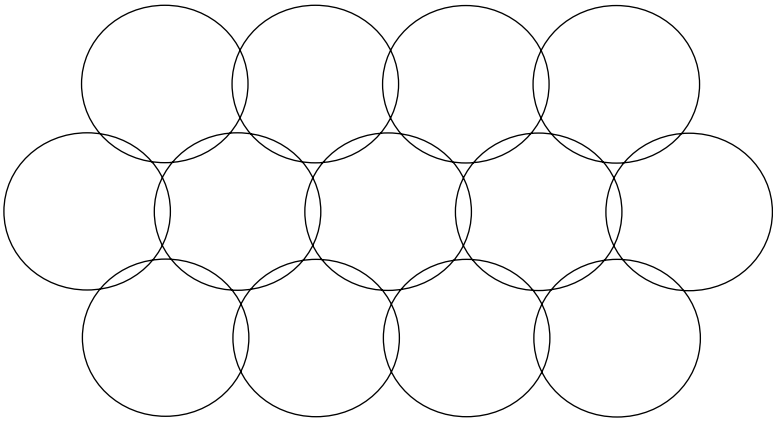


Figure 16.9. Atomic sphere approximation (ASA) in which the muffin-tin spheres are chosen to have the same volume as the Wigner–Seitz cell, which leads to overlapping spheres. The ASA is often a very good approximation for a close-packed solid. Even some open structures (like diamond) can be formed from close-packed spheres (Exercise 16.11) by including "empty spheres" not centered on atoms [669, 670].

16.6 Canonical bands

The simplest version of the MTO equations results if the constant κ is chosen to be $\kappa = 0$, which has been shown to be remarkably accurate for many problems, especially close-packed crystals. The rationale for the freedom to chose κ is that it is finally needed only to represent the variation in the wavefunction in the interstitial between the spheres; if there is only a short distance between the spheres (as in a close-packed solid), the wavefunction will be nearly correct because it has the correct value and slope at the sphere boundary. Many of the applications and much of the motivation for the method [644, 643] is associated with the atomic sphere approximation (ASA) in which the Wigner–Seitz sphere around each atom is replaced by a sphere as shown schematically in Fig. 16.9. It is evident that for close-packed cases the distances between spheres are indeed short; since the spheres overlap, the extrapolation to connect the spheres can be either forward or backward.

For $\kappa = 0$, the wavefunction satisfies the Laplace equation in the interstitial region, i.e. it is equivalent to the electrostatic potential due to a multi-pole moment. The form can be derived from the previous equations with the $\kappa \to 0$ limit of the Bessel and Neumann functions: inside the sphere $j_l \to (r/S)^l$ with logarithmic derivative $D = l$ and outside, $n_l \to (r/S)^{-l-1}$ with $D = -l - 1$. The MTO in (16.37) can be written ([134], Eq. (1-221); see also [643], Eq. (2.1))

$$\chi_L^{\mathrm{MTO}}(\varepsilon, 0, \mathbf{r}) = i^l Y_L(\hat{r}) \psi_l(\varepsilon, S) \begin{cases} \dfrac{\psi_l(\varepsilon, r)}{\psi_l(\varepsilon, S)} - \dfrac{D_l(\varepsilon) + l + 1}{2l + 1} \left(\dfrac{r}{S}\right)^l, & r < S, \\ + \dfrac{l - D_l(\varepsilon)}{2l + 1} \left(\dfrac{S}{r}\right)^{l+1}, & r > S, \end{cases} \tag{16.39}$$

where $D_l(\varepsilon)$ the dimensionless logarithmic derivative of $\psi_l(\varepsilon, r)$ evaluated at the boundary $r = S$. This function is continuous and differentiable everywhere (Exercise 16.12). The expansion theorem can be found as the $\kappa \to 0$ limit of (16.38), which is a well-known multi-pole expansion,

$$\left[\frac{S}{|\mathbf{r}-\mathbf{R}|}\right]^{l+1} i^l Y_L(\widehat{\mathbf{r}-\mathbf{R}})$$
$$= 4\pi \sum_{L'} \left[\frac{r}{S}\right]^{l'} i^{l'} Y_{L'}(\hat{\mathbf{r}}) \left\{ \frac{(2l''-1)!!}{(2l-1)!!(2l'+1)!!} C^L_{L'L''} \left[\frac{S}{|\mathbf{R}|}\right]^{l''+1} i^{-l''} Y^*_{L''}(\hat{\mathbf{R}}) \right\}, \quad (16.40)$$

where $l'' = l' + l$ and $m'' = m' - m$ and the notation $(\ldots)!!$ denotes $1 \times 3 \times 5 \ldots$.

The essential features of the method are illustrated by a crystal with one atom per cell (extension to more atoms per cell is straightforward). Details of the calculation of the structure constants can be found in [643]; we give only limited results to emphasize that they can be cast in closed form using well-known formulas. The structure factor in $\mathbf{k}$ space is found from the Fourier transform of (16.40),

$$\sum_{\mathbf{T}\neq 0} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \left[\frac{S}{|\mathbf{r}-\mathbf{T}|}\right]^{l+1} i^l Y_L(\widehat{\mathbf{r}-\mathbf{T}}) \equiv \sum_{L'} \frac{-1}{2(2l'+1)} \left[\frac{r}{S}\right]^{l'} i^{l'} Y_{L'}(\hat{\mathbf{r}}) S_{L'L}(\mathbf{k}), \quad (16.41)$$

where the factors have been chosen to make $S_{L'L}(\mathbf{k})$ hermitian [643]. The result is

$$S_{L'L}(\mathbf{k}) = g_{l'm',lm} \sum_{\mathbf{T}\neq 0} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \left[\frac{S}{|\mathbf{T}|}\right]^{l''+1} \left[\sqrt{4\pi} i^{l''} Y_{L''}(\hat{\mathbf{T}})\right]^*, \quad (16.42)$$

where $g_{l'm',lm}$ can be expressed in terms of Gaunt coefficients [643].

An MTO basis function with wavevector $\mathbf{k}$ is constructed by placing a localized MTO on each lattice site with the Bloch phase factor, e.g. for $\kappa = 0$,

$$\chi^{\mathrm{MTO}}_{L,\mathbf{k}}(\varepsilon, 0, \mathbf{r}) \;=\; \sum_{\mathbf{T}} \; \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \; \chi^{\mathrm{MTO}}_L(\varepsilon, 0, \mathbf{r} - \mathbf{T}). \quad (16.43)$$

The wavefunction in the sphere at the origin is the sum of the "head function" (Eq. 16.39 for $r < S$) in that sphere plus the tails (Eq. (16.39) for $r > S$) from neighboring spheres, and can be written using Eq. (16.41), as

$$\chi^{\mathrm{MTO}}_{L,\mathbf{k}}(\varepsilon, 0, \mathbf{r}) = \psi_l(\varepsilon, r) i^l Y_L(\hat{r}) \;-\; \frac{D_l(\varepsilon) + l + 1}{2l+1} \; \psi_l(\varepsilon, S) \; \left(\frac{r}{S}\right)^l \; i^l Y_L(\hat{r})$$
$$+ \frac{l - D_l(\varepsilon)}{2l+1} \; \psi_l(\varepsilon, S) \; \sum_{L'} \left(\frac{r}{S}\right)^{l'} \frac{1}{2(2l'+1)} \; i^{l'} Y_{L'}(\hat{r}) \; S_{LL'}(\mathbf{k}). \quad (16.44)$$

The solution can now be found for an eigenstate as a linear combination of the Bloch MTOs Eq. (16.44),

$$\psi_{\mathbf{k}}(\varepsilon, \mathbf{r}) \;=\; \sum_{L} \; a_L(\mathbf{k}) \;\; \chi^{\mathrm{MTO}}_{L,\mathbf{k}}(\varepsilon, 0, \mathbf{r}). \quad (16.45)$$

Since the first term on the right-hand side of (16.44) is already a solution inside the atomic sphere, $\psi_{\mathbf{k}}(\varepsilon, \mathbf{r})$ can be an eigenfunction only if the linear combination of the last two terms

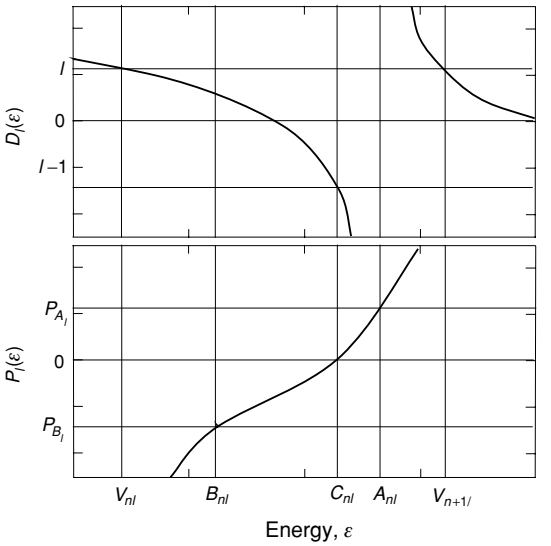


Figure 16.10. Potential function $P_l(\varepsilon)$ (bottom) compared to logarithmic derivative $D_l(\varepsilon)$ (top) versus energy. The functions are related by (16.47) and the energies $\{A, C, B\}$ denote, respectively, the top, center, and bottom of the nth band formed from states of angular momentum l. The energy V denotes the singularities in $P_l(\varepsilon)$ that separate bands. The key point is that $P_l(\varepsilon)$ is a smooth function for all energies in the band so that it can be parameterized as discussed in the text. (Taken from similar figure in [643], Ch. 2.)

on the right-hand side of (16.44) vanishes – called "tail cancellation" for obvious reasons. This condition can be expressed as

$$\sum_L \{S_{LL'}(\mathbf{k}) - P_l(\varepsilon)\,\delta_{LL'}\}\; a_L(\mathbf{k}) = 0, \tag{16.46}$$

where $P_l(\varepsilon)$ is the "potential function"[8]

$$P_l(\varepsilon) \;=\; 2(2l+1)\frac{D_l(\varepsilon)+l+1}{D_l(\varepsilon)-l}. \tag{16.47}$$

Equation (16.46) is a set of linear, homogeneous equations for the eigenvectors $a_L(\mathbf{k})$ at energies $\varepsilon = \varepsilon_k$ for which the determinant of the coefficient matrix vanishes

$$\det\,[S_{LL'}(\mathbf{k}) - P_l(\varepsilon)\,\delta_{LL'}] = 0. \tag{16.48}$$

This is a KKR-type equation, but here $S_{LL'}(\mathbf{k})$ does not depend on the energy.

The potential function $P_l(\varepsilon)$ contains the same information as the phase shift or the logarithmic derivative $D_l(\varepsilon)$, and the relation between them is illustrated in Fig. 16.10. $P_l(\varepsilon)$ provides a convenient description of the effective potential in (16.48) because it varies smoothly as a function of energy in the region of the eigenvalues, as opposed to the

[8] We use the symbol P_l since it is the standard term. It should not be confused with a Legendre polynomial

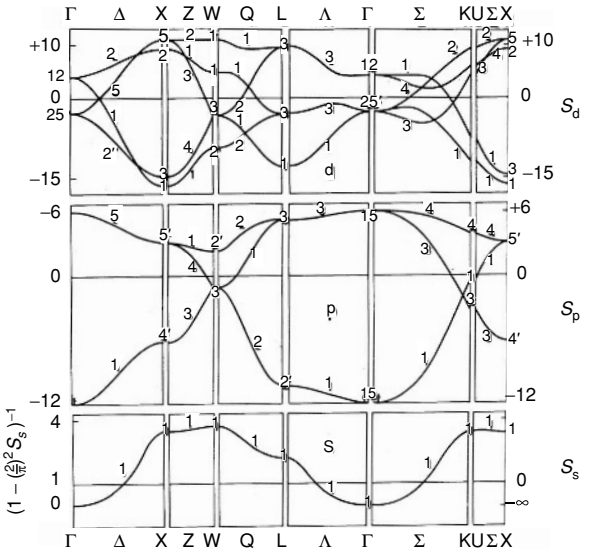


Figure 16.11. Canonical unhybridized bands for an fcc lattice. Comparison with Figs. 2.24, 16.4, and 16.5 shows that this "canonical band" structure has remarkable similarity to the full calculated bands in an elemental fcc crystal. The canonical bands have only one material-dependent factor that scales the overall band width, and even that parameter can be found from simple atomic calculations as discussed in Secs. 10.7 and 17.5. Further improvement can be included through information about the potential function as described in the text. Provided by O. K. Andersen; similar to those in [461], [644], [464] and [643].

logarithmic derivative $D_l(\varepsilon)$ that varies strongly and is very non-linear in the desired energy range. This leads to the simple, but very useful, approximations discussed next.

One of the powerful concepts that arises from (16.46) or (16.48) is "canonical bands," which allow one to obtain more insight into the band structure problem. In essence it is the solution of the problem of states in an atomic sphere (as considered in Sec. 10.7 but here with non-spherical boundary conditions imposed by the lattice through the structure constants, $S_{LL'}(\mathbf{k})$; see also further discussion in Sec. 17.1). Since the potential function, $P_l(\varepsilon)$ does not depend on the magnetic quantum number, m, the structure matrix

$$S_{LL'}(\mathbf{k}) \equiv S_{lm,l'm'}(\mathbf{k}) \tag{16.49}$$

contains $(2l+1) \times (2l+1)$ blocks. If one neglects hybridization, i.e. if one sets the elements of $S_{LL'}(\mathbf{k})$ with $l \neq l'$ equal to zero, the unhybridized bands $[\varepsilon_{li}(\mathbf{k})]$ are simply found as the ith solution of the equation

$$\left| P_l(\varepsilon) - S_{lm,lm'}(\mathbf{k}) \right| = 0. \tag{16.50}$$

This motivates the idea of "canonical bands," which are defined to be the $2l+1$ eigenvalues $S_{l,i}(\mathbf{k})$ of the block of the structure constant matrix, $S_{lm,lm'}(\mathbf{k})$, for angular momentum l. Since each $S_{l,i}(\mathbf{k})$ depends only upon the structure, canonical bands can be defined once and for all for any crystal structure. An example of canonical bands is given in Fig. 16.11 for unhybridized s, p and d canonical bands for an fcc crystal plotted along symmetry lines

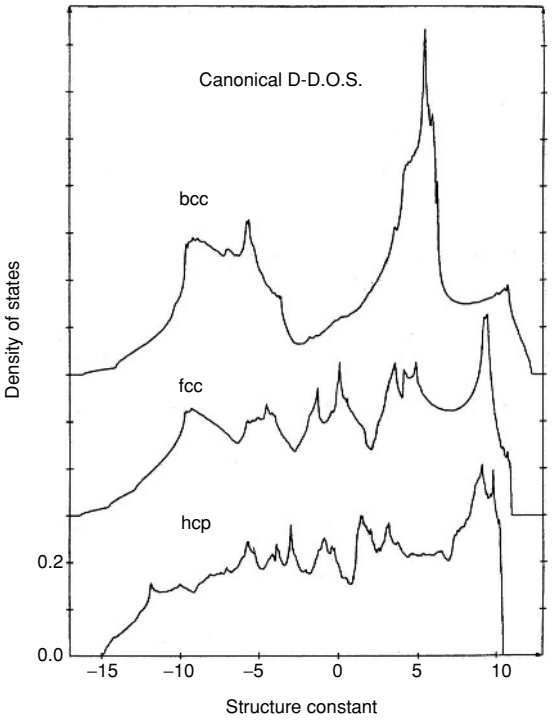


Figure 16.12. Canonical densities of states for unhybridized d bands in the fcc, bcc, and hcp structures. The d states dominate the densities of states for transition metals; the DOS s and p bands are not shown. From [464]. (See also [461], [644], and [643].)

in the Brillouin zone [643, 644, 461, 464]. Similar bands for bcc and hcp can be found in [643, 459, 464] and the canonical bands for hcp along Γ-K are shown in the left panel of Fig. 17.7. The canonical d densities of states for fcc, bcc, and hcp crystals are given in Fig. 16.12. All the information about the actual material-dependent properties are included in the potential function $P(\varepsilon)$.

The potential function $P(\varepsilon)$ captures information about the bands in a material in terms of a few parameters, all of which can be calculated approximately (often accurately) from very simple models. A simple three-parameter form that contains the features shown in Fig. 16.10 is

$$P_l(\varepsilon) = \frac{1}{\gamma}\frac{\varepsilon - C_l}{\varepsilon - V_l}, \tag{16.51}$$

which can be inverted to yield[9]

$$\varepsilon(P_l) = C_l + \gamma(C_l - V_l)\frac{P_l}{1 - \gamma P_l} \equiv C_l + \frac{\hbar^2}{2\mu S^2}\frac{P_l}{1 - \gamma_l P_l}, \tag{16.52}$$

[9] Here we keep the explicit factors of $\hbar$ and m_e to indicate energy units clearly and to avoid confusion with notation in the literature.

where $\hbar^2/2\mu S^2 \equiv \gamma(C_l - V_l)$. This expression has an important physical interpretation with μ an effective mass that sets the scale for the band width. The formulation takes added significance from the fact that μ can be calculated from the wavefunction in the sphere. It is a matter of algebra (Exercise 16.10) to relate μ to the linear energy variation of the logarithmic derivative $D_l(\varepsilon)$ at the band center ($\varepsilon = C_l$, where $D_l(\varepsilon) = -l - 1$), which is then simply related to the value of the wavefunction at the sphere boundary, as given explicitly in (17.10). This expresses the simple physical fact that the band width is due to coupling between sites, which scales with the value of the wavefunction at the atomic sphere boundary as discussed in Sec. 10.7. Expressions can also be derived for γ [643].

Combining (16.52) with (16.50) leads to the unhybridized band structure

$$\varepsilon_{li}(\mathbf{k}) = C_l + \frac{\hbar^2}{2\mu S^2}\frac{S_{l,i}(\mathbf{k})}{1 - \gamma_l S_{l,i}(\mathbf{k})}. \tag{16.53}$$

This formulation provides a simple intuitive formulation of the energy bands in terms of the "canonical bands" $S_{l,i}(\mathbf{k})$, with center fixed by C_l, the width scaled by an effective mass μ, and a distortion parameter γ that is very similar to the effect of a non-orthogonal basis. If γ is small, the canonical bands illustrated in Fig. 16.11 and the DOS in Fig. 16.12 are a scaled version of the actual bands and DOS of a crystal, thus providing a good starting point for understanding the bands. This can be seen from the remarkable similarity to the calculated bands in Figs. 16.4, 2.24, and 16.5 and to the tight-binding bands in Fig. 14.7. (Indeed, the real-space interpretation of canonical bands leads to a new formulation of tight-binding described in Sec. 16.7.) Of course, it is necessary to take into account hybridization to describe the bands fully. This is a notable achievement: *essential features of all five* d *bands are captured by one parameter, the mass in (16.53).* Furthermore, as we shall see in Sec. 17.5, the mass can be determined simply from an atomic calculation. In addition, the bands can be improved by including the parameter γ which distorts the canonical bands as in (16.53), and which can also be calculated from atomic information. Canonical bands can be used to predict tight-binding parameters, which follows from the structure factors in real space and is discussed in the next section. Finally, many important results for real materials can be found simply using the notions of canonical bands; however, the examples are best deferred to Sec. 17.7 to include features of the linear LMTO method.

16.7 Localized "tight-binding" MTO and KKR formulations

The subject of this section is transformations to express the MTO and KKR methods in localized form, with the goal of making possible "first-principles" tight-binding (Ch. 14), localized interpretations, linear scaling methods (Ch. 23), and other developments in electronic structure. The original formulations of the KKR method involve structure constants $B_{LL'}$ in (16.23) that oscillate and decay slowly as a function of distance $|\mathbf{R} - \mathbf{R}'|$. For

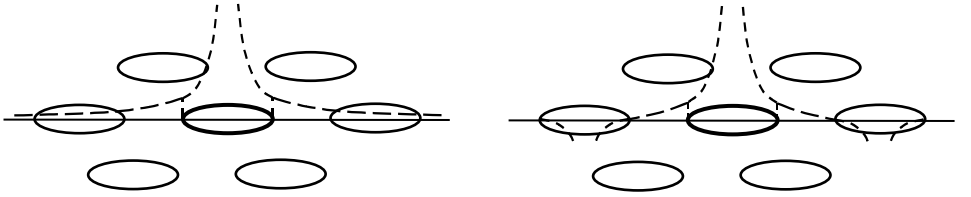


Figure 16.13. Schematic illustration of MTO orbital centered on the site indicated by the dark sphere. Left: A standard MTO. Outside its sphere it decays as a power law with a smooth tail that extends through other spheres. Right: The "screened MTO" that results from linear transformation of the MTOs set of. The essence of the transformation is to add neighboring MTOs with opposite signs as shown; since the tails of the original MTO functions have excatly the same form as the fields due to electronstatic multipoles, the long-range behaviour can be "screened" by a linear combination that cancels each multi-pole field. The transformation to localized functions can also be understood in terms of the construction of Wannier functions; see text.

positive energies[10] the range is so long that it is not possible to make any simple short-range pictures analogous to the local orbital or tight-binding pictures.

The MTO formalism partly remedies this situation to provide a more localized picture. The distance dependence in (16.40) illustrates the important features that emerge from the MTO approach: since $\kappa = 0$ has been shown to be a good approximation in many cases and since all the information about interactions between sites is contained in the structure constants, this identity shows the characteristic feature that interactions between orbitals of angular momenta $L = l, m$ and $L' = l', m'$ decrease as a structure factor

$$S_{LL'}(|\mathbf{R}|) \propto \left[\frac{S}{|\mathbf{R}|}\right]^{l+l'+1} . \tag{16.54}$$

For high angular momenta, the sums converge rapidly, which provides a new formulation of tight-binding [590] in which the matrix elements decay as $(1/r)^{l+l'+1}$ and are derived from the original independent-particle Schrödinger equation.

For $l = 0$ and $l = 1$, however, this does not lead to a simple picture because the sums do not converge rapidly – in fact there are singularities in the longest range terms just as for the Coulomb problem. This is not a pleasant prospect for providing a simple physical picture of electronic states! How can the properties of the MTO basis be interpreted to provide a more satisfying picture? The answer lies in the fact that the long-range terms are Coulomb multi-pole in nature; the distance dependence has inverse power because it is equivalent to the long-range behavior of electrostatic multi-pole fields. By a unitary transformation that is equivalent to "screening of the multipoles" one can transform to a fully localized tight-binding form for all angular momenta [668]. There are many ways of screening multipoles, all having the effect shown in Fig. 16.13 that contributions from neighboring sites are added with opposite sign to give net exponential decay of the basis function. An example

[10] If the problem is transformed to complex energies, e.g. in Green's function method, the range can be short. See Sec. D.4.

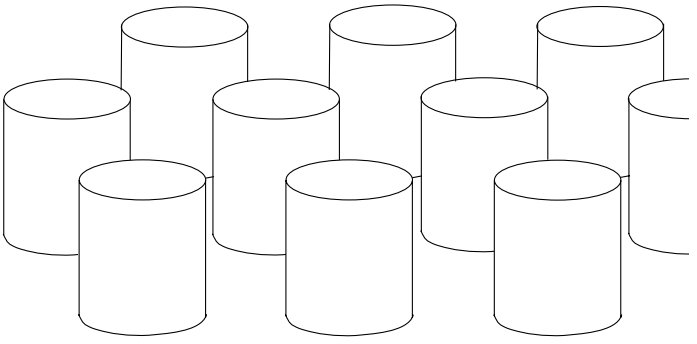


Figure 16.14. Schematic illustration for creation of localized Green's functions G_0 in the KKR method. Because of strong repulsive potentials at every site, G_0 decays exponentially at all energies in the range of the energy bands.

of a transformation choice is given in [668]. The reason that one can transform to a set of exponentially decaying orbitals is not accidental; this properly can be understood using the same ideas as for the construction of Wannier functions. Since the space spanned by the minimal basis MTO hamiltonian is a finite set of bands, bounded both above and below in energy, the transformations given in Ch. 21 can be used to construct localized functions that span this finite basis subspace.

A localized form of KKR also can be generated very straightforwardly, even though the ideas may at first seem counterintuitive. The idea is simply to choose a different reference G_0 instead of the free propagator equation, (16.19), that satisfies (16.18). If G_0 is chosen to be the solution of the Schrödinger equation for a particle in a set of strongly repulsive potentials, as illustrated in Fig. 16.14, then $G_0(\mathbf{r})$ is localized for all energies of interest [652]. Simply inserting this into the Dyson equation, (16.13), leads to a localized form for any of the KKR expressions in Sec. 16.3. The greatest advantage is realized in the Green's function formulation in which the non-linearity is not a problem and the equations can be made fully localized. "Order-N" methods have been developed using this approach (see, e.g. [671] and Ch. 23).

There are important advantages of using augmented localized orbitals over the standard tight-binding-like local orbital approach. A basis of fixed local orbitals has the inherent difficulty that the tails of orbitals extending into the neighboring atoms are far from the correct solution – e.g. they do not obey the correct cusp conditions at the nucleus – and a sufficient number of orbitals (beyond a minimal basis) must be used to achieve the "tail cancellation" that is built into the KKR and MTO methods. In general, local orbitals are non-orthogonal, whereas the transformed MTO basis can be made nearly orthogonal [668, 672].

On the other hand there are disadvantages in the use of KKR and MTO methods. The KKR formalism is more difficult to apply to general low-symmetry problems where the potential is not of muffin-tin form. The localized MTO form has been developed for close-packed systems and application to open structures requires care and often introduction of empty spheres (see also Sec. 17.5). Thus these methods have been applied primarily to

close-packed metals and high-symmetry ionic crystals, but have not been widely applied to molecules, surfaces, and related systems.

16.8 Total energy, force, and pressure in augmented methods

Total energies and related quantities are more difficult to calculate than in the pseudopotential method because of the large energies and strong potentials involved. It is especially important to use appropriate functions for the total energy, such as (9.9) which was derived by Weinert and coworkers [416] explicitly for APW-type methods. Augmented methods have played a key role in total energy calculations since the 1960s when self-consistent calculations became feasible, e.g. for KCl [110], alkali metals [111, 112], and Cu [113]. One of the most complete studies was done by Janak, Moruzzi, and Williams [114, 106], who were pioneers in making Kohn–Sham density functional theory a practical approach to computation of the properties of solids. Their results, shown in Fig. 2.3, were calculated using the KKR method. Many other examples of LAPW calculations are given in Chs. 2 and 17.

Straightforward application of the "force (Hellmann–Feynman) theorem" is fraught with difficulty in any all-electron method. The wavefunctions must be described extremely accurately very near the nucleus in order for the derivative to be accurate, and the wavefunctions must be extremely well converged since the force is not a variational quantity. The problem is in the core electrons. In the atom because of spherical symmetry the force on the nucleus must vanish, which is easy to accomplish since the core states are symmetric. If the nucleus is at a site of low symmetry in a molecule or solid, however, the electric field $\mathbf{E}$ at the nucleus and the net force is non-zero. Even though the core electrons are nearly inert, in fact they polarize slightly and transmit forces to the nucleus. It is only by proper inclusion of the polarized core that one arrives at the correct conclusion that the force due to an electric field on an ion (nucleus plus core) is the "screened" force $\mathbf{F} = Z_{\text{ion}}\mathbf{E}$ instead of the "bare" force $\mathbf{F} = Z_{\text{nucleus}}\mathbf{E}$.

Difficult problems associated with calculation of the force on a nucleus can be avoided by the use of force expression that are alternative to the usual force theorem. As emphasized in App. I, difficult core–nucleus terms can be explicitly avoided by displacing rigidly the core around each nucleus long with the nucleus. The resulting expressions then involve additional terms due to displacement of the core. Although they lack the elegant simplicity of the original force theorem, they can be much more intuitive and appropriate for actual calculations. A method for calculation of forces and stresses in APW (and LAPW) approaches has been developed by Soler and Williams [673] and by Yu, Singh, and Krakauer [674]. The general ideas, given in Sec. I.5, involve finding the force on a sphere in terms of the boundary conditions that transmit forces from the plane waves plus Coulomb forces on the charge in the sphere due to charges outside. The expressions can be found by directly differentiating the explicit APW expressions for the energy.

Within the atomic sphere approximation the pressure can be calculated using the remarkably simple expressions given in Sec. I.3. Only the wavefunctions at the boundary of the sphere are needed. This can be applied in any of the augmented methods, and examples are given using the LMTO approach in Sec. 17.7.

SELECT FURTHER READING

General augmented methods:

Kübler, J., *Theory of Itinerant Electron Magnetism*, Oxford University Press, Oxford, 2001.
Kübler, J., and Eyert, V., in *Electronic and Magnetic Properties of Metals and Ceramics*, edited by K. H. J. Buschow, VCH-Verlag, Weinheim, Germany, 1992, p. 1.
Computational Methods in Band Theory, edited by P. M. Marcus, J. F. Janak, and A. R. Williams, Plenum, New York, 1971.
Ziman, J., in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1971, Vol. 26, pp. 1–101.

APW:

Dimmock, J. O, in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1971, Vol. 26, pp. 104–274.
Loucks, T., *The Augmented Plane Wave Method*, Benjamin, New York, 1967.
Singh, D. J., *Planewaves, Pseudopotentials, and the APW Method*, Kluwer Academic Publishers, Boston, 1994, and references therein.
Slater, J. C., *Symmetry and Energy Bands in Crystals (Corrected and reprinted version of 1965 Quantum Theory of Molecules and Solids, Vol. 2)*, Dover, New York, 1972.
Slater, J. C., *The Self-Consistent Field Theory for Molecules and Solids: Quantum Theory of Molecules and Solids, Vol. 4*, McGraw-Hill, New York, 1974.

Multiple scattering and KKR:

Butler, W. H., Dederichs, P. H., Gonis, A., and Weaver, R. L., *Applications of Multiple Scattering Theory to Material Science*, Materials Reserach Society, Pittsburg, Penn., 1992.
Lloyd, P., and Smith, P. V., "Multiple scattering theory in condensed materials," *Adv. Phys.* 21:29, 1972.

(L)MTO:

Andersen, O. K., "Linear methods in band theory," *Phys. Rev. B* 12:3060–3083, 1975.
Andersen, O. K., and Jepsen, O., "Explicit, first-principles tight-binding theory," *Physica* 91B:317, 1977.
Skriver, H., *The LMTO Method*, Springer, New York, 1984.

16.9 Exercises

16.1 The basic ideas of the APW method can be illustrated by a one-dimensional Schrödinger equation for which the solution is given in Exercise 4.22. In addition, close relations to pseudopotentials, plane wave, KKR, and MTO methods are brought out by comparison with Exercises 11.14, 12.6, 16.7, and 16.13. Consider an array of potentials $V(x)$ spaced by lattice constant a; $V(x)$ is arbitrary except that it is assumed to be like a muffin-tin composed of non-overlapping potentials with $V(x) = 0$ in the interstitial regions. For actual calculations it is useful to treat the case where $V(x)$ is a periodic array of square wells.
(a) Consider the deep well defined in Exercise 11.14 with width $s = 2a_0$ and depth $-V_0 = -12Ha$. Solve for the two lowest states (analogous to "core" states) using the approximation that they are bound states of an infinite well.

(b) Construct APW functions that are e^{ikx} outside the well; inside, the APW is a sum of solutions at energy ε (as yet unknown) that matches e^{ikx} at the boundary. Show that the expansion inside the cell analogous to (16.2), and the plane wave expansion, analogous to (16.5), are sums only over two terms, sine and cosine, and give the explicit form for the APW.
(c) Derive the explicit APW hamiltonian for this case. Include the terms from the discontinuity of the derivative. Show that the equation has the simple interpretation of plane waves in the interstitial with boundary conditions due to the well.
(d) Construct a computer code to solve for the eigenvalues and compare to the results of the general method described in Exercise 4.22.
(e) Use the computer code also to treat the shallow square well defined in Exercise 12.6 and compare with the results found there using the plane wave method.
(f) Compare and contrast the APW, plane wave, and the general approach in Exercise 4.22.

16.2 Derive the form for the contribution to the hamiltonian matrix elements from the kink in the wavefunctions given in Eq. (16.9) using Green's identity to transform to a surface integral.

16.3 Derive the identity given in (16.22)–(16.24) for the expansion of a spherical wave defined about one center in terms of spherical waves about another center. One procedure is through the use of Eq. J.1, which is also given in (16.5).

16.4 Evaluate values for the logarithmic derivatives of the radial wavefunctions for free electrons and compare with the curves shown in Fig. 16.3 for Cu. The expressions follow from (16.5) (also given in Eq. (J.1)) for zero potential and the functions should be evaluated at the radius $S = 2.415a_0$ appropriate for metallic Cu.

16.5 Show that the nearly parabolic band energies for Cu in Fig. 2.24 are well approximated by free-electron values given that Cu has fcc crystal structure with cube edge $a = 6.831a_0$. Show also that the states at the zone boundary labeled would be expected to act like p states (l=1, odd) about each atom. (Quantitative comparisons are given in [134], p. 25).

16.6 As the simplest example of the "s–d" hybridization model, derive the bands for a 2×2 hamiltonian for flat bands crossing a wide band in one dimension: $H_{11}(k) = E_1 + W\cos(2\pi k/a)$, $H_{22}(k) = E_2$, and $H_{12}(k) = H_{21}(k) = \Delta$. Find the minimum gap, and the minimum direct gap in the bands. Show that the bands have a form resembling the bands in a transition metal.

16.7 The KKR method can be illustrated by a one-dimensional Schrödinger equation, for which the solution is given in Exercise 4.22. See [664] for an extended analysis. Close relations to pseudopotentials, plane wave, APW, and MTO methods are brought out by comparison with Exercises 11.6, 12.6, 16.1, and 16.13. As in Exercise 16.1, the KKR approach can be applied to any periodic potential $V(x)$. The KKR solution is then given by (16.29) with the structure constants defined in (16.27). (Here we assume $V(x)$ is symmetric in each cell for simplicity. If it is not symmetric there are also cross terms η^{+-}.)
(a) The phase shifts are found from the potential in a single cell. In Exercise 11.6 it is shown that the scattering is described by two phase shifts η^+ and η^-.
(b) In one dimension the structure constants define a 2×2 matrix $B_{L,L'}(\varepsilon, \mathbf{k})$, with $L = +, -$ and $L' = +, -$. Each term is a sum of exponentials that oscillates and does not converge at large distance. Find physically meaning expressions for $B_{L,L'}(\varepsilon, \mathbf{k})$ by adding a damped exponential convergence factor.

(c) Using the relations from Exercise 11.6, show that the KKR equations lead to the same results as the general solution, (4.49), with $\delta = \eta^+ + \eta^-$ and $|t| = \cos(\eta^+ - \eta^-)$.

16.8 This exercise is to show the relation of the Green's function expression, (16.32), and the Schrödinger equation. This can be done in four steps that reveal subtle features.
(a) Show that application of the free-electron hamiltonian $\hat{H}_0$ to both sides of the equation leads to a Schrödinger-like equation but without the eigenvalue. Hint: Use the fact that $\hat{H}_0\, G_0 = \delta(|\mathbf{r} - \mathbf{r}'|)$.
(b) Show that this is consistent with the Schrödinger equation using the fact that a constant shift in V has no effect on the wavefunction.
(c) Give an auxiliary equation that allows one to find the eigenvalue.
(d) Finally, give the expression for the full Green's function G analogous to (16.13) from which one can derive the full spectrum of eigenvalues.

16.9 Show that $\chi_L^{\mathrm{MTO}}(\varepsilon, 0, \mathbf{r})$, defined in (16.39) is continuous and has continuous derivative (i.e. D is the same inside and outside) at the boundary $r = S$.

16.10 Find the relation of the mass parameter μ to the energy derivative $dD(E)/dE$ evaluated at the band center, assuming P_l has the simple form given in (16.51).

16.11 The diamond structure can be viewed as a dense-packed structure of touching spheres with some spheres not filled with atoms. Show this explicitly, starting with the crystal structure shown in Fig. 4.7 and insert empty spheres in the holes in the structure.

16.12 Show that (16.39) indeed leads to a function that is continuous and has a continuous derivative at its boundary.

16.13 The MTO method can be illustrated by a one-dimensional Schrödinger equation. The purpose of this exercise is to show that the solution in Exercises 4.22 and 16.7 can be viewed as "tail cancellation." (An extended analysis can be found in [675].) This re-interpretation of the equations can be cast in terms of the solutions of the single cell problem given in Exercise 4.22, ψ_l and ψ_r, which correspond to waves incident from the left and from the right; only the part outside the cell is needed. Consider the superposition of waves inside a central cell at $T = 0$ formed by the sum of waves $\psi_l(x)$ and $\psi_r(x)$ from all *other* cells at positions $T \neq 0$ with a phase factor $\mathrm{e}^{\mathrm{i}kT}$. Show that the requirement that the sum of waves from all other cells vanishes at any point x in the central cell (i.e. tail cancellation) and leads to the same equations as in Exercises 4.22 and 16.7.

17

Augmented functions: linear methods

Summary

The great disadvantage of augmentation is that the basis functions are energy dependent, so that matching conditions must be satisfied separately for each eigenstate at its (initially unknown) eigenenergy. This leads to non-linear equations that make such methods much more complicated than the straightforward linear equations for the eigenvalues of the hamiltonian expressed in fixed energy-independent bases such as plane waves, atomic orbitals, gaussians, etc. *Linearization* is achieved by defining augmentation functions as linear combinations of a radial function $\psi(E_\nu, r)$ and its energy derivative $\dot{\psi}(E_\nu, r)$ evaluated at a chosen fixed energy E_ν. In essence, $\psi(E_\nu, r)$ and $\dot{\psi}(E_\nu, r)$ form a basis adapted to a particular system that is suitable for calculation of all states in an energy "window." Any of the augmented methods can be written in linearized form, leading to secular equations like the familiar ones for fixed bases. The simplification has other advantages, e.g. it facilitates construction of full potential methods not feasible in the original non-linear problem. In addition, the hamiltonian thus defined leads to *linear methods* that take advantage of the fact that the original problem has been reduced to a finite basis. This approach is exemplified in the LMTO method which defines a minimal basis that both provides physical insight and quantitative tools for interpretation of electronic structure.

It should be emphasized from the outset that the terms "non-linear" and "linear" have *nothing to do with the fundamental linearity of quantum mechanics*. Linearity of the governing differential equation, the Schrödinger equation, is at the heart of the quantum nature of electrons and any non-linearities would have profoundly undesirable consequences. *Linearization* and *linear methods* have to do with practical matters of solving and interpreting the independent-particle Schrödinger equations.

Formulations in which the wavefunctions are expressed as linear combinations of fixed basis functions, such as plane waves, gaussians, atomic-like orbitals, *etc.*, are manifestly linear. This leads directly to standard linear algebra eigenvalue equations, which is a great advantage in actual calculations. Since the same basis is used for all states, it is simple to express the conditions of superposition, orthogonality, *etc.*, and it is simple to determine many eigenfunctions together in one calculation.

Augmented methods are also linear in the fundamental sense that the wavefunctions can be expressed as linear combinations of basis functions. However, non-linear equations for the eigenstates arise because the basis is energy dependent. This choice has great advantages, effectively representing electronic wavefunctions both near the nucleus and in the interstitial regions between the atoms. But there is a high price. The matching conditions lead to non-linear equations due to the intrinsic energy-dependence of the phase shifts that determine the scattering from the atoms. This results in greatly increased computational complexity since each eigenstate must be computed separately, as described in Ch. 16.

Linearization of non-linear equations around selected reference energies allows the construction of operators that act in the same way as ordinary familiar linear operators, while at the same time taking advantage of the desirable attributes of the augmentation and achieving accurate solutions by choice of the energies about which the problem is linearized. The LAPW approach illustrates clearly the advantages of linearization.

Linear methods result from the same process, but lead to different formulation of the problem. The resulting hamiltonian matrix is expressed in terms of the wavefunctions and their energy derivatives, which are determined from the original independent-particle Schrödinger equation. Thus this defines a hamiltonian matrix in a reduced space. Working with this derived hamiltonian leads to a class of linear methods that provide physically motivated interpretations of the electronic structure in terms of a minimal basis. This is the hallmark of the LMTO method.

In a nutshell the key idea of linearization is to work with two augmented functions, $\psi_l(r)$ and its energy derivative denoted by $\dot{\psi}_l(r)$, each calculated at the chosen reference energy. These two functions give greater degrees of freedom for the augmentation, which allow the functions to be continuous and to have continuous derivatives at the matching boundaries. However, *the basis does not double in size*: the energy dependence is taken into account to first order by the change of the wavefunction with energy. The wavefunction is correct to first order $\propto (\Delta\varepsilon)$, where $\Delta\varepsilon$ is the difference of the actual energy from the chosen linearization energy; therefore, the energies are correct to $(\Delta\varepsilon)^2$ and variational expressions [643] are correct to $(\Delta\varepsilon)^3$, illustrating the "$2n + 1$" theorem (Sec. 3.7) which is important in actual applications. The methods can be used with any augmentation approach, and have led to the widely used LAPW, LMTO, and other methods.

17.1 Energy derivative of the wavefunction: ψ and $\dot{\psi}$

In this section we assume that the potential has muffin-tin form, i.e. spherically symmetric within a sphere of radius S about each atom and flat in the interstitial. The equations can be generalized to non-muffin-tin potentials using the same basis functions. Initially, we consider a single spherical potential. The analysis involves radial equations exactly like those for atoms and scattering problems (Sec. J.1) and the analysis has useful relations to the derivation of norm-conserving pseudopotentials in Sec. 11.4 although the application is quite different.

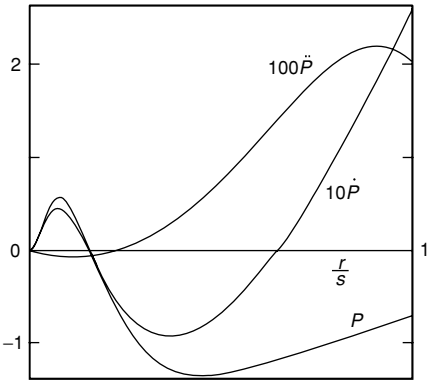


Figure 17.1. Radial d function, $P \equiv S^{1/2}\psi(r)$, and energy derivatives, $\dot{P} \equiv S^{-2}\partial P/\partial E$, and $\ddot{P} \equiv S^{-4}\partial^2 P/\partial E^2$, for ytterbium. From [644].

The goal is to sidestep the problems of the non-linear methods of Ch. 16. Linearized methods achieve this by expanding the solution of the single-sphere Schrödinger equation in terms of $\psi_l(\varepsilon, r)$ belonging to one arbitrarily chosen energy, $\varepsilon = E_\nu$, i.e.[1]

$$\left(-\frac{\hbar^2}{2m_e}\frac{\mathrm{d}^2}{\mathrm{d}r^2} + V_{\text{sphere}} - E_\nu\right) r\psi_l(E_\nu, r) = 0 \tag{17.1}$$

and its energy derivative

$$\dot{\psi}(\varepsilon, r) \equiv \frac{\partial}{\partial\varepsilon}\psi(\varepsilon, r)|_{\varepsilon=E_\nu}. \tag{17.2}$$

If we define the derivative with respect to energy to mean a partial derivative keeping ψ normalized to the same value in the sphere (even though it is not an eigenfunction at an arbitrary E_ν), then it is easy to show that ψ and $\dot{\psi}$ are orthogonal,

$$\langle\psi|\dot{\psi}\rangle \;=\; 0, \tag{17.3}$$

so that the two functions indeed span a larger space. Furthermore, one can readily show that

$$(\hat{H} - \varepsilon)\dot{\psi}(\varepsilon, r) = \psi(\varepsilon, r) \tag{17.4}$$

and similar relations in Eq. (17.38) in Exercise 17.1. It is also straightforward to show that each successive energy derivative of the function $\psi(r)$ is given by simple relations like

$$\langle\dot{\psi}|\dot{\psi}\rangle \;=\; -\frac{1}{3}\frac{\ddot{\psi}(S)}{\psi(S)}. \tag{17.5}$$

The functions ψ, $\dot{\psi}$, and $\ddot{\psi}$ are illustrated in Fig. 17.1 for ytterbium [644], where it is shown that each order of derivative corresponds to a decrease in size by an order of magnitude.

[1] The factor $\hbar^2/2m_e$ is included explicitly to avoid confusion with the equations given in other sources.

The augmentation functions as a function of energy can be specified in terms of the dimensionless logarithmic derivative, which is defined as

$$D(\varepsilon) = \left[\frac{r}{\psi(\varepsilon, r)} \frac{\mathrm{d}\psi(\varepsilon, r)}{\mathrm{d}r} \right]_{r=S}. \tag{17.6}$$

The linear combination of ψ and $\dot{\psi}$ that has logarithmic derivative D is given by

$$\psi(D, r) \;=\; \psi(r) \;+\; \omega(D)\, \dot{\psi}(r), \tag{17.7}$$

where $\omega(D)$ has dimensions of energy and is given by

$$\omega(D) \;=\; -\frac{\psi(S)}{\dot{\psi}(S)} \frac{D - D(\psi)}{D - D(\dot{\psi})}, \tag{17.8}$$

with $D(\dot{\psi})$ denoting the logarithmic derivative of $\dot{\psi}$. If $\psi(r)$ and $\dot{\psi}(r)$ are calculated at a reference energy E_ν, then (17.7) is the wavefunction to first order in the energy $E(D) - E_\nu$. It then follows that the variational estimate of the eigenvalue,

$$E(D) = \frac{\langle \psi(D) | \hat{H} | \psi(D) \rangle}{\langle \psi(D) | \psi(D) \rangle} = E_\nu + \frac{\omega(D)}{1 + \omega(D)^2 \langle \dot{\psi}(D) | \dot{\psi}(D) \rangle}, \tag{17.9}$$

is correct to third order and the simpler expression $E_\nu + \omega(D)$ is correct to second order [459].

The logarithmic derivative at the sphere radius S can also be expressed in a Taylor series in $E - E_\nu$. The first term is given by the analysis in Sec. 11.4, where Eq. (11.28) shows that to first order

$$D(E) - D(E_\nu) = -\frac{m_e}{\hbar^2} \frac{1}{S \psi_l(S)^2} (E - E_\nu), \tag{17.10}$$

where we have substituted $\psi = \phi / r$. (The factor $m_e/\hbar^2 = 1$ in Hartree atomic units and $m_e/\hbar^2 = 1/2$ in Rydberg units.) In deriving (17.10) from (11.28), it is assumed that, the charge $Q_l(S)$ in the sphere is unity. This is not essential to the logic, but it is convenient and valid in the atomic sphere approximation, and is a good approximation in many cases. Higher-order expressions are given in [459] and [643] and related expressions in the pseudopotential literature in [497].

17.2 General form of linearized equations

We are now in a position[2] to define an energy-independent orbital $\chi_j(\mathbf{r})$ everywhere in space for a system of many spheres,

$$\chi_j(\mathbf{r}) = \chi_j^e(\mathbf{r}) + \sum_{L,s} \left[\psi_{l,s}(\mathbf{r} - \tau_s) \Pi_{Lsj} + \dot{\psi}_{l,s}(\mathbf{r} - \tau_s) \Omega_{Lsj} \right] i^l Y_L(\widehat{\mathbf{r} - \tau_s}), \tag{17.11}$$

where $\psi_{l,s}$ and $\dot{\psi}_{l,s}$ are the radial functions in each sphere and Π and Ω are factors to be determined. Between the spheres the function is defined by the envelope function $\chi_j^e(\mathbf{r})$,

[2] This section follows the approach of Kübler and V. Eyert [134].

which is a sum of plane waves in the LAPW method. In the LMTO approach, $\chi_j^e(\mathbf{r})$ is a sum of Neumann or Hankel functions as specified in (16.37) or is proportional to $(r/S)^{-l-1}$ in the $\kappa = 0$ formulation of (16.39).

The explicit form of the linearized equations depends upon the choice of the envelope function, but first we can give the general form. The result is quite remarkable: because of the properties of ψ and $\dot{\psi}$ expressed in (17.3) and (17.4) (see Exercise 17.1), the form of the hamiltonian can be greatly simplified. Furthermore, the "hamiltonian" is expressed in terms of the solution for the wavefunctions; this allows a reinterpretation of the problem as a strictly linear solution of the new hamiltonian expressed as matrix elements in the reduced space of states that span an energy range around the chosen linearization energy.

For a crystal the label s can be restricted to the atoms in one cell, and a basis function with Bloch symmetry can be defined at each $\mathbf{k}$ by a sum over cells $\mathbf{T}$,

$$\psi_{Ls\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{T}} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T}}\psi_{Ls}(\mathbf{r}-\mathbf{T}), \tag{17.12}$$

and similarly for $\dot{\psi}_{Ls}$, so that (17.11) becomes

$$\chi_{j\mathbf{k}}(\mathbf{r}) = \chi^e_{j\mathbf{k}}(\mathbf{r}) + \sum_{L,s}\left[\psi_{Ls\mathbf{k}}(\mathbf{r})\Pi_{Lsj}(\mathbf{k}) + \dot{\psi}_{Ls\mathbf{k}}(\mathbf{r})\Omega_{Lsj}(\mathbf{k})\right] i^l Y_L(\widehat{\mathbf{r}-\tau_s}). \tag{17.13}$$

The wavefunction is defined by the coefficients Π_{Lsj} and Ω_{Lsj} that are constructed at each $\mathbf{k}$ so that the basis function $\chi_{j\mathbf{k}}(\mathbf{r})$ satisfies the continuity conditions, with actual equations that depend upon the choice of basis (see sections below). The construction of the hamiltonian $H_{ij}(\mathbf{k})$ and overlap matrices $S_{ij}(\mathbf{k})$ can be divided into the envelope part and the interior of the spheres at each $\mathbf{k}$, yielding (see Exercise 17.2)

$$S_{ij}(\mathbf{k}) = \langle i\mathbf{k}|j\mathbf{k}\rangle^e + \sum_{L,s}\left[\Pi^\dagger_{Lsi}(\mathbf{k})\Pi_{Lsj}(\mathbf{k}) + \Omega^\dagger_{Lsi}(\mathbf{k})\langle\dot{\psi}_{ls}|\dot{\psi}_{ls}\rangle\Omega_{Lsj}(\mathbf{k})\right]. \tag{17.14}$$

and

$$H_{ij}(\mathbf{k}) - E_v S_{ij}(\mathbf{k}) = \langle i\mathbf{k}|H - E_v|j\mathbf{k}\rangle^e + \sum_{L,s}\Pi^\dagger_{Lsi}(\mathbf{k})\Omega_{Lsj}(\mathbf{k}). \tag{17.15}$$

The secular equation $\sum_j \left[H_{ij}(\mathbf{k}) - \varepsilon S_{ij}(\mathbf{k})\right] a_j(\mathbf{k}) = 0$ becomes

$$\sum_j \left[\langle i\mathbf{k}|H - E_v|j\mathbf{k}\rangle^e + V_{ij}(\mathbf{k}) - \varepsilon' S_{ij}(\mathbf{k})\right] a_j(\mathbf{k}) = 0, \tag{17.16}$$

where $\varepsilon' = \varepsilon - E_v$ is the energy relative to E_v.[3] The potential operator acting inside the spheres is given by

$$V_{ij}(\mathbf{k}) = \frac{1}{2}\sum_{Ls}\left[\Pi^\dagger_{Lsi}(\mathbf{k})\Omega_{Lsj}(\mathbf{k}) + \Pi_{Lsi}(\mathbf{k})\Omega^\dagger_{Lsj}(\mathbf{k})\right], \tag{17.17}$$

which has been made explicitly hermitian [414]. Note that unlike the APW operator V^{APW} in (16.10), there is no energy dependence in $V_{ij}(\mathbf{k})$. The linear energy dependence is absorbed into the overlap term $\varepsilon' S_{ij}(\mathbf{k})$ in (17.16).

[3] For simplicity, a single linearization energy E_v is used here; in general, E_v depends on l and s, leading to expressions that are straightforward but more cumbersome.

As promised, the resulting equations are remarkable, with the "hamiltonian" expressed in terms of Π and Ω, i.e. in terms of the wavefunctions ψ and $\dot{\psi}$ calculated in the sphere at the chosen energy E_v. However, this is not the whole story. It would appear that the basis must be doubled in size by adding the function $\dot{\psi}$ along with each ψ; this is exactly what happens in the usual local orbital formulation where one possible way to improve the basis is by adding $\dot{\psi}$ to the set of basis functions. Similarly, the basis is doubled in the related "augmented spherical wave" (ASW) approach [678], which uses functions at nearby energies $\psi(E_v)$ and $\psi(E_v + \Delta E)$ instead of $\psi(E_v)$ and $\dot{\psi}(E_v)$. However, as we shall see in the following two sections, there is a relation between Π and Ω provided by the boundary conditions. Therefore, it will turn out that *the basis does not double in size*, but nevertheless the wavefunction is correct to linear order in $\varepsilon - E_v$. Thus errors in the energy are $\propto (\varepsilon - E_v)^2$, and variational estimates of the energy ([643], Sec. 3.5) are accurate to $\propto (\varepsilon - E_v)^3$, an example of the "$2n + 1$" theorem, Sec. 3.7.

17.3 Linearized augmented plane waves (LAPWs)

If we choose a plane wave for the envelope function, we obtain the LAPW method [414] (see also [677, 679–682] and on-line information referred to in Ch. 24). The quantum label j becomes a reciprocal lattice vector $\mathbf{G}_m$ and the form of (16.2) for an APW can be adapted

$$\chi^{\mathrm{LAPW}}_{\mathbf{k}+\mathbf{G}_m}(\mathbf{r}) = \begin{cases} \exp(\mathrm{i}(\mathbf{k}+\mathbf{G}_m)\cdot\mathbf{r}), & r > S, \\ \sum_{Ls} C_{Ls}(\mathbf{k}+\mathbf{G}_m)\psi_{Ls}(D_{ls|\mathbf{K}_m|}, \mathbf{r}) i^l Y_L(\widehat{\mathbf{r}-\tau_s}), & r < S, \end{cases} \tag{17.18}$$

where s denotes the site in the unit cell, $L \equiv l, m_l$, $\mathbf{K}_m \equiv \mathbf{k} + \mathbf{G}_m$. The solution inside the sphere of radius S_s is fixed by matching the plane wave, requiring the function to be continuous and have continuous first derivative. This boundary condition leads to $\psi_{ls}(D_{ls|\mathbf{K}_m|}, \mathbf{r})$ as a combination of ψ_{ls} and $\dot{\psi}_{ls}$ as given below. It is this step that includes energy dependence to first order without increasing the size of the basis.

Since the expansion of the plane wave is given by Eq. (J.1), this is accomplished if the logarithmic derivative is the same as for the plane wave,

$$D_{lsK} = \left[x \frac{j_l'(x)}{j_l(x)} \right]_{x=KS_s}, \tag{17.19}$$

which fixes the solution inside the sphere s for a given L and $\mathbf{K}$ to be given by (see Eq. (17.7))

$$\psi_{ls}(D_{lsK}, r) \;=\; \psi_{ls}(r) \;+\; \omega_{lsK} \;\; \dot{\psi}_{ls}(r), \tag{17.20}$$

and the total solution in the sphere is given by (16.6) with

$$j_l(K_m r) \;\rightarrow\; \frac{j_l(K_m S_s)}{\psi_{ls}(D_{lsK}, S_s)} \psi_{ls}(D_{lsK}, r). \tag{17.21}$$

Thus coefficients Π and Ω are given by

$$\Pi_{Ls\mathbf{G}_m}(\mathbf{k}) \;=\; 4\pi\; \mathrm{e}^{\mathrm{i}\mathbf{K}_m\cdot\tau_s}\; \frac{j_l(K_m S_s)}{\psi_{ls}(D_{ls\,K_m},\, S_s)} Y_L(\hat{K}_m) \tag{17.22}$$

and

$$\Omega_{Ls\mathbf{G}_m}(\mathbf{k}) \;=\; \Pi_{Ls\mathbf{G}_m}(\mathbf{k})\;\; \omega_{lsG_m}, \tag{17.23}$$

where $\mathbf{K}_m = \mathbf{k} + \mathbf{G}_m$ and $K_m = |\mathbf{K}_m|$.

The resulting equations have exactly the same form as the APW equations (16.10) and (16.12), with the addition of the overlap term and the simplification that the operator $V^{\mathrm{LAPW}}_{\mathbf{G}',\mathbf{G}}(\mathbf{k})$ is independent of energy. The explicit kinetic energy terms are the same as in (16.10), which is energy-independent. The remaining terms in the secular equation, (17.16), involving Π, Ω, and the overlap can be used conveniently in actual calculations [414] in the form given in (17.14)–(17.16) with relations (17.22) and (17.23). The expressions can also be transformed into a form for $V^{\mathrm{LAPW}}_{\mathbf{G}',\mathbf{G}}(\mathbf{k})$ that is very similar to the APW expression (16.12), with additional terms but with no energy dependence [132, 134].

Major advantages of the LAPW method are its general applicability for different materials and structures, its high accuracy, and the relative ease with which it can treat a general potential (Sec. 17.9). Disadvantages are increased difficulty compared to plane wave pseudopotential methods (so that it is more difficult to develop techniques such as Car–Parrinello simulations based upon the LAPW method) and the fact that a large basis set is required compared to KKR and LMTO methods (so that it is more difficult to extract the simple physical interpretations than for those methods).

More on the LAPW basis

How large is the basis required in realistic LAPW calculations? A general idea can be derived from simple reasoning [414]. The number of plane waves, chosen to have wavevector $G < G_{\max}$, is expected to be comparable to pseudopotential calculation for materials without d and f electrons (since the rapidly varying part of the d and f states are taken care of by the radial functions), e.g. $\approx$100 plane waves/atom, typical for high-quality pseudopotential calculations on Si (see also Sec. 16.2). The size of the basis is somewhat larger than for the APW since each function is continuous in value and slope. The expansion in angular harmonics is then fixed by the requirement that the plane waves continue smoothly into the sphere of radius S: since a Y_{lm} has $2l$ zeros around the sphere, an expansion up to $l_{\max}$ can provide resolution $\approx 2\pi S/l_{\max}$ in real space or a maximum wavevector $\approx l_{\max}/S$. Thus, in order for the angular momentum expansion to match smoothly onto plane waves up to the cutoff $G_{\max}$, one needs $l_{\max} \approx SG_{\max}$, which finally results in $l_{\max} \approx 8$ (see Exercise 17.3) and larger for accurate calculations in complex cases [414, 683].

17.4 Applications of the LAPW method

The LAPW method, including the full-potential generalization of Sec. 17.9, is the most accurate and general method for electronic structure at the present time. The calculations

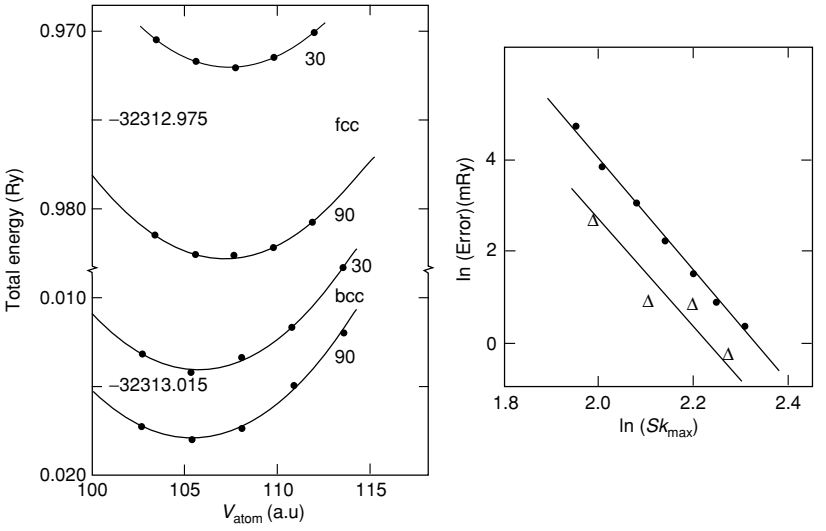


Figure 17.2. Full potential LAPW calculations of the total energy of W in bcc and fcc structures. Note the break in the vertical scale. The two curves for each structure denote integrations over the irreducible BZ with 30 and 90 points, respectively. The absolute value of the energy is given in the figure on the left. On the right is shown the convergence with plane wave cutoff $Sk_{\max}$, where S is the radius of the muffin-tin sphere. From [684].

can be done for structures of arbitrary symmetry with no bias if the basis is extended to convergence. Extensive tests of convergence are illustrated in Fig. 17.2 taken from the work of Jansen and Freeman [684] in the early development of the full-potential LAPW method. The figure shows the total energy of W in the bcc and fcc crystal structures as a function of volume. The total energy is $\approx -16{,}156$ Ha and the energy is converged to less than 0.001 Ha, including the basis set convergence and integration over the BZ. On the right-hand side of Fig. 17.2 is shown the convergence as a function of plane wave cutoff $k_{\max}$ plotted on a logarithmic scale.

Examples of comparisons of LAPW with other methods are given in Tab. 13.1. When done carefully, pseudopotential and PAW agree well with LAPW for most cases. Exceptions are materials in which there is significant core relaxation, such as the Ca core in CaF_2. Although core relaxation can, in principle, be included in pseudopotential methods, it requires special efforts not yet developed. LAPW calculations automatically include all the core states, so that it is straightforward to include relaxation (or exclude it when not needed), relativistic effects [446], nuclear magnetic resonance chemical shifts, electric field gradients at the nucleus, and many other effects.

However, the generality and accuracy of LAPW comes at a price: there is a large basis set of plane waves and high angular momentum functions which in turn means that the potentials must be represented accurately (to twice the cutoffs in wavevectors and angular moments used for the wavefunctions) as described in Sec. 17.9. Other methods are faster, in which case LAPW calculations can serve as a check. Other methods are much more

adaptable for generation of new developments that are the subject of Part V, Chs. 18–23. In fact all the developments of quantum molecular dynamics, polarization and localization, excitations, and O (N) methods were stimulated by other approaches and have been adapted to LAPW in only a few cases.

Examples of total energies and bands have already been shown in Ch. 2. One is the energy versus displacement for the unstable optic mode that leads to the ferroelectric distortion in $BaTiO_3$ shown in Fig. 2.8. The LAPW results [142] are the standard to which the other calculations are compared for this relatively simple structure. As shown in Fig. 2.8, local orbital pseudopotential methods (and also plane wave calculations) give nearly the same results when done carefully. When using a pseudopotential, it has been found to be essential to treat the Ba semicore states as valence states for accurate calculations.

The other examples in Ch. 2 are MgB_2 and graphite, for which LAPW bands are shown in Fig. 2.29. These are cases involving s and p states where pseudopotentials are traditionally applied and, indeed, pseudopotentials give essentially identical results. There are advantages with LAPW, nevertheless, since the same codes and the same level of approximation can be utilized for these open structures with light atoms as for materials with heavy atoms and d and/or f states. Also the DOS for ferromagnetic Fe is shown in Fig. 14.10, where it is compared to tight-binding fit to the LAPW bands.

Perhaps the most important class of application in which the LAPW approach is particularly adapted are compounds involving transition metals and rare earth elements. Understanding many properties of these interesting materials often involves small energy differences due to magnetic order and/or lattice distortions. Linearization simplifies the problem so that one can use full potential methods with no shape approximations. Since the LAPW approach describes the wavefunctions with unbiased spherical and plane waves, it is often the method of choice.

Perhaps the best example are the bands and total energies for the high-temperature superconductors [683]. For, example, the structure of $YBa_2Cu_3O_7$ is shown in Fig. 17.3. There are two CuO_2 planes that form a double layer sandwiching the Y atoms, one CuO chain and two Ba–O layers per cell. The structure must be optimized with respect to all the degrees of freedom and three independent cell parameters. The O atoms in the planes are not exactly in the same plane as the Cu atoms and the "dimpling" has significant effects on the bands. The process of energy minimization with respect to the atomic positions leads to comparison with experiment, including phonon energies that are found to be in very good agreement with experiment, e.g. for $YBa_2Cu_3O_7$ in [685].

Figure 17.4, as an example, shows one of the host of calculations [683] that have led to similar conclusions. The most important conclusion is that the states near the Fermi energy are primarily the one simple band made up of states that involve the Cu $d_{x^2-y^2}$ and O p states in an anti-bonding combination. The number of electrons is just enough to almost fill the bands, and there is one hole per Cu atom in the band that crosses the Fermi energy (Exercise 17.8). The properties of this band on a square lattice representing one plane have been emphasized in Sec. 14.5 and a quantitative description of this band, disentangled from the rest, is given later in Figs. 17.10 and 17.11.

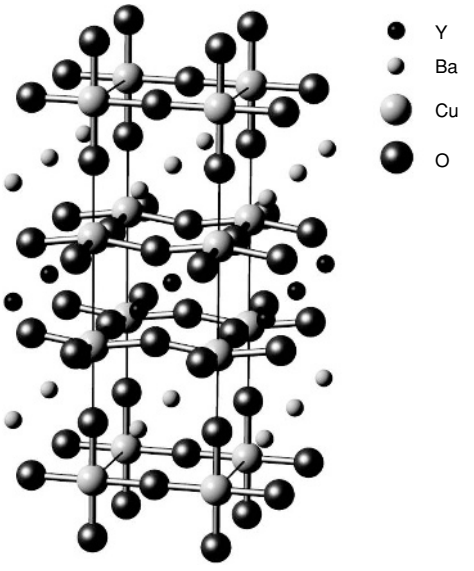


Figure 17.3. Crystal structure of $YBa_2Cu_3O_7$ showing two CuO_2 planes that form a double layer sandwiching the Y atoms, the CuO chain, and the two Ba–O layers per cell. The orthorhombic BZ is shown with the y-direction along the chain axis. Other high-temperature superconductors have related structures all involving CuO_2 planes. Provided by W. Pickett, similar to Fig. 6 in [683].

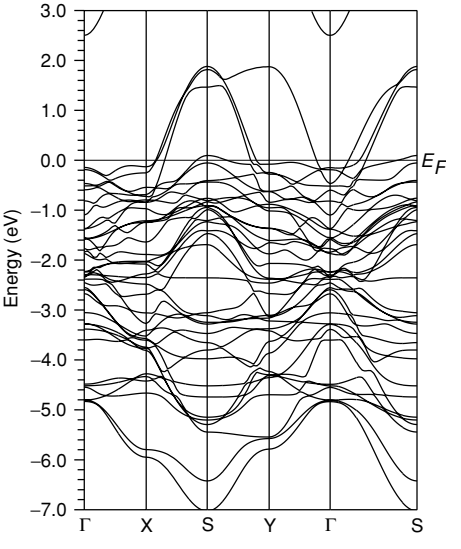


Figure 17.4. Band structure of $YBa_2Cu_3O_7$ computed using the LAPW method [686]. Other calculations [683] with various methods give essentially the same results. The band that protrudes upward from the "spaghetti" of other bands is the antibonding (out-of-phase) Cu–O band that is mainly O 2p in character. Simple counting of electrons (Exercise 17.8) shows that the highest band has one missing electron per Cu, leading to the Fermi level in the band, as shown. Provided by W. Pickett, similar to Fig. 25 in [683].

Thus the Kohn–Sham equations indicate which states are important at the Fermi energy. Yet there is a fundamental failure of the simplest forms of density functional theory, i.e. the LDA and GGA approximations. Experimentally the CuO systems with one hole per Cu are antiferromagnetic insulators, not metals with half-filled bands. It is far beyond the subject of this book to attempt to summarize all the issues. Let it suffice to say that it appears to be essential to describe both non-local exchange (which can be done in Hartree–Fock or exact exchange methods (Ch. 8)) and correlation among the electrons in the band near the Fermi energy.

17.5 Linear muffin-tin orbital (LMTO) method

The LMTO method [643,644] builds upon the properties of muffin-tin orbitals, which have been defined in Sec. 16.5 in terms of the energy ε and the decay constant κ that characterizes the envelope function. For a fixed value of κ an LMTO basis function inside a sphere is defined to be a linear combination of $\psi(\varepsilon, r)$ and $\dot{\psi}(\varepsilon, r)$ evaluated at the energy $\varepsilon = E_\nu$ as in (17.11). The differences from an MTO defined in (16.37) are: (1) inside the "head sphere" in which a given LMTO is centered, it is a linear combination of $\psi_l(E_\nu, r)$ and $\dot{\psi}_l(E_\nu, r)$; and (2) the tail in other spheres is replaced by a combination of $\dot{\psi}_l(E_\nu, r)$. The form of an LMTO can be expressed in a very intuitive and compact form by defining functions J_l and N_l, which play a role analogous to the Bessel and Neumann functions j_l and n_l in (16.37):

$$\chi_L^{\mathrm{LMTO}}(\varepsilon, \kappa, \mathbf{r}) = i^l Y_L(\hat{\mathbf{r}}) \begin{cases} \psi_l(\varepsilon, r) + \kappa \cot(\eta_l(\varepsilon)) J_l(\kappa r), & r < S, \\ \kappa N_l(\kappa r), & r > S, \end{cases} \tag{17.24}$$

The form of J_l is fixed by the requirement that the energy derivative of χ_L^{LMTO} vanishes at $\varepsilon = E_\nu$ for $r \leq S$,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \chi_L^{\mathrm{LMTO}}(\varepsilon, \kappa, \mathbf{r}) = i^l Y_L(\hat{r}) \left[\dot{\psi}_l(\varepsilon, r) + \kappa \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \cot(\eta_l(\varepsilon)) J_l(\kappa, r) \right] = 0, \tag{17.25}$$

which leads to (Exercise 17.4)

$$J_l(\kappa r) = -\frac{\dot{\psi}_l(E_\nu, r)}{\kappa \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \cot(\eta_l(E_\nu))}, \quad r \leq S. \tag{17.26}$$

This defines an *energy-independent* LMTO basis function $\chi_L^{\mathrm{LMTO}}(E_\nu, \kappa, \mathbf{r})$ inside the sphere, given by the first line of (17.24) with $\varepsilon = E_\nu$.

The augmented Neumann functions N_L can be *defined* as the usual n_l in the interstitial, with the tails in other spheres given by the *same expansion* as in (16.38) with $n_l \to N_l$ and $j_l \to J_l$,

$$N_L(\kappa, \mathbf{r} - \mathbf{R}) = 4\pi \sum_{L', L''} C_{LL'L''} n_{L''}^*(\kappa, \mathbf{R} - \mathbf{R}') J_{L'}(\kappa, \mathbf{r} - \mathbf{R}'), \tag{17.27}$$

where $N_L(\kappa, \mathbf{r}) \equiv i^l Y_L(\hat{\mathbf{r}}) N_l(\kappa r)$, *etc.* Thus an LMTO is a linear combination of ψ and $\dot{\psi}$ in the central sphere, which continues smoothly into the interstitial region and joins smoothly to $\dot{\psi}$ in each neighboring sphere.

If we chose $\kappa = 0$ for the orbital in the interstitial region, as was done for an MTO in Sec. 16.6, then the expressions can be simplified in a way analogous to (16.44). The wavefunction inside the sphere is chosen to match the solution $\propto (r/S)^{-l-1}$ in the interstitial; this is accomplished for $r < S$ by choosing the radial wavefunction with $D = -l - 1$ as defined in (17.7), i.e. $\psi_l(D = -l - 1, r) \equiv \psi_{l-}(r)$. In turn this can be expressed in terms of ψ and $\dot{\psi}$ at a chosen reference energy together with ω from (17.8). The tails from other spheres continued into the central sphere must replace the tail $\propto (r/S)^l$ keeping the same logarithmic derivative, i.e. $(r/S)^l \rightarrow \psi_l(D = l, r) \equiv \psi_{l+}(r)$ with the proper normalization. The result is

$$\chi^{\mathrm{LMTO}}_{L,\mathbf{k}}(\mathbf{r}) = \frac{\psi_{L-}(\mathbf{r})}{\psi_{l-}(S)} \; - \; \frac{1}{\psi_{l+}(S)} \sum_{L'} \; \psi_{L'+}(\mathbf{r}) \frac{1}{2(2l'+1)} \; S_{LL'}(\mathbf{k}). \tag{17.28}$$

This defines an energy-independent LMTO orbital, along with the continuation into the interstitial region. The orbital itself contains effects of the neighbors through the structure constants and through a second effect, the requirement on the logarithmic derivative $D = -l - 1$ in the first term needed to make the wavefunction continuous and have continuous slope. *Thus the orbital contains the tail cancellation to lowest order and the energy dependence to linear order has been incorporated into the definition of the LMTO basis function.*

The LMTO method then finds the final eigenvalues using the LMTO basis and a variational expression with the full hamiltonian. This has many advantages: the energy is thus accurate to second order (and third order using appropriate expressions [643]) and the equations extend directly to full-potential methods. This is analogous to the expression for a single sphere and is accomplished by solving the eigenvalue equation,

$$\det \left| \langle \mathbf{k}L | \hat{H} | \mathbf{k}L' \rangle - \varepsilon \langle \mathbf{k}L | \mathbf{k}L' \rangle \right| = 0, \tag{17.29}$$

by standard methods. It is clear from the form of (17.28) that the matrix elements of the hamiltonian and overlap will be expressed as a sum of one-, two- and three-center terms, respectively, involving the structure constants to powers 0, 1, and 2. The expressions can be put in a rather compact form after algebraic manipulation, and we will only quote results [134, 643]. Here we consider only a muffin-tin potential which simplifies the expressions. If we define $\omega_{l-} = \omega_l(-l-1)$, $\omega_{l+} = \omega_l(l)$, $\Delta_l = \omega_{l+} - \omega_{l-}$, and $\tilde{\psi}_l = \psi_{l-}\sqrt{(S/2)}$, then the expression for $\chi_{j\mathbf{k}}(\mathbf{r})$ in (17.13) can be specified by

$$\Pi_{LL''}(\mathbf{k}) \; = \; \tilde{\psi}_l^{-1} \; \delta_{LL''} \; + \; \frac{\tilde{\psi}_{l''}}{\Delta_{l''}} \; S_{LL''}(\mathbf{k}), \tag{17.30}$$

and

$$\Omega_{LL''}(\mathbf{k}) \; = \; \omega_{l''-} \; \tilde{\psi}_l^{-1} \; \delta_{LL''} \; + \; \frac{\tilde{\psi}_{l''}}{\Delta_{l''}} \; \omega_{l''+} \; S_{LL''}(\mathbf{k}). \tag{17.31}$$

The expressions for the matrix elements are, in general, complicated since they involve the interstitial region, but the main points can be seen by considering only the atomic sphere approximation (ASA) as used in Sec. 16.6 in which the interstitial region is eliminated. Also the equations are simplified if the linearization energy E_ν is set to zero, i.e. the energy ε is relative to E_ν; this is always possible and it is straightforward to allow E_ν to depend upon l as a diagonal shift for each l. The resulting expressions have simple forms [134, 643]

$$\langle L\mathbf{k}|H|\mathbf{k}L'\rangle = \frac{\omega_{l-}}{\tilde{\psi}_l{}^2}\delta_{LL'} + \left[\frac{\omega_{l+}}{\Delta_l} + \frac{\omega_{l'+}}{\Delta_{l'}}\right] S_{LL'}(\mathbf{k}) + \sum_{L''} S_{LL''}(\mathbf{k}) \left[\tilde{\psi}_{l''}^2 \frac{\omega_{l''+}}{\Delta_{l''}^2}\right] S_{L''L'}(\mathbf{k}), \tag{17.32}$$

and

$$\langle \mathbf{k}L|\mathbf{k}L'\rangle = \left\{(1 + \omega_-^2\ \langle\dot{\psi}^2\rangle)/\tilde{\psi}^2\right\}_l\ \delta_{LL'} + \left\{\{(1 + \omega_+\,\omega_-\,\langle\dot{\psi}^2\rangle)/\Delta\}_l + \{\cdots\}_{l'}\right\}\ S^k_{LL'} + \sum_{L''} S^k_{LL''}\,[\tilde{\psi}^2\,(\,1 + \omega_+^2\langle\dot{\psi}^2\rangle)/\Delta^2]_{l''}\ S^k_{L''L'}. \tag{17.33}$$

The terms involving $\delta_{LL'}$ are one-center terms (which are diagonal in L for spherical potentials); terms with one factor of $S_{LL''}$ are two-center; and those with two factors are three-center terms. The hamiltonian has the interpretation that the on-site terms involve the energy $\omega_{l-} = \omega_l(-l-1)$ of the state with $D = -l-1$, whereas all terms due to the tails involve the energy $\omega_{l+} = \omega_l(l)$ for the state with $D = l$. Similarly, the overlap terms involve $\langle\dot{\psi}^2\rangle$ and combinations of ω_+ and ω_-.

Thus, within the ASA the LMTO equations have very simple structure, with each term in (17.32) and (17.33) readily calculated from the wavefunctions in the atomic sphere. Within this approximation, the method is extremely fast, and the goal has been reached of a minimal basis that is accurate. Only wavefunctions with l corresponding to the actual electronic states involved are needed. This is in contrast to the LAPW method where one needs high l in order to match the spherical and plane waves at the sphere boundary. Furthermore, the interstitial region and a full potential can be included; the same basis is used but the expressions for matrix elements are more cumbersome. The size of the basis is still minimal and the method is very efficient.

There is a price, however, for this speed and efficiency. The interstitial region is not treated accurately since the LMTO basis functions are single inverse powers or Hankel or Neumann functions as in (17.24). Open structures can be treated only with correction terms or by using "empty spheres." The latter are useful in static, symmetric structures, but the choice of empty spheres is problematic in general cases, especially if the atoms move. Finally, there is no "knob" to turn to achieve full convergence as there is in the LAPW method. Thus the approximations in the LMTO approach are difficult to control and care is needed to ensure robust results.

Improved description of the interstitial in LMTO approaches

One of the greatest problems with the LMTO approach, as presented so far, is the approximate treatment of the interstitial region. The use of a single, energy-independent tail outside each sphere was justified in the atomic sphere approximation (Fig. 16.10) where the distances between spheres is very small (and in the model the interstitial is non-existent). This approximation fails for open structures where the interstitial region is large, e.g. in the diamond structure, and applications of the LMTO method depend upon tricks like the introduction of empty spheres [669, 670]. This can be done for high-symmetry structures, but the method cannot deal with cases like the changing structures that occur in a simulation.

An alternative approach is to generalize the form of the envelop function, generalizing the single Hankel or power law function given in (16.39), (16.37), or (17.24). One approach is to work with multiple Hankel functions with different decay constants κ_i that can better describe the interstitial region and yet keep the desirable features of Hankel functions [687, 688]. Using this approach, a full-potential LMTO method has been proposed [689] that combines features of the LMTO, LAPW, and PAW approaches. Like the LAPW it has multiple functions outside the spheres, but many fewer functions. Like the PAW method, the smooth functions are continued inside the sphere where additional functions are included as a form of "additive augmentation."

The form of the basis function proposed is an "augmented smooth Hankel function." In (17.24), the tail of the LMTO function is a Neumann function, which at negative energy (imaginary κ) becomes a Hankel function, which is the solution of

$$(\nabla^2 + \kappa^2)\, h_0(\mathbf{r}) = -4\pi\, \delta(\mathbf{r}). \tag{17.34}$$

This function decays as $i^{-l} \mathrm{e}^{-|\kappa| r}/|\kappa| r$ at large r and it diverges at small r as illustrated in Fig. 17.5. The part inside the sphere is not used in the usual LMTO approach and it makes the function unsuitable for continuation in the sphere. Methfessel and van Schilfgaarde [689] instead defined a "smooth Hankel function" that is a solution of

$$(\nabla^2 + \kappa^2)\, \tilde{h}_0(\mathbf{r}) = -4\pi\, g(\mathbf{r}). \tag{17.35}$$

If $g(\mathbf{r})$ is chosen to be a gaussian, $g(\mathbf{r}) \propto \exp(r^2/R_{sm}^2)$, then $\tilde{h}_0(\mathbf{r})$ is a convolution of a gaussian and a Hankel function. It has the smooth form shown in Fig. 17.5 and has many desirable features of both functions, including analytic formulas for two-center integrals and an expansion theorem. It is proposed that the form of the smooth function near the muffin-tin radius more closely resembles the true function than the Hankel function, and that the sum of a small number of such functions can be a good representation of the wavefunctions in the interstitial region [689].

17.6 "*Ab initio*" tight-binding

It has been pointed out in Sec. 16.7 that the MTO approach provides a localized basis and tight-binding-type expressions for the Kohn–Sham equations. With a unitary transformation

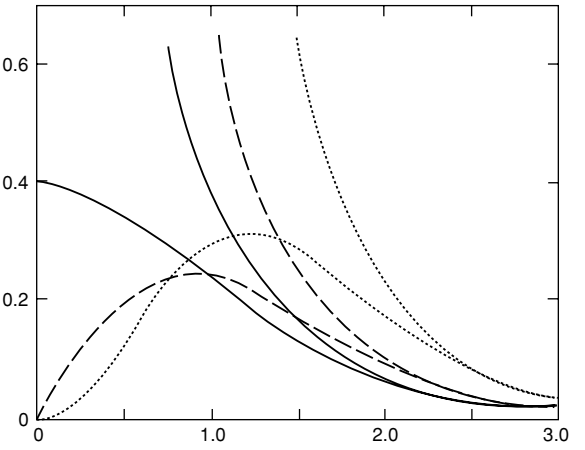


Figure 17.5. Comparison of standard and "smooth" Hankel functions for $l = 0$ (solid lines), $l = 1$ (dashed), and $l = 2$ (dotted) for the case $\kappa = i$ and the smoothing radius $R_{sm} = 1.0$ in the gaussian. From [689].

that is equivalent to "screening of electrostatic multipoles," one can transform to a compact short-range form [668]. The transformation applies in exactly the same way in the LMTO approach since it depends only upon the form of the envelop function outside the sphere. The matrix elements between different MTOs decrease as $R^{-(l+l'+1)}$ which leads to short-range interactions for large l. Matrix elements for $l + l' = 0$, 1, or 3 can be dealt with by suitable transformations [668].

There are two new features provided by linearization. Most important, the linear equations have the same form as the usual secular equations so that all the apparatus for linear equations can be applied. Second, transformation of the equations leads to very simple expressions for the on-site terms and coupling between sites in terms of ψ and $\dot{\psi}$. The short-range LMTO is the ψ function in one sphere coupled continuously to the tails in neighboring spheres which are $\dot{\psi}$ functions. This provides an orthonormal minimal basis tight-binding formulation in which there are only two-center terms, with all hamiltonian matrix elements determined from the underlying Kohn–Sham differential equation. The disadvantage is that all the terms are highly environment dependent, i.e. each matrix element depends in detail upon the type and position of the neighboring atoms.

This "*ab initio*" tight-binding method is now widely used for many problems in electronic structure. Because the essential calculations are done in atomic spheres, determination of the matrix elements can be done very efficiently. Combination of the recursion method (Sec. 23.3) and the tight-binding LMTO [668] provides a powerful method for density-functional calculations for complex systems and topologically disordered matter [690,691]. For example, in Fig. 23.3 is shown the electronic density of states of liquid Fe and Co determined using tight-binding LMTO and recursion [692]. The calculations were done on 600 atom cells with atomic positions, representing a liquid structure generated by classical Monte Carlo and empirical interatomic potentials. Such approaches have been applied to many problems in alloys, magnetic systems, and other complex structures.

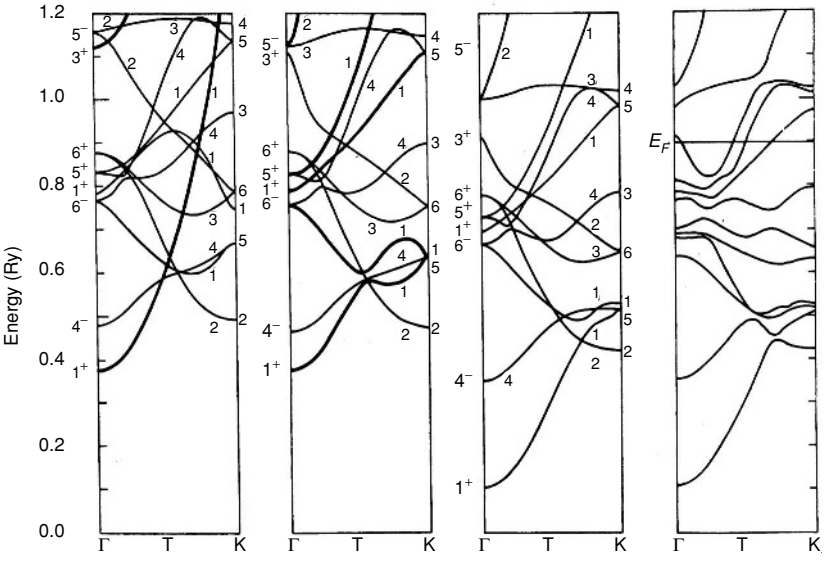


Figure 17.6. Development of the band structure of hcp Os in the LMTO method. From left to right: non-relativistic "canonical" (Sec. 16.6) bands neglecting hybridization of d and s, p bands (shown dark); including hybridization (with dark lines indicating the most affected bands); relativistic bands without spin orbit; fully relativistic bands. From [464]; original calculations in [693].

17.7 Applications of the LMTO method

Here we present several illustrations of understanding and quantitative information gained from the LMTO method in its simplest form. The first is the equilibrium volume and bulk moduli. It is a great advantage to calculate the pressure directly using the formulas valid in the ASA given in Sec. I.3. The equilibrium volume per atom Ω is the volume for which the pressure $P = 0$, and the bulk modulus is the slope $B = -\mathrm{d}P/\mathrm{d}\Omega$. The results for 4d and 5d transition metals [464] compare well with the calculations using the KKR method presented in Fig. 2.3. The results are quite impressive and show the way that important properties of solids can be captured in simple calculations with appropriate interpretation.

A second example is the progression of energy bands from the simplest unhybridized "canonical" form to the full calculation. Figure 17.6 shows this progression for hcp Os along one line in the BZ from unhybridized canonical bands on the left to full hybridized relativistic bands on the right.

It is instructive the give an example of LMTO applied to semiconductors which have open structures – very different from close packed metals for which MTOs were originally designed. By including empty spheres [669, 670] the structure becomes effectively close-packed and accurate calculations can be done with only a few basis functions per empty sphere. An example is the calculation of Wannier functions [694] described in Sec. 21.2 and band offsets of semiconductor structures [580, 581]. As examples of band structure calculations, Figs. 17.7 and 17.8 show calculations for GaAs and Ge including relativistic effects and core relaxation [571]. This was the first work to show that with proper LDA

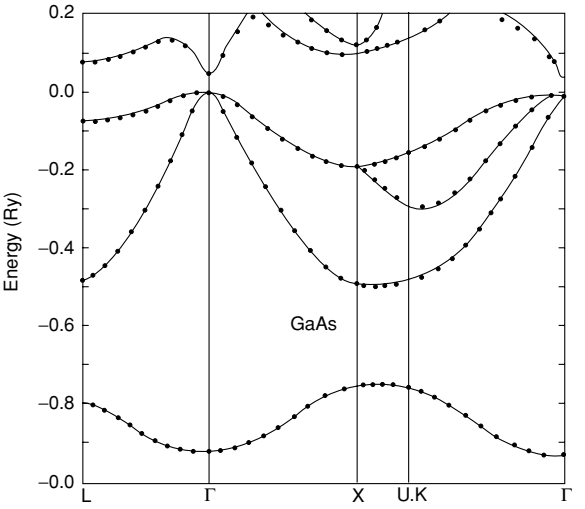


Figure 17.7. Band structure of GaAs (solid lines) calculated using the local density approximation (LDA) including scalar relativistic effects in the LMTO formalism [571]. The dots indicate results of pseudopotential calculations, which are essentially identical. The gap is lower than experiment, as indicated by the LDA results given in Fig. 2.26. From [571].

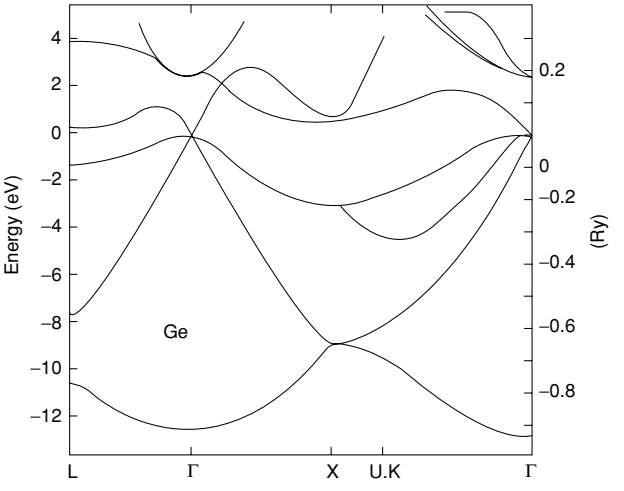


Figure 17.8. Band structure of Ge calculated using the local density approximation (LDA) including scalar relativistic effects in the LMTO formalism [571]. Note that the band gap is essentially zero. Since spin–orbit coupling will split the valence band (not shown), this will cause an overlap of the valence and conduction bands – which means that in the LDA, Ge is predicted to be a metal! The same result is found in pseudopotential calculations done later, for example, as shown in Fig. 2.25. From [571].

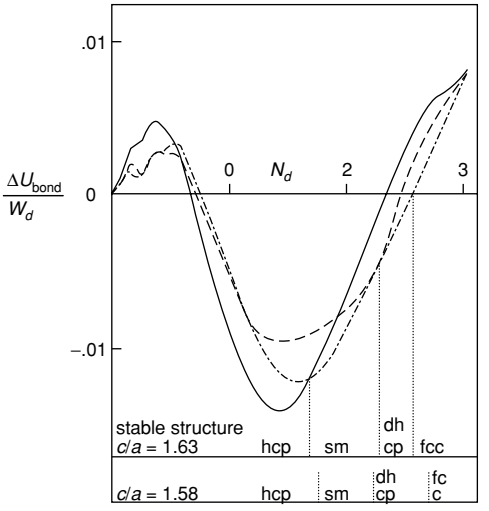


Figure 17.9. Structure sequence for the lanthanide series of trivalent rare earth elements as calculated by Duthi and Pettifor [695]. The notation is "Sm" for the low-symmetry samarium structure and "dhcp" for the double hcp structure, and two different values for the c/a ratio are considered. Stabilization is interpreted to result from filling of the bonding part of the d band following the reasoning of Friedel [697]. From [695].

treatment, Ge is a metal! Also shown in Fig. 17.7 are bands for GaAs calculated using the plane wave pseudopotential method, which give essentially identical results.

Another example that illustrates many different features of the LMTO approach is the work of Duthi and Pettifor [695], which provided a simple explanation for the sequence of structures observed in the series of rare earths elements. Because the energy differences between the structures is very small, these authors made use of the simplification given in Eq. (I.7), which expresses an energy difference between structures in terms of the difference of the sum of eigenvalues. In these expressions they used the atomic sphere approximation (ASA) in which the potential is essentially the same for a given element for the different structures, since the atomic volume is almost the same in the different structures. Finally, the sum of eigenvalues was calculated using the tight-binding form of LMTO and the recursion method of Haydock and coworkers (Sec. M.5 and [696]). The results are presented in Fig. 17.9 which shows the structure sequence hcp, the samarium structure, dhcp (double hcp), and fcc. Stabilization results from filling of the d band and can be considered to be an example of Friedel is argument [697] that stabilization is due to filling of the bonding states, but it took the combination of ideas in [695] to sort out the way in which bonding varies with structure.

17.8 Beyond linear methods: NMTO

Recent developments in MTO methods show how approximations that were introduced during development of the LMTO approach can be overcome. The new NMTO approach

[698, 699] provides a more consistent formalism, treats the interstitial region accurately, and goes beyond the linear approximation.

In the MTO and LMTO approaches, energy-independent orbitals were generated using the approximation of a fixed κ in the envelop function that describes the interstitial region. This breaks the relation of κ and the eigenvalue that causes non-linearities in the KKR method. However, it also is an approximation that is justified only in close-packed solids. In contrast, the wavefunction inside the sphere is treated more accurately through linearization. The NMTO method treats the sphere and interstitial equally by working with MTO-type functions $\psi_L(E_n, \mathbf{r} - \mathbf{R})$ localized around site $\mathbf{R}$ and calculated at fixed energies E_n both inside the sphere and in the interstitial (assumed to have a flat muffin-tin potential). The NMTO basis function is then defined to be a linear combination of N, such functions evaluate at N energies,

$$\chi^{\mathrm{NMTO}}_{\mathbf{R}L}(\varepsilon\mathbf{r}) = \sum_{n=0}^{N} \sum_{\mathbf{R}'L'} \psi_{L'}(E_n, \mathbf{r} - \mathbf{R}') L^N_{nL'\mathbf{R}',L\mathbf{R}}(\varepsilon, \mathbf{r}), \tag{17.36}$$

where $L^N_n(\varepsilon)$ is the transformation matrix that includes the idea of screening (mixing states on different sites) and a linear combination of states evaluated at N fixed energies.

As it stands, the NMTO function is energy dependent and appears to be merely a way to expand the basis. However, Andersen and coworkers [698, 699] have shown a way of generating energy-independent functions $\chi^{\mathrm{NMTO}}_{\mathbf{R}L}(\mathbf{r})$ using a polynomial approximation so that the Schrödinger equation is solved exactly at the N chosen energies. The ideas are a generalization of the transformation given in Sec. 11.9, which were chosen to give the correct phase shifts at an arbitrary set of energies. The basic ideas can be understood, following the steps in Exercise 11.12, where the exact transformation, (11.45), is easily derived. In the present case, the transformation is more general, mixing states of different angular momenta on different sites as indicated in (17.36). The result of the transformation is that each eigenfunction is accurate to order $(\varepsilon - E_0)(\varepsilon - E_1)\cdots(\varepsilon - E_N)$ and the eigenvalue to order $(\varepsilon - E_0)^2(\varepsilon - E_1)^2\cdots(\varepsilon - E_N)^2$.

As an illustration of the NMTO approach, Fig. 17.10 shows the $d_{x^2-y^2}$ orbital centered on a Cu atom in $YBa_2Cu_3O_7$. This orbital is not unique; it is chosen to represent the mixed Cu–O band that crosses the Fermi level, as shown in Fig. 17.4. Note that the state centered on one Cu atom is extended, with important contributions of neighboring O and Cu sites. The band resulting from that single orbital is shown as dark circles in Fig. 17.11, which can be compared with the states near the Fermi energy in Fig. 17.4. (Also shown are the energies at which the state is required to fit the full band structure.) The important point is that the procedure leads to an accurate description of the desired band, without the "spaghetti" of other bands. Such a function is derived by focusing on the energy of interest and by "downfolding" the effects of all the other bands by identifying the angular momentum channels of interest in the transformation, (17.36).

Although it is beyond the scope of this book, we can draw two important conclusions about the promise of the NMTO approach. First, for MTO-type methods, it appears to remove the limitation to close-packed structures, and, second, it allows accurate solution

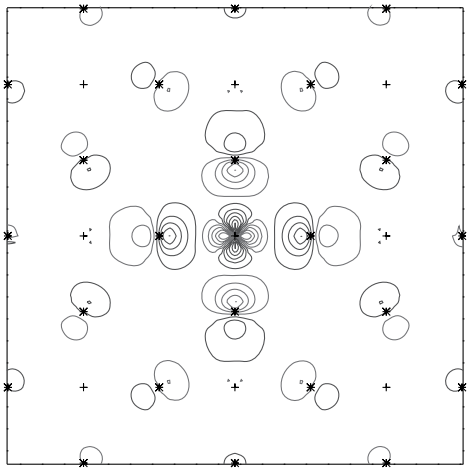


Figure 17.10. Orbital of $d_{x^2-y^2}$ symmetry centered on a Cu atom in $YBa_2Cu_3O_7$ chosen to describe the actual band crossing the Fermi energy and derived using the NMTO method [699]. The resulting band derived from this single orbital is shown in Fig. 17.11. From [699].

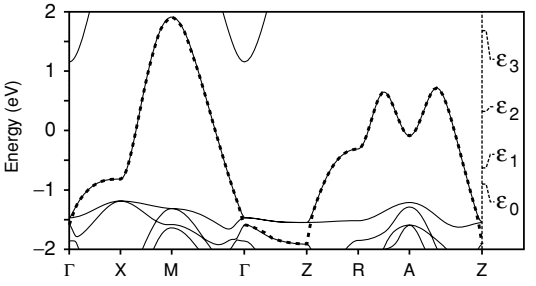


Figure 17.11. Band from the orbital shown in Fig. 17.10 (dark symbols) compared to the full bands (light symbols), which are essentially the same as the bands in Fig. 17.4. Also shown are the energies at which the band is designed to agree. The band is well described even when it has complicated shape and crosses other bands. From [699].

for general structures. This means that the MTO approach can provide a "first-principles tight-binding approach" (see Secs. 16.7 and 17.5) applicable to general structures of crystals and molecules. Furthermore, if calculations can be done efficiently, then NMTO calculations can provide forces and can be used in molecular dynamics. Taking a broader perspective, the NMTO approach is a promising addition to all-band structure methods, potentially providing new approaches beyond the present linearized methods.

17.9 Full potential in augmented methods

One of the most important outcomes of linearization is the development of full-potential augmented methods, e.g. for LAPW [414, 677, 681, 682] and LMTO [700] methods. Although actual implementations may be cumbersome and cannot be described here, the basic ideas can be stated very simply. Since the linearized methods have been derived in terms

of matrix elements of the hamiltonian in a fixed basis, one simply needs to calculate matrix elements of the full non-spherical potential ΔV in the sphere and the full spatially varying potential in the interstitial. The basis functions are still the same APW, PAW, or LMTO functions χ_L, which are derived from a spherical approximation to the full potential. However, the spheres merely denote convenient boundaries defining the regions where the basis functions and the potential have different representations. In principle, there are no approximations on the wavefunctions or the potential except for truncations at some $l_{\max}$ and $G_{\max}$. If the basis is carried to convergence inside and outside the spheres, the accuracy is, in principle, limited only by the linearization.

Inside each sphere the potential is expanded in spherical harmonics,

$$V(r, \theta, \phi) = \sum_L V_l(r) i^l Y_{lm}(\theta, \phi), \tag{17.37}$$

so that matrix elements $\langle L|V|L'\rangle$ can be calculated in terms of radial integrals. Similarly, the interstitial matrix elements are no longer diagonal in plane waves, but they can be found straightforwardly by integrating in real space. In the PAW method and the multiple-κ LMTO method, the smooth functions continue into the sphere and it is convenient also to define the potential as a smooth part everywhere plus a sharply varying part restricted to spheres. In that case, the matrix elements of the smooth part can be calculated by FFT methods just as is done in pseudopotential methods (Sec. 13.1).

Of course, in the self-consistent calculation one also needs to calculate the potential arising from the density. This necessitates a procedure in which the Poisson equation is solved taking into account the sharply varying charge density inside the spheres. This is always possible since the field inside can be expanded in spherical harmonics and outside the spheres can be represented by smooth functions plus multipole fields due to the charge inside the spheres. Perhaps the simplest approach is to define both the density and the potential as smooth functions everywhere, with sharply varying components restricted to spheres [414, 475].

There is a quantitative difference between the LAPW and LMTO approaches in the requirements on the full potential. Since the minimal basis LMTO only involves functions with $l_{\max}$ given by the actual angular momenta of the primary states making up the band (e.g. $l = 2$ for transition metals), only angular momenta up to $2l_{\max}$ are relevant. However, for the LAPW methods, much higher angular momenta in the wavefunctions (typically $l_{\max} \approx 8$ to 12 for accurate calculations) are required to satisfy the continuity conditions accurately. In principle, very large values of $l_{\max}$ are needed for the potential, and in practice accurate numerical convergence can be reached with $l_{\max} \approx 8$ to 12. The difference results from the fact that the LAPW basis is much larger; in order to represent the interstitial region accurately many plane waves are needed, which leads to the need for high angular momenta in order to maintain the continuity requirements (see Exercises 17.5–17.7).

SELECT FURTHER READING

Original papers:

Andersen, O. K., "Linear methods in band theory," *Phys. Rev. B* 12: 3060–3083, 1975.

Andersen, O. K. and Jepsen, O., "Explicit, first-principles tight-binding theory," *Physica* 91B: 317, 1977.

Marcus, P. M., "Variational methods in the computation of enegy bands," *Int. J. Quant. Chem.* 1S: 567–588, 1967. [690]

Summary of various methods:

Blaha, P. Schwarz, K. Sorantin, P. and Trickey, S.B., "Full-potential, linearized augmented plane wave programs for crystalline systems," *Computer Phys. Commun.* 59(2): 399, 1990.

Kübler, J., *Theory of Itinerant Electron Magnetism*, Oxford University Press, Oxford, 2001.

Kübler, J. and Eyert, V., in *Electronic and Magnetic Properties of Metals and Ceramics*, edited by K. H. J. Buschow, VCH-Verlag, Weinheim, Germany, 1992, p. 1.

Singh, D. J., *Planewaves, Pseudopotentials, and the APW Method*, Kluwer Academic Publishers, Boston, 1994, and references therein.

Skriver, H., *The LMTO Method*, Springer, New York, 1984.

Exercises

17.1 Derive Eq. (17.4) from the definition of $\dot{\psi}$. In addition, show the more general relation

$$(\hat{H} - \varepsilon)\psi^{(n)}(\varepsilon, r) = n\psi^{(n+1)}(\varepsilon, r), \tag{17.38}$$

where n is the order of the derivative. Hint: Use the normalization condition.

17.2 Carry out the manipulations to show that the hamiltonian and overlap matrix elements can be cast in the linearized energy-independent form of Eqs. (17.14) to (17.17). Thus the matrix elements are expressed in terms of Π and Ω, which are functions of the wavefunctions ψ and $\dot{\psi}$ calculated in the sphere at the chosen energy E_ν.

17.3 Derive the result that $l_{\max} \approx 8$ in LAPW calculations. Consider a simple cubic crystal with one atom/cell with the volume of the atomic sphere $\approx 1/2$ the volume of the unit cell. The order of magnitude of ≈ 100 planes waves is reasonable since it corresponds to a resolution of $\approx 100^{1/3}$ points in each direction. If the plane waves are in a sphere of radius $G_{\max}$, find $G_{\max}$ in terms of the lattice constant a. This is sufficient to find an estimate of $l_{\max}$ using the arguments in the text. If the number of plane waves were increased to 1,000, what would be the corresponding $l_{\max}$?

17.4 The condition (17.25) requires that the LMTO be independent of the energy to first order and is the key step that defines an LMTO orbital; this removes the rather arbitrary form of the MTO and leads to the expression in terms of $\dot{\psi}$. Show that this condition leads to the expression, (17.26), for the J function proportional to $\dot{\psi}$ inside the sphere.

17.5 If the augmented wavefunction (LAPW or LMTO) is expanded in Y_{lm} up to $l_{\max}$, what is the corresponding $l_{\max}^{\text{density}}$ needed in an exact expansion for the charge density for the given wavefunction? Give reasons why it may not be essential to have $l_{\max}^{\text{density}}$ this large in an actual calculation.

17.6 If the density is expanded in Y_{lm} up to $l_{\max}^{\text{density}}$, what is $l_{\max}$ for the Hartree potential? For V_{xc}?

17.7 What is the maximum angular momentum $l_{\max}^{\text{pot}}$ of the potential (17.37) needed for exact evaluation of matrix elements $\langle L|V|L'\rangle$ if the wavefunction is expended up to $l_{\max}$? Just as in Exercise 17.5, give reasons why smaller values of $l_{\max}^{\text{pot}}$ may be acceptable.

17.8 Consider the compound $YBa_2Cu_3O_7$. Determine the number of electrons that would be required to fill the oxygen states to make a closed-shell ionic compound. Show that for $YBa_2Cu_3O_7$ there is one too few electrons per Cu atom. Thus, this material corresponds to one missing electron (i.e. one hole per Cu).

PART V

PREDICTING PROPERTIES OF MATTER FROM ELECTRONIC STRUCTURE – RECENT DEVELOPMENTS

Imagination is more important than knowledge

Albert Einstein

18

Quantum molecular dynamics (QMD)

Summary

Of all the recent methods for computing the properties of materials from electronic equations, one stands out: i.e. the quantum molecular dynamics (QMD) simulations pioneered by Car and Parrinello in 1985 [156]. This work and subsequent developments have led to a revolution in the capabilities of theory to treat real, complex molecules, solids, and liquids including thermal motion (molecular dynamics), with the forces derived from the electrons treated by (quantum) density functional methods. Altogether, four advances create the new approach to electronic structure. These comprise:

- optimization methods (instead of variational equations),
- equations of motion (instead of matrix diagonalization),
- fast Fourier transforms (FFTs) – (instead of matrix operations), and
- a trace of occupied subspace (instead of eigenvector operations).

Car and Parrinello combined these features into one unified algorithm for electronic states, self-consistency, and nuclear movement. There has also been an explosion of alternative approaches that utilize the force theorem, together with efficient iterative methods described in App. M or simpler tight-binding-type methods. These are described in the present chapter as well as the Car–Parrinello method *per se*.

18.1 Molecular dynamics (MD): forces from the electrons

The basic equations for the motion of classical objects are Newton's equations. For a set of nuclei treated as classical masses with an interaction energy $E[\{\mathbf{R}_I\}]$ dependent upon the positions of the particles $\{\mathbf{R}_I\}$, the equations of motion are

$$M_I \ddot{\mathbf{R}}_I = -\frac{\partial E}{\partial \mathbf{R}_I} = \mathbf{F}_I[\{\mathbf{R}_J\}]. \tag{18.1}$$

Such equations can be solved analytically only in the small-amplitude harmonic approximation. In general, the solution is done by numerical simulations using discrete time steps based upon discrete equations such as the Verlet algorithm, the properties of which are well

established. [437] At each time step t the position of each nucleus is advanced to the next time step $t + \Delta t$ depending upon the forces due to the other nuclei at the present time step:

$$\mathbf{R}_I(t + \Delta t) = 2\mathbf{R}_I(t) + \mathbf{R}_I(t - \Delta t) + \frac{(\Delta t)^2}{M_I}\mathbf{F}_I[\{\mathbf{R}_J(t)\}], \tag{18.2}$$

where the first two terms are just the law of inertia. The key property of the Verlet algorithm, well established in classical simulations, is that *the errors do not accumulate.* Despite the fact that the equations are only approximate for any finite Δt, the energy is conserved and the simulations are stable for long runs.

Of course, the forces on the nuclei are determined by the electrons in addition to direct forces between the nuclei. In the past this has been done by effective potentials (such as the Lennard–Jones potential) that incorporate effects of the electrons. These are adequate for many cases like rare gas atoms, but it is clear that one must go beyond such simple pair potentials for real problems of interest in materials. One approach is to use empirical models which attempt to include additional effects and, usually, are parameterized.

Advances in electronic structure calculations have made molecular dynamics (MD) simulations possible with forces derived directly from the electrons with no parameters. Such simulations are often termed "*ab initio*" or "first principles," but here we shall use the nomenclature "quantum MD (QMD)" simulations. Within the Born–Oppenheimer (adiabatic) approximation[1] the electrons stay in their instantaneous ground state as the nuclei move. Thus the correct forces on the nuclei are given by the force (Hellmann–Feynman) theorem, Eq. (3.18), which is practical to implement in pseudopotential density functional theory as shown by many examples in previous chapters. We repeat here the basic formulas from Chs. 7 and 9 in slightly different notation. Within the Kohn–Sham approach to density functional theory, the total energy of the system of ions and electrons is given by

$$E[\{\psi_i\}, \{\mathbf{R}_I\}] = 2\sum_{i=1}^{N}\int \psi_i^*(\mathbf{r})\left(-\frac{1}{2}\nabla^2\right)\psi_i(\mathbf{r})\mathrm{d}\mathbf{r} + U[n] + E_{II}[\{\mathbf{R}_I\}], \tag{18.3}$$

$$U[n] = \int \mathrm{d}\mathbf{r} V_{\mathrm{ext}}(\mathbf{r})n(\mathbf{r}) + \frac{1}{2}\int\int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + E_{\mathrm{xc}}[n], \tag{18.4}$$

$$n(\mathbf{r}) = 2\sum_{i=1}^{N}|\psi_i(\mathbf{r})|^2, \tag{18.5}$$

$$\mathbf{F}_I = -\frac{\partial E}{\partial \mathbf{R}_I}, \tag{18.6}$$

where ψ_i are the one-electron states, $\mathbf{R}_I$ are the positions of the ions, E_{II} is the ion–ion interaction, $n(\mathbf{r})$ is the electronic charge density, $V_{\mathrm{ext}}(\mathbf{r})$ is the electron–ion interaction,

[1] This is an excellent approximation for many properties such as phonon energies, as discussed in Ch. 3 and App. C.

$E_{\rm xc}[n]$ is the exchange-correlation energy, and $\mathbf{F}_I$ is the force given by the force theorem expressed in (3.18), (13.3), and other forms.

The key problem, however, is that the calculations must be *very efficient*. The development of new efficient algorithms by Car and Parrinello [156] and others [440, 710] has led to the explosion of many forms of QMD simulations since 1985.

18.2 Car–Parrinello unified algorithm for electrons and ions

The essential feature of the Car–Parrinello approach takes advantage of the fact that the total energy of the system of interacting ions and electrons is a function of *both* the classical variables $\{\mathbf{R}_I\}$ for the ions *and* the quantum variables $\{\psi_i\}$ for the electrons. Instead of considering the motion of the nuclei and the solution of the equations for the electrons at fixed $\{\mathbf{R}_I\}$ as separate problems (an approach that is inherent in the flow charts describing the usual approach in Fig. 9.1 and Fig. M.1), the Car–Parrinello approach considers these as *one unified problem*. Within the Born-Oppenheimer (adiabatic) approximation, the problem becomes one of minimizing the energy of the electrons and solving for the motion of the nuclei simultaneously. This applies to relaxation of the nuclei to find stable structures as well as to thermal simulations of solids and liquids using MD methods. In one stroke, calculation of the ground-state electronic structure and simulation of material phenomena have been unified.[2]

In the Car–Parrinello approach, the total Kohn–Sham energy is the potential energy as a function of the positions of the nuclei. Molecular dynamics for the nuclei using forces from this energy is the defining criterion for all forms of so-called "*ab initio* MD" using density functionals. *The special feature of the Car–Parrinello algorithm is that it also solves the quantum electronic problem using MD.* This is accomplished by adding a *fictitious* kinetic energy for the electronic states, which leads to a fictitious lagrangian for both nuclei and electrons [156][3]

$$\mathcal{L} = \sum_{i=1}^{N} \frac{1}{2}(2\mu) \int \mathrm{d}\mathbf{r} |\dot{\psi}_i(\mathbf{r})|^2 + \sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 - E[\psi_i, \mathbf{R}_I] + \sum_{ij} \Lambda_{ij} \left[\int \mathrm{d}\mathbf{r} \psi_i^*(\mathbf{r}) \psi_j(\mathbf{r}) - \delta_{ij} \right]. \tag{18.7}$$

The final term in (18.7) is essential for orthonormality of the electronic states. This lagrangian leads to MD equations for *both* classical ionic degrees of freedom $\{\mathbf{R}_I\}$ *and* electronic degrees of freedom, expressed as independent-particle Kohn–Sham orbitals $\psi_i(\mathbf{r})$.

[2] It is essential to emphasize that the Car–Parrinello algorithm does *not* treat the real dynamics of electrons, which requires a time-dependent Schrödinger equation, (7.22) or (20.11). The algorithm is designed to find the ground state (adiabatic or Born-Oppenheimer) solution for the electrons as the nuclei move.

[3] Note the similarity to the lagrangian in (M.14), except that here the signature ingredient of the Car-Parrinello method, the "fictitious electronic mass," is added. Such fictitious lagrangians are also used in other quantum field theories [704].

The resulting equations of motion are

$$\mu\ddot{\psi}_i(\mathbf{r},t) = -\frac{\delta E}{\delta\psi_i^*(\mathbf{r})} + \sum_k \Lambda_{ik}\psi_k(\mathbf{r},t)$$
$$= -H\psi_i(\mathbf{r},t) + \sum_k \Lambda_{ik}\psi_k(\mathbf{r},t), \tag{18.8}$$

$$M_I\ddot{\mathbf{R}}_I = \mathbf{F}_I = -\frac{\partial E}{\partial \mathbf{R}_I}. \tag{18.9}$$

The equations of motion (18.8) and (18.9), are just Newtonian equations for acceleration in terms of forces, subject to the constraint of orthogonality in the case of electrons. The masses of the ions are their physical masses, and the "mass" for the electrons is chosen for optimal convergence of the solution to the true adiabatic solution. Thus the equations can be solved by the well-known Verlet algorithm (App. L) with the constraints handled using standard methods for holonomic constraints [711]. This can be achieved by solving the equations for Λ_{ik} at each time step so that ψ_i are exactly orthonormal using an iterative method called SHAKE [701, 711]. The resulting discrete equations for time $t^n = n\delta t$, are

$$\psi_i^{n+1}(\mathbf{r}) = 2\psi_i^n(\mathbf{r}) - \psi_i^{n-1}(\mathbf{r}) - \frac{(\Delta t)^2}{\mu}\left[\hat{H}\psi_i^n(\mathbf{r}) - \sum_k \Lambda_{ik}\psi_k^n(\mathbf{r},t)\right],$$
$$\mathbf{R}_I^{n+1} = 2\mathbf{R}_I^{n+1} - \mathbf{R}_I^{n-1} + \frac{(\Delta t)^2}{M_I}\mathbf{F}_I. \tag{18.10}$$

Note the similarity to those equations for minimization of electronic energy, e.g. (M.15). The most time-consuming operation (applying the hamiltonian to a trial vector) is exactly the same in *all* the iterative methods; the only difference is the way in which the wavefunctions are updated as a function of time t or step n.

The stationary solution

The meaning of the equations can be clarified by considering a stationary solution of the equations, which we now show is equivalent to the usual Kohn–Sham variational equations. At steady state, all time derivatives vanish and (18.9) leads to

$$H\psi_i(\mathbf{r},t) = \sum_k \Lambda_{ik}\psi_k(\mathbf{r},t), \tag{18.11}$$

which is the usual solution with Λ_{ik} the matrix of Lagrange multipliers. Taking the matrix elements, (18.11) shows that Λ is the transpose of H ($\Lambda_{ik} = H_{ki}$), where H is the usual Kohn–Sham hamiltonian. Diagonalizing Λ leads to the eigenvalues of the Kohn–Sham equations. Furthermore, this is a self-consistent solution since we have minimized the full Kohn–Sham energy, (18.3). Thus the solution is stationary if, and only if, the Kohn–Sham energy is at a variational minimum (or saddle point). In fact, cooling the system down by reducing the kinetic energy is termed dynamical simulated annealing, which is a way to find the minimum of the non-linear self-consistent Kohn–Sham equations. This is illustrated in the original paper of Car and Parrinello; their results copied here in Fig. 18.1 show the eigenvalues reaching the values that would also be found in a self-consistent calculation.

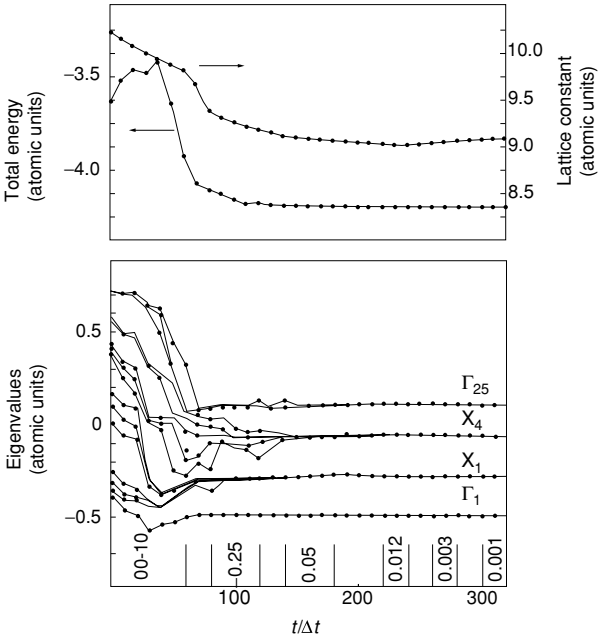


Figure 18.1. Eigenvalues at $\mathbf{k} = 0$ for crystalline Si calculated by quenching the "fictitious kinetic energy" in the lagrangian to reach the steady state [156].

Nuclear dynamics

The real power of the Car–Parrinello method is found in simulations of the coupled motion of nuclei and electrons. This leads to the ability to include the real dynamics of the nuclei in *ab initio* electronic structure algorithms, treating e.g., thermal motion, liquids, thermal phase transitions, etc. Also, by quenching, one can search for stable structures. Examples are given later in Sec. 18.6.

It should be emphasized that the *fictitious kinetic energy* has nothing to do with the real quantum mechanical energy of the electrons and the electron dynamics that results from this lagrangian is also fictitious; *it does not represent the real excitations of the electron system*. The purpose of this fictitious kinetic energy is to allow the ground state of the electrons to move efficiently through the space of basis functions, always staying close to the true ground state. This is in fact realized in many cases, but is also problematic in other cases as discussed below.

Difficulties in the Car–Parrinello unified algorithm

There are three primary disadvantages in using the Car–Parrinello approach.

First, any effects of the fictitious lagrangian must be examined and reckoned with if they are problematic. The method works well for systems with an energy gap (for all steps in the simulation). The characteristic frequency of the fictitious oscillations of the electron degrees of freedom are $\propto E_{\text{gap}}/\mu$ (Exercise 18.2) and, if all such frequencies are much greater than

typical nuclear vibration frequencies, then the electrons follow the nuclei adiabatically as they should. This is tested by checking the conservation of proper energy (not including the fictitious kinetic energy). Simple examples and discussion are given in [702], [703], [704], and the exercises at the end of this chapter. Even in the best cases, however, one still needs to choose the mass so that the adiabatic condition is satisfied to within acceptable accuracy. There has been controversy on this point [712], but a joint study of Car, Parrinello, and Payne [713] concluded that, with care, problems can be avoided.

Second, the time step Δt must be short. It is governed by the fictitious electronic degrees of freedom and must be chosen smaller than in typical simulations for ions alone. A typical "mass" for the electrons is $\mu = 400 m_e$ (as used for carbon [714]). In general, the value depends upon the basis functions, and an issue arises in the plane wave method where (18.14) below reveals a problem for high Fourier components of the wavefunction. Since the diagonal part of the hamiltonian $H(\mathbf{G}, \mathbf{G}) \propto |\mathbf{G}|^2$ for large $|\mathbf{G}|$, a coefficient $c_i^n(\mathbf{G})$ is multiplied by $((\Delta t)^2/\mu)|\mathbf{G}|^2$. This endangers one of the desirable properties of plane waves: that the cutoff can be increased indefinitely to achieve convergence. It has been proposed to integrate the high Fourier components over the time step interval (since they obey a very simple harmonic oscillator equation) instead of taking the linear variation [710]. Another approach is to take different masses for different Fourier components [706, 707].

Finally, it was recognized from the beginning that problems occur with level crossing, where the gap vanishes, and in metals. This leads to unphysical transfer energy to the fictitious degrees of freedom (Exercise 18.2). The problem has been side-stepped by use of "thermostats" that pump energy into the ion system and remove it from the fictitious kinetic energy of the electron system. This has been used, e.g., in calculations of metallic carbon at high pressure [159].[4] However, problems with metals simulations have led to the widespread use of alternative approaches (see Sec. 18.4).

18.3 Expressions for plane waves

The Car–Parrinello equations can be made more transparent by choosing an explicit basis. The equations have exactly the same form for any orthonormal basis and we choose plane waves as the best example. For simplicity of notation we consider Bloch states only at the center of the Brillouin zone, $\mathbf{k} = 0$, in which case the Bloch functions can be written

$$u_i(\mathbf{r}) = \sum_{\mathbf{G}} c_i(\mathbf{G}) \frac{1}{\sqrt{\Omega}} \exp(\mathrm{i}\mathbf{G} \cdot \mathbf{r}), \tag{18.12}$$

where Ω is the volume of the unit cell. Since each band holds one electron per cell (of a given spin) the $c_i(\mathbf{G})$ are orthonormal

$$\sum_{\mathbf{G}} c_i^*(\mathbf{G}) c_j(\mathbf{G}) = \delta_{ij}. \tag{18.13}$$

[4] It is also possible to treat the occupations as dynamical variables in a way related to the ensemble density functional theory method [426], which can potentially allow the Car–Parrinello unified algorithm to apply directly to metals.

The discrete time step equation corresponding to (18.10) becomes

$$c_i^{n+1}(\mathbf{G}) = 2c_i^n(\mathbf{G}) - c_i^{n-1}(\mathbf{G}) - \frac{(\Delta t)^2}{\mu}\left[\sum_{\mathbf{G}'} H(\mathbf{G},\mathbf{G}')c_i^n(\mathbf{G}') - \sum_k \Lambda_{ik}c_k^n(\mathbf{G})\right], \quad (18.14)$$

where Δt denotes the time step. The equation for Λ_{ik} is derived by assuming that $c_i^n(\mathbf{G})$ and $c_i^{n-1}(\mathbf{G})$ are each orthonormal and imposing the condition that $c_i^{n+1}(\mathbf{G})$ is also orthonormal. The complete solution is then found by updating the electron density at each iteration, finding the new Kohn–Sham effective potential, and, if desired, moving the atoms according to Eq. (18.2) using the force theorem. The procedure then starts over with a new iteration. Thus all the operations have been combined into one unified algorithm.

The algorithm as presented is still too slow to be useful because of the matrix multiplication in Eq. (18.14), for which the number of operations scales as the square of the number of plane waves N_{PW}^2. To circumvent this bottleneck, Car and Parrinello used fast Fourier transforms (FFTs) to reduce the scaling to $N_{\mathrm{PW}}\log N_{\mathrm{PW}}$. The ideas have very general applicability and are described in Secs. 12.7 and M.11, where the algorithms are summarized in Figs. 12.4 and M.2. The key steps are the operation $\hat{H}\psi$ and calculation of the density. The kinetic energy operation is a diagonal matrix in Fourier space, whereas multiplication by V is simple in real space where V is diagonal. By the use of the FFT, the operations can be carried out, respectively, in Fourier and real spaces, and the results collected in either space. The limiting factor is the FFT, which scales as $N_{\mathrm{PW}}\log N_{\mathrm{PW}}$. The sequence of steps is described in Fig. M.2.

Finally, at every step the energy and force on each nucleus can be calculated using the force theorem expressed in plane waves, Eq. (13.3). A variant of this form, however, may be more convenient for simulations with large cells. As explained in Sec. F.3 and [705], the force on an ion due other ions (the Ewald term) can be combined with the local pseudopotential term, leading to a combined expression in reciprocal space plus correction terms expressed as short-range forces between ions, Eq. (F.16). The last are easily included in a standard MD simulation.

The method can be extended to "ultrasoft" pseudopotentials [565] and to the PAW method [475, 476], which is very useful for simulations with atoms that require high plane wave cutoffs using norm-conserving pseudopotentials, e.g., transition metals. The basic idea is that in any such approach the same general formulation can be used for updating the plane wave coefficients, calculating forces, *etc.* The difference is that there are additional terms rigidly attached to the nuclei that must be added in the expressions [475, 476, 565].

18.4 Alternative approaches to density functional QMD

As pointed out in Sec. 18.1, quantum molecular dynamics (QMD) and relaxation of atom positions can be carried out by *any method* that derives forces from the electrons. Of course, the force (Hellmann–Feynman) theorem, Eq. (3.18), is well known and has been widely applied in pseudopotential calculations (e.g., Chs. 13 and 19). The problem is that the forces must be calculated *very efficiently* for simulations of large systems. Progress was underway

contemporaneous with the Car–Parrinello work on self-consistent density functional theory methods (e.g., [429]) and simpler tight-binding-type methods (e.g., [715]). There has been great progress in creating new algorithms [440], so that there are now very efficient alternatives to the unified algorithm of Car and Parrinello. *The bottom line is that different approaches can each be used to great advantage; each works well if used with care; and each has particular advantages that can be utilized in individual situations.*

Instead of one unified algorithm, a key feature of alternative approaches is that the motion of the ions and the updating of the electrons are done by different algorithms. Although the elegance of a unified approach is lost, this division gives additional options that can be used to advantage. The numerically intensive steps are essentially the same: the calculation of the energy gradients with respect to atom positions and electron wavefunctions used in the Car–Parrinello equations are exactly the same as in the iterative methods of App. M. The same tricks can be used, e.g., the use of FFTs. There may be a difference, however, in how often these steps have to be applied. By dividing the problem into two parts, the entire algorithm has the following general properties:

- The time step is governed by the nuclear dynamics, i.e. it is the same as in an ordinary classical MD calculation. This is longer (by about an order of magnitude) than the step in the Car–Parrinello algorithm, so that the atoms move further in one step.
- At each step, the electrons must be solved accurately – much more accurately than in the Car–Parrinello method. This requires more cycles of self-consistency at each MD step, roughly an order of magnitude more calculations per MD step than in the Car–Parrinello method. Thus, to a first approximation, the two approaches require similar amounts of computation.
- Different iterative methods (App. M) can be chosen to find the eigenvalues and eigenvectors of all the occupied states, or only the occupied subspace that spans the eigenvectors. The latter is in general faster; the former has the advantage that the eigenvalues can be used in the Fermi function to treat metals with no essential problems.
- Since the computation needed to reach self-consistency is such an important factor in the iterative methods, this is a promising avenue for improvement. Thus there can be significant advantages with algorithms designed to give a better starting guess for the potential and wavefunctions at each MD step and faster convergence to self-consistency. This is a matter of active development, e.g., in [716].

18.5 Non-self-consistent QMD methods

Much simpler (and faster computationally) simulation methods can be devised if there is no requirement for the full self-consistent Kohn–Sham equations to be solved. The simplest approach uses the empirical tight-binding method (Ch. 14) in which the hamiltonian is given strictly in terms of matrix elements that are simple functions of the positions of the atoms. Since the basis set is also small (several orbitals per atom) it may be more efficient simply to diagonalize the matrices rather than use an iterative method. Then eigenvectors, energies, and forces can be calculated for all positions of the atoms, usually much faster

than a typical self-consistent plane-wave algorithm. This approach [715] was developed simultaneously with the Car-Parrinello work and still enjoys widespread use because of its speed and simplicity.

Another approach is to solve the electronic problem within a basis using an approximate non-self-consistent form of the hamiltonian. Such methods are "*ab initio*" since they are not parameterized and the approximation forms have been used effectively for many problems with a total potential that is a sum of atomic-like potentials [601, 603]. Together with the explicit functional for the energy in terms of the input density, Eq. (9.9), and the usual expressions for the forces, this enables a complete, albeit approximate, DFT QMD algorithm. Self-consistency can be added in limited ways that still preserve efficiency, as done e.g., in [717].

18.6 Examples of simulations

It is instructive to apply the Car–Parrinello algorithm to simplified problems. Examples are described in detail in Exercises 18.2 and 18.3 that illustrate the simplest 2-state problem; finding the eigenstates of a simple problem by quenching (the analog of the original calculation of Car and Parrinello in Fig. 18.1), and the equations of motion using the fictitious lagrangian.

Phase stability: carbon, iron, . . .

As an examples of calculations on real materials, carbon is particularly interesting because of its many forms with extreme properties. Simulations were crucial in providing information for understanding of the phase diagram at high pressure and high temperature, which has been the subject of debate and revisions for decades. The most complete phase diagram of carbon proposed to date is given in Fig. 2.10 based upon data at low pressures and Car–Parrinello simulations at high pressure and high temperature. The high P,T regions are unknown in laboratory experiments, but are conditions found in geology and astrophysics. The region above 5,000 K and around 1 Mbar (100 GPa) is very difficult to access experimentally, which has led to controversies, e.g., whether liquid C is a metal or insulator. Since the electrons are treated with quantum Kohn–Sham theory as the atoms move, the simulations also yield information on the nature of liquid carbon and can answer such questions.

Simulations using the Car–Parrinello method were carried out by Galli, et al. [159] using Bachelet–Hamann–Schluter [499] pseudopotentials, a plane wave basis (usually with a small cutoff of 20 Ry that introduces some errors and with checks at +32 Ry), a fictitious mass of $\mu = 200$ a.u., and a time step of $\Delta t = 4$ a.u. Typical calculations involved heating and quenches of the order of 3,000 steps, each followed by an anneal of ≈5,000 steps to create an amorphous carbon structure at room temperature. This was then heated in steps and finally equilibrated for ≈10,000 steps to compute averages in the liquid state at $T = 5{,}000$ K. Because of the energy transfer problem in metals, thermostats were used as described in Sec. 18.2.

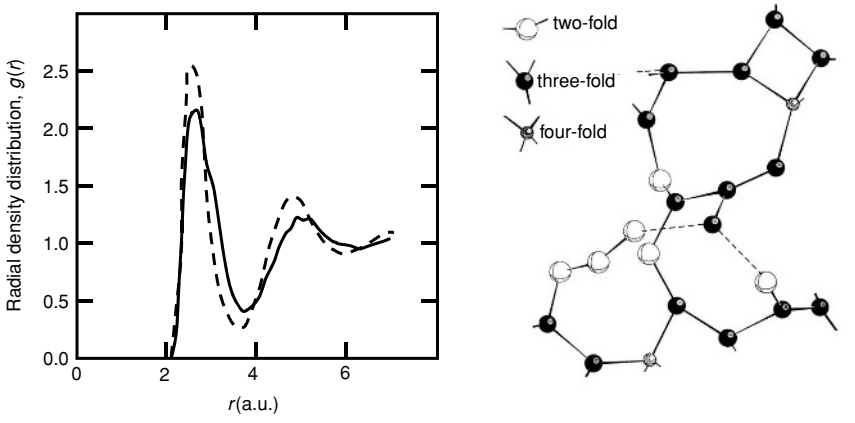


Figure 18.2. Simulations of liquid carbon at $P \approx 0$ and $T = 5{,}000$ K. Left: Radial density distribution $g(r)$ calculated by Car–Parrinello plane wave methods [159] and using tight-binding methods [162]. The definite peak at $r \approx 2.3a_0$ and minimum at $r \approx 2.8a_0$ provide a quantitative definition of bonded neighbors as those with $r < 2.8a_0$. Right: Snapshot of typical configuration showing bonded atoms, neighboring 2-, 3-, and 4-fold coordination present in the low-pressure liquid [159].

The calculated radial density distribution $g(r)$ for liquid C at $P \approx 0$ and $T = 5{,}000$ K is shown in Fig. 18.2, where it can be seen that very similar conclusions result from the Car–Parrinello plane wave calculations and tight-binding calculations of Xu, et al. [162]. The latter uses an environment-dependent form of tight-binding matrix elements that has been shown to describe well the properties of C in two-, three-, and four-fold coordinated structures. The bonded neighbors of each atom are defined to be other atoms within a distance that includes the first peak, $r < 2.8a_0$, as illustrated in a typical snapshot of the liquid shown on the right, all the atoms are found to have two-, three-, and four-fold coordination, with no disconnected atoms or higher coordination. The average coordination is $\approx$2.9 [159], which is consistent with the fact that higher coordinated structures are at much higher energies and become relevant only at much higher pressures as shown in Fig. 2.10.

A great advantage of QMD methods is that both electronic and thermal nuclear properties are accessible in the same calculation. For example, in liquid carbon there is a question: is it insulating (diamond like) or metallic? The time-averaged electronic density of states does not answer the question. As shown on the left-hand side of Fig. 18.3, the average density of states is almost free-electron like. On the other hand, there is very strong scattering of the electrons by the ions which might lead to localization. Theory can avoid semantics and directly calculate the conductivity $\sigma(\omega)$ given by Eq. (E.11) and the well-known Eq. (20.2) in terms of momentum matrix elements. This yields $\sigma(\omega)$ at any configuration of the nuclei and the final results are found by averaging over configurations in the MD simulation. The result shown in Fig. 18.3 is a conductivity at $T = 5{,}000$ K that has typical Drude form [84, 86] with a very short mean free path of the order of the interatomic spacing.

One of the great challenges of geology is to understand the nature of the interior of the Earth. The conditions are very difficult to reproduce in the laboratory, since the pressure

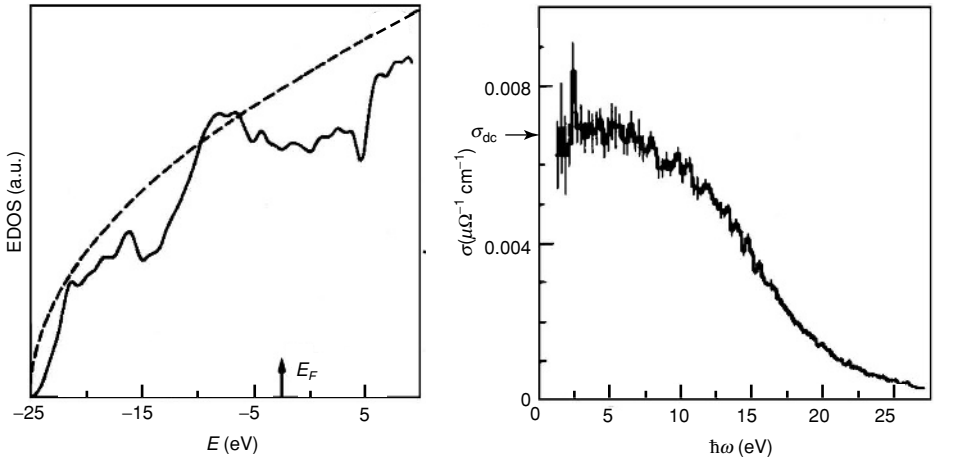


Figure 18.3. Electronic properties of liquid carbon at $P \approx 0$ and $T = 5{,}000$ K. Left: Time-averaged density of states which is close to the free-electron parabola (dashed line). Right: Conductivity $\sigma(\omega)$ calculated for a given ionic configuration using Eq. (E.11) and averaged over configurations. The form of $\sigma(\omega)$ is similar to the Drude form, expected for metal, and the dc conductivity can be estimated from the extrapolation $\omega \to 0$. From [159].

and temperature are estimated to be ≈135 GPa, ≈4,000 K at the mantle/core boundary and ≈330 GPa, ≈5,000 K at the boundary of the outer and inner core. Since the core is made of Fe with undetermined amounts of other elements, there is a great opportunity for simulations to make a major contribution. Toward this end, remarkable achievements have been made by Alfè, Gillan, Kresse, and coworkers, [164], who have carried out simulations on Fe (and Fe–S mixtures [163]) using plane waves and ultrasoft pseudopotentials or the PAW method. The methods were carefully tested on crystalline phases of Fe, demonstrating very good agreement with full-potential LAPW calculations (Ch. 17) for energies, pressure–volume relations, and phonon frequencies. The QMD simulations were performed using the approach of Kresse and Furthmüller [718] in which the electronic equations were solved iteratively to provide forces acting on the nuclei.

The radial density distribution for liquid Fe with fixed density $\rho = 10{,}700$ Kg/m^3 (the value at the core/mantle boundary) is shown in Fig. 18.4 for several temperatures. The calculated pressure ranges from 312 to 172 GPa. The weight of the peak in $g(r)$ corresponds to slightly over 12 neighbors, i.e. to a close-packed liquid. As expected, the peak broadens with increasing T, with no transitions. This represents the first steps in calculations of the melting curve, solid solubilities, *etc.*, that require calculations of the free energy, which is notoriously difficult in simulations [437].

The most important liquid: water

Perhaps the greatest challenge of all is to treat water and aqueous solutions of ions and molecules. Examples of simulation results are shown in Figs. 2.11 and 2.12. The status at

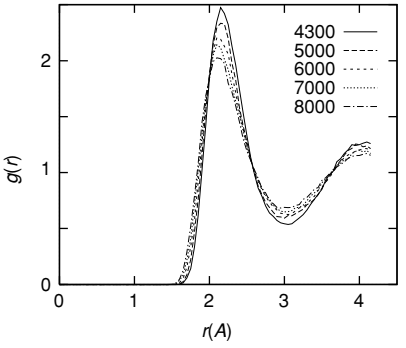


Figure 18.4. The calculated radial density distribution $g(r)$ of liquid Fe at fixed density equal to the value at the core/mantle boundary as a function of temperature [164]. The lowest temperature $T =$ 4,300 K is the expected temperature at the boundary. The simulations were done using the PAW method (Sec. 13.2) and the methods of [476]. The integral under the first peak is $\approx$12 atoms indicating a close-packed liquid at all pressures. From [164].

present is that simulations are making great progress in predicting important properties, but the details are crucially important. This is certain to be among the most important fields of future research.

Reactions and catalysis

QMD methods provide a general approach to calculations of reaction paths and catalysis. This is far too great a subject to attempt to cover in this book. The example of the Ziegler–Natta reaction shown in Fig. 2.13 illustrates the ways that QMD can provide insight into the nature of atomic-scale reactions and clarify proposed mechanisms [172, 173]. However, a word of caution is in order: reaction barriers are particularly sensitive to electronic correlations and present density functionals often are simply not accurate enough for many problems.

Structures of defects, surfaces, clusters, . . .

A vast number of calculations have been performed to predict the atomic-scale structures of molecules, clusters, surfaces, *etc.* Most of these do not involve molecular dynamics, but use force calculations to relax the structures to find (meta)stable minima. Examples are the semiconductor defect and surface structures shown in Figs. 2.15–2.17 and the buckled-dimer structure of the Si (100) surface shown in Fig. 13.7. The latter case has been studied extensively, including MD simulations, because of controversies in comparison of STM experiments with theory that raise the issue of whether thermal motion stabilizes the symmetric state (see references in [586]). The final answer may not yet be determined.

An exciting area of research involves nanoclusters, where theory has much to add since information about such structures is difficult to determine experimentally. Examples of Si clusters shown in Fig. 2.18 have all been determined by relaxation or MD. An instructive

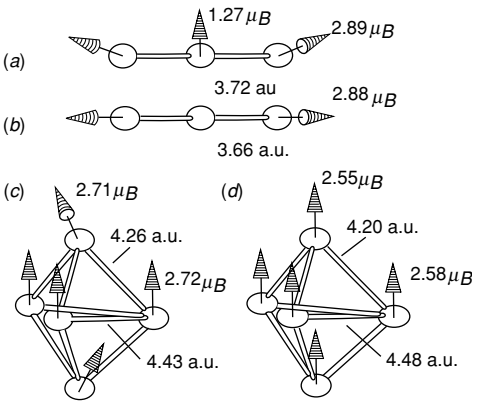


Figure 18.5. Equilibrium structures and magnetic moments of small Fe clusters as calculated [384] using plane waves and ultrasoft pseudopotentials. Of particular note are the predicted non-collinear spin states. Such calculations with the non-colinear spin formalism are needed for treating magnetism at finite temperature [382, 383]. From [384].

example is Si_{13}, where there is competition between a symmetric structure with 12 outer atoms surrounding a central atom [191] and a low-symmetry structure found by quenching finite temperature MD simulations [190]. In fact, quantum Monte Carlo calculations [192, 193] confirm that the low-symmetry structure is lower in energy.

As an example involving magnetism, Fig. 18.5 shows predicted structures of Fe_3 and Fe_5 molecules [384]. This work used ultrasoft pseudopotentials and plane waves, and, most importantly, used non-collinear formalism (Sec. 8.4) to predict the spin density shown in the figure. Such molecules can be treated with other methods that have also found non-collinear spin density. Furthermore, non-collinear formalism is essential for treating bulk magnetism at finite temperature [382, 383].

SELECT FURTHER READING

Original work:

Car, R. and Parrinello, M., "Unified approach for molecular dynamics and density functional theory," *Phys. Rev. Lett.* 55:2471–2474, 1985.

Tutorials:

Remler, D. K. and Madden, P. A. "Molecular dynamics without effective potentials via the Car-Parrinello approach," *Molecular Physics* 70:921, 1990.

Pastore, G. Smargiassi, E. and Buda, F. "Theory of *ab initio* molecular-dynamics calculations," *Phys. Rev.* A 44:6334, 1991.

Thijssen, J. M. *Computational Physics*, Cambridge University Press, Cambridge, England, 2000.

Descriptions of algorithms:

Car, R. and Parrinello, M. in *Simple Molecular Systems at Very High Density*, edited by Polian, A. Loubeyre, P. and Boccara, N. Plenum, New York, 1989, p. 455.

Galli, G. and Parrinello, M. in *Computer Simulations in Material Science*, edited by Meyer, M. and Pontikis, V. Kluwer, Dordrecht, 1991, pp. 283–304.

Payne, M. C. Teter, M. P. Allan, D. C. Arias, T. A. and Joannopoulos, J. D. "Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients," *Rev. Mod. Phys.* 64:1045–1097, 1992.

Tuckerman, M. E. and Parrinello, M. "Integrating the Car-Parrinello equations. I. Basic integration techniques," *J. Chem. Phys.* 101:1302, 1994.

Tuckerman, M. E. and Parrinello, M. "Integrating the Car-Parrinello equations. II. Multiple time scale techniques," *J. Chem. Phys.* 101:1316, 1994.

Simulation techniques:

Special issue, "Techiniques for simulations," *Computational Materials Science* 12, 1998.

Exercises

18.1 In the text it was stated that the "SHAKE" algorithm [701,711] maintains constraints in a holonomic manner, i.e. with no energy loss. An alternative might be the Gram–Schmidt procedure in which one updates the wavefunctions with $\hat{H}\psi_i$ and then orthonormalizing starting with the lowest state.
(a) Show that this will cause energy loss. Hint: One way is to consider the two-state problem in Exercise 18.2. Treat the wavefunctions explicitly and show that there is a difference from the equations given below in which the constraint is imposed analytically.
(b) Read the references for SHAKE [701,711] and summarize how it works.

18.2 Car–Parrinello-type simulation for one electron in a two-state problem is the simplest case and is considered in the tutorial-type paper by Pastore, Smargiassi, and Buda [703]. In this case, the wavefunction can always be written as a linear combination of any two orthonormal states ϕ_1 and ϕ_2,

$$\psi = \cos\left(\frac{\theta}{2}\right)\phi_1 + \sin\left(\frac{\theta}{2}\right)\phi_2.$$

With this definition orthogonality and normalization are explicitly included and we can consider θ to be the variable in the fictitious lagrangian (written for simplicity in the case where ϕ_1 and ϕ_2 are eigenvectors):

$$L = \mu|\frac{\mathrm{d}\theta}{\mathrm{d}t}|^2 + \epsilon_1\cos^2 + \epsilon_2\sin^2$$

Solving the Lagrange equations gives

$$\mu\frac{\mathrm{d}^2\theta(t)}{\mathrm{d}t^2} = (\epsilon_2 - \epsilon_1)\sin\left(\theta(t) - \theta_0\right),$$

which is the equation for a pendulum. For small deviations $\theta - \theta_0$, the solution is simple harmonic oscillations of frequency $\omega_e^2 = \Delta E/\mu$. Thus so long as the oscillations are small, the electronic degrees of freedom act like simple oscillators.

Pastore et al. [703] have analyzed the two-state model and large cell calculations to identify the key features, as illustrated in the figures from their paper. If μ is chosen so that the fictitious electronic frequencies are well above all lattice frequencies and motions are small, then there

is only slow energy transfer and the Car–Parrinello method works well. This can be done in an insulator. But level crossing, metals, etc., give interesting difficulties.

The exercise is to analyze the algorithm for three cases in which the system is driven by an external perturbation of frequency ω_0:

(a) For the case of small amplitudes and μ chosen so that $\omega_e >> \omega_0$, show that the electrons respond almost instantaneously adiabatically following the driving field.
(b) For the more difficult case with ω_e of order ω_0, show that the electrons couple strongly with large non-linear oscillations. (Note: This fictitious dynamics is *not* the correct quantum dynamics.)
(c) For the case where there is a level crossing and ΔE changes sign, show that the electrons can undergo real transitions. (See note in (b).)

18.3 Project for simulation of quantum systems with Car–Parrinello methods. The purpose of this problem set is to write programs and carry out calculations in simple cases for the Car–Parrinello method for simulation of quantum systems by molecular dynamics techniques. Ignore the spin of the electron, which only adds a factor of 2 in paramagnetic cases with even numbers of electrons per cell.
(a) For the case of an "empty lattice" where the potential energy is a constant set equal to zero, write down the Car–Parrinello equations of motion for the electrons. Work in atomic units.
(i) Set up the problem on a one-dimensional lattice, where the wavefunctions are required to be periodic with length L. Write a program which iterates the Verlet equation for a single wavefunction expressed in terms of Fourier coefficients up to $M * (2\pi/L)$.
(ii) Choose $L = 10$ a.u., $\mu = 300$ a.u., and $M = 16$, which are reasonable numbers for solids. Start with a wavefunction having random coefficients, velocities zero, and iterate the equations. Choose a time step and show that the fictitious energy is conserved for your chosen time step. Show that you can carry out the exercise equivalent to the original calculation of Car and Parrinello in Fig. 18.1. Extract energy from the system by rescaling the velocities at each step. Show that the system approaches the correct ground state with energy zero. Make a graph of the energy versus time analogous to Fig. 18.1.
(iii) Now consider several states. Add the orthogonalization constraints, and find the ground state for two, three, and four filled states. Verify that you find the correct lowest states for a line with periodic boundary conditions.

Make a graph of the total energy and fictitious kinetic energy as a function of time. Show the variation in total energy on a fine scale to verify that it is well conserved.
(b) Now add a potential $V(x) = A \sin(2\pi x/L)$. Use an FFT to transform the wavefunction to real space, multiply by the potential, and the inverse FFT to transform back to Fourier space.
(i) For two electrons per cell (up and down) one has a filled band with a gap to the next band. Find the ground wavefunction and electron density for a value of $A = 1$ Hartree, a reasonable number for a solid. (All results can be verified by using the plane wave methods and diagonalization as described in Ch. 12.)
(c) Consider a system with the electrons coupled to slow classical degrees of freedom, let A be coupled to an oscillator, $A = A_0 + A_1 x$, and the energy of the oscillator $E = 0.5M\omega_0^2 x^2$. Choose values typical for ions and phonon frequencies (Ch. 19).
(i) Choose a fictitious mass μ so that all the electronic frequencies are much greater than ω_0. See Exercise 18.2.

(ii) Start the system at $x = 0$, which is not the minimum, and let it evolve. Does the oscillator go through several periods before significant energy is transferred to the electron state? Plot the total energy of the system and the fictitious kinetic energy as a function of time. Show that the total energy is accurately conserved, and the fictitious kinetic energy is much less than the oscillator kinetic energy for several cycles.
(iii) The oscillator should oscillate around the minimum. Check, by calculating the total energy by the quenching method, for fixed x, for several values of x near the minimum. Is the minimum in energy found this way, close to the minimum found from the oscillations of the dynamic system?

19

Response functions: phonons, magnons, . . .

Summary

Many properties of materials – mechanical, electrostatic, magnetic, thermal, etc. – are determined by the variations of the total energy around the equilibrium configuration, defined by formulas such as (2.2)–(2.7). Experimentally, vast amounts of information about materials are garnered from studies of vibration spectra, magnetic excitations, and other responses to experimental probes. This chapter is devoted to the role of electronic structure in providing predictions and understanding of such properties, through the total energy and force methods described in previous chapters, as well as recent advances in efficient methods for calculation of response functions themselves. Through these developments, calculation of full phonon dispersion curves, dielectric functions, infrared activity, Raman scattering intensities, magnons, anharmonic energies to all orders, phase transitions, and many other properties have been brought into the fold of practical electronic structure theory.

The primary properties considered in previous chapters are the *total energy* and (generalized) *forces*. These are sufficient to treat a vast array of problems including stability of structures, phase transitions, surfaces and interfaces, spin polarization, "*ab initio*" molecular dynamics, *etc.* One can also use such direct methods to calculate all the derivatives of the energy with respect to perturbations, by carrying out full self-consistent calculations for various values of the perturbation, and extracting derivatives from finite difference formulas. This has been used very successfully, for example, in the "frozen phonon" method illustrated in Fig. 2.8 and described further in Sec. 19.2.

Is it possible to calculate the derivatives directly? The answer is, of course, "yes," since it is just a matter of well-known perturbation theory and response functions, for which the general theory is summarized in App. D. The subject of this chapter is recent developments that allow the expressions to be re-written in ways that are much more efficient for actual calculations, together with examples for phonons, dielectric functions, and magnons. As an example of the power of the approach, determination of phonon or magnon dispersion curves for a crystal can be done in the "frozen phonon" method only with large "supercell" calculations. In contrast, the perturbation theory approach allows the phonon frequencies at any $\mathbf{k}$ to be found from a much smaller calculation based on one unit cell.

19.1 Lattice dynamics from electronic structure theory

Since nuclear motion is slow compared to typical electron frequencies, it is an excellent approximation to neglect any dependence of electronic energies on the velocities of nuclei, i.e. the *adiabatic or Born and Oppenheimer (BO) approximation* [89] (App. C; see also [90], App. VIII, and [466], pp. 169–172). Then the equations of motions for the nuclei are determined by the total energy $E(\mathbf{R})$ of the system of electrons and nuclei with the nuclear positions $\mathbf{R}$ regarded as parameters. Here $\mathbf{R}_I$ are the coordinate and mass M_I of nucleus I, $\mathbf{R} \equiv \{\mathbf{R}_I\}$ indicates the set of all the nuclear coordinates, and $E(\mathbf{R})$ is the ground state energy, Eq. (3.9) or any of the forms given in Sec. 9.2. $E(\mathbf{R})$ is often referred to as the *Born–Oppenheimer energy surface* and the adiabatic motion of the nuclei is restricted to this surface.

The complete quantum description of the nuclei is determined by the Schrödinger equation for the nuclei, which is a formidable many-body problem [152] important for light atoms. If the nuclei are treated classically, the problem reduces to coupled classical equations of motion for each nuclear position $\mathbf{R}_I(t)$

$$M_I \frac{\partial^2 \mathbf{R}_I}{\partial t^2} = \mathbf{F}_I(\mathbf{R}) = -\frac{\partial}{\partial \mathbf{R}_I} E(\mathbf{R}), \tag{19.1}$$

which can be treated by molecular dynamics. All effects of the electrons are contained in the forces that can be calculated from the electronic structure in simulations described in Ch. 18.

For stable solids at moderate temperature, it is much more useful and informative to cast the expressions in terms of an expansion of the energy $E(\mathbf{R})$ in powers of displacements and external perturbations, as in Eqs. (2.2)–(2.7). Equilibrium positions $\{\mathbf{R}_I^0\} = \mathbf{R}^0$ are determined by the zero-force condition on each nucleus,

$$\mathbf{F}_I(\mathbf{R}^0) = 0. \tag{19.2}$$

Quantum zero-point motion, thermal vibrations, and response to perturbations are described by higher powers of displacements,

$$C_{I,\alpha;J,\beta} = \frac{\partial^2 E(\mathbf{R})}{\partial \mathbf{R}_{I,\alpha} \partial \mathbf{R}_{J,\beta}}, \quad C_{I,\alpha;J,\beta;K,\gamma} = \frac{\partial^3 E(\mathbf{R})}{\partial \mathbf{R}_{I,\alpha} \partial \mathbf{R}_{J,\beta} \partial \mathbf{R}_{K,\gamma}}, \ldots, \tag{19.3}$$

where Greek subscripts $\alpha, \beta, \ldots$, indicate cartesian components.

Within the *harmonic approximation* [90], the vibrational modes at frequency ω are described by displacements

$$\mathbf{u}_I(t) = \mathbf{R}_I(t) - \mathbf{R}_I^0 \equiv \mathbf{u}_I e^{i\omega t}, \tag{19.4}$$

so that (19.1) becomes for each I

$$-\omega^2 M_I u_{I\alpha} = -\sum_{J\beta} C_{I,\alpha;J,\beta} u_{J\beta}. \tag{19.5}$$

The full solution for all vibrational states is the set of independent oscillators, each with vibrational frequency ω, determined by the classical equation

$$\det \left| \frac{1}{\sqrt{M_I M_J}} C_{I,\alpha;J,\beta} - \omega^2 \right| = 0, \tag{19.6}$$

where the dependence upon the masses M_I, M_J has been cast in a symmetric form.

For a crystal, the atomic displacement eigenvectors obey the Bloch theorem, Eqs. (4.32) and (4.33), i.e. the vibrations are classified by $\mathbf{k}$ with the displacements $\mathbf{u}_s(\mathbf{T_n}) \equiv \mathbf{R}_s(\mathbf{T_n}) - \mathbf{R}_s^0(\mathbf{T_n})$ of atom $s = 1, S$ in the cell $\mathbf{T_n}$ given by

$$\mathbf{u}_{s,\mathbf{T_n}} = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T_n}} \mathbf{u}_s(\mathbf{k}). \tag{19.7}$$

Inserting this into (19.6) leads to decoupling of the equations at different $\mathbf{k}$ (just as for electrons – Exercise 19.2), with frequencies $\omega_{i\mathbf{k}}$, $i = 1, 3S$ called dispersion curves that are solutions of the $3S \times 3S$ determinant equation

$$\det \left| \frac{1}{\sqrt{M_s M_{s'}}} \, C_{s,\alpha;s',\alpha'}(\mathbf{k}) \; - \; \omega_{i\mathbf{k}}^2 \right| = 0, \tag{19.8}$$

where the reduced force constant matrix for wavevector $\mathbf{k}$ is given by

$$C_{s,\alpha;s',\alpha'}(\mathbf{k}) = \sum_{\mathbf{T_n}} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{T_n}} \frac{\partial^2 E(\mathbf{R})}{\partial \mathbf{R}_{s,\alpha}(0) \partial \mathbf{R}_{s',\alpha'}(\mathbf{T_n})} = \frac{\partial^2 E(\mathbf{R})}{\partial \mathbf{u}_{s,\alpha}(\mathbf{k}) \partial \mathbf{u}_{s',\alpha'}(\mathbf{k})}. \tag{19.9}$$

Since the vibrations are independent, quantization is easily included as usual for harmonic oscillators: phonons are the quantized states of each oscillator with energy $\hbar\omega_{i\mathbf{k}}$.

A useful analogy can be made between phonons and electrons described in a tight-binding model. Since the nuclei have three spatial degrees of freedom, the equation of motion, (14.7), has exactly the same form as (19.8) for the case with only three states of p symmetry for the electrons. The set of exercises comprising Exercises 19.1–19.7 is designed to show the relationships, derive phonon dispersion curves in simple cases, and explicitly transform a computational code for the tight-binding algorithm into a code for phonon frequencies with force constants described by models analogous to the parameterized models used in tight-binding methods.

Examples of dispersion curves are given in Figs. 2.9, 2.32, 19.2, and 19.4. There are three acoustic modes with $\omega \to 0$ for $\mathbf{k} \to 0$ and the other $3S - 3$ modes are classified as optic. In insulators, there may be non-analytic behavior with different limits for longitudinal and transverse modes, illustrated in Figs. 19.2 and 2.9.

The framework is set for derivation of the lattice dynamical properties from electronic structure so long as attention is paid to certain features:

- Careful treatment of the long-range effects due to macroscopic electric fields in insulators.
- Formulation of the theory of elasticity in ways that facilitate calculations.

The key aspects of dealing with macroscopic electric fields are treated in Sec. E.6. In particular, "Born effective charges" $Z^*_{I,\alpha\beta}$ are defined in (E.20) in terms of the polarization

per unit displacement *in the absence of a macroscopic electric field*. Similarly, proper piezoelectric constants $e_{\alpha,\beta\gamma}$ are defined in terms of polarization per unit strain in the absence of a macroscopic electric field. Fortunately, the theory and practical expressions for polarization are well established due to recent advances (Ch. 22). The result is that $C_{s,\alpha;s',\alpha'}(\mathbf{k})$ can be divided into terms that are analytic, as in the small $\mathbf{k}$ limit, plus non-analytic contributions that can be treated exactly in terms of $Z^*_{I,\alpha\beta}$ and $e_{\alpha,\beta\gamma}$. These considerations are required in any approach that aspires to be a fundamental theory of lattice properties.

The basic definition of stress is given in (G.4) and the elastic constants, defined in (G.5), are the subject of many volumes [86, 88, 721, 722]. The important point to emphasize here is that both the electrons and the nuclei contribute directly to the stress, for which there are rigorous formulations (App. G). The expressions use the (generalized) force theorem which depends upon the requirement that *all internal degrees of freedom be at their minimum energy values*. This includes the electron wavefunctions and the nuclear positions, which are described by "internal strains" that are determined by the zero-force condition, Eq. (G.13). In many simple cases, e.g. in Bravais lattices, the force is zero by symmetry for zero internal strain; however, in general, the positions of the nuclei in a strained crystal are not fixed by symmetry, and any fundamental calculations of elastic properties must find the internal strains from the theory.

19.2 The direct approach: "frozen phonons," magnons, ...

The most direct approach is simply to calculate the total energy and/or forces and stresses as a function of the position of the nuclei, i.e. "frozen phonons", using any of the expressions valid for the electronic structure. Then the relevant quantities are defined by numerical derivatives for displacements

$$C_{I,\alpha;J,\beta} \approx -\frac{\Delta F_{I,\alpha}}{\Delta \mathbf{R}_{J,\beta}}, \qquad Z^*_{I,\alpha\beta}|e| \approx \left.\frac{\Delta \mathbf{P}_\alpha}{\Delta \mathbf{R}_{I,\beta}}\right|_{\mathbf{E}_{\rm mac}}, \tag{19.10}$$

and for strains

$$C_{\alpha\beta;\gamma\delta} \approx -\frac{\Delta \sigma_{\alpha\beta}}{\Delta u_{\alpha\beta}}, \qquad e_{\alpha\beta\gamma} \approx \left.\frac{\Delta \mathbf{P}_\alpha}{\Delta \epsilon_{\alpha\beta}}\right|_{\mathbf{E}_{\rm mac}}. \tag{19.11}$$

Such calculations are widely used since they require no additional computational algorithms; furthermore, this is the method of choice in cases where there are large displacements since the energy is automatically found to all orders. The direct approach played a critical role in early work, for example [143] and [723], where it was shown that phonons in materials (other than the sp-bonded metals) could be derived from the electronic structure.

Transition metals are an excellent example where electronic structure calculations of "frozen phonon" energies can provide much information on the total bonding and the states near the Fermi energy that couple strongly to the phonons. The phonon energies for many transition metals have been shown to be well described, e.g. in [724] and [725], by calculations at wavevectors $\mathbf{k}$ along high-symmetry directions. For example, there is an interesting anomaly in the longitudinal frequency for $\mathbf{k} = \left(\frac{2}{3}, \frac{2}{3}, \frac{2}{3}\right)$ in the bcc structure

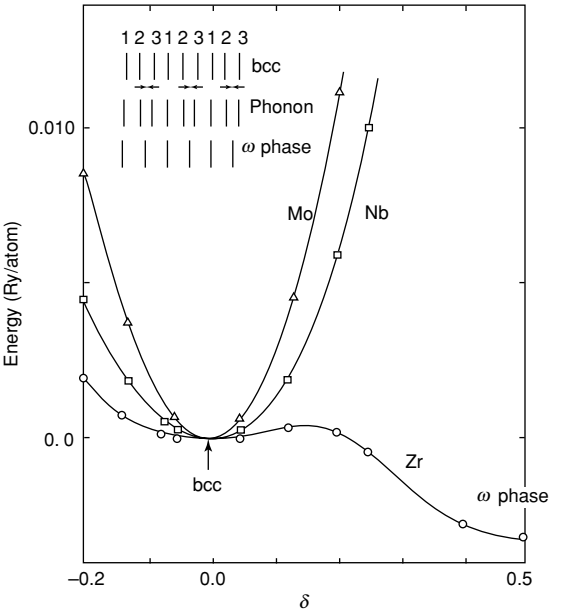


Figure 19.1. Theoretical calculation of the energy versus displacement for longitudinal displacements with wavevector $\mathbf{k} = \left(\frac{2}{3}, \frac{2}{3}, \frac{2}{3}\right)$ in the bcc structure [725]. For Mo and Nb the minimum energy is for bcc structure. The curvature agrees well with measured phonon frequencies and corresponds to a sharp dip in the phonon dispersion curves, which is a precursor to the phase transition that actually occurs in Zr. The minimum energy structure for Zr at low temperature is the "ω phase," which forms by displacements shown in the inset, with each third plane undisplaced and the other two planes forming a more dense bilayer. From [725].

crystals Mo and Nb. This is a precursor to the phase transition that actually occurs in Zr which has bcc structure only above 1,100 K [134, 725]. Figure 19.1 shows the calculated energy versus displacement for this phonon for Mo, Nb, and Zr. The inset shows the displacements that correspond to the "ω phase" with each third plane undisplaced, and the other two planes displaced to form a more dense bilayer. The LDA calculations agree well with the phonon frequencies in Mo and Nb and with the observed low-temperature structure of all three elements. The transition to bcc at high temperature is believed to be an effect of entropy, since it is well known that many metals have bcc structure at high temperature. A simple explanation is provided by the general arguments of Heine and Samson [726] that such a superlattice can occur for a 1/3 filled d band, which is the case for Zr. The electronic structure calculations provide a more complete picture, showing that this is not a delicate Fermi surface effect but is a combination of effects of s–p states and directional (covalent) d bonding involving states in a range around the Fermi energy [725].

Two examples of recent work have been given in Fig. 2.8, which shows the energy versus displacement for the superconductor MgB_2 and the ferroelectric $BaTiO_3$. From this information one can then extract the various orders of the anharmonic terms; for example, all the terms that are required to construct a microscopic model [727] that can be used to construct

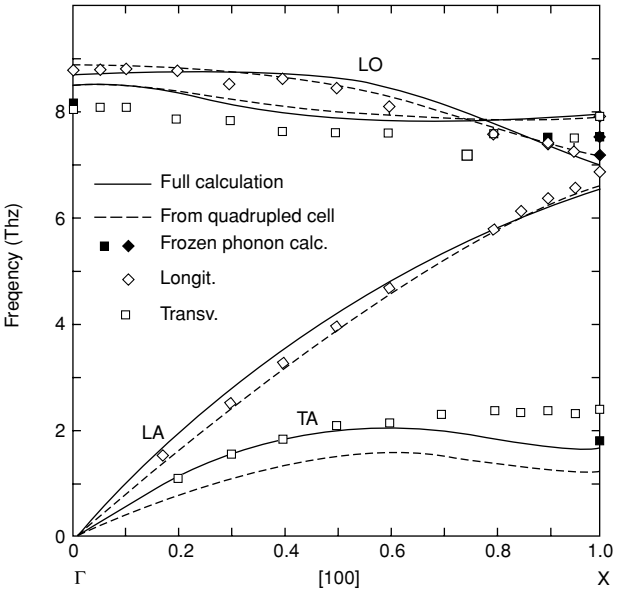


Figure 19.2. Dispersion curves for phonons in GaAs in the [1 0 0] direction calculated using the frozen phonon method. The supercell used for the calculation is a longer version of the cell shown in Fig. 13.5 with displacements as indicated that are either transverse or longitudinal. Compare with Fig. 13.6 and Sec. 13.4, which illustrate the versatility of supercells for many problems. From [574].

free energy models, and to study thermal phase transitions. Direct calculation of displacement energies has been done for hosts of materials, for example the high-temperature superconductors where the results are in good agreement with experiment even though the gaps and magnetic structure may be completely wrong. Since the electrons are treated directly, the information for electron–phonon interactions is intrinsically included – see Sec. 19.8.

It is also possible [574–576] to derive full dispersion curves from the direct force calculations on "supercells" as illustrated in Fig. 13.5 for a zinc-blende crystal. Any given phonon corresponds to a displacement of planes of atoms perpendicular to the wavevector $\mathbf{k}$, as shown on the left-hand side of the figure. All the information needed in Eq. (19.10) *for all phonons with a given direction* $\hat{\mathbf{k}}$ can be derived by displacing each inequivalent plane of atoms and calculating the force of all the atoms [574]. One must do a separate calculation for each inequivalent plane and each inequivalent displacement (four in the case shown in Fig. 13.5, for longitudinal and transverse displacement of Ga and As). If the size of the supercell exceeds twice the range of the forces, then all terms can be identified with no ambiguities. The results [574] for GaAs are shown in Fig. 19.2, compared with experiment. These are early calculations that used a semiempirical local pseudopotential. It is satisfying that better agreement with experiment has been found in more recent calculations [153,728] using *ab initio* norm-conserving potentials and the response function method (Sec. 19.3).

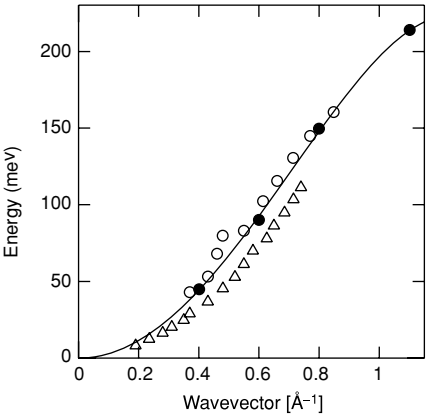


Figure 19.3. Dispersion curves for magnons in Fe: open circles are experimental points, compared to the dark circles that are theoretical results [136] calculated using the Berry's phase approach [135]. Triangles show magnon energies for an alloy with 12% Si. From [136].

The inverse dielectric constant ϵ^{-1} and the effective charges Z_I^* can also be calculated from the change in the potentials due to an induced dipole layer [574].

The practical advantage of this approach is immediately apparent from the fact that exactly the same methods can be used to calculate the properties of superlattices and interfaces. As shown in the middle figure of Fig. 13.5, a superlattice can be created by the theoretical alchemy of replacing Ga with Al in part of the supercell. Furthermore, the same methods apply directly to surfaces where part of the supercell is vacuum. This is illustrated on the right-hand side figure of Fig. 13.5, where the surface may undergo massive reconstruction beyond any perturbation expansion.

Another instructive example of the use of supercells is the calculation [577] of the inverse dielectric constant $\epsilon^{-1}(\mathbf{k})$ by imposing a periodic electrostatic potential of wavevector $\mathbf{k}$, i.e. a "frozen field." If the atoms are held fixed, the resulting potential leads to the static electronic inverse dielectric constant $\epsilon_0^{-1}(\mathbf{k})$; if the atoms are allowed to displace so that the forces are zero, one finds the result including the lattice contribution $\epsilon_\infty^{-1}(\mathbf{k})$. In each case, the $\mathbf{k} \to 0$ limit can be found by extrapolation from a few values of $\mathbf{k}$ and using the fact that $\epsilon_\infty^{-1}(\mathbf{k}) \propto k^2$ at small $\mathbf{k}$.

Spins excitations, or magnons, can be treated in the same way by calculating the energy of "frozen magnons." But how does one freeze the magnetization? Unlike phonons, magnetization is a continuous function that must be calculated. An elegant way of doing this is to generalize Berry's phase idea, Sec. 22.2, for electric polarization to the magnetic case of Niu and Kleinman [135]. An example of results [136] from this method for Fe, calculated using plane waves with a pseudopotential, are shown in Fig. 19.3: this shows excellent agreement with experiment. In this case, a *supercell is not needed*: because the magnon is a spiral excitation, the excitation obeys a generalized Bloch theorem [131] in which each cell is rotated equally with respect to its neighbors.

19.3 Phonons and density response functions

Historically,[1] the first approach for calculation of phonons was based on response functions: since all harmonic force constants, elastic constants, *etc.* involve only second derivatives of the energy, they can be derived using second-order perturbation theory. Furthermore, this builds upon the fact that in many cases one can calculate small changes more accurately than the total energy itself. This section is devoted to the formulation, which is directly useful for simple cases and leads up to the efficient methods of Sec. 19.4.

The general expressions for the response to an external perturbation $V_{\rm ext}(\mathbf{r})$ that varies with parameters λ_i (where λ denotes the position of an atom, the strain, *etc.*) are

$$\begin{aligned} \frac{\partial E}{\partial \lambda_i} &= \frac{\partial E_{II}}{\partial \lambda_i} + \int \frac{\partial V_{\rm ext}(\mathbf{r})}{\partial \lambda_i} n(\mathbf{r}) \mathrm{d}\mathbf{r}, \\ \frac{\partial^2 E}{\partial \lambda_i \partial \lambda_j} &= \frac{\partial^2 E_{II}}{\partial \lambda_i \partial \lambda_j} + \int \frac{\partial^2 V_{\rm ext}(\mathbf{r})}{\partial \lambda_i \partial \lambda_j} n(\mathbf{r}) \mathrm{d}\mathbf{r} + \int \frac{\partial n(\mathbf{r})}{\partial \lambda_i} \frac{\partial V_{\rm ext}(\mathbf{r})}{\partial \lambda_j} \mathrm{d}\mathbf{r}, \end{aligned} \tag{19.12}$$

plus higher-order terms. The first equation is just the force (Hellmann–Feynman) theorem, which involves only the external potential and the unperturbed density. The first two terms in the second equation involve only the unperturbed density; however, the last term requires knowledge of $\partial n(\mathbf{r})/\partial \lambda_i$. Using the chain rule, the last term can be written in symmetric form [152]

$$\int \frac{\partial V_{\rm ext}(\mathbf{r}')}{\partial \lambda_i} \frac{\partial n(\mathbf{r})}{\partial V_{\rm ext}(\mathbf{r}')} \frac{\partial V_{\rm ext}(\mathbf{r})}{\partial \lambda_j} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' = \int \frac{\partial V_{\rm ext}(\mathbf{r}')}{\partial \lambda_i} \chi(\mathbf{r}, \mathbf{r}') \frac{\partial V_{\rm ext}(\mathbf{r})}{\partial \lambda_j} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}', \tag{19.13}$$

where χ is the density response function, Eq. (D.6). The expressions may be written in either $\mathbf{r}$ or $\mathbf{q}$ space as in (D.7), and may be expressed in terms of inverse dielectric function using the relation $\epsilon^{-1} = 1 + V_C \chi$, given explicitly in (E.15).

Following this approach, elegant "textbook" expressions for any second derivative of the energy can be found in two steps: using a relation like (D.11),

$$\chi = \chi^0 [1 - \chi^0 K]^{-1}, \quad \text{or} \quad \chi^{-1} = [\chi^0]^{-1} - \mathrm{K}, \tag{19.14}$$

that expresses χ in terms of the kernel K in (D.10) and the non-interacting response function χ^0. In turn, χ^0 is given by (D.3)–(D.5), which are derived from standard expressions of perturbation theory in terms of sums over the entire spectrum of eigenstates. The form that is most useful for comparison with Green's function methods is Eq. (D.2), repeated here,

$$\Delta n(\mathbf{r}) = 2 \sum_{i=1}^{N} \sum_{j=N+1}^{\infty} \psi_i^*(\mathbf{r}) \psi_j(\mathbf{r}) \frac{\langle \psi_j | \Delta V_{\rm KS} | \psi_i \rangle}{\varepsilon_i - \varepsilon_j}. \tag{19.15}$$

The formulation of response functions in (19.14) and (19.15) is extremely fruitful in cases where the response function can be approximated by a simple form. For example, calculations of phonon dispersion curves in sp-bonded metals are well-described by χ from the homogeneous electron gas, i.e. the static Lindhard dielectric function $\epsilon^{-1}(|\mathbf{q}|)$, Eq. (5.38), which is an analytic, scalar function of the one-dimensional magnitude $|\mathbf{q}|$. The

[1] See references 2–13 in [152].

foundation for the theory is the nearly-free-electron approximation with the ions represented by weak pseudopotentials [467, 468].[2]

However, the approach is very difficult for accurate calculations on general materials. There are two primary problems: The quantities involved are six-dimensional functions $\epsilon^{-1}(\mathbf{q}, \mathbf{q}')$, that become large matrices $\epsilon^{-1}_{\mathbf{GG}'}(\mathbf{k})$ in crystals. Furthermore, calculation of *each of the entries in the matrix* involves a sum over the BZ of the filled and empty bands (as in Eqs. (D.3)–(D.5)) up to some energy sufficient for convergence. This is a very difficult task which we will not belabor; the next section describes a much more efficient approach.

19.4 Green's function formulation

An alternative, much more effective, approach is the recently developed "density functional perturbation theory" (DFPT) [153, 267, 728, 730], which has important relations to earlier classic works [45, 731]. Instead of calculation of the inverse dielectric function $\epsilon^{-1}_{\mathbf{GG}'}(\mathbf{k})$, which gives the response to *all possible perturbations*, DFPT is designed to calculate the needed response to a particular perturbation. Instead of the standard perturbation theory sums over empty states, the expressions are transformed into forms that involve only the occupied states, which can be calculated using efficient electronic structure methods. This leads to two related types of expressions: (1) self-consistent equations for the response function in terms of the change in the wavefunctions to a given order, or (2) variational expressions in which the calculation of a response at any given order of perturbation is cast as a problem of minimizing a functional defined at that order. The theory can be applied to any order (Sec. 3.7) but the main ideas can be seen in the lowest order linear response.

The formulation can be understood by first returning to the fundamentals of perturbation theory. In terms of the wavefunctions, the first-order change in density is

$$\Delta n(\mathbf{r}) = 2\,\mathrm{Re}\sum_{i=1}^{N} \psi_i^*(\mathbf{r})\Delta\psi_i(\mathbf{r}), \tag{19.16}$$

where $\Delta\psi_i(\mathbf{r})$ is given by first-order perturbation theory as

$$(H_{\mathrm{KS}} - \varepsilon_i)|\Delta\psi_i\rangle = -(\Delta V_{\mathrm{KS}} - \Delta\varepsilon_i)|\psi_i\rangle. \tag{19.17}$$

Here H_{KS} is the unperturbed Kohn–Sham hamiltonian, $\Delta\varepsilon_i = \langle\psi_i|\Delta V_{\mathrm{KS}}|\psi_i\rangle$ is the first-order variation of the KS eigenvalue, ε_i, and the change in the effective potential is given by

$$\begin{aligned}\Delta V_{\mathrm{KS}}(\mathbf{r}) &= \Delta V_{\mathrm{ext}}(\mathbf{r}) + e^2\int \mathrm{d}\mathbf{r}'\frac{\Delta n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \int \mathrm{d}\mathbf{r}'\frac{\mathrm{d}V_{\mathrm{xc}}(\mathbf{r})}{\mathrm{d}n(\mathbf{r}')}\Delta n(\mathbf{r}') \\ &\equiv \Delta V_{\mathrm{ext}}(\mathbf{r}) + \int \mathrm{d}\mathbf{r}' K(\mathbf{r},\mathbf{r}')\Delta n(\mathbf{r}').\end{aligned} \tag{19.18}$$

The kernel $K(\mathbf{r}, \mathbf{r}')$ is pervasive in the response function formalism and is discussed further in Ch. 8 and App. D (where K is given in $\mathbf{k}$ space in Eq. (D.10)). It incorporates the effects

[2] The formulas in terms of density response functions do not apply directly to non-local pseudopotentials, but appropriate modifications can be made [729].

of Coulomb interaction and exchange and correlation to linear order through $f_{\rm xc}(\mathbf{r},\mathbf{r}') = \mathrm{d}V_{\rm xc}(\mathbf{r})/\mathrm{d}n(\mathbf{r}')$.

The standard perturbation theory approach is to expand (19.17) in eigenfunctions of the zero-order Schrödinger equation, which leads to expression (19.15). Although this approach is not efficient because it requires knowledge of the full spectrum of the zero-order hamiltonian and sums over many unoccupied states, nevertheless, the expressions are useful for demonstration of important properties. In particular, (19.15) shows that $\Delta n(\mathbf{r})$ involves only mixing of the unoccupied ($j > N$) space into the space of the occupied ($i \leq N$) states because contributions from the occupied states cancel in pairs in the sum.

It is much more effective for actual calculations to view the set of equations, Eqs. (19.16)–(19.18), as a self-consistent set of equations for Δn and $\Delta V_{\rm KS}$ to linear order in $\Delta V_{\rm ext}$. It might appear that there is a problem since the left-hand side of (19.17) is singular because the operator has a zero eigenvalue, for which the eigenvector is ψ_i. However, as shown in (19.15), the response of the system to an external perturbation only depends on the component of the perturbation that couples the manifold occupied states with the empty states. The desired correction to the occupied orbitals can be obtained from (19.17) by projecting the right-hand side onto the empty-state manifold,

$$(H_{\rm KS} - \varepsilon_i)|\Delta\psi_i\rangle = -\hat{P}_{\rm empty}\Delta V_{\rm KS}|\psi_i\rangle, \tag{19.19}$$

where the projection operator is given by (see also Eq. (21.16))

$$\hat{P}_{\rm occ} = \sum_{i=1}^{N} |\psi_i\rangle\langle\psi_i|; \quad \hat{P}_{\rm empty} = 1 - \hat{P}_{\rm occ}. \tag{19.20}$$

In practice, if the linear system is solved by the conjugate-gradient or any other iterative method in which orthogonality to the occupied-state manifold is enforced during iteration, there is no problem with the zero eigenvalue.

The basic algorithm for DFPT consists of solving the set of linear equations (19.19) for $\Delta\psi_i$ given the definition in (19.20) and expression (19.18) for $\Delta V_{\rm KS}$ in terms of Δn, which is given by (19.16). Since Δn is a function of the set of occupied $\Delta\psi_i$, this forms a self-consistent set of equations. Any of the efficient iterative methods (App. M) developed for electronic structure problems can be applied to reach the solution by iteration. This is a more efficient approach than the standard "textbook" approach outlined following (19.13): the equivalent of the matrix inverse in (19.14) is accomplished by the self-consistent solution for $\Delta\psi_i$ and $\Delta V_{\rm KS}$, and the sum over excited states is accomplished by mixing of the unoccupied space into the occupied space using (19.19).

19.5 Variational expressions

Variational expressions in perturbation theory have a long history, for example the ingenious use of the variational principle for accurate solution of the two-electron problem by Hylleraas [45] in the 1930s. The ideas have been brought to the fore recently by Gonze [268, 719] (see also [153]) who has derived expressions equivalent to the Green's function formulas of Secs. 19.4 and 19.6. The variational perspective provides an alternative approach for

solution of electronic structure problems, and is one that involves minimization rather than solution of linear equations.

The basic idea of variational expressions in perturbation theory is very simple. If a system has internal degrees of freedom, that are free to adjust (within any constraints that they must obey), then the static response to an external perturbation requires that all internal degrees of freedom adjust to minimize the energy. Exercise 19.5 gives a simple example of a system made up of two harmonic springs with an internal degree of freedom. In electronic structure the perturbation is a change in the external potential $\Delta V_{\rm ext}$, and the internal degrees of freedom are the density $n(\mathbf{r})$ or the wavefunctions ψ_i. These can be viewed as independent variables, i.e. one can define a functional $E^{(m)}[n]$ or $E^{(m)}[\psi_i]$ valid at a chosen order m of perturbation theory, and the correct solution is found by minimizing the functional. Just as for the variational principle that leads to the Schrödinger or Kohn–Sham equations, the energy, ψ_i, and $n(\mathbf{r})$ are determined by minimizing the energy.

The variational principle in perturbation theory can be derived directly from the same variational principle that leads to the Schrödinger or Kohn–Sham equations, but the new point is that it is applied only to a given order. For example, to second order in the changes $\Delta V_{\rm ext} = V_{\rm ext} - V^0_{\rm ext}$ and $\Delta n = n - n^0$, the Kohn–Sham energy functional, Eq. (7.5) or (9.7), can be written in a form similar to (9.8) with the addition of terms involving $\Delta V_{\rm ext}$,[3]

$$\begin{aligned} E^{(2)}[V_{\rm ext}, n] = E[V^0_{\rm ext}, n^0] &+ \int \mathrm{d}\mathbf{r} n^0(\mathrm{d}\mathbf{r}) \Delta V_{\rm ext}(\mathbf{r}) \\ &+ \frac{1}{2}\int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \left[\frac{\delta^2 E}{\delta V_{\rm ext}(\mathbf{r})\Delta n(\mathbf{r}')}\right] \Delta V_{\rm ext}(\mathbf{r})\Delta n(\mathbf{r}') \\ &+ \frac{1}{2}\int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \left[\frac{\delta^2 E}{\delta n(\mathbf{r})\delta n(\mathbf{r}')}\right] \Delta n(\mathbf{r})\Delta n(\mathbf{r}'), \end{aligned} \tag{19.21}$$

where derivatives are evaluated at the minimum energy solution with $V^0_{\rm ext}$ and n^0. There is no term involving $\Delta V^2_{\rm ext}$ since the functional is linear in $V_{\rm ext}$. The first intergral in (19.21) is the force theorem. The middle line is linear in Δn. The last line is quadratic in Δn and is always positive since the functional is minimum at $V^0_{\rm ext}$ and n^0. Minimization of the functional is, in principle, merely a matter of minimizing a quadratic form, no harder than the harmonic springs in Exercise 19.5. However, practical solutions must be done in terms of the wavefunctions ψ_i since the functionals of density are unknown.

In terms of the orbitals, expression (19.21) becomes [153] (omitting terms that are zero for $\Delta V_{\rm ext} = 0$)

$$\begin{aligned} E^{(2)}[V_{\rm ext}, \psi_i] = E[V^0_{\rm ext}, \psi^0_i] &+ \frac{1}{2}\sum_i \int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \left[\frac{\delta^2 E}{\delta V_{\rm ext}(\mathbf{r})\delta\psi_i(\mathbf{r}')}\right] \Delta V_{\rm ext}(\mathbf{r})\Delta\psi_i(\mathbf{r}') \\ &+ \frac{1}{2}\sum_{ij} \int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \left[\frac{\delta^2 E}{\delta\psi_i(\mathbf{r})\delta\psi_j(\mathbf{r}')}\right] \Delta\psi_i(\mathbf{r})\Delta\psi_j(\mathbf{r}'), \end{aligned} \tag{19.22}$$

[3] Spin indices are omitted for simplicity and the subscript "KS" is omitted because the expressions apply more generally.

where

$$\left[\frac{\delta^2 E}{\delta V_{\rm ext}(\mathbf{r})\delta\psi_i(\mathbf{r}')}\right] = \psi_i^0(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}'),$$
$$\left[\frac{\delta^2 E}{\delta\psi_i(\mathbf{r})\delta\psi_j(\mathbf{r}')}\right] = H_{\rm KS}^0(\mathbf{r},\mathbf{r}') + K(\mathbf{r},\mathbf{r}')\Delta\psi_i(\mathbf{r})\Delta\psi_j(\mathbf{r}')\Delta\psi_i(\mathbf{r})\Delta\psi_j(\mathbf{r}'), \tag{19.23}$$

with K defined in (19.18); (see also (D.10)).

The solution can be found directly by minimizing (19.22) with respect to $\Delta\psi_i$ subject to the orthonormality constraint

$$\left\langle \psi_i^0 + \Delta\psi_i \middle| \psi_j^0 + \Delta\psi_j \right\rangle = \delta_{ij}. \tag{19.24}$$

The steepest descent directions for $\Delta\psi_i$ are found by writing out the equations, which also show equivalence to the Green's function method [153]. Minimizing $E^{(2)}[V_{\rm ext}, \psi_i]$ in (19.22) with condition (19.24) leads to

$$H_{\rm KS}^0\Delta\psi_i - \sum_j \Lambda_{ij}\Delta\psi_j = -\left(\Delta V_{\rm eff} - \varepsilon_i\right)\psi_i + \sum_j \Lambda_{ij}\Delta\psi_j, \tag{19.25}$$

where $\Delta V_{\rm eff}$ is the change in total effective potential given to the second order by (19.18). Taking matrix elements with respect to the orbitals, one recovers (Exercise 19.10) the form given in (19.19).

19.6 Periodic perturbations and phonon dispersion curves

The DFPT equations have a marvellous simplification for the case of a crystal with a perturbation at a given wavevector, e.g. a phonon of wavevector $\mathbf{k}_p$, with displacements given by Eq. (19.7). To linear order, the change in density, external potential, and Kohn–Sham potential all have Fourier components only for wavevectors $\mathbf{k}_p + \mathbf{G}$, where $\mathbf{G}$ is any reciprocal lattice vector. The expressions can be written

$$\Delta V_{\rm ext}(\mathbf{r}) = \Delta v_{\rm ext}^{\mathbf{k}_p}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_p\cdot\mathbf{r}} = \sum_{\mathbf{T}} \frac{V_s[\mathbf{r}-\mathbf{R}_s(\mathbf{T})]}{\partial\mathbf{R}_s(\mathbf{T})}\mathrm{e}^{-\mathrm{i}\mathbf{k}_p\cdot(\mathbf{r}-\mathbf{R}_s(\mathbf{T}))}\mathbf{u}_s(\mathbf{k}_p)\mathrm{e}^{\mathrm{i}\mathbf{k}_p\cdot\mathbf{r}},$$
$$\Delta V_{\rm KS}(\mathbf{r}) = \Delta v_{\rm KS}^{\mathbf{k}_p}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_p\cdot\mathbf{r}}, \tag{19.26}$$
$$\Delta n(\mathbf{r}) = \Delta n^{\mathbf{k}_p}(\mathbf{r})e^{\mathrm{i}\mathbf{k}_p\cdot\mathbf{r}}.$$

The wavefunction for an electron at wavevector $\mathbf{k}_e$ is modified to linear order *only by mixing of states with wavevector* $\mathbf{k}_e + \mathbf{k}_p$, so that (19.19) becomes

$$\left(H_{\rm KS}^{\mathbf{k}_e} - \varepsilon_i^{\mathbf{k}_e}\right)\left|\Delta\psi_i^{\mathbf{k}_e+\mathbf{k}_p}\right\rangle = -\left[1 - \hat{P}_{\rm occ}^{\mathbf{k}_e+\mathbf{k}_p}\right]\Delta V_{\rm KS}^{\mathbf{k}_p}\left|\psi_i^{\mathbf{k}_e}\right\rangle. \tag{19.27}$$

To linear order the density is given by

$$\Delta n^{\mathbf{k}_p}(\mathbf{r}) = 2\sum_{\mathbf{k}_e,i} u_{\mathbf{k}_e,i}^*(\mathbf{r})\Delta u_{\mathbf{k}_e+\mathbf{k}_p,i}(\mathbf{r}), \tag{19.28}$$

where u denotes the periodic part of the Bloch function, and the Kohn–Sham potential is

$$\Delta v_{\rm KS}^{\mathbf{k}_p}(\mathbf{r}) = \Delta v_{\rm ext}^{\mathbf{k}_p}(\mathbf{r}) + \int {\rm d}\mathbf{r}' \left[\frac{1}{|\mathbf{r}-\mathbf{r}'|} + f_{\rm xc}(\mathbf{r},\mathbf{r}') \right] \Delta n^{\mathbf{k}_p}(\mathbf{r}). \tag{19.29}$$

The DFPT algorithm for the calculation of the response to any periodic external perturbation $\Delta v_{\rm ext}^{\mathbf{k}_p}(\mathbf{r})$ is the solution of the set of equations (19.27)–(19.29). Note that the calculation involves only *pairs* of wavevectors, $\mathbf{k}_e$ and $\mathbf{k}_e + \mathbf{k}_p$ for the electrons, in the linear equation (19.27), and a sum over all $\mathbf{k}_e$ and filled states i for the self-consistency. The calculations can be done using the same fast Fourier transform (FFT) techniques that are standard in efficient plane wave methods (Ch. 13 and App. M): Eqs. (19.27) and (19.29) can be solved partly in $\mathbf{r}$ space and partly in $\mathbf{k}$ space, and (19.28) is most efficiently done in $\mathbf{r}$ space, with transformations done by FFTs.

Figure 2.9 shows the results of DFPT calculations for the phonon dispersion curves of GaAs [154], done using the local approximation (LDA). Similar results are found for other semiconductors and it is clear that agreement with experiment is nearly perfect. Calculations have been done for many other materials, and an example is the set of results presented in Fig. 19.4 for the phonon dispersion curves of a set of metals. From top to bottom, these represent increasing electron–phonon interactions and increasing complexity of the Fermi surface. The LDA is essentially perfect for Al (as expected!), but the GGA provides a significant improvement in Nb. Another example are the dispersion curves, phonon density of states, and electron–phonon coupling for MgB_2 shown in Fig. 2.32, calculated [155] using the LMTO method [730, 733]. Similar results are found for many materials using many methods, finding agreement with experimental frequencies within $\approx$5% is typical.

19.7 Dielectric response functions, effective charges, . . .

Electric fields present a special problem due to long-range Coulomb interaction. This arises in any property in which electric fields are intrinsically involved, e.g. dielectric functions, effective charges, and piezoelectric constants. One approach is to formulate the theory of response functions at finite wavelengths and take limits analytically [152, 734]. Is it possible to generate an efficient Green's function approach that involves only the infinite wavelength $\mathbf{q} = 0$ limit? The answer is "yes," but only with careful analysis. The problem is that the limit corresponds to a homogeneous electric field with potential $V_{\rm ext}(\mathbf{r}) = \mathbf{E}\cdot\mathbf{r}$, which leads to an ill-defined hamiltonian in an extended system; see Ch. 22. The saving grace is that *perturbation theory in the electric field* involves matrix elements of the form of Eq. (19.15) or (19.19), i.e. only *off-diagonal* matrix elements of the perturbing potential between eigenfunctions of the unperturbed hamiltonian. These matrix elements are well defined even for a macroscopic electric field, which can be seen by re-writing them in terms of the commutator between $\mathbf{r}$ and the unperturbed hamiltonian,

$$\langle\psi_i|\mathbf{r}|\psi_j\rangle = \frac{\langle\psi_i|[H,\mathbf{r}]|\psi_j\rangle}{\epsilon_i - \epsilon_j}, \quad i \neq j. \tag{19.30}$$

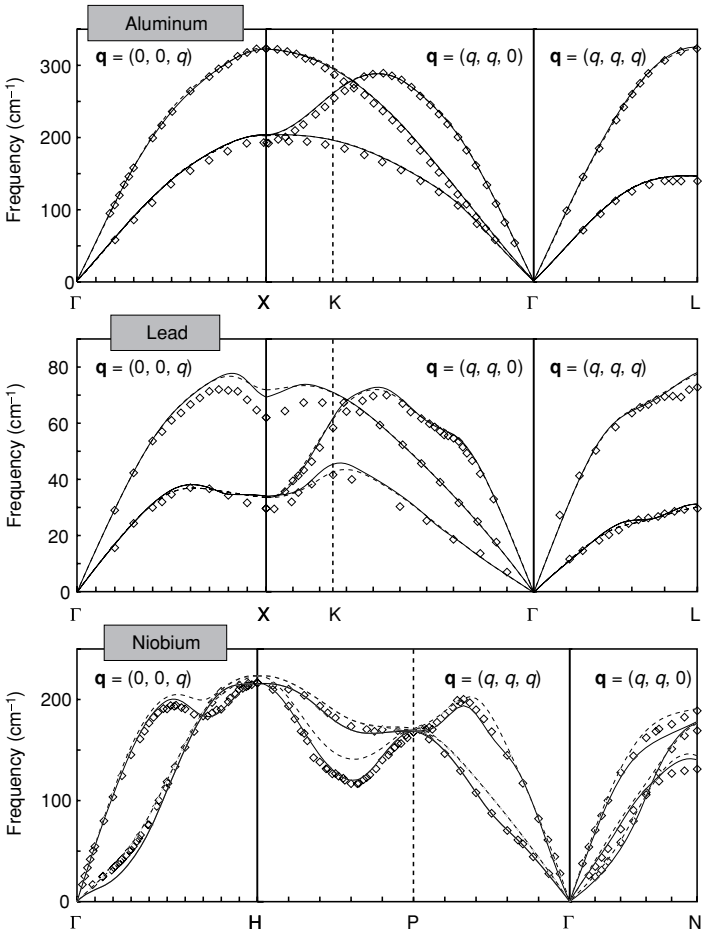


Figure 19.4. Phonon dispersion curves calculated for the metals Al, Pb, and Nb compared with experiment [732]. The agreement using the LDA is excellent for Al and progressively worse for the other metals, which are further from the homogeneous gas, where GGA functionals improve the agreement. The dips in the phonon dispersion curves for Pb and Nb indicate strong electron–phonon interactions and sensitivity to Fermi surface features that are important for superconductivity in these elements.

If the total potential acting on the electrons is local, the commutator is simply proportional to the momentum operator,

$$[H, \mathbf{r}] = -\frac{\hbar^2}{m_e}\frac{\partial}{\partial \mathbf{r}} = i\frac{\hbar}{m_e}\mathbf{p}. \tag{19.31}$$

For non-local potentials, the commutator involves an explicit contribution from the potential [514, 515] as defined in (11.70). Littlewood [735] has used the momentum form for the matrix elements to derive expressions for the dielectric functions of crystals in terms of the

periodic Bloch functions, and the explicit iterative Green's function algorithm corresponding to Eqs. (19.16)–(19.20) is given in [153].

An example of the application of the ideas is the calculation of effective charges Z^* in ionic insulators. Linear response calculations have been done for many materials, e.g. the ferroelectric perovskite materials with formula unit ABO_3. Anomalously large effective charges of the B atoms and the O atoms moving along the line joining them has been found and interpreted as resulting from covalency [736, 737]. Results are essentially the same as those given in Tab. 22.1 from [573] using the Berry's phase method.

19.8 Electron–phonon interactions and superconductivity

Electron–phonon coupling plays a crucial role in the theory of transport and superconductivity in solids, and there are excellent sources showing the relation of the microscopic interactions to the phenomena [199, 242, 243]. In particular, the basic quantity in the Eliashberg [738] equations for phonon-mediated superconductivity is $\alpha^2 F(\omega)$, where $F(\omega)$ is a phonon density of states and α denotes an average over all phonons of energy ω. An example is shown in Fig. 2.32.

The quantities that can be derived from the underlying electronic structure are the electron bands and density of states, the phonon dispersion curves and density of states, and electron–phonon coupling (App. C). The matrix element for scattering an electron from state $i\mathbf{k}$ to $j\mathbf{k}+\mathbf{q}$ while emitting or absorbing a phonon $\nu\mathbf{q}$ with frequency ω is given by [243]

$$g_{i\mathbf{k};j\mathbf{k}+\mathbf{q}}(\nu) = \frac{1}{\sqrt{2M\omega_{\nu\mathbf{q}}}}\langle i\mathbf{k}|\Delta V_{\nu\mathbf{q}}|j\mathbf{k}+\mathbf{q}\rangle, \tag{19.32}$$

where M is the (mode-dependent) reduced mass and $1/\sqrt{(2M\omega_{\nu\mathbf{q}})}$ is the zero-point phonon amplitude.

For scattering at the Fermi surface, one can define the dimensionless coupling to the phonon branch ν, where $\nu = 1, \ldots, 3S$, with S the number of atoms per cell, by [243]

$$\lambda_\nu = \frac{2}{N(0)}\sum_{\mathbf{q}}\frac{1}{\omega_{\nu\mathbf{q}}}\sum_{ij\mathbf{k}}\left|g_{i\mathbf{k};j\mathbf{k}+\mathbf{q}}(\nu)\right|^2\delta(\varepsilon_{i\mathbf{k}})\delta(\varepsilon_{j\mathbf{k}+\mathbf{q}}-\varepsilon_{i\mathbf{k}}-\omega_{\nu\mathbf{q}}), \tag{19.33}$$

where $N(0)$ is the electronic density of states per spin at the Fermi energy and $\omega_{\nu\mathbf{q}}$ is the energy of phonon ν with the wavevector $\mathbf{q}$. The energy of electron band i with the wavevector $\mathbf{k}$ is $\varepsilon_{i\mathbf{k}}$ and $g_{i\mathbf{k};j\mathbf{k}+\mathbf{q}}(\nu)$ is the matrix element between the states $i\mathbf{k}$ and $j\mathbf{k}+\mathbf{q}$ due to the induced potential when phonon $\nu\mathbf{q}$ is excited. The delta functions restrict the electron scattering to the Fermi surface.

A key quantity in the theory is the induced potential ΔV per unit displacement needed in the matrix element (19.32). To linear order it is given by

$$\Delta V_{\nu\mathbf{q}} = \frac{\partial V}{\partial u_{\nu\mathbf{q}}} = \sum_{I\alpha} X_{\nu\mathbf{q}}(I\alpha)\frac{\partial V}{\partial u_{I\alpha}}, \tag{19.34}$$

where $u_{\nu\mathbf{q}}$ is the phonon normal coordinate and $X_{\nu\mathbf{q}}(I\alpha)$ is the eigenvector of the dynamical matrix, Eq. (19.8), expressed in terms of displacements $u_{I\alpha}$ of the nucleus I in the α direction. Since the phonon is low frequency, the potential $\Delta V_{\nu\mathbf{q}}$ is a "screened potential," i.e. the electrons react to contribute to the effective potential.

There are four general methods for calculation of the matrix elements with the properly screened $\Delta V_{\nu\mathbf{q}}$:

- Displacement of rigid atomic-like potentials, e.g. muffin-tin potentials [739]. The justification is that the coupling is primarily local [740] and the potential is very much like a sum of atomic potentials, which has been shown to be appropriate for transition metals, even with displaced atoms. The same ideas apply for displacement of any effective total potential, such as empirical pseudopotentials.
- Calculations using a general expression for the screening which can be conveniently expressed in Fourier space as

$$\frac{\partial V(\mathbf{q}+\mathbf{G})}{\partial u_{\nu\mathbf{q}}} = \sum_{\mathbf{G}'} \epsilon^{-1}(\mathbf{q}+\mathbf{G}, \mathbf{q}+\mathbf{G}') \frac{\partial V_{\text{ion}}(\mathbf{q}+\mathbf{G}')}{\partial u_{\nu\mathbf{q}}}, \tag{19.35}$$

  following the definition of ϵ^{-1}, given, e.g., above Eq. (19.14). This is particularly convenient when one can use a simple model for ϵ^{-1}, e.g., the Lindhard function, Eq. (5.38), which is appropriate for simple metals. In general, however, it more efficient to use the Green's function approach.
- Methods using direct calculation of the linear order screened potential [153, 733]. This can be done conveniently using the Green's function technique described in Secs. 19.4 and 19.6. As is clear from Eq. (19.35), one of the by-products of the calculation of phonons is the screened $\Delta V_{\nu\mathbf{q}}$ itself. This needs to be done for all relevant phonons needed for the integral over the Fermi surface.
- Self-consistent "frozen phonon" calculations [685, 741]. As described in Sec. 19.2, the calculations involve V directly to all orders in the displacement. The linear change in potential can be extracted as the linear fit to calculations with finite displacements of the atoms. Alternatively, one can find the mixing of wavefunctions from which the matrix elements can be found. The "frozen phonons" can be determined on a grid of $\mathbf{k}$ points and interpolated to points on the Fermi surface.

The full calculation requires a double sum over the Brillouin zone and determination of the matrix elements g at each pair of points and for each phonon ν.

19.9 Magnons and spin response functions

Response functions for spin excitations are the fundamental quantities measured in spin-dependent neutron scattering from solids just as lattice dynamical response functions are the fundamental quantities measured in lattice dynamics experiments. Since spin dynamics is strongly affected by magnetic order and excitations tend to be damped by coupling to

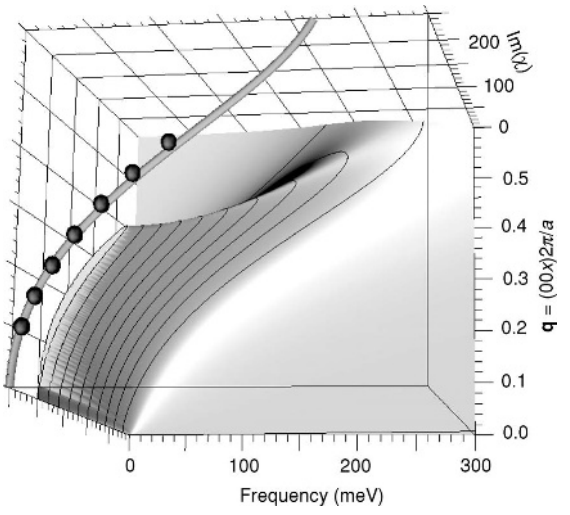


Figure 19.5. Spin response spectral functions $\mathrm{Im}\chi(\mathbf{q}, \omega)$ for Fe. The line shows the dispersion corresponding to the maximum in $\mathrm{Im}\chi$ compared with experimental values (points). Compare also with Fig. 19.3. Note that the excitations broaden at higher frequency and momentum. From [720].

the electron degrees of freedom, the form of the spectral function $\mathrm{Im}\chi(\mathbf{q}, \omega)$ is often of importance. For many years the basic formalism has been known and calculations have been done with standard Green's function techniques and realistic band structures of metals, e.g. in the work of Cooke [742] and many more recent calculations. The method is based on a simple extension of the random-phase approximation of the relevant Green's function equation and an interpolation formalism for treating wave functions and matrix elements.

Recently a new Green's function method has been developed that generalizes the approach used for phonons to treat spin response functions [720]. The formulas are closely related and involve spin density instead of number density. An example of results for Fe using the LMTO approach is shown in Fig. 19.5. Comparison with Fig. 19.3 shows good agreement with calculations done using Berry's phase method and the plane wave pseudopotential method. In addition, however, the results in Fig. 19.5 shown the full spectral function, with its broadening increasing rapidly for wavevectors near the zone boundary, in qualitative agreement with experiment. Such work has been extended to alloys using the KKR–CPA method and for the treatment of disorder within linear response theory [236].

SELECT FURTHER READING

Baroni, S., de Gironcoli, S., Dal Corso, A. and Giannozzi, P., "Phonons and related crystal properties from density-functional perturbation theory," *Rev. Mod. Phys.* 73:515–562, 2000.

Born, M. and Huang, K., *Dynamical Theory of Crystal Lattices*, Oxford University Press, Oxford, 1954.

Gonze, X., "Adiabatic density-functional perturbation theory," *Phys. Rev. A* 52:1096–1114, 1995.

Gonze, X. and Vigneron, J. P., "Density functional approach to non-linear response coefficients in solids," *Phys. Rev. B* 39:13120, 1989.

Pick, R., Cohen, M. H. and Martin, R. M., "Microscopic theory of force constants in the adiabatic approximation," *Phys. Rev. B* 1:910–920, 1970.

Savrasov, S. Y., "Linear response calculations of spin fluctuations," *Phys. Rev. Lett.* 81:2570–2573, 1998.

Exercises

19.1 See many excellent problems (and solutions) on phonons in the book by Mihaly and Martin [248].

19.2 Show that the Bloch theorem for phonons follows from exactly the same logic as for electrons treated in the local orbital representation (Exercise 14.2).

19.3 By comparing expressions (14.7) and (19.8), show that the equations for phonons are exactly the same as a tight-binding formulation in which there are three states of p symmetry for each atom, corresponding to the three degrees of freedom for the atomic displacements. Show explicitly the correspondence of the terms of the two problems, especially the fact that the eigenvalue in the electron problem corresponds to the *square* of the frequency in the phonon problem.

19.4 This exercise is to explain a salient difference in the tight-binding electron and the phonon problems. The force constant has the property of *translational invariance*; show that the fact that the total energy does not change if all the atoms are displaced uniformly leads to the relation $\sum_J C_{I,\alpha;J,\beta} = 0$. This condition can be used to fix the self term $C_{I,\alpha;I,\beta}$ so that, unlike the electron tight-binding problem, the on-site term is not an independent variable. In addition, show that this means that there are three zero frequency phonon modes at $\mathbf{k} = 0$.

19.5 The simplest model for phonons is the central force model in which the energy is a function only of the distance between the nearest-neighbors. Find expressions for a force constant $C_{I,\alpha;J,\beta}$ using the definition as a second derivative of the energy expressed as a sum of pair terms $E = \sum_{I<J} E_{IJ}(|\mathbf{R}_I - \mathbf{R}_J|)$. Show that the resulting expressions are equivalent to a tight-binding problem of electron p states, Exercise 14.10, with the $t_{pp\pi}$ matrix elements equal to zero.

19.6 Find expressions for phonon dispersion curves respectively in elemental simple cubic and fcc crystals using the simplest model for phonons, a nearest-neighbor central potential model with energy given by $E = \frac{1}{2}\sum_{I<J} K|\mathbf{R}_I - \mathbf{R}_J|^2$, where J is restricted to nearest neighbor. Show the relation to tight-binding equations for p bands in Exercise 14.10 as explicit examples of the relationship given in Exercises 19.3–19.5.

(a) Show that there are two dispersion curves that have zero frequency for all k in simple cubic but not in fcc crystals. Explain why this instability occurs in a simple cubic crystal in a central potential model.

(b) There is a corresponding result in the tight-binding model for p bands in a simple cubic crystal if the only non-zero matrix element is $t_{pp\sigma}$ between nearest neighbors. Show that in this case there are two bands with no dispersion.

19.7 Project: Using the properties established in Exercises 19.3–19.5, construct a computer program to evaluate phonon frequencies in a model analogous to a tight-binding model for electrons. The program described in App. N can be used, including only p symmetry states. The input information must include the masses and a model, which could be the central force model described in Exercise 19.6.

19.8 Why is there no term involving $\partial^2 n/(\partial\lambda_i \partial\lambda_j)$ in the expression for $\partial^2 E/(\partial\lambda_i \partial\lambda_j)$ in Eq. (19.12)?

19.9 Derive Eq. (19.15), which is the same as from the basic perturbation expression (D.1) and (19.16). In particular, show that all contributions involving i and j, both occupied, vanish in the expression for the change in the density.

19.10 Show that Eq. (19.19) follows from (19.25) by taking matrix elements of the equation.

19.11 This is an example of the variational principle in perturbation theory. Consider a system composed of three points x_0, x_1, x_2 in a line connected by two springs. The energy is $E = \frac{1}{2}k_1(x_1 - x_0)^2 + \frac{1}{2}k_2(x_2 - x_1)^2$. Suppose forces $f = f_0 = -f_2$ are applied to the two ends.
(a) Identify the functional $F[f, x_1]$ valid for all f and x_1.
(b) If the middle position x_1 is free to move, calculate the change in position $x_1 - x_0$ and total length $x_2 - x_0$ as a function of f.
(c) See Exercise 3.25 for extension to non-linear springs.

20

Excitation spectra and optical properties

Summary

Excitation spectra reveal the properties of matter in terms of the response to time- or frequency-dependent perturbations. Particularly important examples are the dielectric function $\epsilon(\omega)$ and the inverse function $\epsilon^{-1}(\omega)$ defined in App. E. The basic formulas relating the response to the electronic structure are rooted in perturbation theory and response functions (Sec. 3.7 and App. D). This chapter is devoted to dynamic response functions for electrons in self-consistent field methods, such as the Kohn–Sham approach, and to the alternative approach of solving directly the time-dependent Kohn–Sham equation to find the solution to all orders. The formal structure is based upon time-dependent density functional theory (TDDFT) (Sec. 7.6), which provides an exact framework in principle. In practice, simple approximations are remarkably successful in many cases and there is active research to develop new functionals.

As emphasized in the overview, Sec. 2.10, two types of excitations are of primary importance for electronic structure: excitations in which an electron is added or subtracted from the system, and excitations in which the number of electrons remains fixed. The former are of great interest as the "one-electron excitations" in an interacting many-body system; however, in independent-particle theories, such as the Hartree–Fock or Kohn–Sham approaches, these excitations are just the eigenstates of the independent-particle hamiltonian. In a crystal, the eigenvalues form well-defined bands $\varepsilon_{i\mathbf{k}}$ with none of the renormalization and broadening that can only be included in a many-body theory. Many examples of bands are given in other chapters and will not be covered further here.

Even within independent-particle theories, however, a new set of concepts emerges for excitations in which the number of electrons does not change, e.g. optical excitations. The excitation of the system to linear order in the perturbation is described by a dynamic linear response function, as in App. D. This chapter is concerned with three important aspects of the formulation: the general form of linear response functions, expanding upon the expressions in App. D; the particular case of the dielectric function; and applications of time-dependent density functional theory (TDDFT) in Ch. 7. The last provides an in-principle way to treat the response of a many-body system *exactly*; in practice, it has been shown to be very useful in actual calculations on clusters.

20.1 Dielectric response for non-interacting particles

It is useful to first give the form of the dielectric function in the independent-particle approximation. Consider the macroscopic long-wavelength limit treated in the vector gauge. The perturbation can be written in terms of the vector potential $\mathbf{A}$ as in Eq. (E.19)

$$\Delta\hat{H}(t) = \frac{1}{2m_e}\sum_i \left\{\left[\hat{\mathbf{p}}_i - \frac{e}{c}\mathbf{A}(t)\right]^2 - \hat{\mathbf{p}}_i^2\right\}, \tag{20.1}$$

where $\mathbf{E}(t) = -\frac{1}{c}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}t}$ and $\mathbf{E}(\omega) = -\frac{i\omega}{c}\mathbf{A}(\omega)$. The desired response is the macroscopic average current density $\mathbf{j} = -e\langle\mathbf{v}\rangle$, and since $\mathbf{p} = m\mathbf{v} - \frac{e}{c}\mathbf{A}$, it follows that $\mathbf{j} = \frac{-e}{m}\langle\mathbf{p} + \frac{e}{c}\mathbf{A}\rangle$, and the relation of $\mathbf{j}(\omega)$ to $\mathbf{E}(\omega)$ determines the conductivity $\sigma(\omega)$. Using (D.16) and (E.11), it follows that [88]

$$\epsilon_{\alpha\beta}(\omega) = \delta_{\alpha\beta} - \frac{e^2}{m_e\Omega}\frac{1}{\omega^2}\sum_i\left[f_i\delta_{\alpha\beta} + \sum_j \frac{f_i - f_j}{\hbar m_e}\frac{\langle\psi_i|p_\alpha|\psi_j\rangle\langle\psi_j|p_\beta|\psi_i\rangle}{\varepsilon_i - \varepsilon_j + \omega + i\eta}\right], \tag{20.2}$$

where the f_i are occupation numbers and $\eta > 0$ is a small damping factor. (The first term comes from the contribution of $\mathbf{A}$ to the current operator (Exercise 20.1).)

This expression shows the basic reason that measurements of optical spectra are one of the most powerful tools for studies of electronic properties of crystals [470, 531]. Since the $\mathbf{p}$ matrix elements do not vary rapidly as a function of energy for transitions between each pair of bands of electronic states, the imaginary part of $\epsilon(\omega)$ or the real part of $\sigma(\omega)$ directly reveals singularities in the density of states for optical transitions. In the non-interacting approximation, this is a joint density of states for transitions between pairs of filled and empty bands weighted by the matrix elements. Examples of $\epsilon(\omega)$ calculated in the independent-particle approximation are given in Figs. 2.27 and 2.28; see [470] and [531] for many other examples.

The dielectric function can also be calculated by considering scalar potentials at finite wavelength, and taking the long-wavelength limit. For a finite system, such as a cluster, with size much less than the wavelength of light, this is a convenient approach. In a finite system, the external perturbation[1] can be written in terms of the applied electric field $\mathbf{E}_{\mathrm{ext}}(t)$ acting upon an electron at point $\mathbf{r}$,

$$\Delta\hat{H}(t) = \Delta V_{\mathrm{ext}}(\mathbf{r}, t) = -e\mathbf{E}_{\mathrm{ext}}(t)\cdot\mathbf{r}, \tag{20.3}$$

instead of (20.1). The response is the dipole moment given by (see Eq. (22.3))

$$\Delta\mathbf{d}(t) = \frac{1}{\Omega}\int_{\text{all space}} \mathrm{d}\mathbf{r}\ \Delta n(\mathbf{r}, t)\ \mathbf{r}. \tag{20.4}$$

For a driving field of frequency ω the expression for linear response corresponding to (20.2) involves matrix elements of the position operator $\mathbf{r}$. An easy way to derive the resulting expression is to use the transformation given in Eqs. (19.30) and (19.31) to convert to matrix elements of $\mathbf{r}$.

[1] Here the electron charge $-e$ is explicitly included to avoid confusion, since the standard definition of $\mathbf{E}$ is the field acting on a positive charge.

20.2 Time-dependent density functional theory and linear response

Dielectric properties are affected by interactions among the electrons, leading to fundamental changes from expression (20.2), which neglects interactions. The fact that the electron (in the otherwise empty band) and hole (in the otherwise filled band) attract one another leads to bound states in the gap and changes the entire spectrum. Examples are also shown in Figs. 2.27 and 2.28, where the latter case illustrates that the spectrum is dominated by electron–hole interaction in a wide-band-gap insulator like CaF_2. There are two basic approaches for inclusion of interactions: many-body perturbation theory and self-consistent field methods. Both are active fields of research for calculation of optical response in real materials; however, the former is beyond the scope of this volume. The latter is in the realm of independent-particle methods, including well-known methods like the "random phase approximation" (RPA) and Hartree–Fock. Here we emphasize time-dependent Kohn–Sham theory because it holds the promise of exact results[2] – even though at present actual calculations are based upon approximations, often rather severe.

The reader may ask: "Was the fundamental starting point of Kohn–Sham theory, the mapping of the many-body system to an independent-particle problem?" We seem to be coming full circle and claiming more. Is there a contradiction? The answer is that *the Kohn–Sham approach is a theory of independent particles, but an interacting density.* The evolution of the interacting density is cast in terms of the evolution of electrons that obey independent-particle Schrödinger-like equations, with a time-dependent effective potential.

The general form of response functions in self-consistent field theories is given in App. D in terms of the non-interacting response functions χ^0 and the interaction kernel K. In particular, the dynamical density response function is given by (D.20), repeated here,

$$\chi(\omega) = \chi^0(\omega)[1 - \chi^0(\omega)K(\omega)]^{-1}, \tag{20.5}$$

where K is the Fourier transform of the space- and time-dependent kernel given in (D.19). To put flesh on these bare bones, we can give useful explicit expressions following [231]. Expanding the expressions in terms of the time-independent Kohn–Sham orbitals, the needed expressions can be written using

$$\delta n^\sigma(\mathbf{r}, \omega) = \sum_{ij} \psi_i^\sigma(\mathbf{r}) \rho_{ij}^\sigma \psi_j^\sigma(\mathbf{r}), \tag{20.6}$$

and matrix elements of the effective potential $\delta[V_{\rm eff}]_{ij}^\sigma \equiv \langle \psi_i^\sigma | \delta V_{\rm eff}(\mathbf{r}, \omega) | \psi_j^\sigma \rangle$. Thus the non-interacting χ^0 becomes [231]

$$\chi^0_{ij\sigma,i'j'\sigma'} = \frac{\delta\rho_{ij}^\sigma}{\delta[V_{\rm eff}]_{i'j'}^{\sigma'}} = \delta_{ii'}\delta_{jj'}\delta_{\sigma\sigma'} \frac{f_i^\sigma - f_j^\sigma}{\omega - \left(\varepsilon_i^\sigma - \varepsilon_j^\sigma\right)}. \tag{20.7}$$

The interacting response function can be derived from the relation $\delta V_{\rm eff} = \delta V_{\rm ext} + \delta V_H + \delta V_{\rm xc} \equiv \delta V_{\rm ext} + \delta V_{H\rm xc}$ which to linear order is given by $\delta V_{\rm eff} = \delta V_{\rm ext} + K\delta n$. The full

[2] In Sec. 7.6 are given qualifications that an exact theory must involve current functionals and long-range functionals.

expression can be written in the form given in Sec. D.2,

$$\chi_{ij\sigma,i'j'\sigma'} = \frac{\delta\rho^{\sigma}_{ij}}{\delta[V_{\rm ext}]^{\sigma'}_{i'j'}} = \chi^{0}_{ij\sigma,i'j'\sigma'}\left[\delta_{ii'}\delta_{jj'}\delta_{\sigma\sigma'} + \sum_{i''j''\sigma''} K_{ij\sigma,i''j''\sigma''}\ \chi_{i''j''\sigma'',i'j'\sigma'}\right], \quad (20.8)$$

where K is the array of matrix elements of the interaction terms

$$\begin{aligned} K_{ij\sigma,i'j'\sigma'} &= \frac{\delta[V_{H\rm xc}]^{\sigma}_{ij}}{\delta\rho^{\sigma'}_{i'j'}} \quad (20.9)\\ &= \int {\rm d}\mathbf{r} \int {\rm d}\mathbf{r}' \psi^{\sigma\,*}_{i}(\mathbf{r})\psi^{\sigma}_{j}(\mathbf{r})\left[\frac{\delta_{\sigma\sigma'}}{|\mathbf{r}-\mathbf{r}'|} + f^{\sigma\sigma'}_{\rm xc}(\mathbf{r},\mathbf{r}',\omega)\right]\psi^{\sigma'\,*}_{i'}(\mathbf{r}')\psi^{\sigma'}_{j'}(\mathbf{r}'), \end{aligned}$$

where $f_{\rm xc}$ is the second derive of $E_{\rm xc}[n]$ given explicitly in (D.18) and (D.19).

With some algebra (Exercise 20.2), the solution of the equation can be cast in the form of an eigenvalue equation [231]

$$\left[\omega^2_{ij\sigma}\delta_{ii'}\delta_{jj'}\delta_{\sigma\sigma'} + 2\sqrt{f_{ij\sigma}\omega_{ij\sigma}}K_{ij\sigma,i'j'\sigma'}\sqrt{f_{i'j'\sigma'}\omega_{i'j'\sigma'}}\ \right]F_n = \Omega^2_n F_n, \quad (20.10)$$

where $f_{ij\sigma} = f^{\sigma}_i - f^{\sigma}_j$ and $\omega_{ij\sigma} = \varepsilon^{\sigma}_i - \varepsilon^{\sigma}_j$. The TDDFT problem thus becomes a matrix problem with the basis of *pairs of Kohn–Sham states* $ij\sigma$. If there are $N_{\rm occ}$ filled orbitals and one includes $N_{\rm empty}$ empty orbitals of each spin, then the size of the matrices is $N_{\rm pair} \times N_{\rm pair}$, where $N_{\rm pair} = 2N_{\rm occ} \times N_{\rm empty}$. Thus this approach is appropriate for molecules and small clusters where the matrices are of manageable size.

Many calculations have been done using the adiabatic LDA and GGA approximations [231, 234] which improve agreement with experiment compared to the non-interacting approximation, although the agreement is not as good as for static structural properties. These methods are now routinely used for applications to molecules and clusters, and the approach can also be applied to crystals [235] using the periodicity.[3]

Metal clusters are an interesting class of systems that can be varied from atomic size to the bulk [189]. The simplest approximation is spherical jellium ignoring the atomic structure, which leads to the correct general features of the optical spectra dominated by a plasmon-like peak.[4] The question for quantitative calculations is the extent to which the real atomic structure matters. An example of one of the first quantitative calculations on metal clusters, reproduced in Fig. 20.1, shows a two-peak structure very different from the jellium model, with the main peak shifted and in better agreement with experiment. The shift in the main peak from the non-interacting particle response, Eq. (20.2), shows the effects of the Coulomb interaction in this confined geometry. Many such calculations have been reported (e.g. [234]), including quantum molecular dynamics (QMD) simulations [743] that account for thermal broadening due to dynamical motion of the nuclei.

Semiconductor clusters or "quantum dots" terminated by H or other elements are of great interest because of their enhanced optical properties (see, e.g. [745]). Figure 20.2 shows

[3] Note that there are formal problems in infinite systems, as outlined in Sec. 7.6.

[4] In a bulk solid, the plasma peak [84] is due to a longitudinal density response dominated by a peak in the inverse function Im $\epsilon^{-1}(\omega)$ for $\omega \approx \omega_p$; however, in a small confined system, the distinction between longitudinal and transverse is lost and there is a peak in the absorption of light at $\omega \approx \omega_p$.

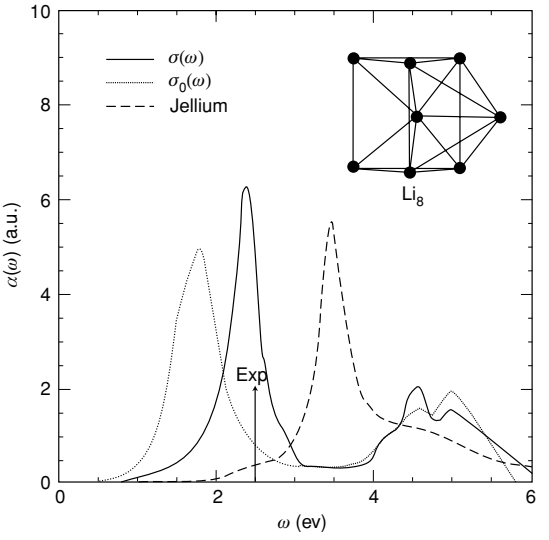


Figure 20.1. Example of optical spectrum of a metal cluster (Li_8) calculated [744] by time-dependent density functional theory (TDDFT) using the adiabatic local density approximation (LDA). The inset shows the structure of the cluster determined from the usual ground state density functional theory. The resulting spectrum is significantly changed from the spherical jellium model, in better agreement with experiment (arrow), and from the non-interacting approximation (dotted line labeled σ_0). From [744].

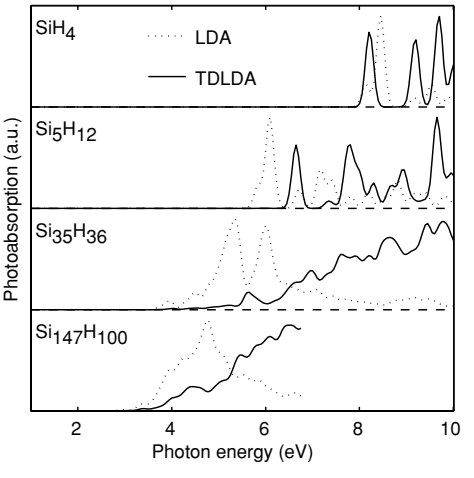


Figure 20.2. Optical spectra of selected hydrogen-terminated Si systems, from the molecule SiH_4 to the $Si_{147}H_{100}$ cluster. The solid lines are spectra from TDLDA calculations, which are compared to uncorrected independent-particle spectra (dotted lines) that follows from Eq. (20.2) applied to a finite system. For SiH_4, TDLDA results in a small shift to lower energy; whereas for the larger clusters, the intensity of the optical absorption shifts to higher energy due to Coulomb interactions and finite size effects (see text). Provided by I. Vasiliev.

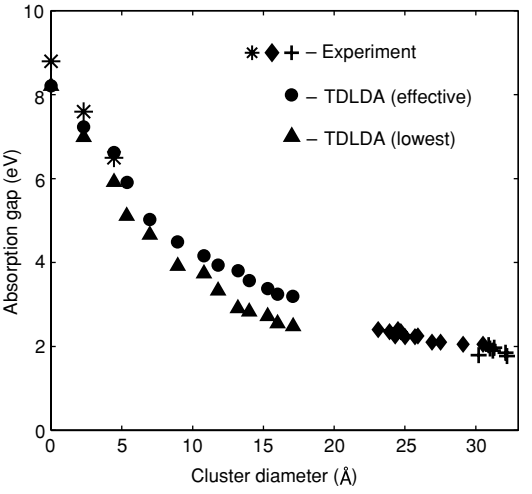


Figure 20.3. Optical properties of Si nanoclusters predicted by time-dependent LDA (TDLDA) compared to experiment. The graph shows results for the lowest gap and for a higher energy, designating a threshold in the optical strength that corresponds to experimental assignments of an effective gap [359]. Provided by I. Vasiliev.

TDLDA spectra for selected structures compared to independent-particle spectra. For the smallest system (the SiH_4 molecule), TDDFT results in a small shift to lower energy due to the fact that the electron and hole are confined to a small volume reducing the energy due to Coulomb attraction. For large clusters, still much smaller than the wavelength of light, the spectra shift to higher energy and have a plasma-like resonance that couples to light because of finite size effects.

The lowest energy excitations and an effective gap corresponding to the onset of strong absorption are shown in Fig. 20.3 for H-terminated Si clusters as a function of diameter. Experimental results are shown for molecules and for large clusters. The gap increases with decreasing cluster size due to quantum confinement effects, and it is evident that the theory explains the trends in general agreement with experiment. For example, the gap of Si nanoclusters is increased above the bulk gap, so that Si clusters can become efficient emitters of blue light [746].

20.3 Variational Green's function methods for dynamical linear response

Dynamical linear-response functions and can also be formulated using techniques closely related to the iterative or variational Green's function methods of Ch. 19 [747]. The approach involves an algorithm like that in Eqs. (19.16)–(19.20) for the static case. In the dynamic case, however, the left-hand side of Eq. (19.17) involves the frequency ω. The method can be applied, at least for frequencies in the absorption gap where there are no resonances, in the same manner as the static algorithm. In addition, the approach can be extended [748] to high order by applying the "$2n + 1$" theorem to the action functional as defined in time-dependent

density functional theory. A first application was for the case of two-photon transitions in the hydrogen atom [747], followed by work on linear and non-linear susceptibilities of a range of semiconductors [748]. However, there has not been widespread use thus far.

20.4 Explicit real-time calculations

The evolution of each one-particle state $\psi_i(t)$ is given by a Schrödinger-like equation, (7.22), repeated here,

$$i\hbar \frac{\mathrm{d}\psi_i(t)}{\mathrm{d}t} = \hat{H}(t)\psi_i(t), \tag{20.11}$$

with an effective hamiltonian that depends upon time t given in (7.23). As pointed out following that equation, there are difficult issues in creating a fully satisfactory functional since it must depend upon the density *at previous times*. Nevertheless, much work has been done with the simplest possible adiabatic approximation in which $V_{\mathrm{xc}}(\mathbf{r}, t)$ is taken to be the usual approximation in terms of the density *at time* t, i.e. neglecting any memory effects. The same approximation is made in the linear response formulas if f_{xc} is assumed to be frequency independent.

There are two important differences from the linear response approach. The evolution is not limited to small perturbations and can be used for non-linear effects, including extreme conditions created by laser pulses [749, 750]. In addition, there is an advantage to using the real-time approach for large systems. Since only the occupied states are evolved, the calculations can be made to scale linearly with the size of the system. In contrast, the linear response matrix formalism of Sec. 20.2 involves diagonalization of matrices of size $N_{\mathrm{pair}} \times N_{\mathrm{pair}}$.

One approach is to calculate the temporal propagation by iteration of the time-dependent Schrödinger equation in steps in real time. This can be done in many ways with the choice of algorithm governed by the fact that the propagation is unitary, which is essential for particle number conservation. Many of the basic ideas originated in nuclear physics [751] and an excellent exposition can be found in the text by Koonin and Meredith [444]. If the functions $\psi_i(t)$ are expanded in a fixed time-independent basis

$$\psi_i(t) = \sum_{\alpha} c_{i,\alpha}(t)\chi_\alpha, \tag{20.12}$$

then the iteration from $c^n_{i,\alpha}$ at time t^n to $c^{n+1}_{i,\alpha'}$ at time $t^{n+1} = t^n + \delta t$ is given by

$$c^{n+1}_{i,\alpha} = \sum_{\alpha'} [\mathrm{e}^{-\mathrm{i}\hat{H}\delta t}]_{\alpha,\alpha'} c^n_{i,\alpha'}, \tag{20.13}$$

where $\hat{H}$ is a matrix in the basis α, α'. Because the complex exponential operator has unity modulus, Eq. (20.13) is a unitary transformation. The size of the time step δt is limited by condition that $\hat{H}(t)$ can be considered constant over the interval δt; nevertheless, one can still choose the time at which $\hat{H}(t)$ is evaluated. Since $\hat{H}$ must be updated as a function of the time-dependent density, this can have important consequences for efficiency. Two examples are the predictor–corrector method [233] and the "railway curve interpolation" [752].

There are four types of approaches in actual calculations.

Explicit operations with the exponential

In general it is not possible to perform the exponentiation of an operator exactly, and one must bring the operators to diagonal form. In that case, $\exp(A) = \sum_i |\zeta_i\rangle \exp(A_{ii}) \langle\zeta_i|$, $A_{ii} = \langle\zeta_i|A|\zeta_i\rangle$, with ζ_i an eigenvector of A, $A_{ii} = \langle\zeta_i|A|\zeta_i\rangle$, and $\exp(A_{ii})$ the ordinary exponential of a scalar (Exercise 20.3). This can be done separately for the potential (V) and kinetic (T) operators, which are diagonal, respectively, in real and reciprocal space. However, the hamiltonian involves both operators, which cannot in general be separated since they do not commute. Nevertheless, one can use a Suzuki–Trotter expansion, of which a simple example is

$$\exp[-\mathrm{i}(T+V)\delta t] \simeq \exp\left(-\mathrm{i}\frac{1}{2}V\Delta t\right)\exp(-\mathrm{i}T\Delta t)\exp\left(-\mathrm{i}\frac{1}{2}V\Delta t\right), \quad (20.14)$$

plus corrections $\propto \delta t^2$ (Exercise 20.4). This approach is well suited for plane-wave or real-space methods, where efficient fast Fourier transform algorithms provide an *exact* transformation between *finite* plane wave expansions and real-space grids. The transformations are exactly the same as used in ground state plane codes and described in Sec. M.11, especially Fig. M.2.

Expansion of the exponential

The simplest approach is to expand the exponential in (20.13) in powers of the hamiltonian, which leads to

$$c^{n+1} = \left[1 - i\hat{H}\delta t - \frac{1}{2}\hat{H}^2\delta t^2 + \cdots\right] c^n. \quad (20.15)$$

The expansion can easily be carried to high orders and the calculation done by iterative application of $\hat{H}$. Just as for other iterative methods (App. M), the operations can be done efficiently so long as the hamiltonian is sparse, e.g. with localized states, or using transformations such the FFT to make all operations sparse (Sec. M.11, especially Fig. M.2). Although the operations are not manifestly unitary, they can be used for practical calculations [233] with small δt.

Unitary expansion of the exponential

The expansion of the exponential can be done in an alternative form using the Crank–Nicholson operator [444],

$$c^{n+1} = \frac{1 - i\hat{H}\frac{\delta t}{2} + \cdots}{1 + i\hat{H}\frac{\delta t}{2} + \cdots}\, c^n. \quad (20.16)$$

This method is unitary, strictly preserving the orthonormality of the states for an arbitrary δt. For time-independent hamiltonians, it is also explicitly time-reversal invariant, and exactly conserves energy. In practice, with a suitable choice of δt, energy is satisfactorily conserved even when the hamiltonian changes with time. The disadvantage is that it involves an inverse operator, which requires a matrix inversion or solution of linear equations [444].

Expansion in Chebyshev polynomials

An alternative to iterative approaches is to expand in Chebyshev orthogonal polynomials [753] that provide a *global* fit to the propagation over the entire time range,

$$\mathrm{e}^{-\mathrm{i}Ht} \simeq \mathrm{e}^{-\mathrm{i}E_{\mathrm{av}}t} \sum_{n=0}^{N} a_n \left(\frac{\Delta E t}{2} \right) T_n(H_{\mathrm{norm}}), \tag{20.17}$$

where T_n are the Chebyshev polynomials (Sec. K.5) and the expansion coefficients $a_n(x)$ can be shown to be analogous to Bessel functions of the first kind of order n. Here $H_{\mathrm{norm}} \equiv (2H - E_{\mathrm{av}})/\Delta E$ is a normalized hamiltonian, where $E_{\mathrm{av}} = (E_{\max} + E_{\min})/2$ and $\Delta E = E_{\max} - E_{\min}$, with $E_{\max}$ and $E_{\min}$ the maximum and minimum eigenvalues of H. The Chebyshev polynomials are chosen because their error decreases exponentially when N is large enough, due to the uniform character of the Chebyshev expansion [753].

TDDFT calculations have proven to be very useful for the optical properties of finite systems with size $\ll$ wavelength of light, such as clusters. The most useful quantity is the dipole strength function, $S(\omega)$, which is proportional to the experimentally measured photoabsorption cross-section. $S(\omega)$ is related to the polarizability by[5]

$$S(\omega) = \frac{2m}{\pi e^2 \hbar} \omega \operatorname{Im} \alpha(\omega), \tag{20.18}$$

and it satisfies the *f sum rule*

$$\hbar \int_0^\infty d\omega S(\omega) = \sum_i f_i = N_e, \tag{20.19}$$

where f_i are the oscillator strengths. The quantity that can be readily calculated is $\alpha(\omega) = d(\omega)/E(\omega)$, where d is the dipole moment and E the applied electric field.

A convenient way of calculating $\alpha(\omega)$ is to find the equilibrium ground state $\psi_i^{\bar{E}}$ of the finite system in a constant applied electric field $\bar{E}$ in the $\hat{x}$ direction, i.e. with time-independent hamiltonian $\hat{H} = \hat{H}_0 - e\bar{E}x$. At time $t = 0$, the field $\bar{E}$ is suddenly removed and for $t > 0$ the system evolves with initial independent-particle states $\psi_i^{\bar{E}}$ and the hamiltonian $\hat{H}_0$. In the TDDFT approach, $\hat{H}_0$ is a function of time for $t > 0$ since the density $n(\mathbf{r}, t)$ is a function of time; however, there is no explicit external time dependence. The electric field $E(\omega)$ is the Fourier transform of a step function $E(t) = \bar{E}\Theta(-t)$, so that $E(\omega) = \bar{E}/(i\omega)$ and $\operatorname{Im} \alpha(\omega) = \omega \operatorname{Re} d(\omega)/\bar{E}$. Finally, $d(\omega)$ can be calculated from $d(t) = e \int d\mathbf{r} n(\mathbf{r}, t) x = e \sum_i \langle \psi_i(t)|x|\psi_i(t)\rangle$ and Fourier transforming to give

$$d(\omega) = \int_0^\infty dt \mathrm{e}^{\mathrm{i}\omega t - \delta t} d(t), \tag{20.20}$$

where the factor $\mathrm{e}^{-\delta t}$ is a damping factor introduced for convergence at large times. An actual example of real-time behavior of the dipole moment of C_{60} is shown in Fig. 20.4, which yields the spectrum shown as the dashed line in Fig. 20.5.

[5] Here m, e and $\hbar$ are written out explicitly to enable the conversion to usual units.

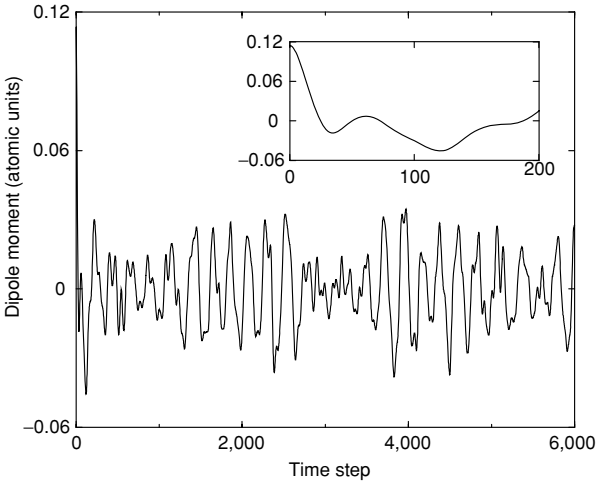


Figure 20.4. Dipole moment $d(t)$ of the C_{60} molecule as a function of time. At $t = 0$ the molecule is in the ground state with a dipole moment in the presence of an applied electric field, and the field is suddenly set to zero after which the dipole oscillates as shown. This is the actual data [754] from which is derived the spectrum shown in Fig. 20.5. The inset shows the short time transient behavior. The oscillations over a longer time are dominated by periods corresponding to the large peaks in Fig. 20.5. The time step is 5.144×10^{-3}/eV or 2.128×10^{-17} s and the total time in the evolution is 130 fs. Provided by A. Tsolakidis.

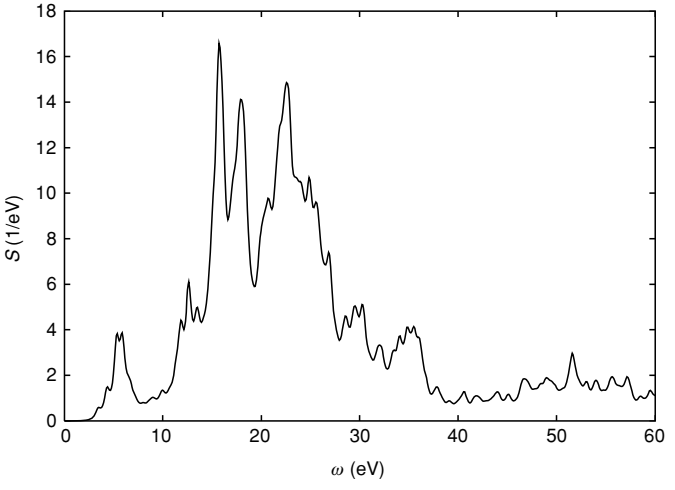


Figure 20.5. Strength function $S(\omega)$ in Eq. (20.18) for light absorption in the C_{60} molecule using the local orbital method of [754] and derived by Fourier transform of the real-time data in Fig. 20.4. (Similar results are found with the real-space, real-time method of [233].) There is good agreement of theoretical results in the lower energy range with experiment (as quoted in [233]). Provided by A. Tsolakidis.

There are now a number examples of real-time calculations that use the adiabatic LDA or GGA approximations. One of the first is that of Yabana and Bertsch [233], who considered large metal clusters and the C_{60} molecule. The spectra for Li_{147} are similar to jellium (as expected), in contrast to the results in Fig. 20.1 for Li_8. The calculations for C_{60} are similar to that shown in Fig. 20.5; there is considerable structure, in general agreement with experiment, although there is more broadening in the experiment. The aspect that can best be compared to experiment is the integrated strength in the low-energy range, which agrees well.

The real-time method can also be implemented [754] for localized bases where it is difficult to exponentiate the hamiltonian operator directly. Instead, it is convenient to use the Crank–Nicholson form in Eq. (20.16). This is explicitly unitary and inversion of the matrix $1 + i\hat{H}\frac{\delta t}{2} + \cdots$ can be done since a local orbital basis can be very small and since there can be efficient inversion methods for matrices close to the unit matrix. An example of the real-time response is illustrated in Fig. 20.4 and the spectrum is shown in Fig. 20.5. Of course, a confined local orbital basis cannot describe well the continuum and in this case the spectrum above $\approx$30 eV is not expected to be accurate.

20.5 Beyond the adiabatic local approximation

A fully satisfactory time-dependent theory must go beyond the adiabatic local (or generalized gradient) approximation [237]. An important step is to include the non-local exchange, which has been shown to be important for band gaps and excitations. For example, Fig. 2.26 shows the improvement in the eigenvalues of the Kohn–Sham equation due to inclusion of "exact exchange" (EXX). This will be reflected in the TDDFT spectra as well. Comparison of spectra of small clusters with various functionals, including EXX, has recently been done in [755]. Functionals with the correct asymptotic behavior [756] outside the system improve the ionization energies, which can greatly affect the spectra, shifting the onset of the continuum. Time dependence in the exchange–correlation functional itself is related to other problems, such as current functionals [332, 333], and is a much more difficult to cast into useful form [237].

SELECT FURTHER READING

Basic expressions for dielectric functions:

Ashcroft, N. W. and Mermin, N. D., *Solid State Physics*, W.B. Saunders Company, Philadelphia, 1976.
Kittel, C., *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1996.
Marder, M., *Condensed Matter Physics*, John Wiley and Sons, New York, 2000.
Pines, D., *Elementary Excitations in Solids*, Wiley, New York, 1964.

Basic formulation of time-dependent density functional theory:

Casida, M. E., in *Recent Developments and Applications of Density Functional Theory*, edited by J. M. Seminario, Elsevier, Amsterdam, 1996, p. 391.
Gross, E. K. U., Ullrich, C. A. and Gossmann, U. J., in *Density Functional Theory*, edited by E. K. U. Gross and R. M. Dreizler, Plenum Press, New York, 1995, p. 149.

Onida, G. Reining, L. and Rubio, A. "Electronic excitations: density-functional versus many-body green's-function approaches," *Rev. Mod. Phys.* 74: 601, 2002.

Runge, E. and Gross, E. K. U. "Density-functional theory for time-dependent systems," *Phys. Rev. Lett.* 52: 997–1000, 1984.

Real-space methods:

Vasiliev, I., Ogut, S. and Chelikowsky, J. R. "First-principles density-functional calculations for optical spectra of clusters and nanocrystals," *Phys. Rev. B* 65: 115416, 2002.

Explicit time-dependent methods:

Koonin, S. E. and Meredith, D. C., *Computational Physics*, Addison Wesley, Menlo Park, CA, 1990, Ch. 7.

Yabana, K. and Bertsch, G. F. "Time-dependent local-density approximation in real time," *Phys. Rev. B* 54: 4484–4487, 1996.

Exercises

20.1 Derive expression (20.2) for the dielectric function for non-interacting particles. Show that the first term in brackets comes from the A^2 term as stated following Eq. (20.2).

20.2 Derive the matrix equation (20.10) for the eigenvalues Ω_n of the density response from the preceding equations. Although there are many indices, this is a straightforward problem of matrix manipulation.

20.3 The general approach for exponentials of operators is described in the text preceding Eq. (20.14); it is also used in the rotation operators in Sec. N.5. Show that for any operator A, $\exp(A) = \sum_i |\zeta_i\rangle \exp(A_{ii}) \langle\zeta_i|$, where A_{ii} and ζ_i are eigenvalues and eigenvectors of A. Hint: Use the power series expansion of the exponential and show the equivalence of the two sides of the equation at every order.

20.4 Show that Eq. (20.14) has error of order $\propto \delta t^2$. Would the error be of the same order if the potential part were not symmetric?

21

Wannier functions

Summary

Wannier functions are enjoying a revival as important, practical tools for electronic structure. They have a long history of providing useful localized functions for formal proofs; however, they are often not regarded as useful because of their inherent non-uniqueness, that is, a dependence upon the choice of a "gauge." This has changed with the realization that Wannier functions can be used effectively to calculate important physical quantities in a gauge-invariant manner. In addition, the particular construction of "maximally localized Wannier functions" provides elegant connections to the Berry's phase formulation of polarization. The subjects of this and the following two chapters are closely related: the expressions given here are useful in understanding localization and polarization, the subject of Ch. 22, and the discussion there brings out the physical meaning of the quantities derived in this chapter. The emergence of "order-N" methods (Ch. 23) has given impetus to the development of useful localized functions closely related to Wannier functions.

21.1 Definition and properties

Wannier functions [338, 759, 763] are orthonormal localized functions that span the same space as the eigenstates of a band or a group of bands. Extensive reviews of their properties have been given by Wannier [338], Blount [759], and Nenciu [339]. Here we consider properties relevant to understanding the electronic properties of materials and to present-day practical calculations.

The eigenstates of electrons in a crystal are extended throughout the crystal with each state having the same magnitude in each unit cell. This has been shown in the independent-particle approximation Sec. 4.3 using the fact that the hamiltonian $\hat{H}$ in (4.22) commutes with the translations operations $\hat{T}_{\mathbf{n}}$ in (4.23). Thus eigenstates of hamiltonian $\hat{H}$ are also eigenstates of $\hat{T}_{\mathbf{n}}$, leading to the Bloch theorem, Eqs. (4.33) or (12.11),

$$\psi_i^{\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} u_i^{\mathbf{k}}(\mathbf{r}), \tag{21.1}$$

which here is taken to be normalized in one cell. Since the overall phase of each eigenstate

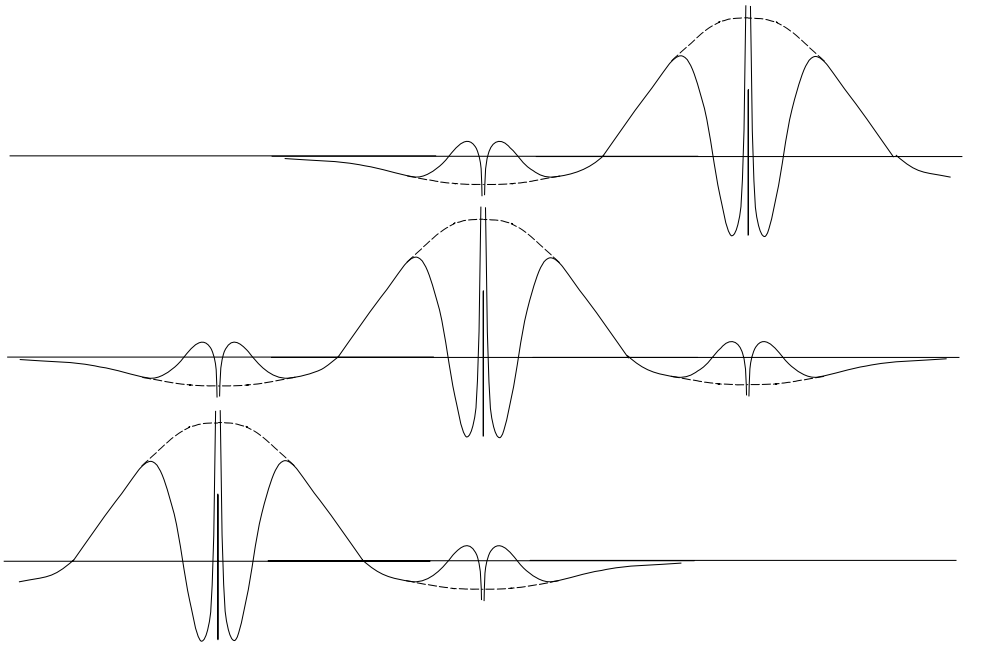


Figure 21.1. Schematic example of Wannier functions that correspond to the Bloch functions in Fig. 4.11. These are for a band made of 3s atomic-like states and the smooth functions denote the smooth part of the wavefunction as illustrated in Fig. 11.2.

is arbitrary, any Bloch function is subject to a "gauge transformation"

$$\psi_i^{\mathbf{k}}(\mathbf{r}) \to \tilde{\psi}_i^{\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}\phi_i(\mathbf{k})}\psi_i^{\mathbf{k}}(\mathbf{r}) \tag{21.2}$$

which leaves unchanged all physically meaningful quantities.[1]

Wannier functions are Fourier transforms of the Bloch eigenstates. For one band i the function associated with the cell labeled by the lattice point $\mathbf{T}_m$ is

$$w_i(\mathbf{r}-\mathbf{T}_m) = \frac{\Omega_{\mathrm{cell}}}{(2\pi)^3}\int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}_m}\psi_i^{\mathbf{k}}(\mathbf{r}) = \frac{\Omega_{\mathrm{cell}}}{(2\pi)^3}\int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{T}_m)}u_i^{\mathbf{k}}(\mathbf{r}), \tag{21.3}$$

as shown schematically in Fig. 21.1. The function w_i associated with a different cell $\mathbf{T}_{m'}$ is the same function, except it is translated by $\mathbf{T}_{m'} - \mathbf{T}_m$. Conversely, Eq. (21.3) leads to

$$\psi_i^{\mathbf{k}}(\mathbf{r}) = \sum_m \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}_m} w_i(\mathbf{r}-\mathbf{T}_m). \tag{21.4}$$

The transformation (21.3) assumes that the Bloch functions $\psi_i^{\mathbf{k}}(\mathbf{r})$ are periodic in reciprocal space, i.e. the "periodic gauge" where

$$\psi_i^{\mathbf{k}}(\mathbf{r}) = \psi_i^{\mathbf{k}+\mathbf{G}}(\mathbf{r}) \tag{21.5}$$

[1] In general, $\phi_i(\mathbf{k})$ is completely arbitrary, but in some cases it is desirable to restrict $\phi_i(\mathbf{k})$ to be a continuous function of $\mathbf{k}$, i.e. a "differentiable gauge." It is then implicitly assumed that $\psi_i^{\mathbf{k}}(\mathbf{r})$ is also a smooth function of $\mathbf{k}$.

for all reciprocal lattice vectors $\mathbf{G}$. This is clearly obeyed by the Bloch functions given by (21.4).

Wannier functions, labeled $i = 1, 2, \ldots$, can be defined for a set of bands $j = 1, 2, \ldots, .$ In general, the functions are defined as a linear combination of the Bloch functions of different bands, so that the definition is an extension of (21.2). Each Wannier function is given by (21.3) with

$$u_{i\mathbf{k}} = \sum_j U^{\mathbf{k}}_{ji} u^{(0)}_{j\mathbf{k}}, \tag{21.6}$$

where $U^{\mathbf{k}}_{ji}$ is a $\mathbf{k}$-dependent unitary transformation. For example, in the diamond or zinc-blende semiconductors, four occupied bands, together, are needed to form Wannier functions with sp^3 character, that have a simple interpretation in terms of chemical bonding.

It is straightforward to show (Exercise 21.3) that the Wannier functions are orthonormal

$$\langle \mathbf{T}_m i | \mathbf{T}_{m'} j \rangle = \int_{\text{all space}} \mathrm{d}\mathbf{r} w^*_i(\mathbf{r} - \mathbf{T}_m) w_j(\mathbf{r} - \mathbf{T}_{m'}) = \delta_{ij}\delta_{nn'}, \tag{21.7}$$

using (21.3) and the fact that the eigenfunctions $\psi^{\mathbf{k}}_i(\mathbf{r})$ are orthonormal. Note that the integral in (21.7) is over all space.

Matrix elements of the position operator $\hat{\mathbf{r}}$ can be defined using notation analogous to (21.7) with the result [759]

$$\langle \mathbf{T}i | \hat{\mathbf{r}} | 0j \rangle = i \frac{\Omega}{(2\pi)^3} \int \mathrm{d}\mathbf{k} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \langle u_{i\mathbf{k}} | \boldsymbol{\nabla}_{\mathbf{k}} | u_{j\mathbf{k}} \rangle, \tag{21.8}$$

and conversely

$$\langle u_{i\mathbf{k}} | \boldsymbol{\nabla}_{\mathbf{k}} | u_{j\mathbf{k}} \rangle = -i \sum_{\mathbf{T}} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \langle \mathbf{T}i | \hat{\mathbf{r}} | 0j \rangle, \tag{21.9}$$

where it is understood that $\boldsymbol{\nabla}_{\mathbf{k}}$ acts only to the right. These expressions can be derived by noting that

$$\langle u_{i\mathbf{k}} | u_{j\mathbf{k}+\mathbf{q}} \rangle = \langle \psi_{i\mathbf{k}} | \mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{r}} | \psi_{j\mathbf{k}+\mathbf{q}} \rangle = \sum_{\mathbf{T}} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \langle \mathbf{T}i | \mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{r}} | 0j \rangle, \tag{21.10}$$

and expanding in powers of $\mathbf{q}$. Similarly to second order in $\mathbf{q}$ this leads to (see Exercise 21.4)

$$\langle \mathbf{T}i | \hat{\mathbf{r}}^2 | 0j \rangle = -\frac{\Omega}{(2\pi)^3} \int \mathrm{d}\mathbf{k} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{T}} \langle u_{i\mathbf{k}} | \nabla^2_{\mathbf{k}} | u_{j\mathbf{k}} \rangle. \tag{21.11}$$

Non-uniqueness of Wannier functions

The most serious drawback of the Wannier representation is that the functions are not uniquely defined. They can vary strongly in shape and range, as opposed to the Bloch functions that are unique (except for an overall phase that is constant in space). This can be seen from (21.3) together with (21.2) or (21.6): the Wannier function changes because variations in $\phi_i(\mathbf{k})$ or $U^{\mathbf{k}}_{ji}$ change the *relative* phases and amplitudes of Bloch functions at different $\mathbf{k}$ and different bands i.

The centers of Wannier functions defined by $\langle \mathbf{T}i|\hat{\mathbf{r}}|\mathbf{T}i\rangle$ are an important example. Blount [759] has shown that the *sums of the centers of all the Wannier functions in a cell*, i.e. the center of mass $\bar{\mathbf{r}}$, is invariant, except that of course $\bar{\mathbf{r}}$ can change by any translation vector to an equivalent point.[2] However, all higher moments are "gauge dependent," i.e. they are not invariant to the choice of "gauge," $\phi_i(\mathbf{k})$ in (21.2) or $U_{ji}^{\mathbf{k}}$ in (21.6).

21.2 "Maximally projected" Wannier functions

The term "maximally projected Wannier functions" is introduced here to describe a simple, intuitive approach for construction of Wannier functions [758, 764]. For simple bands, it is sufficient to choose the phases of the Bloch functions so that the Wannier function is maximum at a chosen point; more generally, one can choose phases so that the Wannier function has maximum overlap with a chosen localized function – hence the term "maximal projection." As will be clear from the discussion below, the construction is valuable for general proofs [758], for construction of functions that allow a localized formulation of the electronic structure problem [765, 766], for actual calculations in materials [694], and for a general approach in linear-scaling order-N methods (Ch. 23).

The simplest example has been analyzed by Kohn [758, 767]. For a crystal with one atom per cell and a single band derived from s-symmetry orbitals, the Wannier function $w_i(\mathbf{r})$ on site $\mathbf{T}_m = 0$ can be chosen to be the sum of Bloch functions $\psi_{\mathbf{k}}(\mathbf{r})$ with phases such that $\psi_{\mathbf{k}}(0)$ is real and positive for each $\mathbf{k}$. The Wannier function thus defined by (21.3) is maximal on site 0 and decays as a function of distance from 0. In the case of a one-dimensional crystal, it has been proven [758, 767] that the decay is exponential and this is the *only* exponentially decaying Wannier function that is real and symmetric about the origin. However, there are no proofs for a general three-dimensional crystal.[3] Similarly, one can construct bond-centered functions by requiring a maximal value at a given bond center.

This approach can be extended to more general cases by requiring that the phase of the Bloch function be chosen to have maximum overlap with a chosen localized function, i.e. maximum projection of the function. An example is a p-symmetry atom-centered Wannier function chosen to have maximal overlap with a p atomic-like state on an atom in the cell at the origin. Maximum projection on any orbital in the basis is easy to accomplish in localized basis representations, simply by choosing the phase of each Bloch function so that the amplitude is real and positive for the given orbital. In a plane wave calculation, for example, it means taking a projection much like the projectors for separable pseudopotentials (Sec. 11.8). For bond-centered functions, one can require maximal overlap with a localized bonding-like function.

[2] This has an important physical interpretation in the theory of polarization in Ch. 22.

[3] The proofs of Kohn have been extended by He and Vanderbilt [768] to show that in the one-dimensional case the Wannier functions and various matrix elements between them decay as an exponential multiplied by $(1/r)^{\alpha}$, where α is fractional and has characteristic values for each of the different quantities. Taraskin, Drabold, and Elliott [769] have shown that the results also apply to a two-band problem on a simple cubic lattice in three dimensions.

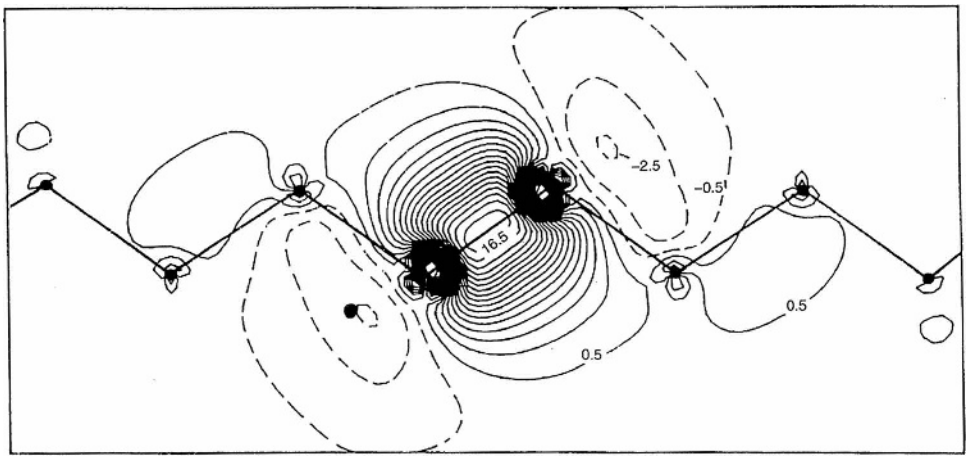


Figure 21.2. Bond-centered Wannier function for Si calculated [694] by requiring the phases of the Bloch functions to have real, positive amplitude at one of the four equivalent band centers. Note the similarity to Fig. 21.4. The decay of the orbital on a log scale is shown in Fig. 21.3.

Stated in this broad way, the construction leads to a transformation of the electronic structure problem into a new basis of localized orthonormal orbitals. This is the basis of the formulations of Bullett [765] and Anderson [766] that provide a fundamental way of deriving generalized Hubbard-type models [392, 393], and used, for example to calculate model parameters for orbitals centered on Cu and O in CuO_2 materials [770] and for orbitals that span a space of d and s symmetry functions in Cu metal [771].

A construction often used in "order-N" calculations is to find functions localized to a sphere within some radius around a given site. This can be interpreted as "maximal overlap" with a function that is unity inside the sphere and zero outside, usually applied with the boundary condition that the function vanish at the sphere boundary.

Bond-centered Wannier in silicon

The construction of bond-centered Wannier functions in diamond structure crystals is discussed by Kohn [758] and careful numerical calculations have been done by Satpathy and Pawlowska [694] for Si – the standard test case, of course. The calculations used the LMTO method (Ch. 17), in which the orbitals are described in terms of functions centered on the atoms (and on empty spheres). A bond-centered Wannier function is generated simply by choosing the phases of the Bloch functions to be positive on one of the four bond centers in a unit cell. The function can then be plotted in real space as shown in Fig. 21.2; note the striking resemblance to the "maximally localized" function shown below on the left-hand side of Fig. 21.4.

Satpathy and Pawlowska [694] also showed numerically that the bond-centered function in Fig. 21.2 decays exponentially, as presented in Fig. 21.3. This is perhaps the first such accurate numerical test of the exponential decay in solids like Si, which has since been found in many other calculations. Presumably the reason for the similarity of the Wannier

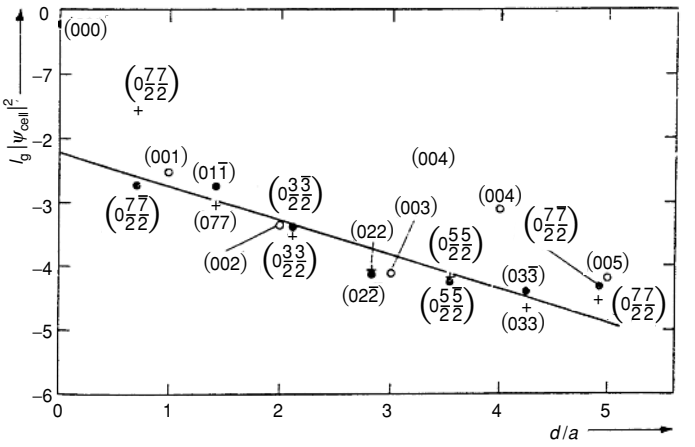


Figure 21.3. The decay of the bond-centered Wannier function for Si (shown in Fig. 21.2) in various directions, plotted on a log scale. It is evident that the decay is consistent with an exponential, although there are no rigorous proofs that this is the case in a real three-dimensional material.

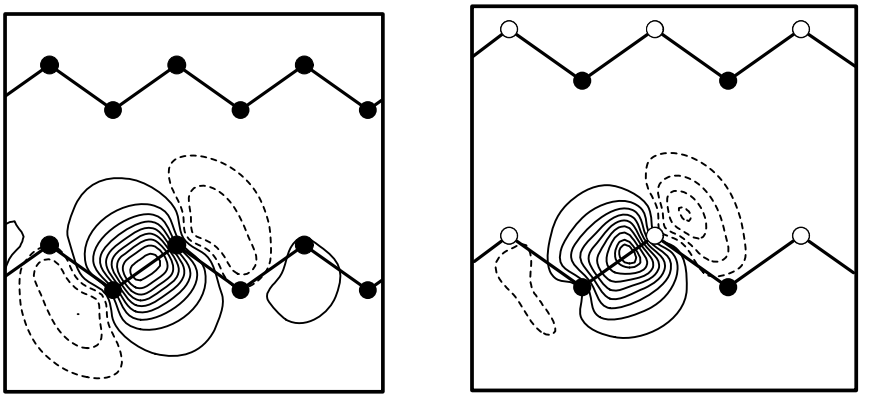


Figure 21.4. "Maximally localized" Wannier functions for Si (left) and GaAs (right) from [762]. Each figure shows one of the four equivalent functions found for the four occupied valence bands. Provided by N. Marzari.

functions for Si in Figs. 21.2 and 21.4 is related to Kohn's proof in one dimension that the function is uniquely fixed by the requirements that the function be real, symmetric, and exponentially decaying; however, there is no general proof in three dimensions at the present time.

21.3 Maximally localized Wannier functions

Finding highly localized Wannier functions (or transforms of Wannier functions) with desired properties is a venerable subject in chemistry [760, 761, 772, 773], where they are called "localized molecular orbitals." Such functions are useful in constructing efficient

methods (see Ch. 23 on "order-N" algorithms) and in providing insight through simple descriptions of the electronic states using a small number of functions [760, 761, 773, 774].

Although there are many possible ways to define "maximally localized,"[4] one stands out: minimization of the mean square spread Ω defined by

$$\Omega = \sum_{i=1}^{N_{\text{bands}}} \left[\langle r^2 \rangle_i - \langle \mathbf{r} \rangle_i^2 \right], \tag{21.12}$$

where $\langle \cdots \rangle_i$ means the expectation value over the ith Wannier function in the unit cell (whose total number N_{bands} equals the number of bands considered). As shown by Marzari and Vanderbilt [762] this definition leads to an elegant formulation, in which a part of the spread, Eq. (21.12), can be identified as an invariant (Eq. (21.14) below). Furthermore, this invariant part leads to a physical measure of localization as shown by Souza, et al. [775] (see Sec. 22.5).

Because the Wannier functions are not unique, the Ω defined in (21.12) is not invariant under gauge transformations of the Wannier functions [762]. Nevertheless, Marzari and Vanderbilt were able to decompose Ω into a sum of two positive terms: a gauge-invariant part Ω_I, plus a gauge-dependent term $\widetilde{\Omega}$:

$$\Omega = \Omega_I + \widetilde{\Omega}, \tag{21.13}$$

$$\Omega_I = \sum_{i=1}^{N_{\text{bands}}} \left[\langle r^2 \rangle_i - \sum_{\mathbf{T}j} |\langle \mathbf{T}j|\hat{\mathbf{r}}|0i\rangle|^2 \right], \tag{21.14}$$

$$\widetilde{\Omega} = \sum_{i=1}^{N_{\text{bands}}} \sum_{\mathbf{T}j \neq 0i} |\langle \mathbf{T}j|\hat{\mathbf{r}}|0i\rangle|^2. \tag{21.15}$$

Clearly, the second term $\widetilde{\Omega}$ is always positive. The clever part of the division in (21.13), however, is that Ω_I is *both invariant and always positive*. Furthermore, it has a simple interpretation that may be seen by identifying the projection operator $\hat{P}$ onto the space spanned by the N_{bands} bands,[5]

$$\hat{P} = \sum_{i=1}^{N_{\text{bands}}} \sum_{\mathbf{T}} |\mathbf{T}i\rangle\langle \mathbf{T}i| = \sum_{i=1}^{N_{\text{bands}}} \sum_{\mathbf{k}} |\psi_{i\mathbf{k}}\rangle\langle \psi_{i\mathbf{k}}|, \tag{21.16}$$

and $\hat{Q} = 1 - \hat{P}$ defined to be the projection onto all other bands. Writing out (21.14) leads to the simple expression (here α denotes the vectors index for $\mathbf{r}$)

$$\Omega_I = \sum_{i=1}^{N_{\text{bands}}} \sum_{\alpha=1}^{3} \langle 0i|\hat{\mathbf{r}}_\alpha \hat{Q} \hat{\mathbf{r}}_\alpha|0i\rangle, \tag{21.17}$$

which is manifestly positive (Exercise 21.5). The presence of the $\hat{Q}$ projection operator leads to an informative interpretation of (21.17) as the quantum fluctuations of the position operator from the space spanned by the Wannier functions into the space of the other bands.

[4] For example, a widely used criterion is to maximize the self-Coulomb interaction [772].

[5] This is the same as defined in Eq. (19.20), except that here the sum need not be over all occupied states.

This can also be viewed as a consequence of the fact that the position operator does not commute with $\hat{P}$ or $\hat{Q}$ (Exercise 21.8), so that expression (21.17) is *not* simply the mean square width of the Wannier function. Instead Ω_I is an invariant, as is apparent in the explicit expressions in **k** space given below and is explained further in [762].[6] Furthermore, the fact that (21.17) represents fluctuations leads to the physical interpretation of Ω_I brought out in Sec. 22.5.

Practical expressions in k space

Expressions for Ω_I and $\widetilde{\Omega}$ in terms of the Bloch states can be derived by substituting the definitions of the Wannier functions, Eq. (21.3), into (21.14) and (21.15). It is an advantage for practical calculations to write the formulas in terms of discrete sums instead of integrals using Eq. (12.14). If one uses a finite difference approximation for the derivatives w.r.t **k** in (21.8) and (21.11), one finds [762]

$$\langle \mathbf{r} \rangle_j = \frac{i}{N_k} \sum_{\mathbf{kb}} w_{\mathbf{b}} \mathbf{b} \left[\langle u_{j\mathbf{k}} | u_{j\mathbf{k}+\mathbf{b}} \rangle - 1 \right], \tag{21.18}$$

and

$$\langle r^2 \rangle_j = \frac{1}{N_k} \sum_{\mathbf{kb}} w_{\mathbf{b}} \left[2 - \mathrm{Re} \langle u_{j\mathbf{k}} | u_{j\mathbf{k}+\mathbf{b}} \rangle \right], \tag{21.19}$$

where **b** denote the vectors connecting the points **k** to neighboring points $\mathbf{k}+\mathbf{b}$ and $w_{\mathbf{b}}$ denotes the weights in the finite difference formula.

Although these formulas reduce to the integral in the limit $\mathbf{b} \to 0$, they are not acceptable because they violate the fundamental requirement of translation invariance for any finite **b**. If one makes the substitution $\psi_i^{\mathbf{k}}(\mathbf{r}) \to \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{T}_m} \psi_i^{\mathbf{k}}(\mathbf{r})$, the expectation values should change by a translation,

$$\begin{aligned} \langle \mathbf{r} \rangle_j &\to \langle \mathbf{r} \rangle_j + \mathbf{T}_m, \\ \langle r^2 \rangle_j &\to \langle r^2 \rangle_j + 2 \langle \mathbf{r} \rangle_j \cdot \mathbf{T}_m + T_m^2, \end{aligned} \tag{21.20}$$

so that Ω is unchanged. These properties are not obeyed by (21.18) or (21.19).

Acceptable expressions can be found [762] that have the same limit for $\mathbf{b} \to 0$ yet satisfy Eq. (21.20). Functions with the desired character are complex log functions that have a Taylor series expansion, $\ln(1+ix) \to ix - x^2 + \cdots$ for small x (similar to Eqs. (21.18) and (21.19) for x real), but are periodic functions for large $\mathrm{Re}\{x\}$. If we define $\langle u_{i\mathbf{k}} | u_{j\mathbf{k}+\mathbf{b}} \rangle \equiv M_{ij}(\mathbf{k}, \mathbf{b})$, (21.18) and (21.19) can be replaced by[7] (Exercise 21.6)

$$\langle \mathbf{r} \rangle_j = \frac{i}{N_k} \sum_{\mathbf{kb}} w_{\mathbf{b}} \mathbf{b} \mathrm{Im} \ln M_{jj}(\mathbf{k}, \mathbf{b}), \tag{21.21}$$

[6] An interesting feature is that Ω_I can be expressed in terms of a "metric" that defines the "quantum distance" along a given path in the Brillouin zone [762]. This "distance" quantifies the change of character of the occupied states $u_{n\mathbf{k}}$ as one traverses the path, leading to the heuristic interpretation of Ω_I as representing a measure of the dispersion throughout the Brillouin zone [762].

[7] These forms are not unique. Alternatives are pointed out in [762] and the expression for the center, Eq. (21.21), is not the same as given earlier in the theory of polarization [147].

and

$$\langle r^2 \rangle_j = \frac{1}{N_k} \sum_{\mathbf{kb}} w_{\mathbf{b}} \left\{ 1 - |M_{jj}(\mathbf{k}, \mathbf{b})|^2 + [\mathrm{Im} \ln M_{jj}(\mathbf{k}, \mathbf{b})]^2 \right\}. \tag{21.22}$$

The invariant part can be found in a way similar to (21.22) with the result

$$\Omega_I = \frac{1}{N_k} \sum_{\mathbf{kb}} w_{\mathbf{b}} \left[N_{\mathrm{bands}} - \sum_{ij}^{N_{\mathrm{bands}}} |M_{ij}(\mathbf{k}, \mathbf{b})|^2 \right], \tag{21.23}$$

which is positive (Exercise 21.7). The meaning of this term and closely related expressions are given in Sec. 22.5.

In one dimension it is possible to choose Wannier functions so that $\widetilde{\Omega} = 0$, i.e. the minimum possible spread. However, in general, it is not possible for $\widetilde{\Omega}$ to vanish in higher dimensions. This follows (Exercise 21.8) from the expression for $\widetilde{\Omega}$ given later in Eq. (21.28) and the fact that the *projected* operators $\{\hat{P}\hat{x}\hat{P}, \hat{P}\hat{y}\hat{P}, \hat{P}\hat{z}\hat{P}\}$ do not commute, i.e. the matrices representing the matrix elements $\langle \mathbf{T}i|\hat{x}|\mathbf{T}'j\rangle$ do not commute.

Minimization by steepest descent

Finding Wannier functions that are maximally localized can be accomplished by minimizing the spread, Eq. (21.22), as a function of the Bloch functions. (This means minimizing $\widetilde{\Omega}$ since Ω_I is invariant.) For a given set of Bloch functions $u^{(0)}_{j\mathbf{k}}$ one can consider all possible unitary transformations given by Eq. (21.6), which can be written as,

$$\mathbf{M}(\mathbf{k}, \mathbf{b}) = [\mathbf{U}^{\mathbf{k}}]^{\dagger} \mathbf{M}^{(\mathbf{0})}(\mathbf{k}, \mathbf{b}) \mathbf{U}^{\mathbf{k}+\mathbf{b}}, \tag{21.24}$$

where $\mathbf{M}$ and $\mathbf{U}$ are understood to be matrices in the band indices. To minimize, one can vary $\mathbf{U}^{\mathbf{k}}$, which is done most conveniently by defining

$$\mathbf{U}^{\mathbf{k}} = \mathrm{e}^{\mathbf{W}^{\mathbf{k}}}, \tag{21.25}$$

where $\mathbf{W}^{\mathbf{k}}$ is an antihermitian matrix (Exercise 21.14). The solution can be found by the method of steepest descent (App. L). The gradient can be found by considering infinitesimal changes, $\mathbf{U}^{\mathbf{k}} \to \mathbf{U}^{\mathbf{k}}(\mathbf{1} + \delta\mathbf{W}^{\mathbf{k}})$. Expressions for the gradient,

$$\frac{\delta\Omega}{\delta\mathbf{W}^{\mathbf{k}}} = \mathbf{G}^{\mathbf{k}}, \tag{21.26}$$

in $\mathbf{k}$ space are given in [762]; we will give equivalent expressions in (21.28) and (21.29) that bring out the physical meaning. Choosing $\delta\mathbf{W}^{\mathbf{k}} = \epsilon\delta\mathbf{G}^{\mathbf{k}}$ along the steepest decent direction leads to a useful minimization algorithm, which corresponds to updating the $\mathbf{M}$ matrices at each step n

$$\mathbf{M}^{(\mathbf{n}+\mathbf{1})}(\mathbf{k}, \mathbf{b}) = \mathrm{e}^{-\delta\mathbf{W}^{\mathbf{k}}} \mathbf{M}^{(\mathbf{n})}(\mathbf{k}, \mathbf{b}) \mathrm{e}^{\delta\mathbf{W}^{\mathbf{k}+\mathbf{b}}}, \tag{21.27}$$

where the exponentiation can be done by diagonalizing $\delta\mathbf{W}$.

Examples of Wannier functions calculated by this "maximal localization" prescription are shown in Fig. 21.4 for Si and GaAs [762]. These functions are derived by considering the

full set of four occupied valence bands, leading to four equivalent bonding-like orbitals. For GaAs there is an alternative possibility: since the four valence bands (see Fig. 17.8) consist of one well-separated lowest band plus three mixed bands at higher energy, Wannier functions can be derived separately for the two classes of bands. The result is one function that is primarily s-like on the As atom and three functions primarily p-like on the As atom [762]. However, these Wannier functions do not lead to the maximum overall localization, so that the bonding orbitals appear to provide the most natural picture of local chemical bonding.

Wannier functions in disordered systems

Up to now the derivations have focused entirly upon crystals and have used Bloch functions. How can one find useful maximally localized functions for non-crystalline systems, such as molecules or disordered materials? This is particularly important for interpretation purposes and for calculations of electric polarization, etc., in large Car-Parrinello-type simulations (Ch. 18) where often calculations are done only for periodic boundary conditions, i.e. for $\mathbf{k} = 0$. Many properties, such as the total dipole moment of the sum of Wannier functions, are invariant (Sec. 22.3), so any approach that finds accurate Wannier functions is sufficient. For other properties, it is desirable to derive maximally localized functions.

The most direct approach is to construct "maximally projected" functions (Sec. 21.2) that are chosen to have weight at a center or maximum overlap with a chosen function. A closely related procedure is actually used in the "order-N" linear-scaling methods described in Sec. 23.5 that explicitly construct Wannier-like functions constrained to be localized to a given region [776]. These methods provide an alternative approach for direct construction of Wannier functions without ever constructing eigenstates. An example of such a Wannier function is shown in Fig. 23.9 for a typical bonding function in a large cell of 4096 atoms that is a model for amorphous Si [777]. The contour plots show the logarithm of the square of the Wannier function which decays exponentially over 20 orders of magnitude. This is a case where the O(N) linear scaling method is much more efficient for construction of the Wannier function than is a method that constructs the function from eigenstates.

It is also useful to construct "maximally localized" functions. For example, they are directly useful in the concept of localization (Sec. 22.5). The functions can be derived by working directly with the definitions in real space and minimizing $\widetilde{\Omega}$ given by Eq. (21.15). It follows from the definitions (as shown in App. A of [762] and further elucidated in [778] and [779]), that (21.15) can be written as

$$\widetilde{\Omega} = \mathrm{Tr}[\hat{X}'^2 + \hat{Y}'^2 + \hat{Z}'^2], \tag{21.28}$$

where $X_{ij} = \langle 0i|\hat{x}|0j\rangle$, $[X_D]_{ij} = X_{ii}\delta_{ij}$, and $X'_{ij} = X_{ij} - [X_D]_{ij}$, with corresponding expressions for $\hat{Y}$ and $\hat{Z}$. For infinitesimal unitary transformation $|i\rangle \rightarrow |i\rangle + \sum_j \delta W_{ji}|j\rangle$, the gradient of (21.28) can be written as (see Exercise 21.15) $\delta\Omega = 2\mathrm{Tr}[\hat{X}'\delta\hat{X} + \hat{Y}'\delta\hat{Y} + \hat{Z}'\delta\hat{Z}]$, where $\delta\hat{X} = [\hat{X}, \delta\hat{W}]$. Finally, one finds $\delta\Omega = \mathrm{Tr}[\delta\hat{W}\hat{G}]$, where

$$\frac{\delta\widetilde{\Omega}}{\delta\hat{W}} = \hat{G} = 2\{[\hat{X}', \hat{X}_D] + [\hat{Y}', \hat{Y}_D] + [\hat{Z}', \hat{Z}_D]\}. \tag{21.29}$$

These forms are the most compact expressions for $\widetilde{\Omega}$ and its gradient. They are directly useful in real-space calculations in terms of Wannier functions $w_i(\mathbf{r})$; the corresponding forms in $\mathbf{k}$ space [762] can be derived using the transformations of Sec. 21.1 and used in (21.26).

An example of Wannier functions calculated at steps in a quantum molecular dynamics simulation of water under high-pressure, high-temperature conditions are shown in Fig. 2.12. Three "snapshots" during a simulation show a sequence that involves a proton transfer and the associated transfer of a Wannier function (only this one of all the Wannier functions is shown) to form H^+ and $(H_3O)^-$.

21.4 Non-orthogonal localized functions

One can also define a set of non-orthogonal localized orbitals $\tilde{w}_i$ that span the same space as the Wannier functions w_i and which can be advantageous for practical applications and for intuitive understanding [772]. Just as for Wannier functions, one must choose some criterion for "maximal localization" to fix the $\tilde{w}_i$. A recent work of Liu, et al. [774] is particularly illuminating since it uses the same mean square radius criterion as in Eq. (21.12) and provides a practical approach for calculating the functions directly related to optimizing functionals in O(N) methods (Sec. 23.5).

The transformation to non-orthogonal $\tilde{w}_i$ can be defined by

$$\tilde{w}_i = \sum_{j=1}^{N_{\text{bands}}} A_{ij} w_j, \tag{21.30}$$

where A is a non-singular matrix that must satisfy

$$\sum_{i=1}^{N_{\text{bands}}} (A_{ij})^2 = 1, \tag{21.31}$$

since the $\tilde{w}_i$ are defined to be normalized. The mean square spread, Eq. (21.12), generalizes to [774]

$$\Omega[A] = \sum_{i=1}^{N_{\text{bands}}} \left[\langle \tilde{w}_i | r^2 | \tilde{w}_i \rangle - \langle \tilde{w}_i | \mathbf{r} | \tilde{w}_i \rangle^2 \right], \tag{21.32}$$

which is to be minimized as a function of the matrix A subject to two conditions: A is non-singular and satisfies Eq. (21.31). It is simple to enforce the latter condition; however, it is not so simple to search only in the space of non-singular matrices. It is shown in [774] that one can use the fact that a non-singular matrix must have full rank, i.e. $\text{rank}(A) = N$. Using $\text{rank}(A) = \text{rank}(A^\dagger A)$ and the variational principle, Eq. (23.33), developed for minimizing the energy functional [780] with $S \to A^\dagger A$, the result is

$$\text{rank}(A) = -\min\{\text{Tr}[(-A^\dagger A)(2X - XA^\dagger AX)]\}, \tag{21.33}$$

which is minimized for all hermitian matrices X. Defining a constraint functional $\Omega_a[A, X] = (N - \text{Tr}((A^\dagger A)(2X - XA^\dagger AX)))^2$, it follows that maximally localized non-

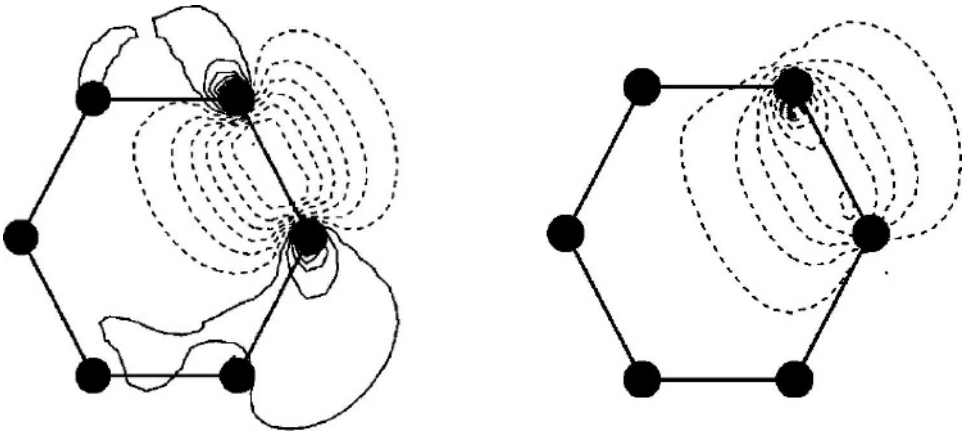


Figure 21.5. Comparison of orthogonal and non-orthogonal maximally localized orbitals for C–C σ bonds (left) and C–H bonds (right) in benzene C_6H_6. The non-orthogonal orbitals are more localized and more transferrable since the extended wiggles in the orthogonal functions depend in detail upon the neighboring atoms. From [774].

orthogonal orbitals $\tilde{w}_i$ can be found by minimizing $\Omega[A] + C_a \min\{\Omega_a[A, X]\}$, for all matrices A that satisfy Eq. (21.31). Here C_a is an adjustable positive constant and the second term ensures that the final transformation matrix A is non-singular (Exercise 21.16).

An example of maximally localized orbitals for a benzene molecule are shown in Fig. 21.5, where we see that the non-orthogonal orbitals are much more localized and much easier to interpret as simple bonding orbitals than the corresponding orthogonal orbitals. The short range of the non-orthogonal orbitals can be used in calculations to reduce the cost, for example, in O(N) methods as discussed in Sec. 23.5.

21.5 Wannier functions for "entangled bands"

The subject this section is construction of Wannier-type functions that describe bands in some energy range *even though they are not isolated and are "entangled" with other bands*. Strictly speaking, Wannier functions as defined in Sec. 21.1 will not be useful; if the bands cannot be disentangled then there will be non-analytic properties resulting from mixing with others bands in the integrals over the Brillouin zone. However, one can define useful functions that have real-space properties like Wannier functions and form an *orthonormal, localized basis for a subspace of bands that spans a desired range of energies.*

There are two basic approaches for construction of functions that span a desired subspace. One approach is to identify the type of orbitals involved and to generate a reduced set of localized functions that describes the energy bands over a given range. Outside that range, the full band structure is, of course, not reproduced: the reduced set of bands has an upper and a lower bound, i.e. they form a set of isolated bands in the reduced space. This is in essence the idea of "maximally projected" functions of Sec. 21.2, but now constrained only to match the bands over some range. An example of such an approach is the "downfolding"

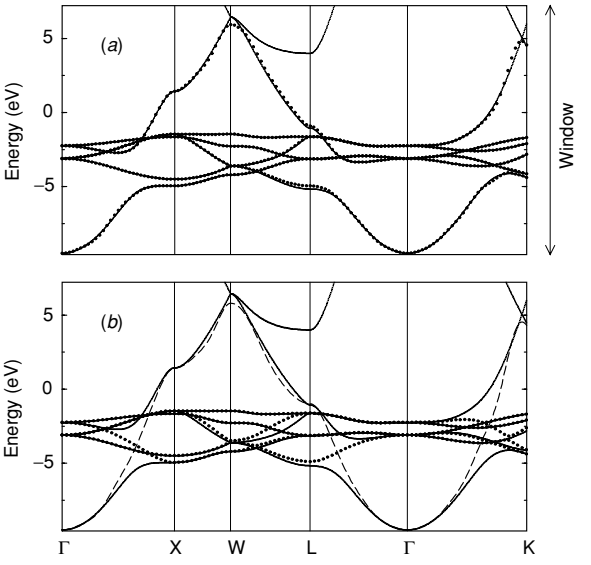


Figure 21.6. Bands of Cu produced by maximally localized Wannier-like functions [781]. Top panel: Functions that span the six-dimensional subspace for the 5 d states and 1 s state compared to the full band structure. Similar results for the bands are found in [771]. The bands are accurately reproduced up to well above the Fermi energy, even though the higher bands are missing. The lower panel shows results if the subspace is decomposed into the 5 d orbitals (chosen as maximally localized with a narrow energy window around the primarily d bands) plus the complement that is the s orbital. From [781].

method [699, 782] the results of which are illustrated in Figs. 17.11 and 17.12. The single orbital centered on a Cu atom is sufficient to describe accurately the main band that crosses the Fermi energy without explicitly including the rest of the "spaghetti" of bands.

The second approach [771, 781] generalizes the idea of "maximally localized" Wannier functions (Sec. 21.3) to maximize the overlap with Bloch functions *only over an energy window*. This also generates a finite subspace of bands that describes the actual bands only within the chosen range. Of course, the functions are not unique since there are many choices for the energy range and weighting functions. Two recent calculations for Cu done using pseudopotentials and plane waves [781] and the LMTO method [771] give very similar results for the desired bands, but with different localized functions. For example, maximally localized functions constructed from 6 d and s bands taken together are each centered in interstitial positions near the Cu atom [781]. The bands for the six-dimensional subspace of orbitals are given in Fig. 21.6, which shows that the band structure is accurately represented for energies extended to well above the Fermi energy, even though the higher bands are missing. The lower panel of the figure shows a different decomposition with the subspace decomposed into 5 d orbitals (which have the expected form of atom-centered d orbitals in a cubic symmetry crystal) plus the complement that is an optimal s-symmetry orbital. Similar results for the bands are found using the LMTO method [771] where the authors also showed that the functions decay exponentially (or at least as a very high power).

SELECT FURTHER READING

Textbooks and extensive review:

Ashcroft, N. W. and Mermin, N. D., *Solid State Physics*, W. B. Saunders Company, Philadelphia, 1976.

Blount, G., in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1962, p. 305.

Weinreich, G., *Solids: Elementary Theory for Advanced Students*, John Wiley and Sons, New York, 1965.

Reviews:

Boys, S. F., "Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another," *Rev. Mod. Phys.* 32:296–299, 1960.

Edmiston, C. and Ruedenberg, K., "Localized atomic and molecular orbitals," *Rev. Mod. Phys.* 35:457–464, 1963.

Kohn, W., "Construction of wannier functions and applications to energy bands," *Phys. Rev. B* 7:4388–4398, 1973.

Nenciu, G., "Dynamics of band electrons in electric and magnetic fields: rigorous justification of the effective hamiltonians," *Rev. Mod. Phys.* 63:91, 1991.

Wannier, G., "Dynamics of band electrons in electric and magnetic fields," *Rev. Mod. Phys.* 34:645, 1962.

Maximally localized Wannier functions:

Marzari, N. and Vanderbilt, D., "Maximally localized generalized wannier functions for composite energy bands," *Phys. Rev. B* 56:12847–12865, 1997.

(See also Ch. 22 and Ch. 23.)

Exercises

21.1 This exercise is to construct a localized Wannier function for the s bands described in Sec. 14.4 and Exercises 14.5 and 14.6. The hamiltonian has only nearest-neighbor matrix elements t and the basis is assumed to be orthogonal. For all cases (line, square, and simple cubic lattices), show that one can choose the periodic part of the Bloch functions $u_i^{\mathbf{k}}(\mathbf{r})$ to be real, in which case they are independent of $\mathbf{k}$. Next, show from the definition, Eq. (21.3), that this choice leads to the most localized possible Wannier function, which is identical to the basis function.

21.2 This exercise is to analyze Wannier functions for s bands as described in Exercise 21.1, except that the basis is non-orthogonal with nearest-neighbor overlap s. Show that one can choose the periodic part of the Bloch functions $u_i^{\mathbf{k}}(\mathbf{r})$ to be real, and find the $\mathbf{k}$ dependence of $u_i^{\mathbf{k}}(\mathbf{r})$ as a function of s. (Hint: For non-orthogonal functions the normalization coefficient given in (14.3) and Exercise 14.3 is $\mathbf{k}$ dependent, which in constructing the Wannier function using Eq. (21.3).) Show that the resulting Wannier function has infinite range; even though it decays rapidly, its amplitude does not vanish at any finite distance.

21.3 Derive Eq. (21.7) using definition (21.3) and properties of the eigenfunctions.

21.4 Show that expression (21.11) to second order in $\mathbf{q}$ follows in analogy to the expansion that leads to Eq. (21.10).

21.5 Show that Ω_I is always positive by noting that $\hat{Q} = \hat{Q}^2$ so that Eq. (21.17) can be written as the sum of expectation values of squares of operators.

21.6 Show that (21.21) and (21.22) have the same limit for $\mathbf{b} \to 0$ as Eqs. (21.18) and (21.19) and that they obey the translation invariance conditions, Eq. (21.20). Show further that this means that Ω is unchanged.

21.7 Show that Ω_I in Eq. (21.23) is positive using the definition of the $\mathbf{M}$ matrices and the fact that the overlap of Bloch functions at different $\mathbf{k}$ points must be less than unity.

21.8 Explain why it is not possible to make $\widetilde{\Omega}$ vanish in higher dimensions.
(a) First show that the *projected* operators $\{\hat{P}\hat{x}\hat{P},\ \hat{P}\hat{y}\hat{P},\ \hat{P}\hat{z}\hat{P}\}$ do not commute. Show this is equivalent to the statement that $\hat{x}$ and $\hat{P}$ do not commute. Then show that $\hat{x}$ and $\hat{P}$ do not commute.
(b) Use the fact that non-commuting operators cannot be simultaneously diagonalized to complete the demonstration.

21.9 Demonstrate that it is possible to find functions with $\widetilde{\Omega} = 0$ in one dimension by explicitly minimizing $\widetilde{\Omega}$ for a one-band, nearest-neighbor tight-binding model with overlap (see definitions in Sec. 14.4):

$$H_{i,i\pm1} = t; \quad S_{i,i\pm1} = s, \tag{21.34}$$

where $S_{i,i} = 1$.
(a) First consider the case with $s = 0$: show that in this artificial model the minimum spread is the spread of the basis function. However, one can also choose more delocalized states, e.g. the eigenstates.
(b) For $s \neq 0$, find the minimum spread Ω_I as a function of t. Show it is greater than in part (a). For explicit evaluation of the spread Ω_I, use (21.23) with the eigenvectors given by analytic solution of the Schrödinger equation and the sum over k done approximately on a regular grid of values in one dimension.

21.10 This exercise is to construct maximally localized Wannier functions for the one-dimensional ionic dimer model in Exercise 14.12 using the fact that the gauge-dependent term in it can be made to vanish.
(a) Let $t_1 = t_2$ so that each atom is at a center of symmetry. Show that the maximally localized Wannier function for the lower band is centered on the atom with lower energy ε_A or ε_B, and the function for the upper band is centered on the atom with higher energy. (Hint: If there is a center of inversion the periodic part of the Bloch functions can be made real.)
(b) Similarly, there are two centers of inversion if $\varepsilon_A = \varepsilon_B$ and $t_1 \neq t_2$. Show that in this case the Wannier functions are centered respectively on the strong and the weak bonds between the atoms.
(c) In each of the cases above, calculate the maximally localized Wannier function as a sum of localized basis functions. The eigenfunctions can be calculated analytically and the Wannier functions constructed using the definition in (21.3) and approximating the integral by a sum over a regular grid of k points. (This can be done with a small computer code. Note that the grid spacing must be small for a small gap between the bands.)

21.11 Using the model of Exercise 14.12 and the methods described in Sec. 21.3, construct a computer code to calculate the centers of the Wannier functions in a general case, $\varepsilon_A \neq \varepsilon_B$ and $t_1 \neq t_2$.

This can be used to find polarization and effective charges as described in Exercises 22.8 and 22.9.

21.12 Construct the maximally localized Wannier function for the lowest band in the one-dimensional continuum model of Exercise 12.5. Show that the function is centered at the minimum of the potential. Calculate the functions using the analytic expressions for the Bloch functions and the same approach as in Exercise 21.10, part (c).

21.13 See Exercise 15.6 for a project to construct Wannier functions in one dimension.

21.14 Show that $\mathbf{U}^{\mathbf{k}}$, defined in Eq. (21.25), is unitary if $\mathbf{W}^{\mathbf{k}}$ is antihermitian, i.e. $W_{ij} = -W_{ji}^{*}$.

21.15 Show that the gradient, Eq. (21.29), follows from the definitions. To do this, verify the operator commutation relations, and note that $\mathrm{Tr}[\hat{X}'\hat{X}_D] = 0$, etc.

21.16 Show that the minimization of the functional $\Omega[A] + C_a \min\{\Omega_a[A, X]\}$ leads to the desired solution of a non-singular transformation to non-orthogonal orbitals. Hint: Use the conditions stated following Eq. (23.33) and the relations given in (21.30)–(21.33).

22

Polarization, localization, and Berry's phases

Summary

Electric polarization is one of the basic quantities in physics, essential to the theory of dielectrics, effective charges in lattice dynamics, piezoelectricity, ferroelectricity, and other phenomena. However, descriptions in widely used texts are often based upon oversimplified models that are misleading or incorrect. The basic problem is that the expression for a dipole moment is ill defined in an extended system, and there is no unique way to find the moment as a sum of dipoles by "cutting" the charge density into finite regions. For extended matter such as crystals, a theory of polarization formulated directly in terms of the quantum mechanical wavefunction of the electrons has only recently been derived, with an elegant formulation in terms of a Berry's phase and alternative expressions using Wannier functions. The other essential property of insulators is "localization" of the electrons. Although the concept of localization is well known, recent theoretical advances have provided new quantitative approaches and demonstrated that localization is directly measurable by optical experiments. This chapter is closely related to Ch. 21 on Wannier functions, in particular to the gauge-invariant center of mass and contribution to the spread of Wannier functions Ω_I of Sec. 21.3.

The theory of electrodynamics of matter [448, 790] (see App. E) is cast in terms of electric fields $\mathbf{E}(\mathbf{r}, t)$ and currents $\mathbf{j}(\mathbf{r}', t')$. (Here we ignore response to magnetic fields.) In metals, there are *real currents* and, in the static limit, electrons flow to screen all macroscopic electric fields. Thus, the description of the metal divides cleanly into two parts: the bulk, which is completely unaffected by the external fields, and surface regions, where there is an accumulation of charge $\delta n(\mathbf{r})$ that adjusts to bring the surfaces to an equipotential. The surface thus determines the *absolute value* of the potential in the interior relative to vacuum (see Sec. 13.4), but this has no affect upon any physical properties intrinsic to the bulk interior of the metal.

The fundamental definition of an insulator, on the other hand, is that it can support a static electric field. In insulators, charge cannot flow over macroscopic distances, but there can be time-dependent currents termed *polarization currents*. The state of the material is

defined by the polarization field $\mathbf{P}(\mathbf{r}, t)$ which satisfies the equation

$$\nabla \cdot \mathbf{P}(\mathbf{r}, t) = -\delta n(\mathbf{r}, t), \tag{22.1}$$

or, using the conservation condition $\nabla \cdot \mathbf{j}(\mathbf{r}, t) = -\mathrm{d}n(\mathbf{r}, t)/\mathrm{d}t$,

$$\frac{\mathrm{d}\mathbf{P}(\mathbf{r}, t)}{\mathrm{d}t} = \mathbf{j}(\mathbf{r}, t) + \nabla \times \mathbf{M}(\mathbf{r}, t), \tag{22.2}$$

where $\mathbf{M}(\mathbf{r}, t)$ is an arbitrary vector field. The theory of dielectrics [448, 790] is based upon the existence of local constitutive relations of $\mathbf{P}(\mathbf{r}, t)$ to the macroscopic electric field, atomic displacements, strain, etc. (see also Ch. 19).

The first part of this chapter addresses the problem of the definition of polarization in condensed matter. This is treated in some detail because there has been great confusion for many years that has been satisfactorily resolved only recently. The issue is the definition of the static macroscopic polarization $\mathbf{P}$, i.e. the average value of $\mathbf{P}(\mathbf{r})$, as an *intrinsic* property of the bulk of an insulating crystal, i.e. with no dependence upon surface termination. The essential question for electronic structure theory is: *can one determine the macroscopic polarization* $\mathbf{P}$ *in terms of the intrinsic bulk ground state wavefunction*? This is the fundamental problem if we want to find proper theoretical expressions for the polarization in a ferroelectric or pyroelectric material. Expressions for energy, force, magnetization,[1] and stress have been given in previous chapters; electric polarization completes the set of properties needed to specify the macroscopic state of insulators. In addition, expressions valid to all orders in perturbation theory provide an alternative to the response function approach of Ch. 19.

Traditional textbooks [84, 86, 88, 448, 790, 791] are little help:[2] ionic crystals are represented by point charge models and polarization is considered only in approximate models, such as the Clausius–Mossotti model of a solid as a collection of polarizable units. However, the electron density $n(\mathbf{r})$ is a continuous function of $\mathbf{r}$ and there is no way of finding a unique value of $\mathbf{P}$ as a sum of dipole moments of units derived by "cutting" the density into parts [786]. Attempts to make such identifications have lead to much confusion and claims that properties such as piezoelectric constants are not true bulk properties (see [787] for a review). The polarization of ferroelectrics or pyroelectrics is even more problematic [784].

The resolution to these issues and an elegant – yet practical – quantum mechanical formulation has only recently been derived. This places polarization firmly in the body of electronic structure theory. As shown in Sec. 22.1, the key steps are to relate *changes* in the polarization to integrals over currents flowing through the interior of the body. This is the basis for the new developments (Sec. 22.2) that express the integrated current as a geometric Berry's phase involving integrals over the *phases of the electronic wavefunctions*. This was realized by King-Smith and Vanderbilt [147] who built upon the earlier work of

[1] Only spin was treated explicitly; orbital magnetization is a difficult problem that requires special treatment as does polarization.

[2] An exception is Marder [88], whose presentation is based upon the recent theoretical advances described in more detail here.

Thouless and coworkers [792–794]. For reviews see Resta [148, 788] and for the extension to interacting many-body systems see Ortiz and Martin [795].

The formulation of polarization in terms of "phases" of the wavefunctions has led to a re-examination of density functional theory, since the density is independent of the phases. The problem was pointed out by Gonze, Ghosez, and Godby [340], and the resolution is very subtle and can only be summarized here. In the absence of a macroscopic electric field, the bulk polarization is, in principle, a *functional* of the bulk density in the spirit of the original Hohenberg–Kohn theorem since the wavefunction is also a functional of the density. But if there is a macroscopic electric field, the state of the bulk is *not determined by the bulk density alone* [796] but can be written in terms of a "density polarization theory"([341,342] and references cited there).

The properties of an insulator are fundamentally related to *localization* [783]. Recent work [775, 789, 797, 798] has shown how to express polarization and localization in a unified way in terms of the ground state wavefunction. A summary of this work is given in Sec. 22.5, including useful explicit formulas for the localization length in an insulator, which are experimentally measurable [775, 798] and which reduce to the invariant part of the spread of the Wannier functions, Eq. (21.17), in the independent-particle approximation.

22.1 Polarization: the fundamental difficulty

In a finite system, as shown in Fig. 22.1 there is no problem in defining the average value of the polarization $\mathbf{P}$. From Eq. (22.1) one can integrate by parts and use the requirement that $\mathbf{P}(\mathbf{r}) = 0$ outside the body (see Exercise 22.1) to express $\mathbf{P}$ in terms of the total dipole moment $\mathbf{d}$,

$$\mathbf{P} \equiv \frac{\mathbf{d}}{\Omega} = \frac{1}{\Omega} \int_{\text{all space}} \mathrm{d}\mathbf{r}\, n(\mathbf{r})\, \mathbf{r}. \tag{22.3}$$

This integral is well defined since the density vanishes outside the finite system and there is no difficulty from the factor $\mathbf{r}$. This expression has the desired properties: in particular, a change in polarization $\Delta\mathbf{P} = \mathbf{P}^{(1)} - \mathbf{P}^{(0)}$ is given strictly in terms of the density difference

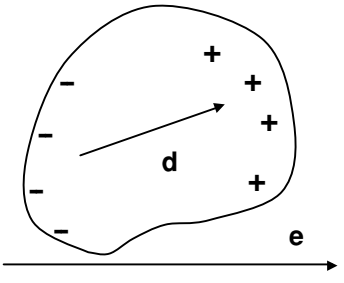


Figure 22.1. Illustration of finite system for which the total dipole moment is well defined. However, the total dipole cannot be used to find the bulk polarization in the large system limit because there is a surface contribution that does not vanish. A bulk theory must be cast solely in terms of bulk quantities.

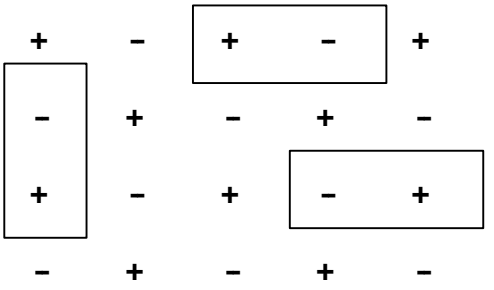


Figure 22.2. Point charge model of an ionic crystal. The dipole is obviously not unique since the cells shown all have different moments. However, a change in moment is the same for all cells so long as charges do not cross boundaries.

$\Delta n = n^{(1)} - n^{(0)}$ independent of the path along which the density changed in going from the starting point 0 to the end point 1.

In an extended system, the goal is to identify an intrinsic bulk polarization $\mathbf{P}^{\text{bulk}}$. The first step is to specify what is meant by "intrinsic bulk." A well-defined thermodynamic reference state of a bulk material can be specified by requiring the macroscopic electric field $\mathbf{E}_{\text{mac}}$ to vanish. This is the *only* well-defined reference state [84, 86, 88, 562] since it is only in this case that the bulk is not influenced by *extrinsic* charges at long distance. With this requirement, the electrons in a crystal are in a periodic potential, which is a fortunate situation for the derivation of the theoretical value for the polarization. (Of course, this is not the whole story: there are real physical effects due to long-range electric fields, which is the subject of the dielectric theory of insulators [448, 790]. The complete description requires a full quantum mechanical theory involving both the thermodynamic reference state, with $\mathbf{E}_{\text{mac}} \equiv \mathbf{0}$, and changes in the presence of macroscopic electric fields that can be treated by perturbation theory (App. E and Ch. 19).

In an extended system, any interpretation of polarization based upon Eq. (22.3) suffers from a fatal difficulty originating in the factor of the position vector $\mathbf{r}$, which is unbounded. This is illustrated for a periodic array of point charges in Fig. 22.2. Suppose we consider a large finite system of charges that repeat the pattern shown. Depending upon the termination of the charges there can be a surface contribution due to factor $\mathbf{r}$ that remains even in the infinite system limit. If we attempt to consider only one unit cell, then the choices shown in Fig. 22.2 illustrate three choices all with different moments. Another approach is the Clausius–Mossotti-type models where the material is assumed to be a set of localized "molecule-like" densities, each of which has a moment and is polarizable (see [84], Chap. 27, and [86], Chap. 13). Even though all such models are at first sight oversimplifications, we shall see that they can be derived from well-defined theoretical approaches.

The first step is to note that if we try to apply the ideas of the simplified models directly to the charge density, it is impossible to find the polarization (or changes in the polarization) simply from the density [786]. This is due to the fact that the electron charge density is continuous and there is no direct way to "cut" it into pieces that is unique. Thus a different approach has to be used. The next step is to note that *changes* in the polarization in the point

charge model are well defined except for jumps when charges cross the boundary. This is true only because of the assumption of point charges, which turns out to be important in the Wannier function interpretation of Sec. 22.3, which will also give ways to construct localized overlapping densities that can play the role of the units in the Claussius–Mosotti model.

A proper definition of polarization in an infinite periodic crystal requires that the expression in terms of the *ill-defined* $\mathbf{r}$ operator be replaced by a different form. This can be done using the relation to the current Eq. (22.2), in which case the static changes in the polarization can be calculated from *adiabatic* evolution of the system with "time" replaced by a parameter λ that characterizes the evolution (e.g. λ might represent positions of atoms). Thus a *change* in polarization $\Delta\mathbf{P}$ can be determined strictly from the polarization current that flows *through* the bulk [786, 799–801]. Since the macroscopic current is a physically measurable *unique* quantity this provides a well-defined procedure for calculating the changes in polarization as a purely bulk property. Indeed the change $\Delta\mathbf{P}$ is the quantity that appears in the fundamental definition of polarization, given in (E.5) and repeated here,

$$\mathbf{P}(\mathbf{r}, t) = \int^{t} \mathrm{d}t' \, \mathbf{j}_{\mathrm{int}}(\mathbf{r}, t'). \tag{22.4}$$

To clarify that the integrated current is the desired quantity, we must carefully specify the experimentally measurable quantity that can be identified as a physical polarization. As shown schematically in Fig. 22.3, the basic experimental measurement is a *current that flows through an external circuit under the conditions that the internal macroscopic field vanishes*. This makes it clear that *changes in polarization* $\Delta\mathbf{P}$ are the quantities measured directly. Since the current is physically measurable this expression eliminates the difficulties that occur if one tries to extend (22.3) to an infinite system. As examples of physical

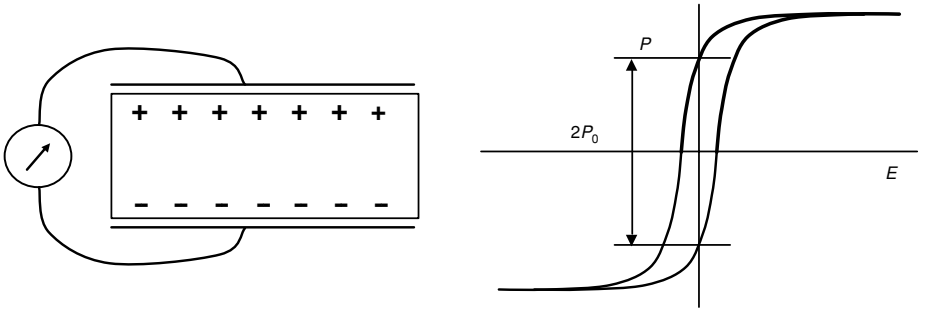


Figure 22.3. Left: Schematic illustration for measurement of a change in polarization $\Delta\mathbf{P}$. In order to keep the two surfaces at the same potential (i.e. zero macroscopic field), the integrated current that flows in the external circuit exactly balances any change in surface charge. Since the surface charge is given by (see Eq. (22.1)) $\Delta n_{\mathrm{surface}} = -\int_{\mathrm{surface}} \boldsymbol{\nabla} \cdot \Delta\mathbf{P}(\mathbf{r}) = \Delta\mathbf{P}_{\mathrm{bulk}}$, this is a direct measure of $\Delta\mathbf{P}$. Right: Schematic hysteresis loop for a ferroelectric showing that a change in polarization is the quantity actually measured to determine the remnant permanent polarization P_0 for zero macroscopic electric field E.

measurements, piezoelectric constants are changes $\Delta\mathbf{P}$ caused by a strain ϵ_{ij} measured under conditions where there is no flow of internal free charges that could "short out" the external circuit [802].[3] Ferroelectrics are materials whose state can be switched by an external field, so that the magnitude of the intrinsic polarization $|\mathbf{P}|$ can be found as one-half the measured change $\Delta\mathbf{P}$ between states of opposite remnant polarization.

There is still one feature missing. Up to now we have only used definitions and dielectric theory. The expression, Eq. (22.4), for polarization is correct but not sufficient. There is no proof or physical reasoning that shows that the value of the polarization is independent of the path in the integral. Thus as it stands, Eq. (22.4) is not acceptable as a definition of an intrinsic bulk property. Quantum theory provides the needed proof: this is the subject of the next section.

22.2 Geometric Berry's phase theory of polarization

Recently [147, 148, 795], there has been a breakthrough providing a new approach for calculation of polarization in crystalline dielectrics. Within independent-particle approaches, all physical quantities can be written as integrals over the filled bands in the complete Brillouin zone taking advantage of periodicity in $\mathbf{k}$ space. The change in polarization can be found when a parameter of the hamiltonian, λ, is changed adiabatically (e.g. when atoms are displaced which leads to a Kohn–Sham potential $V_{\mathrm{KS}}^{\lambda}$) from definition (22.4) with time t replaced by the parameter λ,

$$\Delta\mathbf{P} = \int_0^1 \mathrm{d}\lambda \, \frac{\partial \mathbf{P}}{\partial \lambda}, \tag{22.5}$$

where the macroscopic electric field is required to vanish at all λ.

The analysis starts with the perturbation expression for $\partial\mathbf{P}/\partial\lambda$ in terms of momentum matrix elements that are well defined in the infinite system

$$\frac{\partial \mathbf{P}}{\partial \lambda} = -i\frac{e\hbar}{\Omega m_e}\sum_{\mathbf{k}}\sum_{i}^{\mathrm{occ}}\sum_{j}^{\mathrm{empty}} \frac{\langle\psi_{\mathbf{k}i}^{\lambda}|\,\hat{\mathbf{p}}\,|\psi_{\mathbf{k}j}^{\lambda}\rangle\langle\psi_{\mathbf{k}j}^{\lambda}|\,\partial V_{\mathrm{KS}}^{\lambda}/\partial\lambda\,|\psi_{\mathbf{k}i}^{\lambda}\rangle}{(\varepsilon_{\mathbf{k}i}^{\lambda}-\varepsilon_{\mathbf{k}j}^{\lambda})^2} + \text{c.c.}, \tag{22.6}$$

where the sum over i, j is assumed to include a sum over the two spin states. This expression can be cast in a form involving only the occupied states following the approach of Thouless and coworkers [792] (see also Ch. 19) using the transformed $\mathbf{k}$-dependent hamiltonian, $\hat{H}(\mathbf{k},\lambda)$, whose eigenfunctions are the strictly periodic part of the Bloch functions $u_{\mathbf{k}i}^{\lambda}(\mathbf{r})$ as expressed in (4.37). The relations required are (Exercise 22.2)

$$\langle\psi_{\mathbf{k}i}^{\lambda}|\hat{\mathbf{p}}|\psi_{\mathbf{k}j}^{\lambda}\rangle = \frac{m_e}{\hbar}\langle u_{\mathbf{k}i}^{\lambda}|[\partial/\partial\mathbf{k},\,\hat{H}(\mathbf{k},\lambda)]|u_{\mathbf{k}j}^{\lambda}\rangle \tag{22.7}$$

[3] There is a distinction between "proper" and "improper" piezoelectricity, the latter being the change of moment when a material with a permanent moment is rotated. Only the former is a real response of the material [785,803] as is clear in Fig. 22.3 where obviously nothing happens if the sample and electrodes are rotated. Vanderbilt [803] has shown that the Berry's phase expressions below give the proper terms since the $\mathbf{G}$ vectors rotate with the crystal.

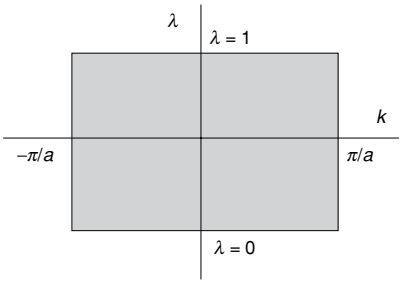


Figure 22.4. Schematic figure of region of integration in (k, λ) space for calculation of $\Delta\mathbf{P}$ using the Berry's phase formula Eq. 22.9.

and

$$\langle \psi^{\lambda}_{\mathbf{k}i} | \partial V^{\lambda}_{\mathrm{KS}} / \partial\lambda | \psi^{\lambda}_{\mathbf{k}j} \rangle = \langle u^{\lambda}_{\mathbf{k}i} | [\partial/\partial\lambda, \hat{H}(\mathbf{k}, \lambda)] | u^{\lambda}_{\mathbf{k}j} \rangle. \tag{22.8}$$

Substituting in (22.6) and using completeness relations (Exercise 22.3) leads directly to the result [147, 795] for the electronic contribution to $\Delta\mathbf{P}$

$$\Delta\mathbf{P}_{\alpha} = -|e| \frac{2}{(2\pi)^3} \,\mathrm{Im} \int_{\mathrm{BZ}} \mathrm{d}\mathbf{k} \int_0^1 \mathrm{d}\lambda \sum_i^{\mathrm{occ}} \left\langle \frac{\partial u^{\lambda}_{\mathbf{k}i}}{\partial k_{\alpha}} \,\middle|\, \frac{\partial u^{\lambda}_{\mathbf{k}i}}{\partial\lambda} \right\rangle. \tag{22.9}$$

The integrals over the two directions of $\mathbf{k}$ perpendicular to k_{α} are done simply as averages and we focus only upon the one-dimensional integral over $k \equiv k_{\alpha}$.

The general nature of Eq. (22.9) is made clear by realizing that it is a "Berry's phase" [149].[4] The key point is that Eq. 22.9 involves a two dimensional integral over parameters k and λ in the $\hat{H}(\mathbf{k}, \lambda)$. These parameters play the role of the slowly changing parameters in the approach of Berry [149]. By defining a reduced dimensionless vector $(a/2\pi)k$ (where a has dimensions of length), the right-hand side is easily shown to be a factor $[ea/\text{volume}]$ multiplied by a dimensionless quantity which is gauge-independent and is precisely the Berry's phase. The two-dimensional region in (k, λ) space is shown in Fig. 22.4, and using Stokes' theorem the surface integral can be converted into a line integral along the closed path defined as the boundary of the region. Because the parameters form a two-dimensional space, the line integral can be defined and the area enclosed by the path defines a phase [148, 149].

Although the complete expression depends upon the path in (k, λ) space [795], a choice of phases of the wavefunctions to obey the "periodic gauge condition," Eq. 21.5, which can be written

$$u^{\lambda}_{\mathbf{k}+\mathbf{G},i}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}\mathbf{G}\cdot\mathbf{r}} u^{\lambda}_{\mathbf{k},i}(\mathbf{r}), \tag{22.10}$$

where $\mathbf{G}$ is a reciprocal lattice vector, leads to a cancellation of the contribution of the two integrals over λ at $\mathbf{k}$ and $\mathbf{k} + \mathbf{G}$. This leads to the simplest "two-point" formula [147, 148]

[4] See discussions in [148], [795], and [342].

that depends only upon the difference of the integrals over $\mathbf{k}$ at the end-points $\lambda = 0$ and $\lambda = 1$,[5]

$$\Delta \mathbf{P}_\alpha = i \frac{-|e|}{(2\pi)^3} \int_{\mathrm{BZ}} \mathrm{d}\mathbf{k} \sum_i^{\mathrm{occ}} \left[\langle u_{\mathbf{k}i}^{\lambda=1} | \partial_{k_\alpha} u_{\mathbf{k}i}^{\lambda=1} \rangle - \langle u_{\mathbf{k}i}^{\lambda=0} | \partial_{k_\alpha} u_{\mathbf{k}i}^{\lambda=0} \rangle \right] + (\text{integer}) \times \frac{-|e|}{A}, \tag{22.11}$$

where A is the cell volume divided by the length of the unit cell in the direction α, i.e. the area of a cell perpendicular to α. The last term in (22.11) represents "quanta of polarization" that originate in integer multiples of 2π in the Berry's phase. Interestingly, they can be interpretated as transport of an integer number of electrons across the entire crystal, leaving the bulk invariant. This is the part of the transport that was emphasized by Thouless et al., for the quantum Hall effect [792] and quantized charge transport in an insulator [793, 794] in the case where the hamiltonian is changed along a closed path returning to the same point, i.e. $\hat{H}(\mathbf{k}, \lambda = 1) = \hat{H}(\mathbf{k}, \lambda = 0)$. In contrast, changes in the Berry's phase by fractions of 2π correspond to polarization of the bulk crystal.

Note that the geometric phase is non-zero only if the periodic functions $u_{\mathbf{k}i}^{\lambda}$ are complex; this occurs if there is no center of inversion, which is of course exactly the condition under which there may be a non-zero polarization. Hence, *the change in macroscopic polarization between two different insulating states can be regarded as a measure of the phase difference between the initial and final wavefunctions*. In all mean-field approaches, this means Slater determinants of single-body functions $u_{\mathbf{k}i}^{\lambda}$, but the formula generalizes directly to correlated many-body wavefunctions [795].

For actual calculations in crystals, it is convenient to express the polarization terms of the Bloch functions calculated on a grid in the Brillouin zone, rather than the derivatives required in Eq. (22.11). The grid can be constructed with lines of J points in the α direction along which the derivative is to be calculated. The Bloch functions at the two sides on the BZ ($\mathbf{k}_0$ and $\mathbf{k}_J$) are required to be the same, i.e. the periodic gauge. It is not sufficient to simply express the derivative as a finite difference because this would not be gauge invariant. As shown in [147], a possible choice is to replace the integral over $\mathbf{k}_\alpha$ by

$$-i \int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}_\alpha \sum_i^{\mathrm{occ}} \left[\langle u_{\mathbf{k}i}^{\lambda} | \partial_{k_\alpha} u_{\mathbf{k}i}^{\lambda} \rangle \right] \to \Im \left[\ln \Pi_{j=0}^{J-1} \det \left(\langle u_{\mathbf{k}_j i}^{\lambda} | u_{\mathbf{k}_{j+1} i'}^{\lambda} \rangle \right) \right], \tag{22.12}$$

where the determinant is that of the $N \times N$ matrix formed by allowing i and i' to range over all occupied states. Note that the overall phase of each $u_{\mathbf{k}_j i}^{\lambda}$ cancels in the sum since each function appears in a bra and a ket in the product. Expression (22.12) involving discrete points $\mathbf{k}_j$, approaches the continuum expression, (22.11), in the limit of $J \to \infty$. This form is now widely used for calculations of polarization, as exemplified below.

[5] Spin is included in the sum over i. Some authors include a factor of 2 for spin, assuming no spin dependence.

22.3 Relation to centers of Wannier functions

The "two-point" expression for a change in polarization, Eq. (22.11), due to the electrons can be immediately cast in terms of functions involving the centers of Wannier functions, using relation (21.8). The result is

$$\Delta \mathbf{P}_\alpha = \frac{-|e|}{\Omega} \sum_i^{\text{occ}} \left[\langle 0i|\hat{\mathbf{r}}|0i\rangle^{\lambda=1} - \langle 0i|\hat{\mathbf{r}}|0i\rangle^{\lambda=0} \right]. \tag{22.13}$$

This has the simple interpretation that the change in polarization is the same as if the electrons were localized at points corresponding to the centers of the Wannier functions. In general, the center of each function is *not* unique, but the sum of moments of all the functions is unique as shown by Blount [759] and which follows from the gauge invariance of the Berry phase.

Thus the derivation in terms of Wannier functions provides a rigorous basis for the simplified models of charges in crystals. For example, the model in terms of point charges can be identified with the centers of the Wannier functions; each charge is not unique, but any change in polarization is well defined, modulo the "quantum of polarization." Furthermore, the "quantum" has the simple interpretation that the electrons can be shifted by a translation between equivalent Wannier functions, which of course is a symmetry operation of the infinite crystal. Similarly, the Clausius–Mossotti model becomes a rigorous theory of polarization in insulators (at least in the independent-particle approximation) if the polarizable units are taken to be overlapping Wannier functions. Even though the units are not unique, the total polarization is well defined.

22.4 Calculation of polarization in crystals

As examples of the calculation of polarization it is appropriate to consider effective charges, given by Eq. (E.20), because the results can be compared with experiments that can measure the charges accurately in terms of the splittings of longitudinal and transverse modes. For example, several groups [573, 736, 737] have derived the anomalous effective charges in perovskites, such as $BaTiO_3$, which have ferroelectric transitions. In such materials there are several infrared (IR) active modes and it is not possible to determine directly from the experiment the individual effective charges of the atoms, because all the IR modes interact with one another and are mixed. They can be decomposed into contributions from different atoms only by using information on the lattice dynamics from a theoretical model. In contrast, the *ab initio* calculations determine the atomic effective charges and the vibrational modes; thus the effective charges for the eigenmodes can be directly predicted and compared with experiment. The prediction is that there is a great mixing of the IR modes, so that the lowest transverce optic (TO) mode is most closely associated with the highest longitudinal optic (LO) mode, giving a very large effective charge for that mode, i.e. the phonon that softens at the ferroelectric phase transition. The anomalously large effective charges of the B

Table 22.1. Born effective charges for the atoms and mode effective charges for the IR active modes in ABO_3 perovskites. The two charges given for the O atoms are, respectively, for displacements in the plane formed by the O atom and 4B neighbors, $Z_1^*(O)$, and in the perpendicular direction toward the 2A neighbors, $Z_2^*(O)$. Taken from Zhong, et al. [573].

Type	$BaTiO_3$	$PbTiO_3$	$NaNbO_3$
Z*(A)	2.75	3.90	1.13
Z*(B)	7.16	7.06	9.11
$Z_1^*(O)$	−5.69	−5.83	−7.01
$Z_2^*(O)$	−2.11	−2.56	−1.61
|Z*(TO1)|	8.95	7.58	6.95
|Z*(TO2)|	1.69	4.23	2.32
|Z*(TO3)|	1.37	3.21	5.21

atoms and the O atoms moving along the line toward the B atoms are interpreted as resulting from covalency. Selected results are shown in Tab. 22.1, taken from [573]; essentially the same results have been found using linear response methods [736, 737]; however, Berry's phase approach has the advantage that it applies directly to the finite polarizations that develop in the ferroelectric states.

Examples of calculations of linear (Gonze et al. [804]) and non-linear susceptibilities (Dal Corso and Mauri [805]) have been done using transformations between the Wannier and Bloch orbitals and the "$2n + 1$ theorem" (Sec. 3.7).

Spontaneous polarization

Spontaneous polarization occurs in any crystal that lacks a center of inversion. In ferroelectrics, the value P_0 can be measured as indicated in Fig. 22.3 because the direction of the polarization can be reversed. Values of the ferroelectric remnant polarization have been calculated for a limited number of ferroelectric materials, with values in general agreement with measured ones [148]. For example, the calculated residual polarization [736] of $KNbO_3$ is $\Delta\mathbf{P} = 0.35$ C/m^2 compared with the measured value [806] of 0.37 C/m^2, although it is very difficult to find the intrinsic moment experimentally [784].

In other crystals, such as wurtzite structure, there is a net asymmetry with the positive direction of the c-axis inequivalent from the negative direction. The two directions are detected experimentally by the fact that their surfaces are inequivalent. It is not so easy to reverse this axis since it requires breaking and remaking of all the bonds in the crystal. Such crystals are pyroelectrics [722] because the change in the polarization with temperature can be measured. How can the absolute value of the polarization be determined? The absolute

value can be expressed in terms of quantities that can be defined theoretically: the difference $\Delta\mathbf{P}$ given the Berry's phase formula between the acual crystal and a crystal in which $\mathbf{P} = 0$ by symmetry. This can be calculated by constructing a path between the two crystals by "theoretical alchemy", in which the charge on the nuclei is varied, or by large displacements of atoms to change the crystal structure. Both of these possibilities are straightforward in calculations. Although there is no corresponding direct experiment, there are experimental consequences which occur at an interface between regions with different polarization, so that $\boldsymbol{\nabla} \cdot \mathbf{P}$ leads to a net charge.

22.5 Localization: a rigorous measure

An insulator is distinguished from a conductor at zero temperature by its vanishing d.c. conductivity and its ability to sustain a macroscopic polarization, with and without an applied electric field [448, 790]. The theory of polarization, presented thus far, has shown the fundamental relation of the latter property to the ground state wavefunction for the electrons. Regarding the former property, the classic paper "Theory of the insulating state" by W. Kohn [783] has clarified that the many-body system of electrons in an insulator is "localized" in contrast to the delocalized state in a metal. However, until recently there was no rigorous quantitative measure of the degree of localization. In fact, such a measure is provided by the theory of polarization: not only the average value, but also the fluctuations of the polarization [775, 788, 789].

The relation between polarization and localization was established by Kudinov [807], who proposed to measure the degree of localization in terms of the *mean square quantum fluctuation of the ground state polarization*. Kudinov considered the quantum fluctuation of the net dipole moment, $\langle \Delta\hat{\mathbf{d}}^2 \rangle = \langle \hat{d}^2 \rangle - \langle \hat{\mathbf{d}} \rangle^2$ in a large, but finite, volume Ω. Using the zero-temperature limit of the fluctuation–dissipation theorem [263, 808–811], the mean square fluctuation is related to the linear response function by

$$\frac{\langle \Delta \hat{d}_\alpha^{\,2} \rangle}{V} = \frac{\hbar}{\pi} \int_0^\infty \mathrm{d}\omega \frac{1}{\omega} \mathrm{Re}\sigma_{\alpha\alpha}(\omega) = \frac{\hbar}{\pi} \frac{1}{4\pi} \int_0^\infty \mathrm{d}\omega \mathrm{Im}\epsilon_{\alpha\alpha}(\omega). \tag{22.14}$$

For a metal with $\sigma(\omega) \neq 0$ for $\omega \to 0$, the right-hand side diverges, i.e. the mean square fluctuation of the dipole moment diverges in the large Ω limit. However, for an insulator $\sigma(\omega = 0)$ is finite and Kudinov proposed that the integral has a well-defined limit for large volume. Since the dipole is a charge times a displacement, $\hat{\mathbf{d}} = -e\hat{\mathbf{X}}$, where $\hat{\mathbf{X}} = \sum_i^N \hat{\mathbf{x}}_i$ is the center of the mass position operator of N electrons in the volume, relation (22.14) can be used to define a mean square displacement of the electrons, and thus a localization length ξ. Souza et al. [775] have shown that the arguments carry over to the infinite system with proper interpretation of the position operator $\mathbf{X}$ consistent with the polarization operator,[6]

[6] That is, the expectation value $\langle \mathbf{X} \rangle$ is equivalent [775] to the Berry's phase expressions given in Sec. 22.2.

with the result for the length in the α direction

$$\xi_\alpha^2 = \lim_{N\to\infty} \frac{1}{N}\left[\langle \hat{X}_\alpha^2\rangle - \langle \hat{X}_\alpha\rangle^2\right] = \lim_{N\to\infty} \frac{\Omega^2}{e^2 N}\left[\langle \hat{P}_\alpha^2\rangle - \langle \hat{P}_\alpha\rangle^2\right]. \tag{22.15}$$

In terms of measurable conductivity, ξ_α^2 is given by

$$\xi_\alpha^2 = \frac{\hbar}{\pi e^2 n}\int_0^\infty \mathrm{d}\omega \frac{1}{\omega}\mathrm{Re}\sigma_{\alpha\alpha}(\omega). \tag{22.16}$$

Bounds can be placed upon the length [775, 798]. Using the inequality

$$\int_0^\infty \mathrm{d}\omega \frac{1}{\omega}\mathrm{Re}\sigma_{\alpha\alpha}(\omega) \le \frac{1}{E_{\mathrm{gap}}^{\min}}\int_0^\infty \mathrm{d}\omega \mathrm{Re}\sigma_{\alpha\alpha}(\omega), \tag{22.17}$$

where $E_{\mathrm{gap}}^{\min}$ is the minimum direct gap and the sum rule, Eq. (E.13), lead to an upper bound

$$\xi_\alpha^2 \le \frac{\hbar^2}{2m_e E_{\mathrm{gap}}^{\min}}. \tag{22.18}$$

On the other hand, arguments based upon standard perturbation formulas for the static susceptibility $\chi = (\epsilon - 1)/4\pi$ lead to a lower bound [798]

$$\xi_\alpha^2 \ge \frac{E_{\mathrm{gap}}^{\min}}{2n}\chi, \tag{22.19}$$

which also illustrates that ξ diverges in a metal where χ necessarily diverges. The inequalities can be derived, as described in more detail in Exercise 22.5, in terms of integrals of the real part of the conductivity $\sigma'(\omega)$ and the average gap defined by Penn [812] in terms of the electron density and the polarizability.

The localization length also relates to important theoretical quantities, establishing rigorous relations with experimental measurables and providing tests for approximate theories. The demonstrations have been done in several ways [775, 813] (see a review in [788]). For particles that are independent except that they are indistinguishable, there is a general relation, Eq. (3.55), between the correlation function and the density matrix, repeated here for fermions

$$\Delta n_{\mathrm{ip}}(\mathbf{x};\mathbf{x}') = -|\hat{\rho}_\sigma(\mathbf{x},\mathbf{x}')|^2. \tag{22.20}$$

As shown by Sgiarovello et al. [813], transformations of the density matrix lead to the relation [775]

$$\sum_{\alpha=1}^{3}\xi_\alpha^2 = \frac{\Omega_I}{m_e}, \tag{22.21}$$

in terms of the invariant part of the spread of the Wannier functions, Eq. (21.14), thus giving a physical meaning to the invariant spread.

The localization length can be determined directly from the ground state wavefunction using the expressions in Sec. 21.3 for Eq. (21.14) or alternative forms given in [813].

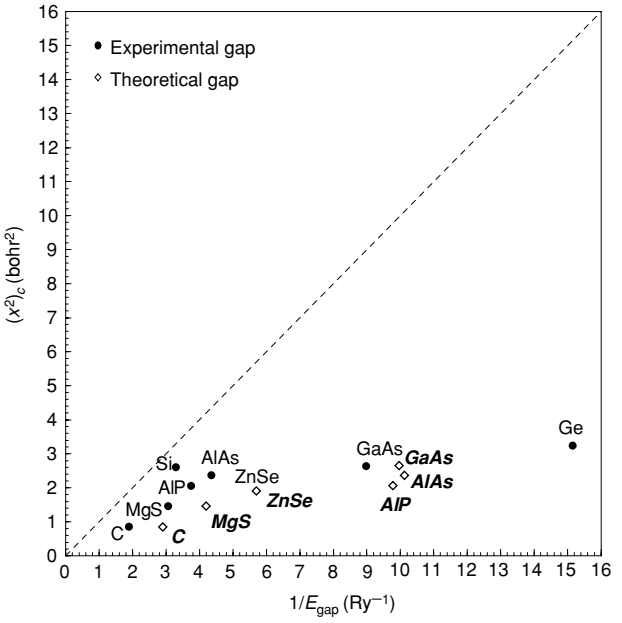


Figure 22.5. Calculated [813] mean square fluctuation of the electron center of mass $\langle x^2 \rangle = \xi_\alpha^2$ plotted versus the inverse of the minimum direct gap in various semiconductors. Two points for each material (except Si and Ge, see text) are for experimental (solid dots) and theoretical (open circles) gaps. The line indicates the upper bound that would apply if all the oscillator strength were associated the minimum gap. It is clear that inequality (22.18) is well obeyed. Calculations were done using expressions for $\langle x^2 \rangle$ that can be expressed in terms of Eq. (22.21) and a variation of the expressions given in Sec. 21.3 for Eq. (21.14). From [813].

Quantitative calculations for many semiconductors have been carried out by [813] with the results shown in Fig. 22.5, compared to the bounds in Eq. (22.18). Note that the figure has two points for each material, representing the experimental minimum direct gap and the theoretical gap in the actual density functional theory calculations. (The theoretical points for Si and Ge are not shown because they are too far off scale due to the "gap problem.") The inequality must be obeyed for the theoretical gap and the figure shows that it is well obeyed for the experimental gap as well.

22.6 Geometric Berry's phase theory of spin waves

The Berry's phase approach for electric polarization can be extended to spin as well [135, 136]. The basic idea is that the spin wave is assumed to be adiabatic and the wavefunction for the electrons is considered to evolve adiabatically in time with a position-dependent Berry's phase. The derivatives of the phase contribute to the energy, as well as to all other spatial variations of the wavefunction. Practical expressions can be worked out in spin-dependent density functional theory (including non-collinear spins) for the energy and

detailed spin distribution of a spin wave. An example of a calculation [136] is shown in Fig. 19.5 from [720].

SELECT FURTHER READING

Kohn, W., "Theory of the insulating state," *Phys. Rev.* 133:A171–181, 1964. A classic paper on the insulating state.

Lines, M. E., and Glass, A. M., *Principles and Applications of Ferroelectrics and Related Materials*, Clarendon Press, Oxford, 1977. A general reference on ferroelectricity.

Issue involved in defining polarization in extended matter:

Martin, R. M., "Comment on: Piezoelectricity under hydrostatic pressure," *Phys. Rev.* B 6:4874, 1972.

Martin, R. M., "Comment on: Calculation of electric polarization in crystals," *Phys. Rev.* B 9:1998, 1974.

Tagantsev, A. K., "Review: Electric polarization in crystals and its response to thermal and elastic perturbations," *Phase Transitions* 35:119, 1991.

Geometric Theory:

King-Smith, R. D., and Vanderbilt, D., "Theory of polarization in crystalline solids," *Phys. Rev.* B 47:1651–1654, 1993.

Martin, R. M., and Ortiz, G., "Recent developments in the theory of polarization in solids," *Solid State Commun.* 102:121–126, 1997.

Resta, R., "Macroscopic polarization in crystalline dielectrics: the geometric phase approach," *Rev. Mod. Phys.* 66:899–915, 1994.

Resta, R., "Why are insulators insulating and metals conducting?" *J. Phys.: Condens. Matter* 14:R625–R656, 2002.

See also:

Resta, R., "The quantum mechanical position operator in extended systems," *Phys. Rev. Lett.* 80:1800–1803, 1998.

Souza, I., Wilkens, T. J., and Martin, R. M., "Polarization and localization in insulators: generating function approach," *Phys. Rev.* B 62:1666–1683, 2000. (Gives general formulation of polarization and localization.)

Exercises

22.1 Verify that the well-known expression for a dipole moment, Eq. (22.3), follows from the definition of the polarization field, Eq. (22.1), with the boundary condition given.

22.2 Show that expressions (22.7) and (22.8) for the expectation values in terms of the commutators follow from the definition of $\hat{H}(\mathbf{k}, \lambda)$.

22.3 Show that Eq. (22.9) follows from the previous equations as stated in the text. Hint: Use completeness relations to eliminate excited states.

22.4 Show that the dipole moment averaged over all possible cells vanishes in any crystal.

22.5 Define a "localization gap" $E_{\rm L}$ by turning Eq. (22.18) into an equality: $\xi^2 \equiv \hbar^2/(2m_e E_{\rm L})$.

(a) Using the f sum rule and Eq. (22.16), show that E_L can be expressed as the first inverse moment of the optical conductivity distribution:

$$E_L^{-1} = \frac{1}{\hbar} \frac{\int \omega^{-1}\sigma'(\omega)\,\mathrm{d}\omega}{\int \omega^0\sigma'(\omega)\,\mathrm{d}\omega}.$$

(b) Use the f sum rule and the Kramers–Krönig expression for $\epsilon(0)$ to show that the Penn gap [812] $E_{\rm Penn}$ defined via the relation $\epsilon(0) = 1 + (\hbar\omega_p/E_{\rm Penn})$, where ω_p is the plasma frequency, can be expressed as the second inverse moment:

$$E_{\rm Penn}^{-2} = \frac{1}{\hbar^2} \frac{\int \omega' - 2\sigma'(\omega)\,\mathrm{d}\omega}{\int \omega^0\sigma'(\omega)\,\mathrm{d}\omega}.$$

(c) Using the results of (a) and (b), show that inequalities (22.18) and (22.19) can be recast in a compact form as follows:

$$E_{\rm Penn}^2 \geq E_{\rm L}\, E_{\rm gap}^{\rm min} \geq (E_{\rm gap}^{\rm min})^2.$$

22.6 Find a reasonable estimate and upper and lower bounds for $\sum_{\alpha=1}^{3} \xi_\alpha^2$ using gaps and dielectric constants of typical semiconductors and the expressions given in Exercise 22.5. The lowest direct gaps can be taken from Fig. 22.5 and values of the dielectric functions can be found in texts such as [84, 86, 88], e.g. $\epsilon \approx 12$ in Si.

22.7 It is also instructive to calculate values of the average "Penn gap" [812] which is defined in Exercise 22.5. The Penn gap is an estimate of the average gap in the optical spectrum and is directly related to the inequalities in Exercises 22.5 and 22.6. As an example, find the gap in Si with $\epsilon \approx 12$ and compare with the minimum direct gap. Find values for other semiconductors as well, using standard references or [812].

22.8 Construct a small computer code to calculate the electronic contribution to the polarization from the Berry's phase expressions given in Sec. 22.2 for the one-dimensional ionic dimer model of Exercise 14.12 in a general case, $\varepsilon_A \neq \varepsilon_B$ and $t_1 \neq t_2$. Compare with the calculations of the centers of the Wannier function found in Exercise 21.11.

22.9 The effective charge, Eq. (E.20), is defined by the change in the polarization induced by displacement on an atom. An important part of the charge is the "dynamical" electronic contribution that results from changes in the electronic wavefunctions in addition to rigid displacements. A simple model for this is given by the one-dimensional ionic dimer model of Exercise 14.12. Consider $\varepsilon_A \neq \varepsilon_B$ and let $t_1 = t + \delta t$ and $t_2 = t - \delta t$. A change in δt causes a change in polarization in addition to any change due to rigid displacement of ionic charges. For a given $\Delta\varepsilon \equiv \varepsilon_A - \varepsilon_B$ calculate $\delta\mathbf{P}/\delta t$ for small δt using computer codes for the Berry's phase (Exercise 22.8) or the centers of the Wannier functions (Exercise 21.11). Show that the contribution to the effective charge can be large and have either sign (depending upon the variation of t with displacement), which can explain large "anomalous effective charges" as described in Sec. 22.4.

22.10 Consider the one-dimensional continuum model of Exercise 12.5 for which Wannier functions are found in Exercise 21.12. The polarization is zero since the crystal has a center of inversion and the eigenfunctions can be chosen to be real. If the entire crystal is shifted rigidly a

distance Δx, so that $V(x) \to V_0 \cos(2\pi(x - \Delta x)/a)$, the origin is not at the center of inversion and the eigenfunctions are not real. Using the Berry's phase expressions in Sec. 22.2, show that the change in the electronic contribution to the polarization is $\Delta P = -2|e|\Delta x/a$. The interpretation of this simple result is that the electrons simply move rigidly with the potential. Give the reasons that this shift does not actually lead to a net polarization since in a real crystal this electronic term is exactly cancelled by the contribution of the positive nuclei which shift rigidly.

23

Locality and linear scaling O(N) methods

Nearsightedness

W. Kohn

Throwing out k-space

V. Heine

Summary

The concept of localization can be imbedded directly into the methods of electronic structure to create new algorithms that take advantage of locality or "nearsightedness" as coined by W. Kohn. As opposed to the textbook starting point for describing crystals in terms of extended Bloch eigenstates, many physical properties can be calculated from the *density matrix* $\rho(\mathbf{r}, \mathbf{r}')$, which is exponentially localized in an insulator or a metal at finite T. For large systems, this fact can be used to make "order-N" or O(N) methods where the computational time scales linearly in the size of the system. There are two aspects of the problem: "building" the hamiltonian and "solving" the equations. Here we emphasize the second aspect, which is more fundamental, and describe representative O(N) approaches that either treat $\rho(\mathbf{r}, \mathbf{r}')$ directly or work in terms of Wannier-like localized orbitals.

The reader should be aware (beware) that O(N) methods are under development; there are problems and shortcomings in actual practice.

Every textbook on solid state physics begins with the symmetry of crystals and the entire subject of electronic structure is cast in the framework of the eigenstates of the hamiltonian classified in k space by the Bloch theorem. So far this volume is no exception. However, the real goal is to understand the properties of materials from the fundamental theory of the electrons and this is not always the best approach, either for understanding or for calculations.

What does one do when there is no periodicity? At a surface? In an amorphous solid or a liquid? In a large molecule? One approach is to continue to use periodic boundary conditions on artificial "supercells" chosen to be large enough that the effects of the boundary conditions are small or can be removed from the calculation by an analytic extrapolation procedure.

This approach can be very effective and is widely used, especially with efficient plane wave methods (Ch. 13) and "*ab initio*" simulations (Ch. 18).

The subject of this chapter is an alternative approach built upon the general principle of *locality* or *nearsightedness*, i.e. that properties at one point can be considered independent of what happens at distant points. If the theory is cast in a way that takes advantage of the locality, then this will lead to algorithms that are "linear scaling" ("order-N" or O(N)), i.e. the computational time and computer memory needed is proportional to N as the number of particles N is increased to a large number. O(N) methods emerge naturally in classical mechanics: if there are only short-range interactions, the forces on each particle depend only upon a small number of neighboring particles. One step in a molecular dynamics calculation can be done updating the position of each particle in time $\propto N$.[1] The same conclusion holds even if there are long-range Coulomb forces, since there are various methods to sum the long-range forces in time $\propto N$ [820].

The problem is that quantum mechanics inherently is *not* O(N). The full description of the states of a quantum system is *not* local: the solutions of the wave equation, in general, depend upon the boundary conditions; eigenstates in extended systems, in general, are extended;[2] and the indistinguishability of identical particles requires that the wavefunctions obey the symmetry or antisymmetry conditions among all particles, whether they are nearby or far away. Features such as critical points in the band structure, a sharp Fermi surface in k space, Kohn anomalies [84], *etc.*, all require extended quantum mechanical waves. The simplest case is the ground state of N electrons in the independent-particle approximation, which is an antisymmetric combination of N eigenstates, each of which is, in general, extended through a volume also proportional to N. Working in terms of the independent-particle eigenstates leads to scaling at least $\propto N^2$, and specific methods are often worse. Full matrix diagonalization scales as N_{basis}^3, where the number of basis functions $N_{\text{basis}} \propto N$. The widely used Car–Parrinello-type algorithms in a plane wave basis scale as $N^2 N_{\text{basis}}$, with $N_{\text{basis}} >> N$ in the orthogonalization step. Solutions of correlated many-body problems, in general, scale much worse, growing exponentially in N for exact solutions and as high powers for practical configuration interaction calculations [819]. Widely used variants of quantum Monte Carlo simulations [81, 95] of ground state or equilibrium properties have the same scaling as independent-particle methods, but with a larger prefactor in the computational time. Recently, a linear-scaling version [821] has been introduced.

23.1 Locality and linear scaling in many-particle quantum systems

Despite the inherent non-locality in quantum mechanics, many important properties can be found without calculating the eigenstates, using information that is only "local" (as defined below). For example, the density and total energy are integrated quantities that are invariant

[1] It is a more difficult question if all desired properties can be found in time proportional to the total size of the system; for example, there could be slow relaxation modes that become increasingly difficult to determine as the system size increases. We will not deal with such issues here.

[2] In disordered systems, some or all eigenstates may be localized, in which case the usual eigenstate methods can be O(N). Even then it may be possible to transform to more localized representation.

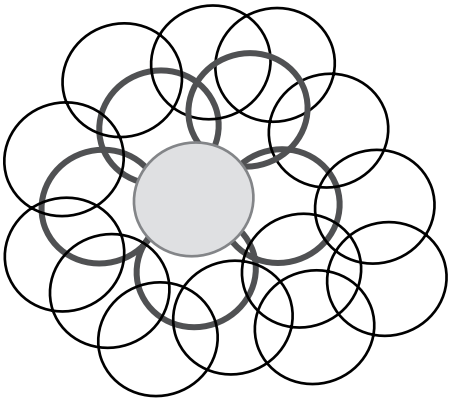


Figure 23.1. Schematic diagram of treating a quantum system in terms of overlapping regions.

to unitary transformations of the states, and these quantities are sufficient to determine the stable ground state and the force on every nucleus. In this chapter we discuss algorithms that actually calculate such quantities with computational time $\propto N$.

The term "nearsightedness" has been coined by Walter Kohn [822] for such integrated quantities that can be calculated at one point $\mathbf{r}$ in terms only of information at points $\mathbf{r}'$ in a neighborhood of $\mathbf{r}$. This embodies the ideas developed by Friedel [697] and Heine and coworkers [465] on locality and other work such as the 1964 paper by Kohn, "The nature of the insulating state," in which the key idea is the localization of electronic states in insulators. Nearsightedness is a property of a many-body system of particles: the density of an individual eigenstate at any point is dependent upon the boundary conditions and the potential at all other points; however, for systems of many particles, the net effect is reduced due to interference between the different independent-particle eigenstates (i.e. in the sum in Eqs. (23.1) and (23.2) below). In insulators and metals at non-zero temperature, the one-electron density matrix decays exponentially, and in metals at $T = 0$ as a power law ($1/R^3$ in three-dimensions), with Gibbs oscillations due to the sharp cutoff at the Fermi surface. Interactions introduce correlations with the longest range being van-de-Waals-type that decay as $1/R^6$ in the energy and $1/R^3$ in the wavefunction [819]. We will use the term "local" in this sense to mean "dependent upon only distant regions to an extent which decays rapidly:" exponential decay is sufficient to ensure convergent algorithms; however, the power law decay in metals at $T = 0$ is problematic so that special care is needed.

There are a number of different linear-scaling O(N) approaches, all of which take advantage of the decay of the density matrix with distance, and which truncate it at some point. The schematic idea is shown in Fig. 23.1. One of the great advantages of the density matrix formalism is that it is a general approach applicable at finite temperature, where all correlations become shorter range. Therefore, in general, the range of the density matrix is decreased and O(N) algorithms should become more feasible and efficient. In addition, the possibility of continuous variation of fractional occupation of states is of great advantage in problematic cases where occupation must be fractional by symmetry or jumps discontinuously

at $T = 0$. This problem is well known in standard electronic structure algorithms and is discussed in Ch. 9, where a fictitious temperature is often added to improve calculations by smoothing details of the state distribution.

For non-interacting particles, the density matrix can be written (see Sec. 3.5) as

$$\hat{\rho} = \sum_{i=1}^{M} |\psi_i\rangle f_i \langle\psi_i| \text{ or } \rho(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{M} \psi_i^*(\mathbf{r}) f_i \psi_i(\mathbf{r}'), \tag{23.1}$$

where $f_i = 1/(1 + \exp(\beta(\epsilon_i - \mu))$ is the Fermi function and $\beta = 1/k_B T$. For $T \neq 0$, the number of states M must be greater than the number of electrons N, which is related to the Fermi energy μ by $N = \sum_{i=1}^{M} f_i$. At $T = 0$ this becomes

$$\hat{\rho} = \sum_{i=1}^{N} |\psi_i\rangle\langle\psi_i| \text{ or } \rho(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{N} \psi_i^*(\mathbf{r})\psi_i(\mathbf{r}'). \tag{23.2}$$

Even if the independent-particle eigenfunctions are extended, the density matrix is exponentially localized and vanishes at large relative separation $|\mathbf{r} - \mathbf{r}'|$ in all cases for $T \neq 0$ (see discussion following Eq. (5.10)) and at $T = 0$ in an insulator.

The sum of independent-particle energies E_s, written in terms of eigenstates, is given by

$$E_s = \sum_{i=1}^{M} \frac{1}{1 + \exp(\beta(\epsilon_i - \mu))} \epsilon_i, \tag{23.3}$$

where the sum is over all eigenvalues. (This is also sufficient for Kohn–Sham theory (Sec. 9.2) where the total energy can always be found from E_s plus terms that involve only the density $n(\mathbf{r}) = \rho(\mathbf{r}, \mathbf{r})$.) In general, one can rewrite the sum over all eigenvectors as a trace, so that energy can be written

$$E_s = \mathrm{Tr}\left\{ \frac{1}{1 + \exp(\beta(\hat{H} - \mu))} \hat{H} \right\} = \mathrm{Tr}\{\hat{\rho}\hat{H}\}. \tag{23.4}$$

Similarly, the grand potential that describes the energy for different numbers of particles is given by

$$\Omega_s = \mathrm{Tr}\left\{ \frac{1}{1 + \exp(\beta(\hat{H} - \mu))} (\hat{H} - \mu) \right\} = \mathrm{Tr}\{\hat{\rho}(\hat{H} - \mu)\}. \tag{23.5}$$

The difficulty in using the Fermi function is the exponentiation of the operators that are non-diagonal, if one wishes to avoid diagonalization, e.g. in an O(N) scheme.

At $T = 0$ in an insulator, the expressions are closely related to the construction of Wannier functions. The N eigenstates can be transformed to N localized orthogonal Wannier-like functions w_i (Ch. 21)

$$\hat{\rho} = \sum_{i=1}^{N} |w_i\rangle\langle w_i|, \quad T = 0. \tag{23.6}$$

It may also be advantageous [780, 823] to work in terms of non-orthogonal functions $\hat{w}_i$

that are transformations of the Wannier functions

$$\hat{\rho} = \sum_{i=1,j}^{N} |w_i\rangle S^{-}_{ij} \langle w_j|, \quad T = 0, \tag{23.7}$$

where S^{-} is the inverse of the overlap matrix.[3] For $T \neq 0$ the form of the decay can be derived for model problems, such as the free-electron gas discussed in Sec. 5.1.

For all cases, the challenge is to determine efficient, robust ways to find the density matrix $\hat{\rho}$ or the generalized Wannier functions w_i or $\tilde{w}_i$. In the latter case, the functions are not unique, which leads to possible advantages that can accrue by using particular choices and possible problems due to approximations that violate the invariance of the functionals.

There are two aspects to the creation of linear-scaling methods, both relying upon sparsity of the hamiltonian and overlap matrices, assumed to have non-zero elements only for a finite range as shown schematically in Fig. 23.1:

- "Building" the hamiltonian, i.e. generating the non-zero matrix elements in a sparse form that is linear in the size of the system. For many approaches, this is the rate limiting step for sizes up to hundreds of atoms, and therefore it is relevant even if the solution is done with traditional $O(N^2)$ or $O(N^3)$ methods.
- "Solving" the equations. This is the more fundamental aspect that is necessary for O(N) scaling in the large N limit. We will consider approaches that treat $\rho(\mathbf{r}, \mathbf{r}')$ (or a Green's function) or work in terms of Wannier-like localized orbitals. The key division is between "non-variational methods" that truncate well-known expansions, and "variational methods" that approximate the solution of variational functionals.

23.2 Building the hamiltonian

The hamiltonian can be constructed in a sparse form in real space in any case in which the basis functions are localized to regions smaller than the system size. The most obvious basis for such an approach are local orbitals, as discussed in Chs. 14 and 15. The tight-binding model approach is ideal for this purpose since the matrix elements of $\hat{H}$ and the overlap matrix $\hat{S}$ are defined to be short range. In the full local orbital method, matrix elements of the hamiltonian vanish beyond some distance only if the orbitals are strictly localized. This can be accomplished in a basis of numerical orbitals purposefully chosen to be strictly localized as described in Sec. 15.4. In general, analytic bases such as gaussians decay exponentially but never vanish entirely, so they must be used with care.

The augmented approaches can also be cast in localized forms. LMTOs can be transformed to an orthogonal tight-binding form, in which the hamiltonian has a power law decay (Sec. 17.5) that is much shorter range than in the original method. In Green's function methods, such as KKR, G is generated in terms of G_0, which is very long range for positive energies, but decays exponentially for negative energies. This has been used to construct an O(N) KKR method [671] in which G_0 is the numerically calculated Green's

[3] Here we use the notation of [780] which generalizes the definition of S^{-1} as shown later in Eq. (23.28).

function for an electron in an array of repulsive centers, as described in Sec. 16.7. In these methods one must invert a matrix labeled by the orbital quantum numbers at the atomic sites, and linear scaling is accomplished in constructing the hamiltonian or Green's function matrix including only neighbors in a "local interaction zone." The approach, termed "locally self-consistent multiple scattering" (LSMS), involves a calculation on a finite cell around each site; an alternative approach to use Lanczos or recursion methods to solve the multiple scattering problem around each site [824].

It is, however, *not essential* that the basis be localized. One of the original ideas is due to Galli and Parrinello [705], who combined the plane wave Car–Parrinello algorithm (Ch. 18) with transformations of the wavefunctions to a localized form as in Eq. (23.5). Physical properties are unchanged due to the invariance of the trace. By constraining orbitals to be localized to regions, they showed that one can construct an algorithm that automatically generates localized functions and a sparse hamiltonian, even though the plane wave basis is not localized.

23.3 Solution of equations: non-variational methods

Green's functions, recursion, and moments

The original ideas for electronic structure methods that take advantage of the locality grew out of Green's function approaches, using the facts that the density matrix and the sum of eigenvalues are directly expressible in terms of integrals over the Green's function. The basic relations are given in Sec. D.4 and in Eqs. (16.34)–(16.36), which we re-write here in slightly different notation. If the basis states are denoted χ_m (here assumed to be orthonormal for simplicity), then local density of states projected on state m is given by

$$n_m(\varepsilon, \mathbf{R}) = -\frac{1}{\pi}\mathrm{Im}G_{m,m}(\varepsilon + i\delta). \tag{23.8}$$

For example, m might denote a site and basis orbital on that site. The sum of eigenvalues of occupied states projected on state m can be found from the relation

$$\sum_i \varepsilon_i |\langle i|\chi_m\rangle|^2 = -\frac{1}{\pi}\int_{-\infty}^{E_F} \mathrm{d}\varepsilon\ \varepsilon\ \mathrm{Im}G_{m,m}(\varepsilon + i\delta, 0), \tag{23.9}$$

and the total sum of eigenvalues is given by

$$\sum_i \varepsilon_i = -\frac{1}{\pi}\sum_m \int_{-\infty}^{E_F} \mathrm{d}\varepsilon\ \varepsilon\ \mathrm{Im}G_{m,m}(\varepsilon + i\delta, 0). \tag{23.10}$$

The left-hand sides of Eqs. (23.9)–(23.10) are in the standard eigenstate form, whereas the right-hand sides are in the form of Green's functions that can be evaluated locally. This provides all the information needed to determine the total energy and related quantities from integrals over the Green's functions.

Elegant methods have been devised to calculate the Green's functions as local quantities. The basic idea can be illustrated by Fig. 23.1 in which we desire to determine the Green's

function $G_{0,0}(\varepsilon + i\delta)$ for the orbital 0 shown as dark gray. This can be accomplished by repeated applications of the hamiltonian, called recursion [696], which has close relations to the Lanczos algorithm (Sec. M.5). The ideas are used in tight-binding approaches (summarized, e.g. in [593]), KKR Green's function methods (see Sec. 16.3 for the basic ideas and references such as [671] and [824] for O(N) algorithm developments), and LMTO Green's function methods that use recursion (see, e.g. [691]). The diagonal elements of the Green's function can be used to find the charge density and the sum of single particle energies; however, there are difficulties in finding the off-diagonal elements of the Green's function needed for forces. This problem is addressed by "bond-order" potential methods [825–827].

The basic idea of the recursion method [696, 828] is to use the Lanczos algorithm (Sec. M.5) as a method to construct a Green's function, using the properties of a tridiagonal matrix. The Lanczos recursion relation, Eq. (M.9), for the hamiltonian applied to a sequence of vectors,

$$\psi_{n+1} = C_{n+1}[\hat{H}\psi_n - H_{nn}\psi_n - H_{nn-1}\psi_{n-1}], \tag{23.11}$$

generates the set of vectors ψ_n and the coefficients in the tridiagonal matrix (M.10), $\alpha_n = H_{nn}$ and $\beta_n = H_{n-1\,n}$. The normalization constant is readily shown (Exercise 23.1) to be $C_{n+1} = 1/\beta_n$, $n \geq 1$. If the starting vector ψ_0 is a basis state localized at a site,[4] and the hamiltonian is short range, then the algorithm generates a sequence of states in "shells" around the central site. The diagonal part of the Green's function for state 0 is given by the continued fraction [593, 696, 828]

$$G_{0,0}(z) = \cfrac{1}{z - \alpha_0 - \cfrac{\beta_1^2}{z - \alpha_1 - \cfrac{\beta_2^2}{z - \alpha_2 - \cfrac{\beta_3^2}{\ddots}}}} \tag{23.12}$$

where z is the complex energy.

The properties of the continued fraction and a proper termination are the key features of the recursion method, and an introductory discussion can be found in [829]. If the fraction is terminated at level N, then the density of states consists of N delta functions that is useful only for integrations over $G(z)$. If a constant imaginary energy δ is used as a terminator, this is equivalent to a lorentzian broadening of each delta function. However, there are other approaches that are just as simple but much more elegant and physical. One follows from the observation [829] that the coefficients α_n and β_n tend to converge quickly to asymptotic values that can be denoted α_∞ and β_∞. If one sets the coefficients $\alpha_n = \alpha_\infty$ and $\beta_n = \beta_\infty$ for all $n > N$, then one can evaluate the remainder of the fraction analytically. This leads

[4] Here 0 denotes the starting state which can be any state in the basis, e.g. any of the localized atomic-like states on site i.

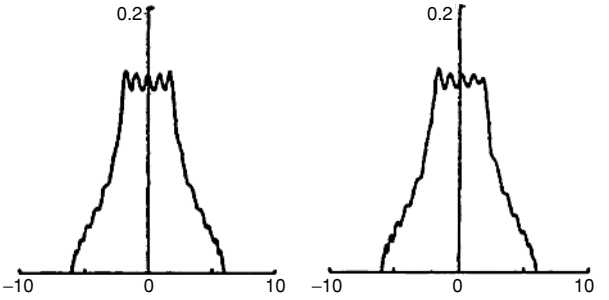


Figure 23.2. Density of states (DOS) for an s band in a simple cubic lattice with nearest-neighbor interactions. The recursion is up to $N = 20$ levels on cells of size $(21)^3$ with open boundary conditions (left) and periodic boundary conditions (right). Compared to the schematic DOS in Fig. 14.3 recursion yields correct features, although there are oscillations. The similarity of the two results demonstrates the insensitivity of local properties to the boundary conditions. (The small asymmetry results from odd-length paths to its image in an adjacent cell.) From [829].

to replacement of the β_{N+1}^2 coefficient by $t(z)$ given by

$$t(z) = \cfrac{1}{z - \alpha_{N+1} - \cfrac{\beta_{N+2}^2}{z - \alpha_{N+2} - \cfrac{\beta_{N+3}^2}{z - \ldots}}} = \frac{1}{z - \alpha_\infty - \beta_\infty^2 t(z)}, \tag{23.13}$$

which has the solution [829] (Exercise 23.3)

$$t(z) = \frac{1}{2\beta_\infty^2}\left\{(z - \alpha_\infty) - \left[(z - \alpha_\infty)^2 - 4\beta_\infty^2\right]^{1/2}\right\}. \tag{23.14}$$

This has the appealing form of a terminator that is an analytic square root function, which is real outside the range $\alpha_\infty \pm 2|\beta_\infty|$ and has a branch cut corresponding to a band width $4|\beta_\infty|$. The ideal choice is that which yields the correct band width. In actual practice it is important not to choose $|\beta_\infty|$ too small in which case there are spurious delta functions in the range between the real band width and the approximate interval $\alpha_\infty \pm 2|\beta_\infty|$. If the range is overestimated, there is broadening but no serious problems.

An example of the use of the recursion method is shown in Fig. 23.2 for the density of states (DOS) for an s band in a simple cubic lattice with nearest-neighbor interactions. The bands are given analytically in Sec. 14.4 and the DOS is shown schematically in Fig. 14.3. In comparison, the DOS calculated with the recursion method up to $N = 20$ levels shows correct features, but with added oscillations. Other types of terminators can improve the convergence to the exact result [696]. The two results shown in Fig. 23.2 are for open and periodic boundary conditions, showing the basic point of the insensitivity of local properties to the boundary conditions.

The recursion method is very powerful and general. It is not limited to tight-binding and it can be applied to *any* hermitian operator to generate a continued fraction form for the Green's function. This is a stable method to generate the density of states so long as care is taken in constructing a proper terminator for the continued fraction [593, 696, 828]. An

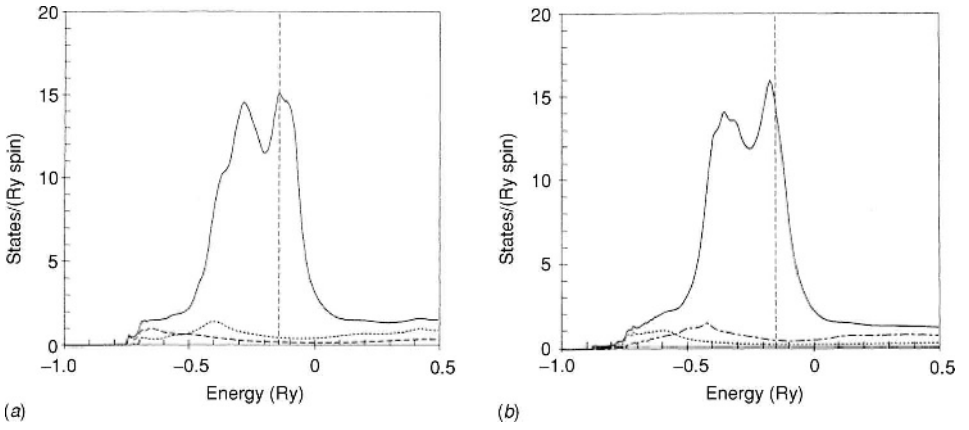


Figure 23.3. Electronic density of states (DOS) for liquid Fe (a) and Co (b) calculated with the tight-binding LMTO method (Sec. 17.5) and recursion [692]. The liquid was simulated by 600 atom cells generated by classical Monte Carlo and empirical interatomic potentials. The density is similar to the crystal (see, for example, the canonical DOS in Fig. 16.13) broadened by the disorder. From [692].

example is shown in Fig. 23.3, which shows the electronic density of states of liquid Fe and Co determined using the tight-binding LMTO method (Sec. 17.5) and recursion [692]. The actual calculations were done on 600 atom cells with atomic positions generated by classical Monte Carlo methods with empirical interatomic potentials. This illustrates a powerful combination of the recursion approach with a basic electronic structure method that is now widely used for many problems in complicated structures and disordered systems.

Determination of the moments of the density of states is also a powerful tool to extract information in an O(N) fashion. Moments of the local DOS for basis function m

$$\mu_m^{(n)} \equiv \langle m|[\hat{H}]^n|m\rangle, \tag{23.15}$$

in principle, contain all the information about the local DOS, and thus about all local properties. The basic ideas are described in Sec. L.7 where the two issues, generating the moments efficiently and inverting the moment information to reconstruct the DOS, are emphasized.

There are a number of ways to generate the moments; one, very stable, approach is to construct the moments in terms of the tridiagonal matrix elements generated by the Lanczos algorithm. For the state labeled 0, the moments can be written [593]

$$\begin{aligned}
\mu_0^{(0)} &= 1, \\
\mu_0^{(1)} &= \alpha_0, \\
\mu_0^{(2)} &= \alpha_0^2 + \beta_1^2, \\
\mu_0^{(3)} &= \alpha_0^3 + 2\alpha_0\beta_1^2 + \alpha_1\beta_1^2, \\
\mu_0^{(4)} &= \alpha_0^4 + 3\alpha_0^2\beta_1^2 + 2\alpha_0\alpha_1\beta_1^2 + \alpha_1^2\beta_1^2 + \beta_1^2\beta_2^2 + \beta_1^4,
\end{aligned} \tag{23.16}$$

and so forth.

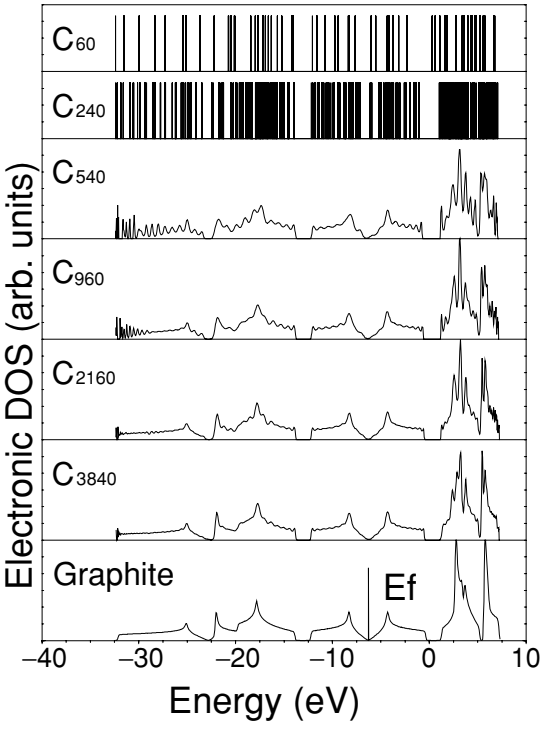


Figure 23.4. Density of states (DOS) for electrons in the fullerenes from C_{60}–C_{3840}. The structures are shown later in Fig. 23.7 and the same tight-binding model of [162] is used in both cases. The figure illustrates the evolution of the DOS from delta functions of the C_{60} molecule to the spectrum of graphene shown in the bottom panel which has the critical point features of a two-dimensional DOS, as shown in Fig. 14.3. The smaller molecules and graphene were computed using diagonalization methods; the DOS for the larger clusters was calculated by the method of moments as implemented in [832] (see Sec. L.7). The results demonstrate the detailed structure that can be resolved using ≈150 moments. From [831].

Inversion of the moments to find the DOS is the well-known, difficult, "classical moment problem" [822] discussed in Sec. L.7. As an example of the construction of the density of states using moments for an actual problem done in O(N) fashion, Fig. 23.4 shows the calculated [831] density of states (DOS) of the series of fullerenes (the structures are shown later in Fig. 23.7) using the orthogonal tight-binding model of [162]. The bottom panel shows the DOS for a sheet of graphene and the progression is evident from the molecule to the continuous spectrum with the critical points of a two-dimensional crystal. The results are found with usual diagonalization for calculations the small molecules and are constructed for the large fullerenes up to C_{3840} from ≈150 moments using the maximum entropy method [832] discussed in Sec. L.7. The results show the exquisite detail that can be achieved, including the sharp features and the approach to the spectrum of graphene calculated exactly for the same tight-binding model. Similarly, the phonon spectrum can be calculated in an O(N) manner as reported for these systems in [833].

Bond order and forces in recursion

A key problem with the recursion method is the difficulty in computing off-diagonal matrix elements of G that are essential for forces. The expression in terms of the density matrix is given in (14.25); omitting the simple two-body term, the contribution from the sum of eigenvalues is given by

$$\mathbf{F}_I = -\mathrm{Tr}\left\{\hat{\rho}\frac{\partial \hat{H}}{\partial \mathbf{R}_I}\right\} = -\sum_{m,m'}\rho_{m,m'}\frac{\partial H_{m,m'}}{\partial \mathbf{R}_I}. \tag{23.17}$$

General expressions in terms of Green's functions with localized bases are given by Feibelman [634]. The generic problem is the calculation of 'bond order," which is the off-diagonal components of the density matrix $\rho_{m,m'}$ that correspond to bonding and are given by integrals over $G_{m,m'}(z)$ analogous to (23.8)–(23.10). There has been considerable work to derive efficient recursion-type expressions for the bond order [825–827]. The basic idea is that off-diagonal terms $G_{m,m'}(z)$ can be calculated using recursion with a starting vector

$$\psi_0 = \frac{1}{\sqrt{2}}[\chi_m + \mathrm{e}^{\mathrm{i}\theta}\chi_{m'}], \tag{23.18}$$

and computing $G_{m,m'}(z)$ from

$$G_{m,m'}(z) = \frac{\partial G^{\lambda}_{0,0}(z)}{\partial \lambda}, \tag{23.19}$$

with $\lambda = \cos(\theta)$. Further generalizations of this idea have been derived as described in [827] with a summary in [593].

"Divide and conquer" or "fragment" method

One of the first O(N) methods is based directly upon the argument that the interior of a large region depends only weakly upon the boundary conditions. The procedure termed "divide and conquer" [834] or "fragment molecular orbital method" [835] is to divide a large system into small subsystems each of size N_{small}, for example the central orbital (gray) plus the set of orbitals shown as heavy circles in Fig. 23.1. For each of these systems one can solve for the electronic eigenstates using ordinary N^3 methods. For each small system one must add "buffer regions" of size N_{buffer} (the outer orbitals in Fig. 23.1) large enough so that the density and energy in the original small subsystem converges and is independent of the buffer termination. The solution for the density and other properties is then kept only for the interior of each small region. In many ways, the "divide and conquer" approach is the counterpart of using supercells: although there is wasted computational effort, the approach is attractive because it uses standard methods and there is much experience in extracting information from small systems with well-chosen terminations. Using traditional methods, the cost is of order $(N_{\mathrm{small}} + N_{\mathrm{buffer}})^3$ for each subsystem, which may be prohibitive, especially for three-dimensional systems where N_{buffer} may need to be very large. Nevertheless, the method is O(N) for large enough systems and it may be particularly applicable for long

linear molecules with large energy gaps (i.e. small localization lengths (see Ch. 22)) that are important in biochemistry (see, e.g. [835]).

Polynomial expansion of the density matrix

One class of methods uses the form in Eq. (23.4) directly and expands the Fermi function in (23.4) in powers of the hamiltonian. This is the approach used, e.g. in [836] and [837]. The basic requirement is for the expansion to have the same properties as the Fermi function, i.e. that all states with eigenvalues far above the Fermi energy μ have vanishing weight, all those well below μ have unity weight, and the variation near μ is reproduced accurately. This is difficult to accomplish in an expansion; however, if the eigenvalues are limited to the range $[E_{\min}, E_{\max}]$, then the expansion can be done efficiently using Chebyshev polynomials T_n defined in Sec. K.5. The advantage of the Chebyshev polynomials is that they are orthogonal and fit the function over the entire range $[-1, +1]$ (see Sec. K.5) and they can be generated recursively using the relation (K.19). Let us define $\Delta E = E_{\max} - E_{\min}$, the scaled hamiltonian operator $\tilde{H} = (\hat{H} - \mu\hat{I})/\Delta E$, and the scaled temperature $\tilde{\beta} = \beta\Delta E$. Then we can express the expansion of the Fermi operator as

$$\hat{F}[\hat{H}] = \frac{1}{1 + \mathrm{e}^{\beta(\hat{H}-\mu)}} \to \frac{c_0}{2}\hat{I} + \sum_{j=1}^{M_p} c_j T_j(\tilde{H}). \tag{23.20}$$

The highest power needed in the expansion depends upon the ratio of the largest energy in the spectrum to the smallest energy resolution required. The higher the temperature the lower the power needed, since the Fermi function is smoother. For a metal, T must be of the order of the actual temperature (or at least smaller than the energy scale on any variations in the states near μ). For insulators, a larger effective T can often be chosen so long as states below (above) the gap are essentially filled (empty). For hamiltonians that are bounded (such as in tight-binding), the ratio $E_{\max}/T$ can be estimated and powers $\approx$10 to 100, are needed for realistic cases.

The key idea is that all operations can be done by repeated applications of $\hat{H}$ to a basis function, which amounts to repeated multiplication of a matrix times a vector. Each of the basis functions shown in Fig. 23.1 is treated one at a time.[5] In the general case, this procedure scales as N^2_{basis}, since it involves multiplication of a vector by the matrix. However, if $\hat{H}$ is sparse, only a few matrix elements are non-zero (e.g. the non-zero elements in a tight-binding hamiltonian that involve only a few neighbors) and the multiplication, times one localized basis function, is independent of the size of the system. Furthermore, if we invoke the localization property of the density matrix, all matrix operations can be made sparse, so that the calculation scales as order $N_{\text{basis}} \propto N$. This method has two great advantages: (1) the computation scales linearly with the number of other basis functions included in the

[5] The Chebyshev polynomial expansion can also be used with random vectors to generate a statistical estimate of extensive quantities, such as the total energy [832, 838]. "maximum entropy" or related methods. This is useful for extremely large matrices where it is not feasible to take the trace.

maximum range, compared to cubic scaling in the "divide and conquer" and the variational function methods; and (2) the algorithm is perfectly parallel. Major disadvantages are: (1) that the results of the expansion are not variational since the truncation errors can be of either sign; and (2) the fact that an independent calculation must be done for each orbital means that information is discarded, in comparison to the variational methods that exchange information between the subparts of the system.

The Chebyshev expansion method can be a very efficient procedure if the basis set is small, e.g. in tight-binding models, where M is only a factor of order 2 larger than N. However, it fails for unbounded spectra, because the polynomial expansion must be taken to higher and higher orders as the range of the spectrum is increased. For a typical plane wave calculation, high powers would be required sine the energy range is large and $N_{\text{basis}} >> N$.

Inverse power expansion of the density matrix

There are several approaches that work with an unbounded spectrum employing operators that properly converge at high energies. One is the inverse power method, which is in essence a Green's function approach, Sec. D.4, and is closely related to the recursion method, Sec. M.5. In this approach, one expands the Fermi function as follows:

$$\hat{F}[\hat{H}] = \frac{1}{1+\exp[\beta(\hat{H}-\mu)]} \rightarrow \sum_{i=1}^{M_{\text{pole}}} \frac{w_i}{\hat{H}-z_i}. \tag{23.21}$$

Using the well-known relations for contour integration, the sum over poles on the real axis can be converted into an integral in the complex plane enclosing the poles [96, 671, 839] as discussed in Sec. D.4 and shown schematically in Fig. D.1. For each of the terms with z_i in the complex plane, the inverse operator $\hat{G}_i(z_i)$ can be found in a basis by solving linear equations $(\hat{H}-z_i)\hat{G}_i(z_i) = I$. In terms of basis functions in real space, the operators $\hat{G}_i(z_i)$ are more localized for large complex z_i, which is advantageous for calculations (Exercise 23.5). Their maximum range is where the contour crosses the real axis. In order to describe the contour integral accurately in an insulator one needs the number of poles $M_{\text{pole}} \propto (\mu - E_{\min})/E_{\text{gap}}$; in a metal there is no gap and an accurate evaluation of the integral near the axis requires poles with a more dense spacing $\propto T$. An advantage of this approach is that, unlike the power expansion, it always converges independent of the high-energy spectrum. Furthermore, it corresponds to the physical picture that high-energy processes are more localized, and low-energy ones more delocalized. An example of multiple-scattering Green's function calculations [671] is shown in Fig. 23.5.

Exponential operators

Perhaps the most fundamental approach of all is to work directly with the time-dependent Schrödinger equation. The density matrix for quantum statistics is $\exp(-\beta\hat{H})$, which is equivalent to the imaginary time propagator. Thus essentially the same techniques can be used as in the real-time methods of Sec. 20.4. As $\beta \rightarrow \infty$, the operation of $\exp(-\beta\hat{H})$ on

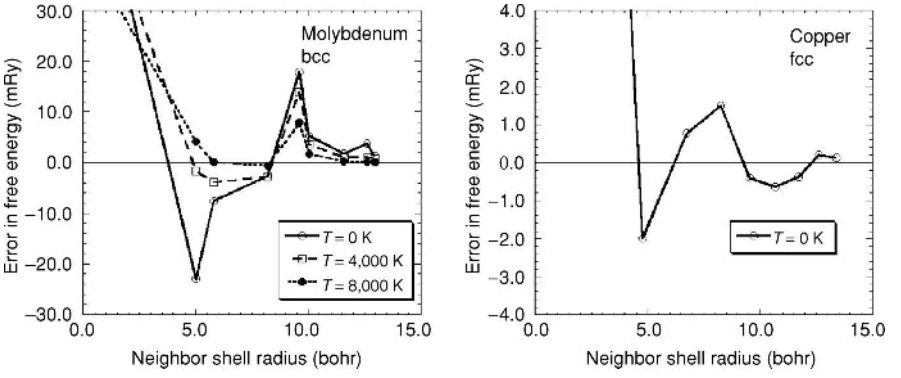


Figure 23.5. Total energy of Cu and Mo calculated using multiple-scattering theory with localized regions, plotted as a function of the radius of the localization region [671]. These are, in fact, hard cases, since the density matrix is most delocalized in perfect crystals, leading to the sharp variations shown. Provided by Y. Wang; similar to [671].

any wavefunction Ψ projects out of the ground state provided it is not orthogonal to Ψ. The expressions can be evaluated using the fact that $\exp(-\beta\hat{H}) = (\exp(-\delta\tau\hat{H}))^n$, where $\delta\tau = \beta/n$ represents a temperature that is higher by a factor n. In the high-temperature, short-time regime, the operations can be simplified using the Suzuki–Trotter decomposition, as in Eq. (20.14) rewritten here,

$$\exp[-\delta\tau(T+V)] \simeq \exp\left(-\frac{1}{2}\delta\tau V\right)\ \exp(-\delta\tau T)\ \exp\left(-\frac{1}{2}\delta\tau V\right), \qquad (23.22)$$

which is factored into exponentials of kinetic and potential terms. One approach for the kinetic term is to use an implicit method that solves linear equations (see Sec. M.10 and [840]). One can also use FFTs to transform from real space (where V is diagonal) to reciprocal space (where T is diagonal) as in Sec. M.11. The former is widely used for quantum dots [840] and the latter has been used in simulations, e.g. of hydrogen fluid at high temperature and pressure [841].

23.4 Variational density matrix methods

Two properties must be satisfied[6] by the density matrix $\hat{\rho}$ at $T = 0$:

- "Idempotency," which literally means $\hat{\rho}^2 = \hat{\rho}$, which is equivalent to requiring all eigenvalues of $\hat{\rho}$ to be 1 or 0.
- The eigenvectors of the density matrix with eigenvalue 1 are the occupied eigenvectors of the hamiltonian.

[6] The density matrix minimization approach can also be extended to finite temperature. Corkill and Ho [842] have used the Mermin functional to allow a continuous variation of the occupation instead of the strict idempotency requirements of the original method.

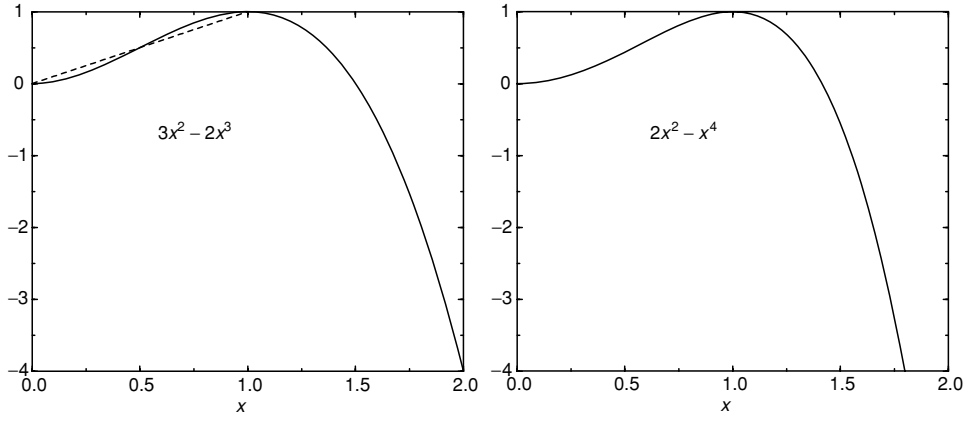


Figure 23.6. Functional forms for the cubic "McWeeny" purification algorithm [843] for the density matrix (left) and the quadratic Mauri–Ordejon–Kim functional [844–846] for the wavefunctions (right). "Purification" of the density matrix results because the function $3x^2 - 2x^3$ is always closer to 0 or 1 than the input value x, as shown by the dashed line ($= x$). The form $2x^2 - x^4$ applied to the localized wavefunctions is minimized for occupied functions normalized to 1, empty functions to 0.

Li et al. [843] showed how to use a minimization method to drive the density matrix to its proper $T = 0$ form. See also Exercise 23.7 for further details. The starting point is the "McWeeny purification [261]" idea: if $\tilde{\rho}_{ij}$ is an approximate trial density matrix with eigenvalues between 0 and 1, then the matrix $3\tilde{\rho}_{ij}^2 - 2\tilde{\rho}_{ij}^3$ is always an improved approximation to the density matrix with eigenvalues closer to 0 or 1. This is illustrated in Fig. 23.6 (left panel) which shows the function $y = 3x^2 - 2x^3$. It is easy to see that for $x < 1/2$, $y < x$, i.e. the occupation is closer to zero, whereas for $x > 1/2$, $y > x$, i.e. the occupation is closer to one. However, if one iterates the matrix using the purification equation alone, there is no reason for the eigenvectors to satisfy the second requirement, i.e. to correspond to the lowest energy states. In order to make a functional which when minimized yields the proper idempotent density matrix that also minimizes the total energy, one can modify the usual expression, Eq. (23.5), for the grand potential at $T = 0$ to use the "purified" form,

$$\Omega_s = \mathrm{Tr}\hat{\rho}(\hat{H} - \mu) \rightarrow \mathrm{Tr}(3\hat{\rho}^2 - 2\hat{\rho}^3)(\hat{H} - \mu). \tag{23.23}$$

Since the functional is minimum for the true density matrix, the energy given by (23.23) is variational.

The functional can be minimized by iteration using the gradients

$$\frac{\partial \Omega_s}{\partial \hat{\rho}} = 3[\hat{\rho}(\hat{H} - \mu) + (\hat{H} - \mu)\hat{\rho}] - 2[\hat{\rho}^2(\hat{H} - \mu) + \hat{\rho}(\hat{H} - \mu)\hat{\rho} + (\hat{H} - \mu)\hat{\rho}^2], \tag{23.24}$$

which denotes the matrix expression for the derivative with respect to each of the elements of the density matrix $\hat{\rho}$. So long as the density matrix is never allowed to go into the unphysical region where the eigenvalues are <0 or >1, then the algorithm is stable and

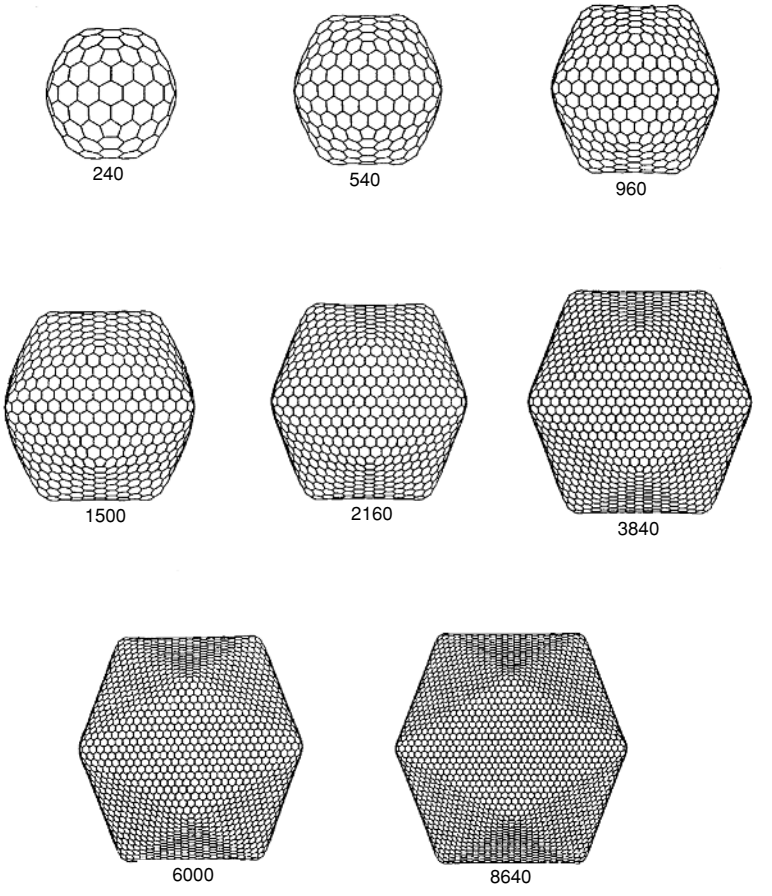


Figure 23.7. Example of giant fullerenes studied by order-N calculations [848]. The shapes are determined by minimizing the tight-binding energy [162] using density matrix purification method [843] and clearly indicate a progression from near-spherical shape for smaller fullerenes to a faceted (polyhedral) geometrical shape made up of graphite-like faces, sharply curved edges, and 12 pentagons at the vertices. Similar results were found by Itoh et al. [849] using the Wannier function method.

the gradients always point toward a lower energy and an improved density matrix. Any matrix satisfying the physical conditions can be used as a starting point. A possible choice is $\rho_{ij} = \delta_{ij}(N_{\text{elec}}/N_{\text{basis}})$, and more optimal choices can be found for any particular problem. The method can be extended to non-orthogonal bases [847]. A principle difficulty with this method is that it requires explicit operations of multiplying matrices which are of the size of the basis N_{basis}. Thus it is appropriate for small bases such as in tight-binding, but not for large bases, such as plane waves where $N_{\text{basis}} >> N_{\text{elec}}$ (but see Sec. 23.7 for alternative methods).

As an example of calculations using the density matrix purification algorithm of [843], Fig. 23.7 shows selected large icosahedral fullerenes (up to C_{3840}) for which the structures

were optimized [848] using the tight-binding potential of Xu et al. [162]. The calculations clearly indicate a progression from near-spherical shape for smaller fullerenes to a faceted (polyhedral) geometrical shape made up of graphite-like faces, sharply curved edges, and 12 pentagons at the vertices. The same conclusions were reached independently [849] using the linear-scaling Wannier function approach (Sec. 23.5).

The density matrix methods [843] can also be employed for MD calculations using the force theorem for the forces in terms of the density matrix and the derivative of the hamiltonian, as given in Sec. 14.8. The same tight-binding potential [162] as employed for the giant fullerenes has been used in simulations of liquid and amorphous carbon. The calculated radial density distribution $g(r)$ of liquid carbon at ordinary pressure agrees well with plane wave Car–Parrinello calculations as shown in Fig. 18.2 and have been done with both usual matrix diagonalization [162] and linear-scaling density matrix methods [850]. In addition, the results are essentially the same as found in [836], which used the Fermi projection operator approach of Sec. 23.3. The combination of tight-binding and linear-scaling methods has allowed calculations on larger sizes and longer times than is feasible using other methods.

23.5 Variational (generalized) Wannier function methods

A different approach is to work with the localized wavefunctions in (23.2) rather than with the density matrix itself. However, in searching for the Wannier-like functions it is not convenient to require the constraint of orthonormality implicit in (23.2). One possibility is to work with non-orthogonal functions and use the correct general expression

$$E_{\text{total}} = \text{Tr}\hat{\rho}\hat{H} = \sum_{i,j=1}^{N} S_{ij}^{-1} H_{ji}, \tag{23.25}$$

where S is the overlap matrix. Here matrices are defined by $H_{ji} = \langle \tilde{w}_j | \hat{H} | \tilde{w}_i \rangle$ (or with $\tilde{w} \to w$ for orthogonal functions), *etc.* However, the entire problem can be rewritten in a more advantageous form through the invention of a new class of functionals [844, 845], the simplest of which is

$$\tilde{E}_{\text{total}} = \sum_{i=1}^{N} H_{ii} - \sum_{i,j=1}^{N} (S_{ij} - \delta_{ij}) H_{ji} = \sum_{i,j=1}^{N} (2\delta_{ij} - S_{ij}) H_{ji}, \tag{23.26}$$

where the terms $S_{ij} - \delta_{ij}$ are like Lagrange multipliers that replace the constraint of orthonormality. The special property of $\tilde{E}$ is that $\tilde{E}_{\text{total}} \geq E_{\text{total}}$ for all wavefunctions whether or not they are normalized or orthogonal. Since $\tilde{E}_{\text{total}} = E_{\text{total}}$ for the orthonormal functions, it follows that one can minimize $\tilde{E}_{\text{total}}$ with respect to the wavefunctions with no constraints, leading to orthonormal functions at the minimum.

The behavior of $\tilde{E}_{\text{total}}$ considered as a functional of the wavefunctions can be seen by expressing (23.26) in terms of the eigenvectors. (Added discussion can be found in Exercise 23.8.) If the coefficient of the jth eigenvector is c_j, then the contribution to $\tilde{E}$ is $\epsilon_j(2c_j^2 - c_j^4)$. Figure 23.6 (right panel) plots the function $y = 2x^2 - x^4$, which shows the

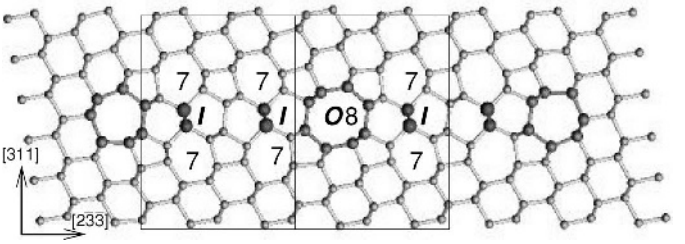


Figure 23.8. Projection on the $\{0\,\bar{1}\,1\}$ plane of a rodlike $\{3\,1\,1\}$ defect containing four interstitial chains I. The structure of $\{3\,1\,1\}$ defects commonly observed in ion-implanted silicon is characterized by random combinations of $/I\,I/$ and $/I\,O/$ units indicated by the boxes. The calculations find the structure of the defects by O(N) MD simulations using a tight-binding method [605], with the result that the extended defects are significantly lower in energy ($\approx$2 eV per interstitial) than for isolated interstitials in the bulk. Analogous work has been done using the density matrix approach [851, 852]. Figure provided by J. Kim.

consequences graphically. For eigenvalues that are negative there is an absolute minimum at $|c_j| = 1$, i.e. a properly normalized function. The same holds for the general case where we have many states, and minimization leads to an orthonormal set of functions that spans the space. For positive eigenvalues there is a local minimum at $c_j = 0$, but also a runaway solution in the unphysical $|c_j| > 1$ that must be avoided.

The functional has been extended by Kim et al. [846] to include more states than electrons, and to minimize the grand potential $\Omega_s = \mathrm{Tr}\{\hat{\rho}(\hat{H} - \mu)\}$, which can be written in matrix form analogous to (23.26) (Exercise 23.9)

$$\tilde{E}'_s = \mathrm{Tr}\{(2 - S)(H - \mu I)\}, \tag{23.27}$$

where S and H denote matrices S_{ij} and H_{ij} and I is the unit matrix. From the above reasoning it follows that (23.27) is minimized by filling all states below the Fermi energy, and leaving empty all those above. (It is not difficult to restrict the variations to avoid the runaway solution.) One can work with a limited number of states (somewhat larger than the number of electrons) and all operations scale as the number of these states. Thus this functional achieves the additional desired feature that the incorporation of extra states makes it easier to reach the minimum, and one can avoid being trapped in the wrong states at level crossings, etc.

An example of calculations using the Kim et al. functional is shown in Fig. 23.8, which shows results for the structure of {311} defects commonly observed in ion-implanted silicon [608, 609]. Similar tight-binding calculations have been done with the density matrix approach [851, 852]. The structures in Fig. 23.8 were determined by O(N) molecular dynamics tight-binding calculations using the model of Kwon et al. [605], and checks were performed on smaller cells in the relaxed geometries with full density functional theory calculations using plane wave DFT calculations with VASP codes [718]. For an interstitial in the bulk the two methods give similar formation energies, 3.4 eV in the planewave LDA method compared with 3.9 eV from the tight-binding calculation. The use of tight-binding methods made possible molecular-dynamics simulations for up to 1 psec at 300 to 600 K to

observe atomic rearrangements. The calculations find extended {311} defects are formed by condensation of interstitial chains with successive rotations of pairs of atoms in the {011} plane. After determination of the extended structure, DFT calculations of the total energy are used to establish its stability. The calculations show that the rodlike {311} defects are greatly favored with formation energies per interstitial much lower than in the bulk (as low as 0.7 and 1.2 eV calculated by the two methods). This is an example of the combination of methods in which O(N) calculations greatly enhance studies of complex problems in materials; at the same time it illustrates the need to calibrate tight-binding models carefully against more accurate methods.

Non-orthogonal orbitals

Finally, generalizing the functional to non-orthogonal localized orbitals $\tilde{w}_i$, as in (23.7), has the advantage that the $\tilde{w}_i$ can be shorter range and more transferable than orthogonal Wannier functions. The latter property is illustrated in Sec. 21.4; it means that good guesses can be made for the orbitals initially and as the atoms move in a simulation. The difficulty is two-fold: finding an efficient way to construct the inverse and dealing with the singular S matrix that results if one allows extra orbitals that have zero norm as in the Kim functional, Eq. (23.27). In the latter case, the size of S is the number of orbitals, but the rank (see below) is the number of electrons, i.e. $N = \text{rank}\{S\}$.

An elegant formulation has been given in [780] and [774], building upon earlier work [823,853], and using the same ideas as for generating non-orthogonal functions in Sec. 21.4. The first step is to define the inverse S^- of a singular matrix S by the relation [780]

$$SS^-S = S, \tag{23.28}$$

so that $S^- = S^{-1}$ if S is non-singular. Next, a functional can be defined that accomplishes the inverse in a way similar to the "Hotelling" method [854]

$$\text{Tr}\{BS^-\} = \text{minTr}\{B(2X - XSX)\}, \tag{23.29}$$

where B is any negative definite matrix and the minimization is for all possible hermitian matrices X (Exercise 23.6). Putting this together with the generalization of expression (23.25) for the energy

$$E_{\text{total}} = \text{Tr}\hat{\rho}\hat{H} = \sum_{i,j=1}^{N} S^-_{ij} H_{ji}, \tag{23.30}$$

leads to the functional

$$\tilde{E}'_{\text{total}}(N) \rightarrow \text{Tr}[2X - XSX][H - \eta S] + \eta N, \tag{23.31}$$

where the constant η is added to shift the eigenvalues to negative energies. Finally, even the constraint on the rank of S can be removed by defining the functional in terms of the Fermi energy μ as

$$\tilde{E}'_{\text{total}}(N) \rightarrow \text{Tr}[2X - XSX][H - \eta S] + (\eta - \mu)\text{rank}(S) + \mu N, \tag{23.32}$$

where the same trick can be used for rank(S)

$$\text{rank}(S) = \text{Tr}[SS^{-}] = -\min\{\text{Tr}[(-S)(2X - XSX)]\}. \tag{23.33}$$

The energy $\tilde{E}'_{\text{total}}(N)$ is the minimum for all non-orthogonal wavefunctions $\tilde{w}_i$ that define the S and H matrices and all hermitian matrices X.

If the orbitals are confined, then the minimum of functionals (23.31) or (23.32) is for non-orthogonal orbitals, since they are more compact than orthogonal ones. Furthermore, one can constrain the orbitals further to require maximal localization, in which case the functionals still give the same energy, and the orbitals are more physical and intuitive, as discussed in Sec. 21.4.

Combining minimization and projection

There are advantages to both the projection methods and the variational methods. An example of a calculation that combines these methods is the calculation of Wannier functions in disordered systems [776, 777, 855]. The variational approaches using Wannier functions suffer from the problem that they scale as M^3, where M is the size of the localization region. Furthermore, it has been found in practice [845] that convergence is slow due to fact that the energy function is only weakly dependent upon the tails of the function. Of the other hand, the projection method scales as M and can be used to improve the functions in the tails. Stephan and coworkers [776, 855] combined the methods to: (1) project Wannier-like functions in a large region; (2) find the largest coefficients, which leads to the best "self-adaptive" functions instead of imposing an arbitrary cutoff; and (3) use minimization method functions to improve the final functions. An example of the density of a bonding-type Wannier function calculated for a model of amorphous Si containing 4,096 atoms is shown in Fig. 23.9, plotted on a logarithmic scale extending over 20 orders of magnitude [777]. The distinct dark lines in the figure represent the zeros of the Wannier function.

23.6 Linear-scaling self-consistent density functional calculations

Full self-consistent density functional theory calculations can be cast in an O(N) linear-scaling form, since the charge density can be computed in O(N) fashion and thus one can "build" the hamiltonian (Sec. 23.2) with effort that scales as O(N). The difference from the tight-binding calculation is that one must explicitly represent the orbitals and one must deal with self-consistency. Because of the difficulties in carrying out such full calculations on very large systems, much less work has been done than with the simpler tight-binding methods. One such approach is illustrated in Fig. 23.5 using the KKR multiple-scattering approach [671].

The local orbital approach is perhaps the most direct implementation since it has exactly the same form as the tight-binding method, except that it involves calculations of all the matrix elements from a basis of orbitals. Any of the approaches in Ch. 15 can be used. An example of a calculation [856] for a complete turn of a selected DNA molecule is

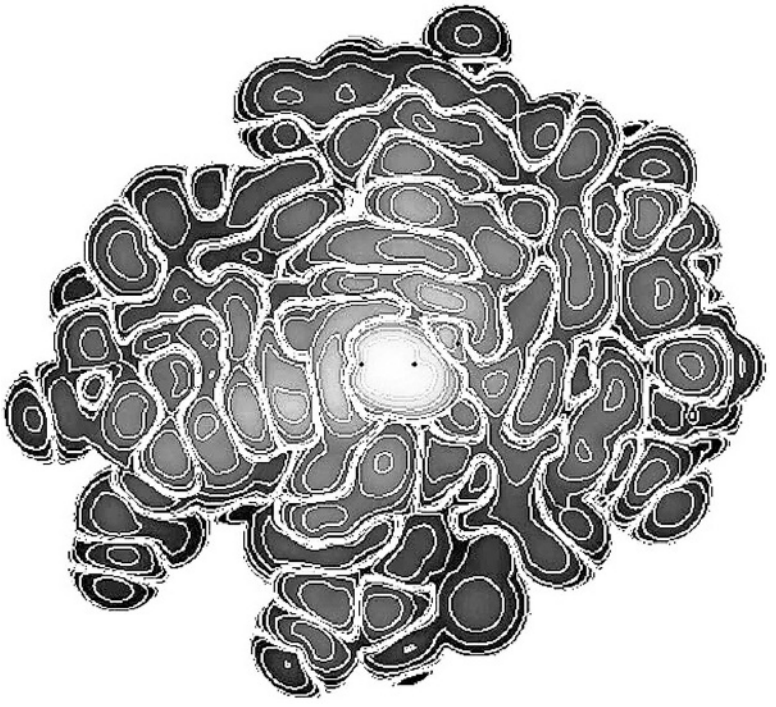


Figure 23.9. Electron density of a Wannier function calculated for a model of amorphous Si containing 4,096 atoms. The example shown is a function centered on a bond. The density is plotted on a logarithmic scale extending over 20 orders of magnitude. Figure provided by D. Drabold, similar to those in [777].

shown in Fig. 23.10. The density, potential, and thermal simulations of the atoms were calculated using O(N) procedures in the SIESTA code [617]. For a given structure, the resulting potential can then be used in an ordinary N^3 diagonalization (or a more efficient inverse iteration procedure such as the RMM-DIIS method; see Sec. 23.8 and App. M) to find selected eigenstates, in particular the fundamental gap between the lowest unoccupied (LUMO) and highest occupied (HOMO) orbitals, which are shown in Fig. 23.10. Further information is given in [856], which investigated the effects of disorder (a mutation) upon localization of the states and electrical conductivity. Calculations on similar size systems, including a fragment of an RNA molecules with 1,026 atoms, have been done using linear-scaling gaussian density matrix methods [619].

23.7 Factorized density matrix for large basis sets

Large basis sets such as plane waves and grids are desirable in order to have robust methods that always converge to the correct answer. For such approaches, however, straightforward application of linear-scaling approaches may not be feasible. In particular, the density matrix formalism leads to unwieldy expressions since it requires matrices of size $N_{\text{basis}} \times N_{\text{basis}}$.

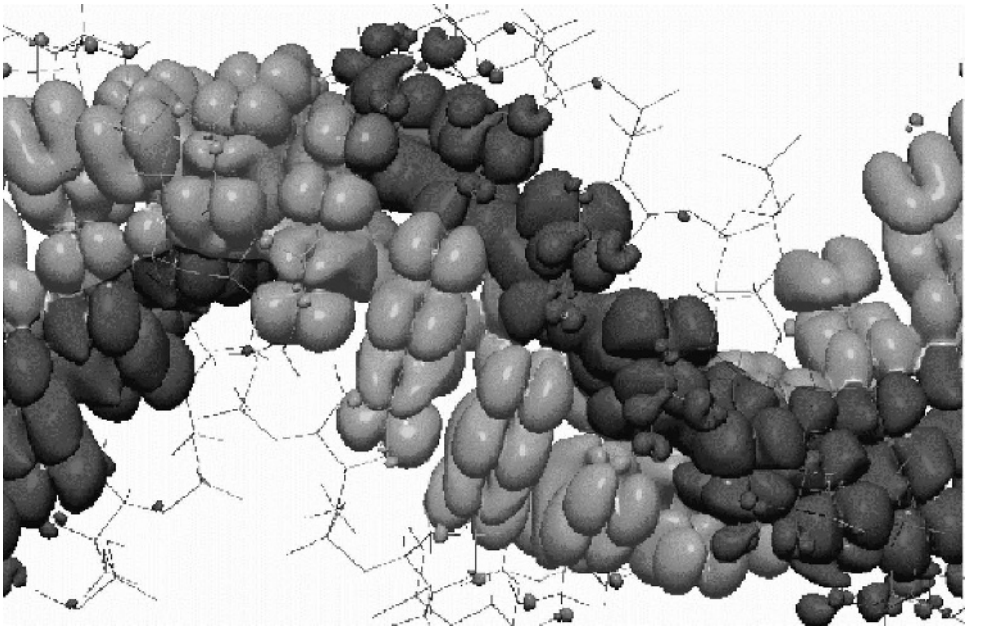


Figure 23.10. Electron density of selected eigenstates in a DNA molecule calculated one complete turn using the self-consistent local orbital SIESTA code and a GGA functional. The density contour shown is $5 \times 10^{-4} a_0^{-3}$ the lighter shaded region is the sum of densities of the 11 highest occupied (HOMO) states; and the darker region represents the 11 lowest unoccupied (LUMO) states. Calculations of comparable complexity have been done using gaussian bases and linear-scaling density matrix methods, e.g. for a fragment of an RNA molecule with 1,026 atoms [619]. Figure provided by E. Artacho, similar to results in [856].

Even if the matrix is sparse, it still becomes very large in the limit of finely spaced grids. How can it be feasible to use density matrix methods with such bases? The answer is remarkably simple: only a limited number of orbitals are needed, but each orbital needs to have many degrees of freedom. This is the basis for replacing (23.23) with a factorized form [857]

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{ij} \Phi_i^*(\mathbf{r}) K_{ij} \Phi_j(\mathbf{r}'), \tag{23.34}$$

where $\Phi_j(\mathbf{r}) = \sum_\nu c_j^\nu \phi_\nu(\mathbf{r})$ denotes an orbital expanded in a (large) basis set $\phi_\nu(\mathbf{r})$ and K_{ij} is a matrix of a size comparable to that needed in a tight-binding calculation. Thus the "purified" density matrix in the reduced space of the orbitals can be written

$$K = 3LOL - 2LOLOL, \tag{23.35}$$

where L is a trial density matrix and O is the overlap matrix of the orbitals. In this form, one minimizes the functional with respect to L_{ij} and the coefficients c_j^ν. Different choices lead to different tradeoffs between the number of orbitals involved in the matrix operations in (23.35) and the number of basis functions needed for each orbital in (23.34). Clearly, the general non-orthogonal formulation in (23.31) or (23.32) can be even more advantageous as

a general approach to making the orbitals as confined as possible, and therefore as optimal as possible for representation in a large basis.

The same idea can, of course, be applied to the Wannier-function-type methods. In fact, the local orbital approach is already in this form, since a limited number of localized Wannier functions are each expanded in a basis of atomic-like orbitals. One can also expand each localized function in a representation on a grid, as is done, e.g. in [858] where non-orthogonal functions are used in a method similar to that proposed by Stechel and coworkers [823, 853]. An expansion in overlapping spherical waves has been proposed by Haynes and Payne [859].

23.8 Combining the methods

Most of the O(N) methods described in this chapter are based upon the localization of the density matrix in space: they capture the physics and perform well for wide-band-gap insulators, highly disordered materials, and metals at (very) high temperature. But they fail on all counts whenever the localization length is large, e.g. in a good metal at ordinary temperatures. Such states can be isolated by methods that separate the physics by localization in energy, i.e. spectral methods (App. M), rather than by localization in space. How can one combine the methods to take advantage of the properties of each? There can be many approaches, but all have the general feature that O(N) methods can be used for solution at one level, e.g. a metallic system at very high temperature T, and the spectral method for the difference between the high- and low-T solutions that depends only upon states near the Fermi energy.

Spectral methods described in App. M are designed to find selected states efficiently. If the hamiltonian operator can be cast in sparse form (i.e. the hamiltonian is localized in real space or one can use transforms as in Sec. M.11) each state can be found with effort $\propto N_{\rm basis}$ or $\propto N_{\rm basis} \ln N_{\rm basis}$. Of course, the usual approach in which all states are desired requires effort that scales as $\propto N \times N_{\rm basis} \propto N^2$ or $\propto N^2 \ln N_{\rm basis}$. The effort needed to treat only the states near the Fermi energy is $\propto N^2 \times {\rm k_B T}/\mu$, where T is the needed smearing for an efficient O(N) scheme and the Fermi energy μ represents the characteristic total energy range of the electronic states. Although the effort scales as $\propto N^2$, there is a small prefactor, so that such spectral methods can be effective. Finally, first attempts [860] have been made to take advantage of the fact that states vary smoothly near the Fermi energy and create a linear-scaling spectral resolution approach that is applicable to metals.

SELECT FURTHER READING

Bowler, D. R. and Gillan, M. J. "Recent progress in first principles O(N) methods," *Molecular Simulations* 25: 239–255, 2000.

Fulde, P. *Electron Correlation in Molecules and Solids,* 2nd Edn., Springer-Verlag, Berlin, 1993. Discusses localization in a many-body context.

Galli, G. "Large scale electronic structure calculations using linear scaling methods," *Phys. Stat. Sol.* 217: 231–249, 2000.

Goedecker, S. "Linear scaling electronic structure methods," *Rev. Mod. Phys.* 71: 1085–1123, 1999. Compares many methods.

Goringe, C. M. Bowler, D. R. and Hernandez, E. "Tight-binding modelling of materials," *Rep. Prog. Phys.* 60: 1447–1512, 1997.

Haydock, R. in *Recursion Method and Its Applications*, edited by D. G. Pettifor and D. L. Weaire, Springer-Verlag, Berlin, 1985.

Ordejon, P. "Linear scaling *ab intio* calculations in nanoscale materials with SIESTA," *Phys. Stat. Sol.* 217: 335–356, 2000.

Exercises

23.1 Derive the result stated in the text that the normalization constant in (23.11) is given by $C_{n+1} = 1/\beta_n$, $n \geq 1$. Show this by directly calculating the normalization of ψ_{n+1} assuming ψ_n is normalized.

23.2 Derive the continued fraction representation of (23.12) using the Lanczos algorithm for the coefficients. It follows that the spectrum of eigenvalues is given by the poles of the continued fraction (i.e. the zeros of the denominator) in (23.12). (Thus the spectrum of eigenvalues is given either by the continued fraction form or by the zeros of the polynomial of the previous problem.)

23.3 Derive the form of the terminator given in (23.13) and (23.14). Show that imaginary part is non-zero in the band range indicated, so that no poles and only continuous DOS can result in this range. In fact, there is another solution with a plus sign in the square root in (23.14); show that this is not allowed since $t(z)$ must vanish for $|z| \to \infty$.

23.4 Show that the square root form for terminator (23.14) satisfies Kramers–Kronig relations, Eq. (D.15), as it must if $G(z)$ is a physically meaningful Green's function.

23.5 Show that the Green's function $G(z) = 1/(\hat{H} - z)$ becomes more localized for large z for the case where $\hat{H}$ is a short-range operator in real space. This is the essence of localization in both the recursion and Fermi function expansion methods. Hint: First consider the $z \to \infty$ limit and then terms in powers of $\hat{H}/z$.

23.6 Derive Eq. (23.29) by showing that the variation around $X = S^{-}$ is quadratic and always positive for matrices B that are negative definite.

23.7 This exercise is to derive the form of the "purification" functional, Eq. (23.24), that leads to idempotency of the density matrix.
(a) The first step is to demonstrate that the function $y = 3x^2 - 2x^3$ has the form shown in the left panel of Fig. 23.6 and that the result y is always closer to 0 or 1 than the input x.
(b) Next, generalize this to a matrix equation for any symmetric matrix leading to eigenvalues closer to 0 or 1.
(c) Finally, show that the functional (23.23) minimized using the gradients (23.24) leads to the desired result.

23.8 This exercise is designed to provide simple examples of the properties of the unconstrained functional, Eq. (23.26).
(a) Consider a diagonal 2×2 hamiltonian with $H_{11} = \epsilon_1 < 0$, $H_{22} = \epsilon_2 > 0$, and $H_{12} = H_{21} = 0$ in an orthonormal basis, ψ_1 and ψ_2. Show that minimization of the functional leads to the ground state ψ_1 properly normalized.

(b) Now consider the same basis as in part (a) but with a hamiltonian that is not diagonal: $H_{11} = H_{22} = 0$ and $H_{12} = H_{21} = \epsilon_0$. Show that in this case the functional also leads to the properly normalized ground state $\psi = \frac{1}{\sqrt{2}}(\psi_1 + \psi_2)$.

23.9 Show that the functional, Eq. (23.27), has the property that it leads to orthonormal eigenvectors for states below the Fermi energy μ and projects to zero the amplitude of any states with eigenvalue above the Fermi energy.

24

Where to find more

It is not appropriate to summarize or conclude this volume on the basic theory and methods of electronic structure. The field is evolving rapidly with new advances in basic theory, algorithms, and computational methods. New developments and applications are opening unforeseen vistas. Volumes of information are now available on-line at thousands of sites.

It is more appropriate to provide a resource for information that will be updated in the future, on-line information available at a site maintained at the University of Illinois:

http://ElectronicStructure.org

A link is maintained at the Cambridge University Press site:

http://books.cambridge.org/0521782856.htm

Resources for materials computation are maintained by the Materials Computation Center at the University of Illinois, supported by the National Science Foundation:

http://www.mcc.uiuc.edu/

Additional sites include the Department of Physics, the electronic structure group at the University of Illinois, and the author's home page:

http://www.physics.uiuc.edu/
http://www.physics.uiuc.edu/research/ElectronicStructure/
http://w3.physics.uiuc.edu/∼rmartin/homepage/

The on-line information includes:

- Additional material coordinated with descriptions in this book, as well as future updates, corrections, additions, and convenient feedback forms.
- Information related to many-body methods beyond the scope of this volume.
- Links to courses, tutorials, and codes maintained at the University of Illinois. Specific codes are meant for pedagogical use, teaching, or individual study, and are coordinated with descriptions in this book, for example, the general purpose empirical pseudopotential and tight-binding code (TBPW) in App. N.
- Links to codes for electronic structure calculations. This will include resources at the Materials Computation Center and many other sites.
- Links to many other sites around the world that provide codes, tutorials, courses, data, descriptions, and other information related to electronic structure.

Appendix A

Functional equations

Summary

A *functional* $F[f]$ is a mapping of an entire function f onto a value. In electronic structure, functionals play a central role, not only in density functional theory, but also in the formulation of most of the theoretical methods as functionals of the underlying variables, in particular the wavefunctions. This appendix deals with the general formulation and derivation of variational equations from the functionals.

A.1 Basic definitions and variational equations

The difference between a *function* $f(x)$ and a *functional* $F[f]$ is that a function is defined to be a mapping of a variable x to a result (a number) $f(x)$; whereas a functional is a mapping of an entire function f to a resulting number $F[f]$. The functional $F[f]$, denoted by square brackets, depends upon the function f over its range of definition $f(x)$ in terms of its argument x. Here we a describe some basic properties related to the functionals and their use in density functional theory; more complete description can be found in [93], App. A. A review of functional derivatives or the "calculus of variations" can be found in [861] and [862].

To illustrate functionals $F[f]$ we first consider two simple examples:

- A definite integral of $w(x)f(x)$, where $w(x)$ is some fixed weighting function,

$$I_w[f] = \int_{x_{\min}}^{x_{\max}} w(x)f(x)\mathrm{d}x. \tag{A.1}$$

- The integral of $(f(x))^\alpha$, where α is an arbitrary power:

$$I_\alpha[f] = \int_{x_{\min}}^{x_{\max}} (f(x))^\alpha \mathrm{d}x. \tag{A.2}$$

A functional derivative is defined by a variation of the functional

$$\delta F[f] = F[f+\delta f] - F[f] = \int_{x_{\min}}^{x_{\max}} \frac{\delta F}{\delta f(x)} \delta f(x)\mathrm{d}x, \tag{A.3}$$

where the quantity $\delta F/\delta f(x)$ is the functional derivative of F with respect to variation of $f(x)$ at the point x. In Eq. (A.1), the fact that the functional is linear in $f(x)$ leads to a simple result for the functional derivative

$$\frac{\delta I_w}{\delta f(x)} = w(x). \tag{A.4}$$

The variational derivation of the many-body Schrödinger equation in (3.10) and (3.12) is an example of this simple form.

The second example of a non-linear functional is of the form needed to minimize the Thomas–Fermi expression, Eq. (6.4). From definition (A.3) one can also show (Exercise A.1) that

$$\frac{\delta I_\alpha}{\delta f(x)} = \alpha(f(x))^{\alpha-1}, \tag{A.5}$$

following the same rules as normal differentiation. In general, however, the functional derivative at point x depends also upon the function $f(x)$ at all other points. Clearly, the definition can be extended to many variables and functions $F[f_1, f_2, \ldots]$.

A.2 Functionals in density functional theory including gradients

In Kohn–Sham density functional theory, the potential, Eq. (7.13), is a sum of functional derivatives. The external term has the linear form of Eq. (A.1); the Hartree term is also simple since it is bilinear; and $V_{\rm xc}^{\sigma}(\mathbf{r})$ is found by varying the more complex functional having the form

$$E_{\rm xc}[n] = \int n(\mathbf{r})\epsilon_{\rm xc}(n(\mathbf{r}), |\boldsymbol{\nabla} n(\mathbf{r})|)\mathrm{d}\mathbf{r}. \tag{A.6}$$

Variations of the gradient terms can be illustrated by the general form:

$$I[n] = \int g(f(\mathbf{r}), |\boldsymbol{\nabla} f(\mathbf{r})|)\mathrm{d}\mathbf{r}, \tag{A.7}$$

so that varying the function f leads to

$$\delta I[g, f] = \int \left[\frac{\delta g}{\delta f}\delta f(\mathbf{r}) + \frac{\delta g}{\delta|\boldsymbol{\nabla} f|}\delta|\boldsymbol{\nabla} f(\mathbf{r})|\right]\mathrm{d}\mathbf{r}. \tag{A.8}$$

Now using

$$\delta|\boldsymbol{\nabla} f(\mathbf{r})| = \delta\boldsymbol{\nabla} f(\mathbf{r}) \cdot \frac{\boldsymbol{\nabla} f(\mathbf{r})}{|\boldsymbol{\nabla} f(\mathbf{r})|} = \frac{\boldsymbol{\nabla} f(\mathbf{r})}{|\boldsymbol{\nabla} f(\mathbf{r})|} \cdot \boldsymbol{\nabla}[\delta f(\mathbf{r})] \tag{A.9}$$

and integrating by parts, one finds a standard form of variations of gradients

$$\delta I[g, f] = \int \left\{\frac{\delta g}{\delta f} - \boldsymbol{\nabla}\cdot\left[\frac{\delta \mathbf{g}}{\delta|\boldsymbol{\nabla}\mathbf{f}|}\frac{\boldsymbol{\nabla}\mathbf{f}(\mathbf{r})}{|\boldsymbol{\nabla}\mathbf{f}(\mathbf{r})|}\right]\right\}\delta f(\mathbf{r})\mathrm{d}\mathbf{r}. \tag{A.10}$$

This form is used in Sec. 8.3, where other forms for the functional derivative are also given that may be advantageous in actual calculations.

SELECT FURTHER READING

A compact description can be found in:

Parr, R. G., and Yang, W., *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989. App. A.

Basic theory of functionals can be found in:

Evans, G. C., *Functionals and Their Applications*, Dover, New York, 1964.
Matthews, J., and Walker, R. L., *Mathmatical Methods of Physics*, W. A. Benjamin, Inc., New York, 1964. Ch. 12.

Exercises

A.1 Show that Eq. (A.5) follows from (A.3), and that application of the expression to the Thomas–Fermi approximation leads to expression (6.4).

A.2 Derive the form of the variational expression in (A.10) involving the gradient terms.

Appendix B

LSDA and GGA functionals

Summary

In this appendix are given representative forms for the exchange–correlation energy and potential in the LSDA and GGA approximations. The forms given here are chosen because they are widely used and are relatively simple. Actual programs that provide energies and potentials for these and other forms can be found on-line (see Ch. 24).

B.1 Local spin density approximation (LSDA)

The local density approximation is based upon the exact expressions for the exchange energy, Eq. (5.15), and various approximations and fitting to numerical correlation energies for the homogeneous gas. Comparison of the forms is shown in Fig. 5.4. The first functions were the Wigner interpolation formula, Eq. (5.22), and the Hedin–Lundqvist [220] form; the latter is derived from many-body perturbation theory and is given below. As described in Ch. 5, the quantum Monte Carlo (QMC) calculations of Ceperley and Alder [297], and more recent work [298, 299, 303] provide essentially exact results for unpolarized and fully polarized cases. These results have been fitted to analytic forms for $\epsilon_c(r_s)$, where r_s is given by Eq. (5.1), leading to two widely used functionals due to Perdew and Zunger (PZ) [300] and Vosko, Wilkes, and Nusiar (VWN) [301], which are very similar quantitatively. Both functionals assume an interpolation form for fractional spin polarization, and Ortiz and Balone [298] report that their QMC calculations at intermediate polarization are somewhat better described by the VWN form. In all cases, the correlation potential is given by

$$V_c(r_s) = \epsilon_c(r_s) - \frac{r_s}{3}\frac{\mathrm{d}\epsilon_c(r_s)}{\mathrm{d}r_s}. \tag{B.1}$$

Here are listed selected forms for the unpolarized case; complete expressions can be found in references [224], [368], and [413].

1. Hedin–Lundqvist (HL) [220].

$$\epsilon_c^{\mathrm{HL}}(r_s) = -\frac{C}{2}\left[(1+x^3)\ln\left(1+\frac{1}{x}\right) + \frac{x}{2} - x^2 - \frac{1}{3}\right], \tag{B.2}$$

where $A = 21$, $C = 0.045$, and $x = r_s/A$. The correlation potential is

$$V_c^{\mathrm{HL}}(r_s) = -\frac{Ce^2}{2}\ln\left(1+\frac{1}{x}\right), \tag{B.3}$$

2. Perdew–Zunger (PZ) [300]

$$\begin{aligned}\epsilon_c^{\mathrm{PZ}}(r_s) &= -0.0480 + 0.031\ln(r_s) - 0.0116 r_s + 0.0020 r_s \ln(r_s), \qquad r_s < 1\\ &= -0.1423/(1 + 1.0529\sqrt{r_s} + 0.3334 r_s), \qquad r_s > 1.\end{aligned} \tag{B.4}$$

The expression [300] for V_c^{PZ} is not given here since it is lengthy, but it is straightforward. For fractional spin polarization, the interpolation for $\epsilon_c^{\mathrm{PZ}}(r_s)$ is assumed to have the same function form as for exchange, Eq. (5.17), with f given by (5.18).

3. Vosko–Wilkes–Nusiar (VWN) [301]

$$\begin{aligned}\epsilon_c^{\mathrm{VWN}}(r_s) = \frac{Ae^2}{2}\Bigg[&\log\left[\frac{y^2}{Y(y)}\right] + \frac{2b}{Q}\tan^{-1}\left(\frac{Q}{2y+b}\right)\\ &- \frac{by_0}{Y(y_0)}\left\{\log\left[\frac{(y-y_0)^2}{Y(y)}\right] + \frac{2(b+2y_0)}{Q}\tan^{-1}\left(\frac{Q}{2y+b}\right)\right\}\Bigg]\end{aligned} \tag{B.5}$$

Here $y = r_s^{1/2}$, $Y(y) = y^2 + by + c$, $Q = (4c - b^2)^{1/2}$, $y_0 = -0.10498$ $b = 3.72744$, $c = 12.93532$, and $A = 0.0621814$. The corresponding potential can be obtained from Eq. (B.1) with [413]

$$r_s\frac{\mathrm{d}\epsilon_c^{\mathrm{VWN}}(r_s)}{\mathrm{d}r_s} = A\frac{e^2}{2}\frac{c(y-y_0) - by_0 y}{(y-y_0)(y^2+by+c)}. \tag{B.6}$$

B.2 Generalized gradient approximation (GGAs)

There are many different forms for gradient approximations; however, it is beyond the scope of the present work to give the formulas for even the most widely used forms. The reader is referred to papers and books listed as "Select further reading."

B.3 GGAs: explicit PBE form

The PBE form is probably the simplest GGA functional. Hence we give it as an explicit example. The reader is referred to other sources such as the paper on "Comparison shopping for a gradient-corrected density functional," by Perdew and Burke [367]. The PBE functional

[373] for exchange is given by a simple form for the enhancement factor F_x defined in Sec. 8.2. The form is chosen with $F_x(0) = 1$ (so that the local approximation is recovered) and $F_x \to$ *constant* at large s,

$$F_x(s) = 1 + \kappa - \kappa/(1 + \mu s^2/\kappa), \tag{B.7}$$

where $\kappa = 0.804$ is chosen to satisfy the Lieb–Oxford bound. The value of $\mu = 0.21951$ is chosen to recover the linear response form of the local approximation, i.e. it is chosen to cancel the term from the correlation. This may seem strange, but it is done to agree better with quantum Monte Carlo calculations. This choice violates the known expansion at low s given in Eq. (8.7), with the rationale of better fitting the entire functional.

The form for correlation is expressed as the local correlation plus an additive term both of which depend upon the gradients and the spin polarization. The form chosen to satisfy several conditions is [373]

$$E_c^{\mathrm{GGA-PBE}}[n^\uparrow, n^\downarrow] = \int \mathrm{d}^3 r \; n \; \left[\epsilon_c^{\mathrm{hom}}(r_s, \zeta) + H(r_s, \zeta, t)\right], \tag{B.8}$$

where $\zeta = (n^\uparrow - n^\downarrow)/n$ is the spin polarization, r_s is the local value of the density parameter, and t is a dimensionless gradient $t = |\boldsymbol{\nabla} n|/(2\phi k_{\mathrm{TF}} n)$. Here $\phi = ((1+\zeta)^{2/3} + (1-\zeta)^{2/3})/2$ and t is scaled by the screening wavevector k_{TF} rather than k_F. The final form is

$$H = \frac{e^2}{a_0}\gamma\phi^3 \log\left(1 + \frac{\beta}{\gamma}t^2 \frac{1 + At^2}{1 + At^2 + A^2t^4}\right), \tag{B.9}$$

where the factor e^2/a_0, with a_0 the Bohr radius, is unity in atomic units. The function A is given by

$$A = \frac{\beta}{\gamma}\left[\exp\left(\frac{-\epsilon_c^{\mathrm{hom}}}{\gamma\phi^3 \frac{e^2}{a_0}}\right) - 1\right]^{-1}. \tag{B.10}$$

SELECT FURTHER READING

Summaries of functionals are given in:

Koch, W. and Holthausen, M. C. *A Chemists' Guide to Density Functional Thoery*, Wiley-VCH, Weinheim, 2001.

Parr, R. G. and Yang, W. *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989.

Pickett, W. E. "Pseudopotential methods in condensed matter applications," *Computer Physics Reports* 9:115, 1989.

Towler, M. D., Zupan, A. and Causa, M. "Density functional theory in periodic systems using local gaussian basis sets," *Computer Physics Commun.* 98:181–205, 1996.

Further information and codes can be found on-line (see Ch. 24).

Appendix C

Adiabatic approximation

Summary

The only small parameter in the electronic structure problem is the inverse nuclear mass $1/M$, i.e. the nuclear kinetic energy terms. The adiabatic or Born-Oppenheimer approximation is a systematic expansion in the small parameter that is fundamental to all electronic structure theory. It comes to the fore in the theory of phonons, electron-phonon interactions, and superconductivity (Ch. 19).

C.1 General formulation

The fundamental hamiltonian for a system of nuclei and electrons, Eq. (3.1), can be written

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{U}, \tag{C.1}$$

where U contains all the potential interaction terms involving the set of all-electron coordinates $\{\mathbf{r}\}$ (which includes spin) and the set of all nuclear coordinates $\{\mathbf{R}\}$. Since the only small term is the kinetic energy operator of the nuclei $\hat{T}_N$, we treat it as a perturbation upon the hamiltonian, Eq. (3.2), for nuclei fixed in their instantaneous positions. The first step is to define the eigenvalues and wavefunctions $E_i(\{\mathbf{R}\})$ and $\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\})$ for the electrons which depend upon the nuclear positions $\{\mathbf{R}\}$ as parameters. This is the same as Eq. (3.13) except that the positions of the nuclei are indicated explicitly, and $i = 0, 1, \ldots,$ denotes the complete set of states at each $\{\mathbf{R}\}$.

The full solutions for the coupled system of nuclei and electrons[1]

$$\hat{H}\Psi_s(\{\mathbf{r}, \mathbf{R}\}) = E_s\Psi_s(\{\mathbf{r}, \mathbf{R}\}), \tag{C.2}$$

where $s = 1, 2, 3, \ldots,$ labels the states of the coupled system, can be written in terms of $\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\})$,

$$\Psi_s(\{\mathbf{r}, \mathbf{R}\}) = \sum_i \chi_{si}(\{\mathbf{R}\})\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}), \tag{C.3}$$

since $\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\})$ defines a complete set of states for the electrons at each $\{\mathbf{R}\}$.

[1] Adapted from notes of K. Kunc and the author.

The states of the coupled electron–nuclear system are now specified by $\chi_{si}(\{\mathbf{R}\})$, which are functions of the nuclear coordinates and are the coefficients of the electronic states $\Psi_m i$. In order to find the equations for $\chi_{si}(\{\mathbf{R}\})$, insert expansion (C.3) into (C.2), multiply the expression on the left by $\Psi_i(\{\mathbf{r}, \mathbf{R}\})$, and integrate over electron variables $\{\mathbf{r}\}$ to find the equation

$$[T_N + E_i(\{\mathbf{R}\}) - E_s]\, \chi_{si}(\{\mathbf{R}\}) = -\sum_{i'} C_{ii'} \chi_{si'}(\{\mathbf{R}\}), \tag{C.4}$$

where $T_N = -\frac{1}{2}(\sum_J \nabla_J^2 / M_J)$ and the matrix elements are given by $C_{ii'} = A_{ii'} + B_{ii'}$, with

$$A_{ii'}(\{\mathbf{R}\}) = \sum_J \frac{1}{M_J} \langle \Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) | \boldsymbol{\nabla}_J | \Psi_{i'}(\{\mathbf{r}\} : \{\mathbf{R}\}) \rangle \boldsymbol{\nabla}_J, \tag{C.5}$$

$$B_{ii'}(\{\mathbf{R}\}) = \sum_J \frac{1}{2M_J} \langle \Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) | \nabla_J^2 | \Psi_{i'}(\{\mathbf{r}\} : \{\mathbf{R}\}) \rangle. \tag{C.6}$$

Here $\langle \Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) | \mathcal{O} | \Psi_{i'}(\{\mathbf{r}\} : \{\mathbf{R}\}) \rangle$ means integrations over only the electronic variables $\{\mathbf{r}\}$ for any operator $\mathcal{O}$.

The adiabatic or Born-Oppenheimer approximation [89] is to ignore the off-diagonal $C_{ii'}$ terms, i.e. the electrons are assumed to remain in a given state m as the nuclei move. Although the electron wavefunction $\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\})$ and the energy of state m change, the electrons do not change state and no energy is transferred between the degrees of freedom described by the equation for the nuclear variables $\{\mathbf{R}\}$ and *excitations* of the electrons, which occurs only if there is a change of state $i \to i'$. The diagonal terms can be treated easily. First, it is simple to show (see Exercise C.1) that $A_{ii} = 0$ simply from the requirement that Ψ is normalized. The term $B_{ii}(\{\mathbf{R}\})$ can be grouped with $E_i(\{\mathbf{R}\})$ to determine a modified potential function for the nuclei $U_i(\{\mathbf{R}\}) = E_i(\{\mathbf{R}\}) + B_{ii}(\{\mathbf{R}\})$. Thus, in the adiabatic approximation, the nuclear motion is described by a purely nuclear equation for each electronic state i

$$\left[-\sum_J \frac{1}{2M_J} \nabla_J^2 + U_i(\{\mathbf{R}\}) - E_{ni} \right] \chi_{ni}(\{\mathbf{R}\}) = 0, \tag{C.7}$$

where $n = 1, 2, 3, \ldots$, labels the nuclear states. Within the adiabatic approximation, the full set of states $s = 0, 1, \ldots$, is a product of nuclear and electronic states.

Equations (C.7) with the neglect at the B_{ii} term is the basis of the "frozen phonon" or perturbation methods for calculation of phonon energies in the adiabatic approximation (Ch. 19). So long as we can justify neglecting the off-diagonal terms that couple different electron states, we can solve the nuclear motion problem, Eq. (C.7), given the function $U_i(\{\mathbf{R}\})$ for the particular electronic state i that evolves adiabatically with nuclear motion. (The term B_{ii} is typically very small due to the large nuclear mass.) In general, this is an excellent approximation except for cases where there is degeneracy or near degeneracy of the electronic states. If there is a gap in the electronic excitation spectrum much larger than typical energies for nuclear motion, then the nuclear excitations are well determined by the adiabatic terms. Special care must be taken for cases such as transition states in molecules where electronic states become degenerate, or in metals where the lack of an energy gap leads to qualitative effects.

C.2 Electron-phonon interactions

Electron-phonon interactions result from the off-diagonal matrix elements $C_{ii'}$ that describe transitions between different electronic states due to the velocities of the nuclei. The dominant terms are given in Eq. (C.5), which involves a gradient of the electron wavefunctions with respect to the nuclear positions and the gradient operator acting on the phonon wavefunction χ. Combination of these operators leads to an electronic transition between states i and i' coupled with emission or absorption of one phonon.

The steps involved in writing the formal expressions are to express the nuclear kinetic operator ∇_J in Eq. (C.5) in terms of phonon creation and annihilation operators [96] and to write out the perturbation expression for the matrix element. The latter step can be accomplished by noting that the variation in the electron function due to the displacement of nucleus J is caused by the change in potential V due to the displacement. To linear order the relation is

$$\langle \Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) | \nabla_J | \Psi_{i'}(\{\mathbf{r}\} : \{\mathbf{R}\}) \rangle = \frac{\langle \Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) | \nabla_J V | \Psi_{i'}(\{\mathbf{r}\} : \{\mathbf{R}\}) \rangle}{E_{i'}(\{\mathbf{R}\}) - E_i(\{\mathbf{R}\})}. \quad \text{(C.8)}$$

This leads to the form of the electron-phonon matrix elements in Sec. 19.8 and treated in references such as [243].

SELECT FURTHER READING

Born, M. and Huang, K. *Dynamical Theory of Crystal Lattices*, Oxford University Press, Oxford, 1954.

Ziman, J. M. *Principles of the Theory of Solids*, Cambridge University Press, Cambridge, 1989.

Pines, D. *Elementary Excitations in Solids*, Wiley, New York, 1964.

Exercises

C.1 Show that the requirement that Ψ is normalized is sufficient to prove $A_{ii} = 0$. Hint: Use the fact that any derivative of $\langle \Psi | \Psi \rangle$ must vanish.

C.2 Derive the equation for nuclear motion, (C.7), from (C.4) using the assumption of the adiabatic approximation as described before (C.7).

C.3 For small nuclear displacements about their equilibrium positions, show that (C.7) leads to harmonic oscillator equations.

C.4 For a simple diatomic molecule treated in the harmonic approximation, show that (C.7) leads to the well-known result that the ground state energy of the nuclear–electron system is $E_{min} + \frac{1}{2}\hbar\omega$, where ω is the harmonic oscillator frequency.

Appendix D

Response functions and Green's functions

Summary

Response functions are the bread and butter of theoretical physics and the connection to important experimental measurements. The basic formulas are rooted in well-known perturbation expressions given in Ch. 3. This appendix is devoted to characteristic forms and properties of response functions, sum rules, and Kramers–Kronig relations. The most important example is the dielectric function described in App. E. Useful expressions are given for self-consistent field methods, which leads to "RPA" and other formulas needed in Chs. 5, 19, and 20.

D.1 Static response functions

Static response functions play two important roles in electronic structure. One is the calculation of quantities that directly relate to experiments, namely the actual response of the electrons to static perturbations such as strain or applied electric fields, and the response at low frequencies that can be considered "adiabatic" (App. C) that governs lattice dynamics, etc. This is the subject of Ch. 19. The other role is the development of methods in the theory of electronic structure to derive improved solutions utilizing perturbation expansions around more approximate solutions. This is the basis of the analysis in Ch. 9.

The basic equations follow from perturbation theory, which was summarized in Sec. 3.7, in particular in Eq. (3.62) which is repeated here:

$$\Delta\langle\hat{O}\rangle = \sum_{i=1}^{\text{occ}}\langle\psi_i|\hat{O}|\psi_i\rangle = \sum_{i=1}^{\text{occ}}\sum_{j}^{\text{empty}}\frac{\langle\psi_i|\hat{O}|\psi_j\rangle\langle\psi_j|\Delta\hat{H}_{\text{eff}}|\psi_i\rangle}{\varepsilon_i-\varepsilon_j} + \text{c.c.} \tag{D.1}$$

The sum over j is be restricted to empty states only, since contributions of pairs of occupied states i, j and j, i cancel in the sum.

The most relevant quantity for static perturbations is the density, for which Eq. (3.62) becomes

$$\Delta n(\mathbf{r}) = \sum_{i=1}^{\text{occ}}\sum_{j}^{\text{empty}}\psi_i^*(\mathbf{r})\psi_j(\mathbf{r})\frac{\langle\psi_j|\Delta V_{\text{eff}}|\psi_i\rangle}{\varepsilon_i-\varepsilon_j} + \text{c.c.} \tag{D.2}$$

The response to a variation of the *total potential* $V_{\rm eff}(\mathbf{r})$ at point $\mathbf{r} = \mathbf{r}'$ (see App. A for definition of functional derivatives) defines the density response function

$$\chi_n^0(\mathbf{r}, \mathbf{r}') = \frac{\delta n(\mathbf{r})}{\delta V_{\rm eff}(\mathbf{r}')} = 2\sum_{i=1}^{\rm occ}\sum_{j}^{\rm empty} \frac{\psi_i^*(\mathbf{r})\psi_j(\mathbf{r})\psi_j^*(\mathbf{r}')\psi_i(\mathbf{r}')}{\varepsilon_i - \varepsilon_j}, \tag{D.3}$$

which is symmetric in $\mathbf{r}$ and $\mathbf{r}'$ since it is the response of $n(\mathbf{r})$ to a perturbation $V_{\rm eff}(\mathbf{r}')n(\mathbf{r}') \propto n(\mathbf{r}')$. Equation (D.3) may also be written in a convenient form

$$\chi_n^0(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{\rm occ} \psi_i^*(\mathbf{r})G_0^i(\mathbf{r}, \mathbf{r}')\psi_i(\mathbf{r}'), \quad G_0^i(\mathbf{r}, \mathbf{r}') = \sum_{j\neq i}^{\infty} \frac{\psi_j(\mathbf{r})\psi_j^*(\mathbf{r}')}{\varepsilon_i - \varepsilon_j}, \tag{D.4}$$

where G_0^i is an independent-particle Green's function (Sec. D.4).

The Fourier transform of $\chi_n^0(\mathbf{r}, \mathbf{r}')$ is the response to particular Fourier components, which is often the most useful form. If we define $\Delta V_{\rm eff}(\mathbf{r}) = \Delta V_{\rm eff} e^{i\mathbf{q}\cdot\mathbf{r}}$ and $n(\mathbf{q}') = \int d\mathbf{r} n(\mathbf{r}) e^{i\mathbf{q}'\cdot\mathbf{r}}$ in (D.2), then one finds (Exercise D.1)

$$\chi_n^0(\mathbf{q}, \mathbf{q}') = \frac{\delta n(\mathbf{q}')}{\delta V_{\rm eff}(\mathbf{q})} = 2\sum_{i=1}^{\rm occ}\sum_{j}^{\rm empty} \frac{M_{ij}^*(\mathbf{q})M_{ij}(\mathbf{q}')}{\varepsilon_i - \varepsilon_j}, \tag{D.5}$$

where $M_{ij}(\mathbf{q}) = \langle\psi_i|e^{i\mathbf{q}\cdot\mathbf{r}}|\psi_j\rangle$. This can be a great simplification, for example, in a homogeneous system, $\chi_n^0(\mathbf{q}, \mathbf{q}') \neq 0$ only for $\mathbf{q} = \mathbf{q}'$ (Ch. 5), or in crystals (Chs. 19 and 20).

The response function χ^0 plays many important roles in electronic structure theory. The simplest is in approximations in which the electrons are considered totally non-interacting; then, $\Delta V_{\rm eff} = \Delta V_{\rm ext}$ and χ^0 represents the response to an external perturbation. However, in an effective mean-field theory, like the Hartree–Fock or Kohn–Sham theories of Chs. 7 and 9, the internal fields also vary and the effective hamiltonian must be found in a self-consistent procedure. This leads to the following section in which χ^0 still plays a crucial role.

D.2 Response functions in self-consistent field theories

In a self-consistent field theory, the total effective field depends upon the internal variables; e.g. in the Kohn–Sham approach, $V_{\rm eff} = V_{\rm ext} + V_{\rm int}[n]$. Since the electrons act as independent particles in the potential $V_{\rm eff}$, χ_n^0 is still given by (D.3)–(D.5). However, the relation to the external field is changed. To linear order, the response to an external field is given by

$$\chi = \frac{\delta n}{\delta V_{\rm ext}}, \tag{D.6}$$

which is shorthand for the functional form that can be written in $\mathbf{r}$ space or $\mathbf{q}$ space,

$$\chi(\mathbf{r}, \mathbf{r}') = \frac{\delta n(\mathbf{r})}{\delta V_{\rm ext}(\mathbf{r}')} \quad \text{or} \quad \chi(\mathbf{q}, \mathbf{q}') = \frac{\delta n(\mathbf{q})}{\delta V_{\rm ext}(\mathbf{q}')}. \tag{D.7}$$

Similarly, the linear response of the spin density $m = n^\uparrow - n^\downarrow$ to an external Zeeman field $\Delta\hat{H} = V_{\rm ext}^m$ has the same form

$$\chi = \frac{\delta m}{\delta V_{\rm ext}^m}, \tag{D.8}$$

so that the analysis applies to both total density and spin density.

The response function can be written (omitting indices for simplicity)

$$\chi = \frac{\delta n}{\delta V_{\rm eff}} \frac{\delta V_{\rm eff}}{\delta V_{\rm ext}} = \chi^0 \left[1 + \frac{\delta V_{\rm int}}{\delta n} \frac{\delta n}{\delta V_{\rm ext}} \right] = \chi^0 \left[1 + K\chi \right], \tag{D.9}$$

where the kernel K given in $\mathbf{r}$ space in (9.12) or in $\mathbf{q}$ space as

$$K(\mathbf{q}, \mathbf{q}') = \frac{\delta V_{\rm int}(\mathbf{q})}{\delta n(\mathbf{q}')} = \frac{4\pi}{q^2} \delta_{\mathbf{q},\mathbf{q}'} + \frac{\delta^2 E_{\rm xc}[n]}{\delta n(\mathbf{q}) \delta n(\mathbf{q}')} \equiv V_C(q) \delta_{\mathbf{q},\mathbf{q}'} + f_{\rm xc}(\mathbf{q}, \mathbf{q}'). \tag{D.10}$$

Solving (D.9) (Exercise D.2), leads to the ubiquitous form [96, 284, 865]

$$\chi = \chi^0 [1 - \chi^0 K]^{-1} \ \text{ or } \ \chi^{-1} = [\chi^0]^{-1} - \mathrm{K}, \tag{D.11}$$

that appears in many contexts. The approximation $f_{\rm xc} = 0$ is the famous "random phase approximation" (RPA) [225] for the Coulomb interaction; many approximations for $f_{\rm xc}$ have been introduced and any of the exchange–correlation functionals implies a form for $f_{\rm xc}$. The density response function $\chi(\mathbf{r}, \mathbf{r}')$ or $\chi(\mathbf{q}, \mathbf{q}')$ is central in the theory of phonons (Ch. 19), dielectric response in App. E, and other response functions. The extension to dynamical response leads to the theory for much of our understanding of electronic excitations, Ch. 20. For spin response, the Coulomb term V_C is absent and the kernel $f_{\rm xc}^m$ leads to the Stoner response function, Eq. (2.5), and the RPA expressions for magnons.

The classic approach for finding χ is to calculate χ^0 from (D.3)–(D.5) and solve the inverse matrix equation (D.11). Despite the elegant simplicity of the equations, the solution can be a laborious procedure except in the simplest cases. An equally elegant approach much more suited for calculations in real electronic structure problems is described in Ch. 19.

D.3 Dynamic response and Kramers–Kronig relations

Harmonic oscillator

The basic ideas of linear response can be appreciated starting with the simple classical driven harmonic oscillator as described eloquently by P. C. Martin [811]. The equation for the displacement x is

$$M \frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} = -K x(t) - \Gamma \frac{\mathrm{d}x(t)}{\mathrm{d}t} + F(t), \tag{D.12}$$

where $F(t)$ is the driving force and Γ is a damping term. If the natural oscillation frequency is denoted $\omega_0 = \sqrt{K/M}$, the response to a force $F(t) = F(\omega)\mathrm{e}^{-\mathrm{i}\omega t}$ with frequency ω is

$$\chi(\omega) \equiv \frac{x(\omega)}{F(\omega)} = \frac{1}{M} \frac{1}{\omega_0^2 - \omega^2 - i\omega\Gamma/M}. \tag{D.13}$$

Note that for real ω the imaginary part of $\chi(\omega)$ is positive since $\Gamma > 0$ corresponds to energy loss. Furthermore, as a function of complex ω, $\chi(\omega)$ is analytic in the upper half-plane, $\Im\omega > 0$; all poles in the response function $\chi(\omega)$ are in the lower half-plane. This

leads to the causal structure of $\chi(\omega)$ that implies the Kramers-Kronig relations below (Exercise D.4).

Frequency–dependent damping

The well-known form for the harmonic oscillator response with a constant Γ suffers from a fatal problem: a constant Γ violates mathematical constraints on the moments of $\chi(\omega)$ and it violates physical reasoning since loss mechanisms vary with frequency. If one introduces a more realistic $\Gamma(\omega)$, there is a simple rule: it is also a response function and must obey the laws of causality, i.e. $\Gamma(\omega)$ must also be a causal function that obeys Kramers-Kronig relations. For example, it might be modeled by a form like (D.13),

$$\Gamma(\omega) = \frac{1}{\omega_1^2 - \omega^2 - i\omega\gamma_1}, \tag{D.14}$$

and so forth. Clearly, this can continue, leading to a continued fraction that is an example of the general memory function formulation of Mori [866].

Kramers–Kronig relations

Because the response functions represent the causal response of the system to external perturbations, they must obey analytic properties illustrated for the harmonic oscillator in (D.13). That is, the response function $\chi(\omega)$ continued into the complex plane is analytic for all $\Im\omega > 0$ in the upper half-plane and has poles only in the lower half-plane. By contour integrations in the complex plane [88, 225] (Exercise D.5), one can than derive the Kramers-Kronig relations that allow one to derive the real and imaginary parts from one another in terms of principle value integrals:

$$\begin{aligned} \mathrm{Re}\chi(\omega) &= -\frac{1}{\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega' \frac{\mathrm{Im}\chi(\omega')}{\omega - \omega'}, \\ \mathrm{Im}\chi(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega' \frac{\mathrm{Re}\chi(\omega')}{\omega - \omega'}. \end{aligned} \tag{D.15}$$

Dynamic response of a quantum system

The response to a time-dependent perturbation is given by Eq. (3.6), which is conveniently solved for a periodic perturbation $\propto \mathrm{e}^{-\mathrm{i}\omega t}$. The analysis is given in original references [867–870] and in many texts [84, 88, 96, 225, 246, 863], leading to the Kubo–Greenwood formula. A general response function in the non-interacting approximation[1] can be written as a complex function, with a small imaginary damping factor $\eta > 0$,

$$\chi^0_{a,b}(\omega) = 2\sum_{i=1}^{\mathrm{occ}}\sum_{j}^{\mathrm{empty}} \frac{[M^a_{ij}]^* M^b_{ij}}{\varepsilon_i - \varepsilon_j + \omega + i\eta}, \tag{D.16}$$

[1] The full many-body expressions can also formally be written in exactly the same form with $M^a_{ij} = \langle\Psi_i|\hat{O}^a|\Psi_j\rangle$ and $\varepsilon_i \to E_i$, which shows that properties such as the Kramers-Kronig relations apply in general and are not restricted to independent-particle approximations.

where the $M^a_{ij} = \langle\psi_i|\hat{O}^a|\psi_j\rangle$ and M^b_{ij} are matrix elements of appropriate operators, e.g. the Fourier components defined following Eq. (D.5) or the momentum matrix elements in the expression for the dielectric function in Eq. (20.2). The real and imaginary parts can be written explicitly as

$$\begin{aligned}
\mathrm{Re}\chi^0(\omega)_{a,b} &= \sum_{i=1}^{\mathrm{occ}} \sum_{j}^{\mathrm{empty}} \frac{[M^a_{ij}]^* M^b_{ij}}{(\varepsilon_i - \varepsilon_j)^2 - \omega^2}, \\
\mathrm{Im}\chi^0(\omega)_{a,b} &= \sum_{i=1}^{\mathrm{occ}} \sum_{j}^{\mathrm{empty}} |[M^a_{ij}]^* M^b_{ij} \delta(\varepsilon_j - \varepsilon_i - \omega).
\end{aligned} \tag{D.17}$$

An important result from Eq. (D.17) is that the imaginary part of the response function $\chi^0(\omega)$ is just a joint density of states (Sec. 4.7) as a function of $\omega = \varepsilon_j - \varepsilon_i$, weighted by the matrix elements.

Dynamical response in self-consistent field theories

The generalization of the independent-particle expressions to self-consistent field approaches is straightforward using the expressions derived in Sec. D.2. The only change is that the effective field is itself time- or frequency-dependent, $V_{\mathrm{eff}} \to V_{\mathrm{eff}}(t)$ or $V_{\mathrm{eff}}(\omega)$. Within the linear response regime, the relevant quantity is the kernel K given in $\mathbf{r}$ space in Eq. (9.12) or in $\mathbf{q}$ space by (D.10), generalized to include time dependence. The explicit expression in $\mathbf{q}$ space is

$$\begin{aligned}
K(\mathbf{q}, \mathbf{q}', t - t') &= \frac{\delta V_{\mathrm{int}}(\mathbf{q}, t)}{\delta n(\mathbf{q}', t')} \\
&= \frac{4\pi}{q^2}\delta_{\mathbf{q},\mathbf{q}'}\delta(t - t') + \frac{\delta^2 E_{\mathrm{xc}}[n]}{\delta n(\mathbf{q}, t)\delta n(\mathbf{q}', t')},
\end{aligned} \tag{D.18}$$

where the Coulomb interaction is taken to be instantaneous and we have used the fact that K can only depend upon a time difference. Fourier transforming leads to the form

$$K(\mathbf{q}, \mathbf{q}', \omega) = V_C(q)\delta_{\mathbf{q},\mathbf{q}'} + f_{\mathrm{xc}}(\mathbf{q}, \mathbf{q}', \omega), \tag{D.19}$$

and a similar expression in $\mathbf{r}$ space. Thus the dynamical generalization of (D.11) can be written in compact form as

$$\chi(\omega) = \chi^0(\omega)[1 - \chi^0(\omega)K(\omega)]^{-1}. \tag{D.20}$$

Note that K itself is a response function, so that it also must have the analytical properties required by causality, it must vanish at high frequency, *etc.* Specific expressions that illustrate how to use this general expression are given in Sec. 20.2.

D.4 Green's functions

Green's functions are widely used in theoretical physics [96, 654, 863]. For independent-particle hamiltonians, the most important Green's function is the spectral function in terms

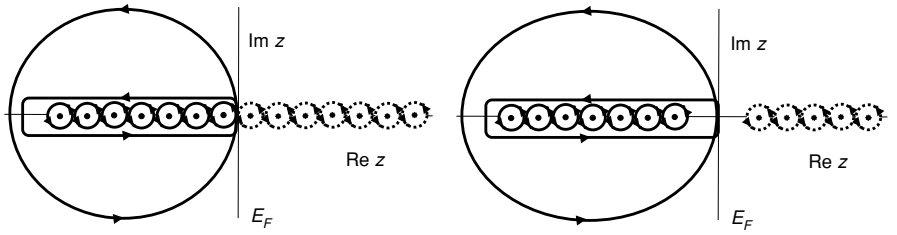


Figure D.1. Contours for line integration over the spectral function to derive integrated quantities. The contours shown enclose all poles below the Fermi energy. (Dotted contours indicate empty states not included.) The integral of the trace $\mathrm{Tr}(G(z)) = \sum_\alpha G_{\alpha,\alpha}(z)$ is the total number of particles. The sum of independent-particle energies is $\mathrm{Tr}\hat{H}[G(z)]$, etc. The left-hand figure indicates a metal where the contour necessarily passes arbitrarily close to a pole. The right-hand figure indicates an insulator where the contour passes through a gap. Whenever z is far from any pole, G decays as a function of distance and therefore can be considered localized.

of the time-independent eigenstates of the hamiltonian

$$G(z, \mathbf{r}, \mathbf{r}') = \sum_i \frac{\psi_i(\mathbf{r})\psi_i(\mathbf{r}')}{z - \varepsilon_i}, \tag{D.21}$$

where z is a complex variable. This may be written in a more general form in terms of any complete set of basis states $\chi_\alpha(\mathbf{r})$,

$$G(z, \mathbf{r}, \mathbf{r}') = \sum_{\alpha,\beta} \chi_\alpha(\mathbf{r}) \left[\frac{1}{z - \hat{H}}\right]_{\alpha,\beta} \chi_\beta(\mathbf{r}'), \tag{D.22}$$

or

$$G_{\alpha,\beta}(z) = \left[z - \hat{H}\right]^{-1}_{\alpha,\beta}. \tag{D.23}$$

The density of states per unit energy projected on the basis function α is given by

$$n_\alpha(\varepsilon) = -\frac{1}{\pi}\mathrm{Im}G_{\alpha,\alpha}(z = \varepsilon + i\delta), \tag{D.24}$$

where δ is a positive infinitesimal, and the total density of states is given by

$$n(\varepsilon) = -\frac{1}{\pi}\mathrm{ImTr}G(z = \varepsilon + i\delta). \tag{D.25}$$

Total integrated quantities at $T = 0$ can be derived by contour integrations in the complex z plane as illustrated in Fig. D.1. Integration of $G(z)$ around each pole in a counterclockwise direction gives $2\pi i$. The contour C can be any closed line that encircles the poles, so that the density matrix is given by

$$\rho(\mathbf{r}, \mathbf{r}') = \frac{1}{2\pi i}\int_C \mathrm{d}z\ G(z, \mathbf{r}, \mathbf{r}'); \tag{D.26}$$

the density by $n(\mathbf{r}) = \rho(\mathbf{r}, \mathbf{r})$; the total number of electrons by

$$N = \int_{-\infty}^{E_F} \mathrm{d}\varepsilon n(\varepsilon) = \frac{1}{2\pi i}\int_C \mathrm{d}z\ \mathrm{Tr}G(z); \tag{D.27}$$

and sum of occupied eigenvalues by

$$\sum_{i\ occ} \varepsilon_i = \int_{-\infty}^{E_F} \mathrm{d}\varepsilon\ \varepsilon\ n(\varepsilon) = \frac{1}{2\pi i} \int_C \mathrm{d}z\ z\ \mathrm{Tr}G(z). \tag{D.28}$$

Since the total energy in the Kohn–Sham method can be derived from the sum of eigenvalues and the density, it follows that all quantities related to total energy can be derived from the independent-particle Kohn–Sham Green's function. Expressions for energies and forces are given in Ch. 23.

SELECT FURTHER READING

Doniach, S., and Sondheimer, E. H., *Green's Functions for Solid State Physicists (Reprinted in Frontiers in Physics Series, No. 44)*, W. A. Benjamin, Reading, Mass., 1974.

Fetter, A. L., and Walecka, J. D., *Quantum Theory of Many-particle Systems*, McGraw-Hill, New York, 1971. [862]

Mahan, G. D., *Many-Particle Physics, 3rd Ed.*, Kluwer Academic/Plenum Publishers, New York, 2000.

Martin, P. C., *Measurement and Correlation Functions*, Gordon and Breach, New York, 1968.

Pines, D., *Elementary Excitations in Solids*, Wiley, New York, 1964.

Exercises

D.1 Derive the general form of the density response function χ_n in Fourier space, (D.5). This applies to any function, periodic or non-periodic.

D.2 Derive the second form given in (D.11) from the first expression. Hint: Move all terms involving χ to the left-hand side, solve for χ in terms of χ^0 and K, and invert both sides of the equation.

D.3 See Exercise 9.7 for the way in which the response function can be used to analyze the form of the energy functionals near the minimum.

D.4 Show that the response of a harmonic oscillator, Eq. (D.13), obeys the KK relations. Hint: The key point is the sign of the damping term that corresponds to energy loss, i.e. $\Gamma > 0$. See Exercise D.5 for an explanation.

D.5 Derive the KK relations, Eq. (D.15), from the analytic properties of the response functions. Causality requires that all poles as a function of complex frequency z be in the lower plane $\Im z < 0$. Hint: An integral along the real axis can be closed in the upper plane with a contour that is at $|z| \to \infty$. Since the contour encloses no poles, the line integral vanishes; also the integral at infinity vanishes. The integral along the axis can be broken into the principal value parts and the residue parts leading to Eq. (D.15). See [88, 225].

Appendix E

Dielectric functions and optical properties

Summary

Dielectric functions are the most important response functions in condensed matter physics: photons are perhaps the most important probe in experimental studies of matter; electrical conductivity and optical properties are among the most important phenomena in technological applications as well as everyday life. Dielectric functions can be defined in terms of currents and fields, which is most appropriate for conductivity and optical response, or in terms of densities and scalar potentials, which is most appropriate for static problems. The needed expressions follow from Maxwell's equations; however, care must be taken in defining the polarization in extended matter, which is treated here and in Ch. 22. This appendix provides the phenomenological definitions; the role of electronic structure is to provide the fundamental foundations in terms of the underlying quantum theory of the electrons, which is the subject of Chs. 19, 20, and 22.

E.1 Electromagnetic waves in matter

Maxwell's equations for electromagnetic fields interacting with particles having charge Q ($Q = -e$ for electrons) and number density n

$$\begin{aligned} \boldsymbol{\nabla} \cdot \mathbf{E} &= 4\pi Qn, \qquad & \boldsymbol{\nabla} \times \mathbf{E}(t) &= -\frac{1}{c}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}, \\ \boldsymbol{\nabla} \cdot \mathbf{B} &= 0, & \boldsymbol{\nabla} \times \mathbf{B}(t) &= \frac{4\pi}{c}\mathbf{j} + \frac{1}{c}\frac{\mathrm{d}\mathbf{E}}{\mathrm{d}t}, \end{aligned} \tag{E.1}$$

are the fundamental equations that describe the interactions of particles in matter. The arguments $\mathbf{r}, t$ have been omitted for simplicity and $\mathbf{j}$ is the charge current density that satisfies the continuity equation

$$\boldsymbol{\nabla} \cdot \mathbf{j} = -Q\,\frac{\mathrm{d}n}{\mathrm{d}t}. \tag{E.2}$$

The basic equations of electronic structure, in particular, the hamiltonian, Eq. (3.1), are based upon (E.1) in the non-relativistic limit where the speed of light $c \to \infty$ in which case it is sufficient to take $\mathbf{B} = 0$ and work with the scalar potential V, which satisfies the

Poisson equation

$$\nabla^2 V = -4\pi Qn\ , \quad \text{with} \quad \mathbf{E} = -\boldsymbol{\nabla} V. \tag{E.3}$$

However, in order to describe important physical phenomena, such as the propagation of electromagnetic waves in matter and the response to external fields, it is essential to return to the full equations in (E.1). Here we summarize[1] the phenomenological theory of matter interacting with external time-dependent fields, defining the appropriate quantities carefully to set the stage for proper derivation from electronic structure theory (see especially Chs. 20 and 22).

Two steps are crucial for defining the structure of the theory:

- In order to derive properties of matter under the influence of external fields, the charges and currents in Maxwell's equations must be divided into "internal" and "external,"

$$n = n_{\text{int}} + n_{\text{ext}}; \quad \mathbf{j} = \mathbf{j}_{\text{int}} + \mathbf{j}_{\text{ext}}. \tag{E.4}$$

  Although such a division can be made for any perturbation, electromagnetic interactions are of special importance because the long-range interactions lead to effects that extend over macroscopic distances into the interior of bodies.
- It is useful to *define* polarization $\mathbf{P}$ by

$$\mathbf{P}(\mathbf{r}, t) = \int^t \mathrm{d}t'\ \mathbf{j}_{\text{int}}(\mathbf{r}, t'), \tag{E.5}$$

  which together with (E.2) yields

$$\boldsymbol{\nabla} \cdot \mathbf{P}(\mathbf{r}, t) = -Qn_{\text{int}}(\mathbf{r}, t). \tag{E.6}$$

  Note that each equation leaves the value of $\mathbf{P}$ *defined only to within an additive constant.* This is easily remedied in a finite system, but is an issue in quantum theory of extended matter that has been fully resolved only recently as summarized in Ch. 22.

In terms of the displacement field $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$, Maxwell's equations can be written in the form

$$\begin{aligned} \boldsymbol{\nabla} \cdot \mathbf{D} &= 4\pi Qn_{\text{ext}}, & \boldsymbol{\nabla} \times \mathbf{E}(t) &= -\frac{1}{c}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}, \\ \boldsymbol{\nabla} \cdot \mathbf{B} &= 0, & \boldsymbol{\nabla} \times \mathbf{B}(t) &= \frac{4\pi}{c}\mathbf{j}_{\text{ext}} + \frac{1}{c}\frac{\mathrm{d}\mathbf{D}}{\mathrm{d}t}. \end{aligned} \tag{E.7}$$

The advantage of this form is that all source terms are "external." In the interior of a sample, n_{ext} and $\mathbf{j}_{\text{ext}}$ vanish even though they can lead to fields inside the sample. As shown by (E.1) and (E.7), $\mathbf{E}$ is the *total field* in the material, whereas $\mathbf{D}$ is the field due only to external sources. Thus the value of $\mathbf{D}$ at any point is independent of the material and is the same as if the material were absent.

[1] Following the clear presentation of [88], Sec. 20.2.

E.2 Conductivity and dielectric tensors

Solution of the equations requires the material relation of $\mathbf{j}_{\rm int}$ or $n_{\rm int}$ to the total fields $\mathbf{E}$ and $\mathbf{B}$. To linear order, the most general relation is

$$\mathbf{j}_{\rm int}(\mathbf{r}, t) = \int d\mathbf{r}' \int^{t} dt' \sigma(\mathbf{r}, \mathbf{r}', t - t')\mathbf{E}(\mathbf{r}', t'), \tag{E.8}$$

where $\sigma(\mathbf{r}, \mathbf{r}', t - t')$ is the microscopic conductivity tensor. For a perturbation with time dependence $\propto \exp(i\omega t)$, (E.8) becomes

$$\mathbf{j}_{\rm int}(\mathbf{r}, \omega) = \int d\mathbf{r}' \sigma(\mathbf{r}, \mathbf{r}', \omega)\mathbf{E}(\mathbf{r}', \omega), \tag{E.9}$$

which implies

$$\mathbf{D}(\mathbf{r}, \omega) = \int d\mathbf{r}' \epsilon(\mathbf{r}, \mathbf{r}', \omega) \cdot \mathbf{E}(\mathbf{r}', \omega) \quad \text{or} \quad \mathbf{E}(\mathbf{r}, \omega) = \int d\mathbf{r}' \epsilon^{-1}(\mathbf{r}, \mathbf{r}', \omega)\mathbf{D}(\mathbf{r}', \omega), \tag{E.10}$$

where

$$\epsilon(\mathbf{r}, \mathbf{r}', \omega) = \mathbf{1}\, \delta(\mathbf{r} - \mathbf{r}') + \frac{4\pi i}{\omega}\sigma(\mathbf{r}, \mathbf{r}', \omega). \tag{E.11}$$

Note that ϵ and σ are the response to the *total field* $\mathbf{E}$, whereas ϵ^{-1} is the response to an *external field*. Interestingly, $\sigma(\omega)$, $\epsilon(\omega) - 1$, and $\epsilon^{-1}(\omega) - 1$ are all response functions and each satisfies Kramers-Kronig relations, (D.15).

The macroscopic average functions $\bar{\epsilon}(\omega)$ or $\bar{\sigma}(\omega)$ are directly measured by the index of refraction for photons and response to macroscopic electric fields, e.g. conductivity and dielectric response where the measured voltage is the line integral of the internal electric field $\mathbf{E}$. On the other hand, scattering of charged particles directly measures $\epsilon^{-\mathbf{1}}(\mathbf{q}, \omega)$([225] p. 126), where $\mathbf{q}$ and ω are the momentum and energy transfers.

E.3 The *f* sum rule

The dielectric functions satisfy the well-known "f sum rule," for which Seitz [1] attributes the original derivation to Wigner [872] and Kramers [873]. A simple way to derive the sum rule ([225], p. 136) is to note that in the $\omega \to \infty$ limit, the electrons act as free particles, from which it follows that (Exercise E.1)

$$\epsilon_{\alpha\beta}(\omega) \to \delta_{\alpha\beta}\left[1 - \frac{\omega_p^2}{\omega^2}\right], \tag{E.12}$$

where ω_p is the plasma frequency $\omega_p^2 = 4\pi(NQ^2/\Omega m_e)$, with N/Ω the average density. (As a check note that this is the first term in square brackets in Eq. (20.2).) Combining this with the Kramers-Kronig relations, Eq. (D.15) leads to (Exercise E.2)

$$\int_0^\infty d\omega\ \omega\ \mathrm{Im}\epsilon_{\alpha\beta}(\omega) = \frac{\pi}{2}\omega_p^2\delta_{\alpha\beta}, \quad \text{or} \quad \int_0^\infty d\omega\ \omega\ \mathrm{Re}\sigma_{\alpha\beta}(\omega) = \frac{\pi}{2}\frac{Q^2}{m_e}\frac{N}{\Omega}\delta_{\alpha\beta}. \tag{E.13}$$

A similar sum rule is satisfied by $\epsilon^{-1}_{\alpha\beta}(\omega)$. Finally, all the versions of the f sum rule apply to the exact many-body response as well as to the simple non-interacting approximation, because the sum rule depends only upon the Kramers-Kronig relations and the high $\omega \to \infty$ limit, in which the electrons always act as uncorrelated free particles.

E.4 Scalar longitudinal dielectric functions

The dielectric relations, Eq. (E.10), can also be written in terms of scalar potentials. This is sufficient for static problems and is convenient for many uses, especially applications in density functional theory, which is cast in terms of potentials and densities. This is called "longitudinal" because it only applies to electric fields that can be derived from a potential $\mathbf{E}(\mathbf{r}) = -\boldsymbol{\nabla} V(\mathbf{r})$. Thus the electric field in Fourier space $\mathbf{E}(\mathbf{q}) = i\mathbf{q}V(\mathbf{q})$ is longitudinal, i.e. parallel to $\mathbf{q}$. Combining (E.3), (E.4), and (E.7), it follows that [152]

$$\epsilon^{-1}(\mathbf{q}, \mathbf{q}', \omega) = \frac{\delta V^C_{\text{total}}(\mathbf{q}, \omega)}{\delta V_{\text{ext}}(\mathbf{q}', \omega)} \quad \text{or} \quad \epsilon(\mathbf{q}, \mathbf{q}', \omega) = \frac{\delta V_{\text{ext}}(\mathbf{q}, \omega)}{\delta V^C_{\text{total}}(\mathbf{q}', \omega)}, \tag{E.14}$$

where the total Coulomb potential is denoted V^C_{total}, i.e. the potential acting on an infinitesimal test charge which does not include the effective exchange–correlation potential V_{xc} that acts upon an electron.

Expressions for ϵ and ϵ^{-1} in terms of electronic states can be derived from the general formulas for response functions χ^0 (Eqs. (D.3)–(D.5)) and χ (Eq. (D.11)). In particular, it follows that (Exercise E.3)

$$\epsilon^{-1}(\mathbf{q}, \mathbf{q}', \omega) = \delta(\mathbf{q} - \mathbf{q}') + V_C(q)\chi(\mathbf{q}, \mathbf{q}', \omega), \tag{E.15}$$

where $V_C(q) = 4\pi e^2/q^2$ is independent of ω (the same as in Eq. (D.10) and we have set $Q = -e$). For a theory in which the electrons interact via an effective field, as in the Kohn–Sham approach, χ is most readily calculated using the expression (D.11)

$$\epsilon^{-1} = 1 + \frac{V_C\chi^0}{1 - (V_C + f_{\text{xc}})\chi^0} = \frac{1 - f_{\text{xc}}\chi^0}{1 - (V_C + f_{\text{xc}})\chi^0}. \tag{E.16}$$

The equation appears simple because the arguments have been omitted, but actual evaluation can be tedious since products such as $f_{\text{xc}}\chi^0$ stand for convolutions over all internal wavevectors and frequencies.

The simplest case is the electron gas (Ch. 5), where χ is non-zero only for $\mathbf{q} = \mathbf{q}'$ and the expressions can be evaluated analytically. The Lindhard expressions, Eq. (5.38), for $\chi^0(q, \omega)$ are given in Sec. 5.4, from which can be derived all the other response functions.

In a crystal, the wavevectors can always be written as $\mathbf{q} = \mathbf{k} + \mathbf{G}$ and $\mathbf{q}' = \mathbf{k} + \mathbf{G}'$, where $\mathbf{k}$ is restricted to the first Brillouin zone, so that $\epsilon(\mathbf{k} + \mathbf{G}, \mathbf{k} + \mathbf{G}', \omega)$ is a matrix $\epsilon_{\mathbf{G}\mathbf{G}'}(\mathbf{k}, \omega)$ and, similarly, for the inverse matrix, $\epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{k}, \omega)$. Optical phenomena involve long wavelengths, $\mathbf{G} = 0$ and $\mathbf{G}' = 0$, and are described by the macroscopic dielectric function $\epsilon(\mathbf{k}, \omega)$, defined by the ratio of internal to external macroscopic fields ($\mathbf{G} = \mathbf{G}' = 0$) keeping the short wavelength ($\mathbf{G}' \neq 0$) *external* fields fixed, it follows that [152, 871]

(Exercise E.4)

$$\epsilon(\mathbf{k},\omega)=\frac{\delta V_{\rm ext}(\mathbf{k},\omega)}{\delta V^{C}_{\rm total}(\mathbf{k},\omega)}=\frac{1}{\epsilon^{-1}_{00}(\mathbf{k},\omega)}. \tag{E.17}$$

Finally, the full dielectric tensor can be recovered considering different directions $\hat{\mathbf{k}}$ using the fact [152] that for long wavelengths, the scalar dielectric function is related to the dielectric tensor by [152]

$$\epsilon(\mathbf{k},\omega)=\lim_{|\mathbf{k}|\to 0}\hat{\mathbf{k}}_\alpha\epsilon_{\alpha\beta}(\mathbf{k},\omega)\hat{\mathbf{k}}_\beta. \tag{E.18}$$

In a cubic crystal $\epsilon_{\alpha\beta}=\epsilon\delta_{\alpha\beta}$, but in general (E.18) depends upon the direction in which the limit is taken.

E.5 Tensor transverse dielectric functions

The general cases of time-dependent electric and magnetic fields can conveniently be treated by calculation of the current response to the vector potential $\mathbf{A}$. The perturbation can be written in terms of $\mathbf{A}$ as

$$\Delta\hat{H}(t)=\frac{1}{2m_e}\sum_i\left\{\left[\mathbf{p}_i-\frac{e}{c}\mathbf{A}(t)\right]^2-\mathbf{p}_i^2\right\}, \tag{E.19}$$

where $\mathbf{E}(t)=-(1/c)(\mathrm{d}\mathbf{A}/\mathrm{d}t)$ or $\mathbf{E}(\omega)=-(i\omega/c)\mathbf{A}(\omega)$, and the magnetic field is given by $B=\boldsymbol{\nabla}\times\mathbf{A}$. The desired response is the current density $\mathbf{j}$. For a transverse electromagnetic wave this is the appropriate response function.

Formulas for response function in the independent-particle approximation have the general form given in App. D and are given explicitly in Sec. 20.1. Self-consistent field expressions have exactly the same form as for the scalar dielectric function except that they involve an effective "exchange–correlation vector potential" that is the fundamental quantity in "current functional theory" [333, 335, 362].

E.6 Lattice contributions to dielectric response

In an ionic insulator, the motion of the ions contributes to the low-frequency dielectric response [90, 152, 874], where the electronic contribution can be considered constant as a function of frequency ω. All quantities are properly defined *holding the macroscopic electric field* $\mathbf{E}_{\rm mac}$ *constant*, which gives the intrinsic response. The macroscopic field is controlled by external conditions, boundary conditions, etc., and such effects should be taken into account in the specific solution. The Born effective charge tensor for each ion I is defined by

$$Z^*_{I,\alpha\beta}|e|=\left.\frac{\mathrm{d}\mathbf{P}_\alpha}{\mathrm{d}\mathbf{R}_{I,\beta}}\right|_{\mathbf{E}_{\rm mac}}, \tag{E.20}$$

where the macroscopic electric field is held constant. The effective charge is non-zero for some displacements in any ionic crystal, and it has been shown that in all elemental crystals with three or more atoms per cell [875] (with the exception [876] of two special cases out of the 230 space groups) there must also non-zero effective charges. In fact, there are large measured effective charges and infrared absorption known in elemental crystals such as trigonal Se [875]. The polarization caused by the effective charges leads to non-analytic terms in the force constant matrix defined in Eq. (19.9), which has the form (see Eq. 4.7 of [152]).

$$C_{s,\alpha;s',\alpha'}(\mathbf{k}) = C^N_{s,\alpha;s',\alpha'}(\mathbf{k}) + \frac{4\pi e^2}{\Omega}\left[\sum_\gamma \hat{\mathbf{k}}_\gamma Z^*_{I,\gamma\alpha}\right]^\dagger \frac{1}{\epsilon(\mathbf{k})}\left[\sum_\gamma \hat{\mathbf{k}}_\gamma Z^*_{I,\gamma\beta}\right], \quad \text{(E.21)}$$

where C^N is the normal analytic part of C and $\epsilon(\mathbf{k})$ is the low-frequency electronic dielectric constant. The full dielectric function including the lattice contribution is given by Cochran and Cowley [874] and the low-frequency limit is given in [152], Eq. (7.1).

Similarly, one can define proper piezoelectric constants [722, 877, 878] in the absence of macroscopic fields,

$$e_{\alpha\beta\gamma} = \left.\frac{\mathrm{d}\mathbf{P}_\alpha}{\mathrm{d}u_{\alpha\beta}}\right|_{\mathbf{E}_{\mathrm{mac}}}, \quad \text{(E.22)}$$

where $u_{\alpha\beta}$ denotes the strain tensor of Eq. (G.2). The effect can be separated into a pure strain effect and an internal displacement contribution,

$$e_{\alpha,\beta\gamma} = e^0_{\alpha,\beta\gamma} + |e| \sum_{s,\delta} Z^*_{s,\alpha\delta}\Gamma_{s,\delta,\beta\gamma}, \quad \text{(E.23)}$$

where $Z^*_{s,\alpha\delta}$ is the same effective charge tensor that governs infrared response of optic modes and Γ is defined in (G.14). This division facilitates calculations and clarifies relations of measurable physical quantities. Crystals with permanent moments present a particular problem, in that a rotation of the moment might be termed a piezoelectric effect. This is an "improper effect" and it has been shown that "proper" expressions for the polarization, such as the Berry's phase form in Sec. 22.2, do not contain such terms [803].

SELECT FURTHER READING

Definitions of dielectric functions:

Pick, R., Cohen, M. H. and Martin, R. M., "Microscopic theory of force constants in the adiabatic approximation," *Phys. Rev. B* 1:910–920, 1970.

Wiser, N., "Dielectric constant with local field effects included," *Phys. Rev.* 129:62–69, 1963.

Pines, D. *Elementary Excitations in Solids*, Wiley, New York, 1964.

A modern formulation of the dielectric equations in terms of polarization currents:

Marder, M. *Condensed Matter Physics*, John Wiley and Sons, New York, 2000.

Exercises

E.1 Derive (E.12) for the dielectric tensor at high frequency using only the fact that electrons respond as free particles at sufficiently high frequency. It may be helpful to relate to the high-frequency limit of the harmonic oscillator response function give in Sec. D.3.

E.2 Show that the f sum rule, (E.13), follows from the high-frequency behavior in (E.12) and the Kramers-Kronig relations, (D.15).

E.3 Show that (E.15) results from the definition of internal and external charges in Eq. (E.4) and the definition of ϵ^{-1} in (E.14).

E.4 The expression for the macroscopic dielectric function, (E.17), can be derived by carefully applying the definition that it is the ratio of external to total internal fields given in (E.17) *for the case where the short wavelength external fields vanish*, and using the definition that the inverse function is the response to external fields. Use these facts to derive (E.17).

Appendix F

Coulomb interactions in extended systems

Summary

The subject of this appendix is formulations and explicit equations for the total energy that properly take into account the long-range effects of Coulomb interactions. We emphasize the Kohn–Sham independent-particle equations and expressions for total energy; however, the ideas and many of the equations also apply to many-body calculations. There are three main issues:

- Identifying various convenient expressions that each yield properly the intrinsic total energy per formula unit for an extended bulk system.
- Understanding and calculating the effect upon the average potential in a bulk material due to dipole terms at surfaces and interfaces.
- Treating finite systems, where there is no essential difficulty, but where it is convenient to carry out the calculations in a periodic "supercell" geometry.

F.1 Basic issues

There is a simple set of guiding principles that must be followed to properly treat long-range Coulomb interactions in extended systems. If the calculations are carried out in a cell that represents an infinite system, i.e. the unit cell of a crystal, or a "supercell" constructed so that its limiting behavior represents a macroscopic system, then:

- The cell must be chosen to be neutral;
- The neutral cell can be used to define a proper thermodynamic "reference state" if in addition we require that there is no average (macroscopic) electric field;
- The average electrostatic potential is *not* an intrinsic property of condensed matter. The value is ill defined in an infinite system. In a large (but finite) sample, the value relative to vacuum depends upon surface conditions.

The first condition is obvious because otherwise the Coulomb energies of the extended system diverge. The second is less obvious, but is clearly required because there is no lower bound to the energy in an infinite system with an electric field. In a metal there is no problem since there can be no uniform electric field in equilibrium. However, in general, in an insulator, the total energy is the energy of this "reference state" plus changes in energy

due to the presence of long-range electric fields. This is the essence of a dielectric in which the energy is a function of applied fields [448, 790], which can be described in terms of dielectric response functions derived by perturbation theory (Ch. 19).[1]

The expressions for the total energy given in Sec. 3.2 and in the chapters on density functional theory (see e.g., Eqs. (7.5), (9.7), (9.9), and (9.13)) are organized into neutral groupings so that they are in the proper form to define the intrinsic, extensive properties of condensed matter in the large size (or thermodynamic) limit. These classical Coulomb contributions to the total energy given in Eq. (3.14) are determined solely by the charge density of the electrons, the nuclei, and any external charges. All effects of quantum mechanics on the electrons and correlations among the electrons can be separated into the other terms in the total energy as expressed in Eq. (3.16) and the "xc" terms in density functional theory; these are short-range in nature and not subject to convergence problems.

A comment is in order regarding terminology. In density functional theory, the "external potential" has a central role. However, the external potential due to the charged nuclei diverges in an infinite system. This nomenclature should cause no difficulty so long as one maintains the principle that the long-range part of the Hartree potential is grouped with the nuclear potential in order to have a well-defined "external potential" and total energy. For example, in the bulk of a crystal the potential regarded as external may include effects of electrons at a large distance which are not part of the intrinsic bulk system.

There are three typical ways to specify the Coulomb energy and the potential of extended systems. One is to add and subtract a uniform background: then the energy can be expressed as the sum of the classical energy of nuclei (or ions) in a compensating negative background plus the total energy of the system of electrons in a compensating positive background. This has the advantage of simplicity and may be close to the real situation in materials where the electrons are nearly uniform. However, it leads to expressions for the total energy that often involve small differences between large numbers that are difficult to interpret physically. The second approach is to "smear" the ions, which allows a convenient rearrangement of terms that is especially useful in Fourier space expressions. A third method is a variation in which one finds the *difference* from isolated neutral atoms (or neutral spherical atomic-like species). Then we only deal with the *difference* between two neutral systems, which has obvious advantages since it relates to the real physical problem of the binding energy relative to atoms. However, it requires that we either specify properties of the real atom or define an arbitrary neutral reference density.

F.2 Point charges in a background: Ewald sums

The Ewald method for summing the Coulomb interactions of point charges is based upon transformation of the potential due to an infinite periodic array of charges. The result is two sums, one in reciprocal space and one in real space, each of which is absolutely convergent. The approach is intimately connected to the expressions for total energy, which must be

[1] Special care must be taken if there is a polarization in the absence of an average electric field, i.e. in pyroelectrics or polled ferroelectrics. See Ch. 22.

evaluated in a consistent way to eliminate the divergent terms in both the total energy and the Kohn–Sham potential. This is the approach used in Sec. 13.1, in particular in Eq. (13.1). The arguments given here are the justification for the exclusion of the $\mathbf{G} = 0$ Fourier components in the expressions for the Hartree term in the Kohn–Sham potential and the $\mathbf{G} = 0$ term in the Hartree energy.

The first step is the identification of appropriate neutral groupings by adding and subtracting a uniform positive background charge density n^+, which is equivalent to adding n^+ and a uniform negative density $n^- = -n^+$. This allows us to rewrite the total energy, Eq. (3.14), (or any of the expressions in the density functional theory chapters) as the classical Coulomb energy

$$E^{CC} = E'_{\text{Hartree}}[n(\mathbf{r}) + n^+] + \int \mathrm{d}^3 r\, V'_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) + E'_{II}, \tag{F.1}$$

where *each term is neutral*. The effects of n^+ are incorporated in E'_{Hartree}, which is the Hartree-like energy having exactly the same form as Eq. (3.15)

$$E'_{\text{Hartree}}[n] = \frac{1}{2} \int \mathrm{d}^3 r \mathrm{d}^3 r' \frac{[n(\mathbf{r}) + n^+)(n(\mathbf{r}') + n^+]}{|\mathbf{r} - \mathbf{r}'|} = \frac{1}{2} 4\pi {\sum_{\mathbf{G}\neq 0}}' \frac{|n(\mathbf{G})|^2}{G^2}, \tag{F.2}$$

with n replaced by the neutral density $n + n^+$. In Fourier space, the addition of n^+ simply amounts to omitting the $\mathbf{G} = 0$ term since $n + n^+$ has zero average value. In (F.1), V'_{ext} is the potential due to the nuclei (or ions) plus the negative background n^-; again, in Fourier space, one simply omits the $\mathbf{G} = 0$ term. The final term is the sum of all interactions involving the nuclei (or ions) and n^-, which is defined to be the Madelung energy and can be evaluated by the Ewald transformation.

The Ewald transformation[2] is based upon the fact that expressions for lattice sums can be written in either real or reciprocal space, or a combination of the two. The explicit formulas utilize the relation ([88], p. 271),

$$\begin{aligned} \sum_{\mathbf{T}} \frac{1}{|\mathbf{r} - \mathbf{T}|} \rightarrow\ & \frac{2}{\sqrt{\pi}} \sum_{\mathbf{T}} \int_\eta^\infty \mathrm{d}\rho \mathrm{e}^{-|\mathbf{r}-\mathbf{T}|^2 \rho^2} \\ & + \frac{2\pi}{\Omega} \sum_{\mathbf{G}\neq 0} \int_0^\eta \mathrm{d}\rho \frac{1}{\rho^3} \mathrm{e}^{-|\mathbf{G}|^2/(4\rho^2)} \mathrm{e}^{\mathrm{i}\mathbf{G}\cdot\mathbf{r}}, \end{aligned} \tag{F.3}$$

where $\mathbf{T}$ are the lattice translation vectors and $\mathbf{G}$ are reciprocal lattice vectors. The integrals can be computed in terms of error functions, $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}u \mathrm{e}^{-u^2}$ and $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$, leading to (see Exercise F.1, [88], p. 271 and [886]),

$$\begin{aligned} \sum_{\mathbf{T}} \frac{1}{|\mathbf{r} - \mathbf{T}|} \rightarrow\ & \sum_{\mathbf{T}} \frac{\mathrm{erfc}(\eta|\mathbf{r} - \mathbf{T}|)}{|\mathbf{r} - \mathbf{T}|} \\ & + \frac{4\pi}{\Omega} {\sum_{\mathbf{G}\neq 0}}' \frac{1}{|\mathbf{G}|^2} \mathrm{e}^{\frac{-|\mathbf{G}|^2}{4\eta^2}} \cos(\mathbf{G}\cdot\mathbf{r}) - \frac{\pi}{\eta^2 \Omega}. \end{aligned} \tag{F.4}$$

[2] The formulas were originally given by Ewald [882], Kornfeld [883] and Fuchs [884] and can be found in extensive reviews, e.g., [879] and [885].

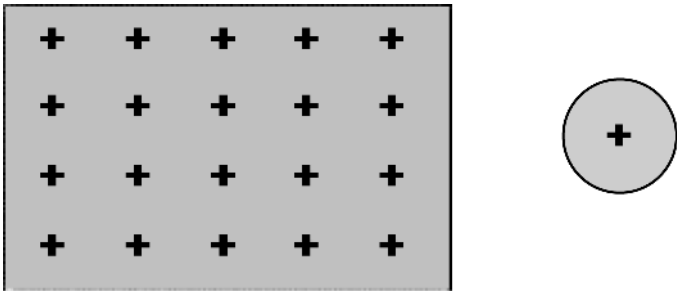


Figure F.1. A lattice of point charges in a uniform compensating background as considered in the Ewald calculation. On the right single a nucleus in a compensating sphere, which provides a good approximation for the Coulomb energy in a close-packed lattice (see Eq. (F.9)).

By dividing the Coulomb sum into two terms in real and reciprocal space, each term in (F.3) and (F.4) is absolutely convergent. The value of η determines the way the sum is apportioned in real and reciprocal space: the result must be independent of η if carried to convergence, and a choice of $\eta \approx |\mathbf{G}|_{min}$ allows each sum to be computed with only a few terms.

The sum on the left-hand sides of these two equations is the electrostatic potential at a general point $\mathbf{r}$ due to a lattice of unit charges, *which is an ill-defined sum.* The arrow in each equation denotes two definitions required to specify the right-hand side. First, the sum is made finite by including a compensating background, which is accomplished by the omission of the $\mathbf{G} = 0$ term. Second, even with the compensating term, the sum is only conditionally convergent, which reflects the fact that the absolute value of the potential is not defined in an infinite system. In (F.4), the last term is chosen so that the average value of the potential is zero [886]. Since the absolute value of the potential does not affect the total energy of a neutral system, this is sufficient for the total energy given in (F.5) below. However, the average potential is required for other properties as discussed in Sec. F.5; this is not specified by the conditions given so far.

The total Coulomb energy per unit cell of a periodic array of point charges plus the compensating uniform negative background (see Fig. F.1) can be expressed using the potential at each site from (F.4), omitting the self-term for each ion. Assuming that the system is neutral and has no net polarization, the expressions are absolutely convergent (with none of the arbitrariness that occurs in the potential) and can be written for charges Z_s at positions τ_s, $s = 1, \ldots, S$ as

$$\begin{aligned}
\gamma_E &= \frac{e^2}{2} \sum_{s,s'} Z_s Z_{s'} {\sum_{\mathbf{T}}}' \frac{1}{|\tau_{s,s'} - \mathbf{T}|} \\
&= \frac{e^2}{2} \sum_{s,s'} Z_s Z_{s'} \left[{\sum_{\mathbf{T}}}' \frac{\operatorname{erfc}(\eta|\tau_{s,s'} - \mathbf{T}|)}{|\tau_{s,s'} - \mathbf{T}|} + \frac{4\pi}{\Omega} {\sum_{\mathbf{G}\neq 0}}' \frac{1}{|\mathbf{G}|^2} \mathrm{e}^{\frac{-|\mathbf{G}|^2}{4\eta^2}} \cos(\mathbf{G} \cdot \tau_{s,s'}) \right] \\
&\quad - \frac{e^2}{2} \left[\sum_s Z_s^2 \right] \frac{2\eta}{\sqrt{\pi}} - \frac{e^2}{2} \left[\sum_s Z_s \right]^2 \frac{\pi}{\eta^2 \Omega},
\end{aligned} \tag{F.5}$$

Table F.1. Typical values of the Madelung constant α for simple ionic crystals and for simple elemental crystals where the background term has been included.

CsCl	NaCl	wurtzite	zinc-blende	
1.762,68	1.747,57	1.638,70	1.638,06	
bcc	fcc	hcp	sc	diamond
1.791,86	1.791,75	1.791,68	1.760,12	1.670,85

where $\tau_{s,s'} = \tau_{s'} - \tau_s$ and the primes on the sums indicate that the divergent terms are omitted. Self-terms for the ions are excluded by the omission of the $\mathbf{T} = 0$ term for $s = s'$ and by the first term in the last line that cancels a self-term included in the reciprocal space term. The $\mathbf{G} = 0$ term is omitted and the correct effects are taken into account by the second term in the last line, which is the analytic limit for $\mathbf{G} \to 0$. This term is absent in the calculation of Madelung energy for an ionic crystal where $\sum_s Z_s = 0$, but it must be included for evaluating the energy of positive ions in a background of density $n^- = -\sum_s Z_s e/\Omega$. Expression (F.5) can be used to compute the E_{II} term in the total energy in (7.5) and the needed term in (13.1), (13.2), and other expressions.

Finally, the real- and reciprocal space sums in Eq. (F.5) can be written in a different form. The reciprocal space sum can be transformed to the square of a single sum over nuclei I (Exercise F.4),

$$\sum_{s,s'} Z_s Z_{s'} \sum_{\mathbf{G}\neq 0}{}' \frac{1}{|\mathbf{G}|^2} \mathrm{e}^{\frac{-|\mathbf{G}|^2}{4\eta^2}} \cos(\mathbf{G}\cdot\tau_{s,s'}) = \sum_{\mathbf{G}\neq 0}{}' \frac{1}{|\mathbf{G}|^2} \left[\sum_s Z_s \mathrm{e}^{\mathrm{i}\mathbf{G}\cdot\tau_s} e^{\frac{-|\mathbf{G}|^2}{8\eta^2}}\right]^2, \qquad \text{(F.6)}$$

which is the Coulomb energy of a charge distribution consisting of gaussian charges at the ion sites. The real-space sum in Eq. (F.5) involving complementary error functions is a short-range sum over neighbors of the *difference* of the interaction of point charges and gaussian distributed charges.

Madelung constant

The Madelung constant α is a dimensionless constant that characterizes the energy per cell of point charges in a lattice γ_E

$$\gamma_E = -\alpha \frac{(Ze)^2}{2R}. \qquad \text{(F.7)}$$

Representative values of α for are given in Tab. F.1, where $2R$ is the nearest-neighbor distance for ionic crystals (top line of Tab. F.1.) and $R = R_{\mathrm{ws}}$ for elemental crystals (bottom line of Tab. F.1). The neutralizing background is included in the calculation of γ_E as in Eq. (F.5) for all cases where the sum of point charges is not zero; however, it does not enter for ionic crystals with neutral cells of positive and negative charges (see additional comments in Exercise F.2).

For close-packed metals, the energies in Tab. F.1 are very close to the energy of single-point charge Ze in a sphere of uniform compensating charge, where the volume of the

sphere equals that of the Wigner–Seitz cell and its radius is $R = R_{\rm WS}$, as illustrated on the right-hand side of Fig. F.1. This can be understood simply because the cell is nearly spherical and there are no interactions between neutral spherical systems so that only internal energies need to be considered. The electrostatic potential at radius r due to the background is (Exercise F.3)

$$V(r) = Ze\left[\frac{r^2}{2R^3} - \frac{3}{2R}\right], \quad r < R, \tag{F.8}$$

where the constant is chosen to cancel the Ze/r potential from the ion at $r = R$. The total energy is the interaction of the ion with the background, plus the self-interaction of the uniform distribution, (Exercise F.3)

$$E_{\rm sphere} = (Ze)^2\left[-\frac{3}{2R} + \left(\frac{3}{2R} - \frac{9}{10R}\right)\right] = -0.90\frac{(Ze)^2}{R} = -1.80\frac{(Ze)^2}{d}, \tag{F.9}$$

which is very close to the Madelung energies for the close-packed metals in Tab. F.1.

Force and stress

The part of the force on any atom due to the other nuclei or ions, treated as point charges, is easy to calculate from the analytic derivative of the Ewald energy, Eq. (F.5). The background is irrelevant in the derivative and one finds

$$\begin{aligned} -\frac{\partial \gamma_{\rm Ewald}}{\partial \tau_s} = &-\frac{e^2}{2} Z_s \sum_{s'} Z_{s'} {\sum_{\mathbf{T}}}' \left[\eta H(\eta D)\frac{\mathbf{D}}{D^2}\right]_{\mathbf{D}=\tau_{s,s'}-\mathbf{T}} \\ &+ \frac{4\pi}{\Omega}\frac{e^2}{2} Z_s \sum_{s'} Z_{s'} {\sum_{\mathbf{G}\neq 0}}' \left[\frac{1\mathbf{G}}{|\mathbf{G}|^2} \mathrm{e}^{\frac{-|\mathbf{G}|^2}{4\eta^2}} \sin(\mathbf{G}\cdot\tau_{s,s'})\right], \end{aligned} \tag{F.10}$$

where $H'(x)$ is

$$H'(x) = \frac{\partial \mathrm{erfc}(x)}{\partial x} - x^{-1}\mathrm{erfc}(x). \tag{F.11}$$

The contribution of the Ewald term to the stress can be found using the forms in App. G. The sum in real space involves short-range two-body terms that can be expressed in the form of Eq. (G.7), and the sum in reciprocal space has the form of (G.8). The final result is (appendix of [104])

$$\begin{aligned} \frac{\partial \gamma_{\rm Ewald}}{\partial \epsilon_{\alpha\beta}} = &\frac{\pi}{2\Omega\eta^2} \sum_{G\neq 0} \frac{\mathrm{e}^{-G^2/4\eta^2}}{G^2/4\eta^2} \left|\sum_s Z_s \mathrm{e}^{\mathrm{i}\mathbf{G}\cdot\tau_s}\right|^2 \left[\frac{2G_\alpha G_\beta}{G^2}(G^2/4\eta + 1) - \delta_{\alpha\beta}\right] \\ &+ \frac{1}{2}\eta \sum_{s,s'\mathbf{T}}{}' Z_s Z_{s'} H'(\eta D)\, \frac{D_\alpha D_\beta}{D^2}\Bigg|_{(D=\tau_{s'}-\tau_s+\mathbf{T}\neq 0)} \\ &+ \frac{\pi}{2\Omega\eta^2}\left[\sum_s Z_s\right]^2 \delta_{\alpha\beta}. \end{aligned} \tag{F.12}$$

F.3 Smeared nuclei or ions

The terms in the total energy can also be rearranged in a form that is readily applied in pseudopotential calculations.[3] The long-range part of the ion pseudopotential is in the local term $V_I^{\rm local}(\mathbf{r})$ defined for each ion I. If we define the charge density that would give rise to this potential as

$$n_I^{\rm local}(\mathbf{r}) \equiv -\frac{1}{4\pi}\nabla^2 V_I^{\rm local}(\mathbf{r}), \tag{F.13}$$

then the total energy for electrons in the presence of the smeared ion density can be written in terms of the total charge density

$$n^{\rm total}(\mathbf{r}) \equiv \sum_s n_I^{\rm local}(\mathbf{r}) + n(\mathbf{r}). \tag{F.14}$$

One can also define a model ion density different from Eq. (F.13); the ideas remain the same and equations given here are easily modified.

With this definition of $n^{\rm total}$, the ion–ion, the Hartree, and local external terms can be combined to write the total energy, Eq. (7.5), in the form

$$E_{\rm KS} = T_s[n] + \langle \delta \hat{V}_{\rm NL}\rangle + E_{\rm xc}[n] + E'_{\rm Hartree}[n^{\rm total}] - \sum_I E_I^{\rm self} + \delta E_{II}, \tag{F.15}$$

where the non-local pseudopotential term has been added, as has also been done in Eq. (13.1). The Hartree-like term $E'_{\rm Hartree}$ is defined as in Eq. (F.2) with $n \to n^{\rm total}$; the "self" term subtracts the ion self-interaction term included in $E'_{\rm Hartree}$; and the last term δE_{II} is a short-range correction to remove spurious effects if the smeared ion densities $n_I^{\rm local}(\mathbf{r})$ overlap.

The correspondence with the Ewald expression can be seen by choosing the densities $n_I^{\rm local}(\mathbf{r})$ to be Gaussians, in which case this analysis is nothing but a rearrangement of the total energy using the Ewald expression, (F.5). The Fourier sum in (F.5) is included with the electron Hartree and external terms to define $\tilde{E}_{\rm Hartree}[n^{\rm total}]$; the real-space sum in Eq. (F.5) is simply the short-range corrections termed δE_{II}, and the constants in (F.5) are the "self" terms.

Force and stress

The force can be found by differentiating the energy, Eq. (F.15), and the force theorem, keeping in mind that $n^{\rm total}$ explicitly depends upon the ion positions. One finds an expression analogous to (13.3) and (F.10), with the Ewald and local terms rearranged,

$$\mathbf{F}_j^\kappa = -\sum_m \mathrm{i}\mathbf{G}_m \mathrm{e}^{i\mathbf{G}_m\cdot\tau_{\kappa,j}} V^\kappa_{\rm local}(\mathbf{G}_m) n^{\rm total}(\mathbf{G}_m) - \frac{\partial \delta E_{II}}{\partial \tau_{\kappa,j}} + \left[\mathbf{F}_j^\kappa\right]^{\rm NL}, \tag{F.16}$$

[3] See [705] and [617] for description of the ideas and practical expressions for calculations. This form is especially suited for Car–Parrinello simulations, as discussed in Sec. 18.3.

where $\left[\mathbf{F}_j^{\kappa}\right]^{\mathrm{NL}}$ are the non-local final terms on the right-hand side of Eq. (13.3), and the contributions due to δE_{II} are simple short-range two-body terms. Stress is found in a form analogous to the expressions in Sec. F.2.

F.4 Energy relative to neutral atoms

It is appealing and useful to formulate expressions for the total energy relative to atoms.[4] This can be viewed as a reformulation of the expressions in the previous section. The total energy relative to separated atoms is the difference of Eq. (F.15) from the sum of corresponding energies for the separated atoms. There is no simple expression for the difference in kinetic, non-local, and exchange–correlation energies which must be calculated separately.

However, there is a simplification in the Coulomb terms that can be used to advantage. Let us define a neutral density for each atom $n_I^{\mathrm{NA}}(\mathbf{r})$ as the sum of its electronic density $n_I(\mathbf{r})$ and the local density representing the positive ion, just as in (F.14). Then the total density can be written as

$$n^{\mathrm{total}}(\mathbf{r}) \equiv \sum_I n_I^{\mathrm{NA}}(\mathbf{r}) + \delta n(\mathbf{r}), \tag{F.17}$$

where $\delta n(\mathbf{r}) = n(\mathbf{r}) - n^{\mathrm{atom}}(\mathbf{r})$, with $n^{\mathrm{atom}}(\mathbf{r})$ the sum of superimposed atomic densities. Substituting (F.17) into (F.2) leads directly to

$$E'_{\mathrm{Hartree}}[n^{\mathrm{total}}] = E'_{\mathrm{Hartree}}[n^{\mathrm{NA}}] + \int \mathrm{d}\mathbf{r}\; V^{\mathrm{NA}}(\mathbf{r})\delta n(\mathbf{r}) + E'_{\mathrm{Hartree}}[\delta n], \tag{F.18}$$

where $V^{\mathrm{NA}}(\mathbf{r})$ is the sum of Coulomb potentials due to the neutral ion densities.

Since both n^{NA} and δn are neutral densities, i.e. having zero average value, each of the individual terms in (F.18) is well defined and can be treated individually using the Hartree-like expression, Eq. (F.2). One approach is to evaluate the first term using the fact that $n^{\mathrm{NA}}(\mathbf{r})$ is a periodic charge density and transforming to Fourier space. However, this does not take advantage of the construction of $n^{\mathrm{NA}}(\mathbf{r})$ as a sum of neutral, spherical densities. Using this fact, the first term can be written as a sum of intra-atom terms plus short-range interactions between the neutral atomic-like units; subtracting the unphysical self-term for the nucleus (or ion) as in (F.15), we have

$$E'_{\mathrm{Hartree}}[n^{\mathrm{NA}}] - \sum_I E_I^{\mathrm{self}} = \sum_I U_I^{\mathrm{NA}} + \sum_{I<J} U_{IJ}^{\mathrm{NA}}(|\mathbf{R}_I - \mathbf{R}_J|), \tag{F.19}$$

where

$$U_I^{\mathrm{NA}} = \int \mathrm{d}\mathbf{r} V_I^{\mathrm{local}}(\mathbf{r}) n_I(\mathbf{r}) + \frac{1}{2}\int \mathrm{d}\mathbf{r} V_I^{\mathrm{Hartree}}(\mathbf{r}) n_I(\mathbf{r}), \tag{F.20}$$

and the interaction $U_{IJ}^{\mathrm{NA}}(|\mathbf{R}_I - \mathbf{R}_J|)$ is non-zero only for overlapping densities. If the density is strictly zero beyond a cutoff radius, then the interactions also vanish for any

[4] Such a form is particularly useful in local orbital methods where an atomic or atomic-like density is readily available. Informative analysis is given in [601] and [617].

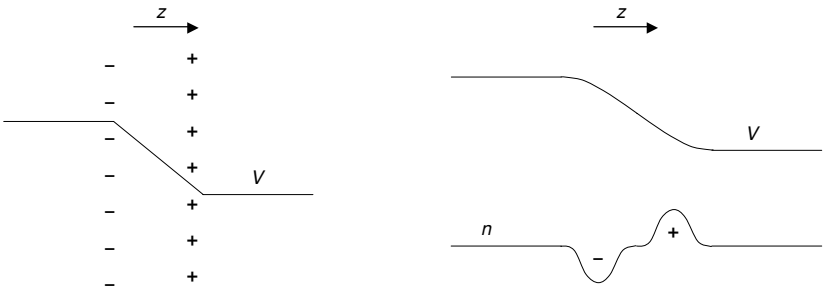


Figure F.2. Schematic for a dipole layer of charge $\sigma(z)$ and the resulting offset of the average potential. On the left is the well-known problem of a parallel-plate capacitor and on the right a schematic illustration of a realistic smooth interface density like that at a surface or interface.

non-overlapping spherical densities [601]. These expressions for the energy are used in expression (15.14), that is particularly useful for local orbital approaches.

F.5 Surface and interface dipoles

Planar distributions of charge are an important special case of the effects of long-range Coulomb interactions that play a major role in surface and interface phenomena. The average electrostatic potential is shifted due to a surface or interface dipole, which gives rise to interface-dependent band-offsets and surface-dependent work functions (see Secs. 2.8 and 13.4). The underlying cause is the long-range Coulomb interaction and the key point is that in the bulk of condensed matter *the absolute energy of a charged particle (*e.g. *an electron) is not an intrinsic bulk property.* One can specify energy relative to some other state (for example, the vacuum) only if the charge state of the entire system is known.

The physical problem is specified by charge density $n(\mathbf{r})$, which is non-zero only near a planar surface or interface. The density $n(\mathbf{r})$ includes both electrons and nuclei, and must be neutral for the energy per particle to be finite. If the coordinate system is fixed with $\hat{z}$ perpendicular to the plane and $\hat{x}$, $\hat{y}$ in the plane, then $n(\mathbf{r})$ can be divided into an average density per unit area $\sigma(z)$ plus $\delta n(\mathbf{r})$, where the latter can vary in the $\hat{x}$, $\hat{y}$ plane. The variations in the plane $\delta n(\mathbf{r})$ give rise to potentials that decrease exponentially [887] as a function of $|z|$. The typical decay length is proportional to the length L_{xy} over which $\delta n(\mathbf{r})$ varies. Thus the only long-range effects are due to $\sigma(z)$.

This leaves us with the problem shown in Fig. F.2, which is equivalent to the planar capacitor shown on the left. The electrostatics is very simple and the *only* effect for z outside the region of the surface or interface is a constant shift of the electrostatic potential that is given by integrating the electric field, which is equivalent to the dipole term

$$\Delta \bar{V}_{\text{Coulomb}} = \int \mathrm{d}z \; z \; \sigma(z). \tag{F.21}$$

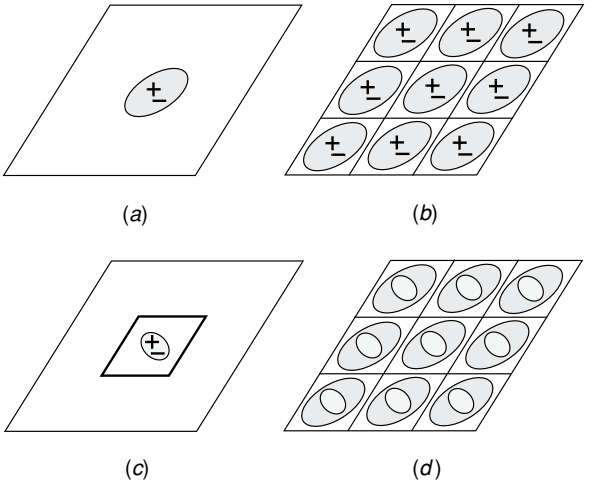


Figure F.3. Schematic illustration of the use of periodic boundary conditions to find the electrostatic potential and to solve Kohn–Sham equations for the isolated system shown in (*a*). The periodically repeated density shown in (*b*) leads to artificial interactions with the images. Subtracting a model density $n_{\rm LM}(\mathbf{r})$ in (*c*) from (*a*) (F.23) leaves the density $n'(\mathbf{r})$ that has no moments for $M \leq M_{\rm max}$. Calculations are done using the periodically repeated $n'(\mathbf{r})$ in (*d*) with expression (F.24) for the electrostatic potential. Provided by P. Schultz; equivalent to Fig. 1 of [880].

This is the dipole that must be calculated from the electronic structure in order to predict band-offsets at interfaces and work functions at surfaces, as referred to in Secs. 2.8 and 13.4.

F.6 Reducing effects of artificial image charges

It is often convenient to apply periodic boundary conditions in calculations of isolated molecules, clusters, or defects in solids. The advantage is that all the machinery developed for crystals is immediately applicable. The disadvantage is unwanted effects due to the use of artificial periodic boundary conditions. There are two types of effects: artificial bands due to overlapping wavefunctions and potentials due to "image charges" from periodically repeated units. Since the bound state wavefunctions are exponentially localized, the longest range effects are Coulomb interactions. Thus it is very useful to identify ways of performing the calculations that minimize effects of the image potentials.

A transparent approach to the problem, with practical equations, can be found in a paper by Schultz [880], as illustrated in Fig. F.3. The goal is to find the properties of the isolated system in part (*a*) using calculations with periodic cells of volume $\Omega \equiv 1/L^3$. If the density is merely repeated periodically as in (*b*) using the usual expressions relating the potentials and densities valid in crystals, then this artifice introduces spurious potentials due to interactions between the system and its periodic images. The effects can be understood in terms of the multi-pole moments of the charge density of one cell (we omit tensor indices for simplicity)

$$\langle n \rangle_M = \int \mathrm{d}\mathbf{r} r^M n(r). \tag{F.22}$$

If the cell is charged (monopole $M = 0$ moment), the sums diverge for any Ω; if there is a dipole ($M = 1$) moment, the limit as $\Omega \to \infty$ depends upon the shape of the cell; quadrupole ($M = 2$) moments lead to convergent expressions for energy (with an error $\propto 1/L^5 = 1/\Omega^{5/3}$), but the potential is only conditionally convergent; the sums are convergent for higher multi-poles.

A general approach to the problem [880] is to divide the density into two parts,

$$n(\mathbf{r}) \equiv n'(\mathbf{r}) + n_{\mathrm{LM}}(\mathbf{r}), \tag{F.23}$$

where $n_{\mathrm{LM}}(\mathbf{r})$ is a model "local moment counter charge" density chosen to reproduce the moments, Eq. (F.22), of $n(\mathbf{r})$ for $M \leq M_{\max}$. One isolated model density $n_{\mathrm{LM}}(\mathbf{r})$ is illustrated in Fig. F.3(c) and the remaining $n'(\mathbf{r})$, which has vanishing moments for $M \leq M_{\max}$, is periodically repeated in (d). The resulting Coulomb potential can be represented as the sum of two terms

$$V_{\mathrm{Coulomb}}(\mathbf{r}) = V'_{\mathrm{Coulomb}}(\mathbf{r}) + V_{\mathrm{Coulomb,LM}}(\mathbf{r}), \tag{F.24}$$

where $n'(\mathbf{r})$ and $V'_{\mathrm{Coulomb}}(\mathbf{r})$ can easily be treated in reciprocal space, whereas $V_{\mathrm{Coulomb,LM}}(\mathbf{r})$ is determined by the model density $n_{\mathrm{LM}}(\mathbf{r})$ with correct boundary conditions for an isolated unit, as shown in Fig. F.3(c). Note that this is *not merely a post-processing step after a usual periodic cell calculation*; the potential calculated during the self-consistency iterations is determined from Eq. (F.24) and not from the first term alone.

There is an additional consideration in the case of a defect in a solid. Since the medium is polarizable, the change in density due to a defect is not localized and, in general, the integrals for moments (F.22) do not converge within the cell. This can be overcome by another application of the general idea of adding model densities, since the long-range terms can be found from perturbation theory for the polarization of the given material due to the slowly varying long-range electric fields.[5]

An important example deserves special mention: an atom, molecule, or defect with charge Z [881]. Periodically repeated charged units can be treated by adding a constant neutralizing background density $n_B = -Z/\Omega$, as in the Ewald method of Sec. F.2. The total energy $E(\Omega)$ can then be calculated as in any other periodic system; however, it includes spurious interaction among the units and the background $\propto 1/L$. This leading term can be cancelled by subtracting the energy of point charges Z in the background, i.e. $Z^2\alpha/(2L)$, where α is the Madelung constant (Sec. F.2). However, there is a difference between the interaction of the background with a point charge and with the real density of the unit. This is a local effect $\propto 1/\Omega$ since the background density varies as $\propto 1/\Omega$. Correcting for this term leads to a more convergent formula for the energy valid for cubic cells [881]

$$E(L) = E_\infty - \alpha\frac{Z^2}{2L} - \frac{2\pi ZQ}{3L^3} + O(1/L^5), \tag{F.25}$$

where Q is the isotropic quadrupole moment $Q = \langle n \rangle_2 = \int \mathrm{d}\mathbf{r} r^2 n(r)$. A different approach has been proposed by Kantorovich [888] that applies for cells of arbitrary shape.

[5] This approach also applies to stress and strain due to defects which obey relations analogous to those for electrostatics.

SELECT FURTHER READING

Summary:

Kittel, C., *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1996.

Extensive review:

Coldwell-Horsfall, R. A. and Maradudin, A. A., "Zero-point energy of an electron lattice," *J. Math. Phys.* 1:395, 1960.

Forms involving "smeared ions" and energy relative to neutral atoms:

Galli, G. and Parrinello, M., in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis, Kluwer, Dordrecht, 1991, pp. 283–304.

Sankey, Otto F. and Niklewski, David J., "*Ab initio* multicenter tight-binding model for molecular dynamics simulations and other applications in covalent systems," *Phys. Rev.* B 40:3979–3995, 1989.

Soler, J. M., Artacho, E., Gale, J., Garcia, A., Junquera, J., Ordejon, P. and Sanchez-Portal, D., "The SIESTA method for *ab intio* order-N materials simulations," *J. Phys.: Condens. Matter* 14:2745–2779, 2002.

General method for reduction of effects of artificial periodic boundary conditions:

Schultz, P. A., "Local electrostatic moments and periodic boundary conditions," *Phys. Rev.* B 60:1551–1554, 1999.

Makov, G., and Payne, M. C., "Periodic boundary conditions in *ab initio* calculations," *Phys. Rev.* B 51:4014–4022, 1995.

Exercises

F.1 Show that the potential in (F.4) has zero average value as claimed. As a hint in the reasoning, the final term can be considered as the limit $G \to 0$ of the middle term.

F.2 Discuss the values of the Madelung constant in Tab. F.1. Compare these with the result of the previous problem. Why are the values larger or smaller? Rationalize the variation of α among the structures.

F.3 The problem of a point charge at the center of a sphere with a neutralizing uniform charge density can be solved analytically. Derive the expressions given for the potential, (F.8) and energy, (F.9). Hint: Use the knowledge that the potential due to the uniform distribution must vary as r^2 (Why?) and that the last term in (F.8) has been chosen to make $V = 0$ at the boundary for the neutral cell (Why?). (Related analysis is given for the Wigner interpolation formula for electron correlation energy by Pines [225], p. 92–94.)

F.4 Show that the two expressions for the Ewald energy, (F.5) and (F.6), are equivalent. As a first step in the proof show that the right-hand side of (F.6) is real. Hint: Expand exponentially and use the cosine addition formula $\cos(A - B) = \cos A \cos B + \sin A \sin B$.

F.5 Explain the meanings of the terms in real and reciprocal space in (F.5) in terms of the physical interactions of gaussian charge distributions, and verify the statements made in the interpretation following (F.6).

F.6 Construct a program to perform Ewald sums in (F.4) and (F.5). (A code is available at the URL given in Ch. 24.)

F.7 Use your program for Ewald sums in Exercise F.6 to check the values of the Madelung constant in Tab. F.1.

F.8 For a chosen simple crystal structure calculate the energy versus lattice constant a. Show that it varies as $1/a$. From the slope of energy vesus volume, calculate the pressure. Check that this agrees with the pressure given by the stress theorem, Eq. (F.12).

F.9 Show analytically that in the simple crystal structures in Tab. F.1, the force on each atom vanishes. Verify this numerically using the force theorem.

F.10 Construct a crystal with two atoms per cell, e.g. diatomic molecules with spacing d placed on an fcc lattice with lattice constant a. Calculate the energy for several values of d; from the slope calculate the force on an atom and compare with the force found using the force theorem, (F.10).

F.11 Following the previous problem, calculate the stress using the stress theorem, Eq. (F.12), and compare with the slope of the energy versus lattice constant a. Give the analytic proof that the stress is given by scaling *both* d and a, and also show this numerically by direct calculation.

F.12 Consider a molecule represented by plus and minus charges so that it has a dipole moment. Place the molecules on a simple cubic lattice and evaluate the Ewald energy. Now make the cell long in one direction so that it is orthorhombic with $a = b \ll c$. (Be sure that the program sums over sufficient vectors in both real and reciprocal space for this anisotropic case.) Find the energy for dipoles along the c direction and for dipoles oriented along a. Are they different? Why? What does this have to do with Ch. 22?

F.13 Modify the program to calculate the potential at an arbitrary point. For the case in the problem above with dipoles along the c direction, show that the potential has the dipole offset given by Eq. (F.21). Vary the in-plane lattice constant $a = b$ (but still with $a = b \ll c$) and show the point stated in Sec. F.5 that variation of the fields in the plane decreases exponentially as a function of distance from the plane of dipoles.

Appendix G

Stress from electronic structure

Summary

The subject of this appendix is the macroscopic stress that enters mechanical properties of matter in the form of stress–strain relations. The stress tensor is the generalization of pressure to all the independent components of dilation and shear, and the "stress theorem" is the generalization of the virial theorem for scalar pressure to all components of the stress tensor. In condensed matter, the state of the system is specified by the forces on each atom and the stress, which is an independent variable. The conditions for equilibrium are: (1) that the total force vanishes on each atom, and (2) that the macroscopic stress equals the externally applied stress.

G.1 Macroscopic stress and strain

Stress and strain are important concepts in characterizing the states of condensed matter [177, 721, 722, 890]. A body is in a state of stress if it is acted upon by external forces or if one part of the body exerts forces upon another part. If we consider two types of forces as illustrated in Fig. G.1: those acting interior to a volume element and those that act upon (or through) the surface of the element due to the surrounding material, which are shown as arrows in the figure. The latter forces (per unit area) are the stresses transmitted throughout the interior of the volume. Since these forces balance on any surface in equilibrium, the stress can be determined in terms of only the intrinsic internal forces; i.e. stress is an intrinsic property of a material in a given state. This brings stress into the realm of "electronic structure" as one of the properties of a body determined by the quantum state of the system of electrons and nuclei.

For condensed matter in which the stress is homogeneous, averaged over volumes of macroscopic dimensions, the state of the system is specified by the forces on each atom and the stress, which is an independent variable. The conditions for equilibrium are that the total force vanishes on each atom, whereas the macroscopic stress is fixed by externally applied forces. The equation of state is the relation of stress to the internal variables, such as the density and temperature. For example, in a homogeneous liquid, the state of the system is fully specified by the volume, pressure, and temperature, and the relation to the underlying hamiltonian is given by the virial theorem that relates pressure to the expectation value of

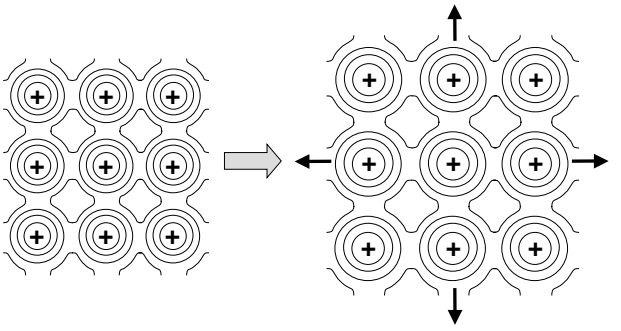


Figure G.1. Illustration of a crystal in equilibrium with no applied forces and with a strain induced by tensile forces (arrows). A uniform strain of all space including the ion cores is shown; this is the essence of the concept of "Streckung des Grundgebietes" ("stretching of the ground state") employed by Fock [259] to derive the virial theorem. Of course this is not what really happens, but it is sufficient for calculation of macroscopic stress from the generalized force theorem (G.4). An alternative approach is shown in Fig. H.2.

the operators for the kinetic energy and the virial of the interaction between particles. This was first proven in quantum mechanics by Born, Heisenberg, and Jordan [257] and later by Finkelstein [258], Hylleraas [44], Fock [259], and Slater [260]. In a crystal, however, there can be shear stress $\sigma_{\alpha\beta}$ in equilibrium, and the equation of state is specified in terms of stress–strain relations. The stress tensor in quantum systems was considered by Schrödinger [891], Pauli [254], Feynman [892], and others (e.g. [893]), and a fundamental relation in terms of the intrinsic hamiltonian has been formulated in the form of the "stress theorem" [104, 129], which is a generalization of the elegant scaling arguments of Fock [259].

Strain is a deformation of a material that causes a displacement of a point $\mathbf{r}_i \to \mathbf{r}'_i$, i.e. a displacement $\mathbf{u} = \mathbf{r}' - \mathbf{r}$. The displacement $\mathbf{u}$ as a function of the coordinate $\mathbf{r}$ specifies the deformation (see Chap. 1 of [721]). Consider two nearby points joined by the vector $\mathrm{d}\mathbf{r}$ which is deformed to $\mathrm{d}\mathbf{r}'$. The distance between the points changes from $\mathrm{d}l = \sqrt{(\mathrm{d}r_1^2 + \mathrm{d}r_2^2 + \mathrm{d}r_3^2)}$ to the corresponding $\mathrm{d}l'$. To lowest order in $\mathbf{u}$, $\mathrm{d}l'$ is given by

$$(\mathrm{d}l')^2 = \mathrm{d}l^2 + 2u_{\alpha,\beta}\mathrm{d}r_\alpha \mathrm{d}r_\beta, \tag{G.1}$$

where summation over repeated cartesian indices α, β is assumed, and where

$$u_{\alpha,\beta} = \frac{1}{2}\left(\frac{\partial u_\alpha}{\partial r_\beta} + \frac{\partial u_\beta}{\partial r_\alpha}\right) \tag{G.2}$$

is the strain tensor. Note that this is equivalent to a *metric tensor* that gives lengths in the deformed system in terms of undeformed coordinates [894]

$$(\mathrm{d}l')^2 = \mathrm{d}r_\alpha g_{\alpha,\beta}\mathrm{d}r_\beta;\ g_{\alpha,\beta} = \delta_{\alpha,\beta} + 2u_{\alpha,\beta}. \tag{G.3}$$

It is also convenient to define the *unsymmetrized strain tensor* $\epsilon_{\alpha\beta}$, which is a scaling of space, $r_\alpha \to (\delta_{\alpha\beta} + \epsilon_{\alpha\beta})r_\beta$. This is often simpler to use, but we must always remember that

it is the symmetric form, Eq. (G.2), that relates to internal energies; antisymmetric terms are rotations that have no effect upon relative internal coordinates.

If the strain is homogeneous over macroscopic regions,[1] then the macroscopic average stress tensor $\sigma_{\alpha\beta}$ is the derivative of the energy with respect to the strain tensor, per unit volume,

$$\sigma_{\alpha\beta} = -\frac{1}{2\Omega}\frac{\partial E_{\text{total}}}{\partial g_{\alpha\beta}} \quad \text{or} \quad \sigma_{\alpha\beta} = -\frac{1}{\Omega}\frac{\partial E_{\text{total}}}{\partial u_{\alpha\beta}}. \tag{G.4}$$

The sign of the stress is chosen as in [721] and [129]: since definition (G.4) applies to the *internal* forces in the system, a negative value indicates that the internal energy decreases for positive (expansive) strain, i.e. it is under compression. For example, under hydrostatic compression, pressure is given by $P = -(1/3)\sum_\alpha \sigma_{\alpha\alpha}$.

Elastic phenomena are described by stress–strain relations, e.g. to linear order the elastic constants are given by

$$C_{\alpha\beta;\gamma\delta} = \frac{1}{\Omega}\frac{\partial^2 E_{\text{total}}}{\partial u_{\alpha\beta}\partial u_{\gamma\delta}} = -\frac{\partial \sigma_{\alpha\beta}}{\partial u_{\gamma\delta}}. \tag{G.5}$$

Symmetry [86, 272, 722] can be used to specify $C_{\alpha\beta;\gamma\delta}$ as a 6×6 array C_{ij} for a general crystal. For a cubic crystal, there are only three independent constants: $C_{11} = C_{xx,xx}$, $C_{12} = C_{xx,yy}$, and $C_{44} = C_{xy,xy}$. (See [722] or the solid state texts [84, 86, 88] for other cases.)

The theory of finite strains can be treated directly from basic theory since the stress is defined by the derivative (G.4), which applies for any state with arbitrary magnitude of strain. In addition, the positions of the atoms in the unit cell are fixed by the zero-force relation (Sec. G.4) at any strain. Thus calculation of stress as a function of strain can be used to find linear and non-linear stress–strain relations [104, 129]. However, care must be taken in defining stress–strain relations because *strain is not unique* since it is defined relative to a reference state.

Using the generalized force theorem, Eq. (3.21), the expression for stress, Eq. (G.4), can be evaluated using various ways of distorting the system. The example of uniform infinitesimal strain of all space (including core states) is shown in Fig. G.1; an alternative, illustrated in Fig. H.2, can be considered if the expression (G.4) is generalized to a non-uniform strain. The derivative in Eq. (G.4) can be evaluated using any of the various expressions that relate total energy E_{total} to fundamental electronic energies. The resulting expressions can appear to be very different and, indeed, even within one approach different contributions to E_{total} may be treated differently. The various types of expressions can be grouped into categories that reveal physical insight and suffice for important applications in electronic structure.

G.2 Stress from two-body pair-wise forces

In electronic structure all fundamental forces are two-body central interactions $V_{kk'} \equiv V(|\mathbf{r}_k - \mathbf{r}_{k'}|)$, where k and k' denote any pair of particles with the relative coordinates

[1] In general, strain $u_{\alpha,\beta}$ or metric $g_{\alpha,\beta}$ is a *tensor field* that is a function of position $\mathbf{r}$. Fields will be considered in App. H.

$\mathbf{r}_{kk'} = \mathbf{r}_k - \mathbf{r}_{k'}$. In any case in which the particles are *explicitly represented* by such terms in the total energy, then the stress is given by the generalized virial (Exercise G.1)

$$\sigma_{\alpha\beta} = -\frac{1}{2\Omega}\sum_{k\neq k'}\frac{\mathrm{d}}{\mathrm{d}\mathbf{r}_k}V_{kk'}\frac{\mathrm{d}\mathbf{r}_k}{\mathrm{d}\epsilon_{\alpha\beta}} = \frac{1}{2\Omega}\sum_{k\neq k'}\mathbf{F}_{kk',\alpha}\mathbf{r}_{k,\beta}, \tag{G.6}$$

which can be written in the manifestly symmetric form

$$\sigma_{\alpha\beta} = \frac{1}{2\Omega}\sum_{k\neq k'}\frac{(\mathbf{r}_{kk'})_\alpha(\mathbf{r}_{kk'})_\beta}{r_{kk'}}\left(\frac{\mathrm{d}}{\mathrm{d}r_{kk'}}V\right). \tag{G.7}$$

Here the sum over k and k' is over all particles considered. Note that $\mathbf{F}_{kk',\alpha}$ is the *contribution to the force* on particle k due to particle k'; it is *not* the total force $\mathbf{F}_{k,\alpha}$ on particle k, which vanishes in equilibrium.

Equation (G.7) provides the stress due to classical particles directly in terms of the potentials and forces; it can also be viewed as a quantum mechanical operator which leads to the most general form of the potential part of the stress in a many-body system, Eq. (3.26). The formulation in (G.7) or (G.6) also provides the needed expressions for any terms in the equation for total energy that depend upon the distance between particles or parameters in the energy. This is the useful form for the real-space terms in the Ewald stress given in (F.12) and for the total energy terms in tight-binding or local orbital approaches that are expressed as a function of distances (see Eq. (14.26) and related terms in Sec. 15.5).

G.3 Expressions in Fourier components

Although it might appear that Eqs. (G.7) and (3.26) are the end of the story for potential interactions, this is not the case. Even in the general many-body expression, Eq. (3.26), the long-range classical Coulomb term should be treated with special care, e.g. using expressions in Fourier space. Mean-field approaches like density functional theory do not represent particle positions directly, and the effective potential is *not* represented in terms of a potential due to specific other particles. Instead, $V_{\mathrm{KS}}(\mathbf{r})$ is defined only by the condition that it reproduces the correct density. How does one proceed? The practical approach is simply to differentiate all the terms in E_{total}.

Expressions in Fourier space can be treated straightforwardly by using the fact that strain also scales in reciprocal space: $\mathbf{q}_\alpha \to (\delta_{\alpha\beta} - \epsilon_{\alpha\beta})\mathbf{q}_\beta$, where $\mathbf{q}$ is any vector in reciprocal space. The derivation is simplified by the fact that structure factors $S^\kappa(\mathbf{G})$, Eq. (12.17), and $\Omega n(\mathbf{G})$ are invariant. For example, the Hartree term (F.2) (which appears in the total energy expressions (3.14), term (9.3), and specific expressions in other chapters), leads to the stress contribution (Exercise G.2)

$$-\frac{1}{\Omega}\frac{\partial E_{\mathrm{Hartree}}}{\partial\epsilon_{\alpha\beta}} = \frac{1}{2}4\pi e^2{\sum_{\mathbf{G}\neq 0}}'\frac{n(\mathbf{G})^2}{G^2}\left[2\frac{\mathbf{G}_\alpha\mathbf{G}_\beta}{G^2} - \delta_{\alpha\beta}\right], \tag{G.8}$$

which is clearly symmetric, as it should be.

Kinetic contributions

Scaling also applies to kinetic terms using $\mathrm{d}/\mathrm{d}\mathbf{r}_\alpha \to (\delta_{\alpha\beta} - \epsilon_{\alpha\beta})(\mathrm{d}/\mathrm{d}\mathbf{r}_\beta)$. This leads directly to a general expression, Eq. (3.26), valid in both many-body and independent-particle formulations. The expressions are particularly simple for wavefunctions expressed in Fourier space: the energy given in Eq. (13.1),

$$T_s = \frac{\hbar^2}{2m_e}\frac{1}{N_k}\sum_{\mathbf{k},i}\sum_m c^*_{i,m}(\mathbf{k})c_{i,m}(\mathbf{k})|\mathbf{k}+\mathbf{G}_m|^2, \tag{G.9}$$

leads to the kinetic contribution to the stress (Exercise G.3)

$$-\frac{1}{\Omega}\frac{\partial T_s}{\partial \epsilon_{\alpha\beta}} = \frac{\hbar^2}{m_e}\frac{1}{N_k}\sum_{\mathbf{k},i}\sum_m c^*_{i,m}(\mathbf{k})c_{i,m}(\mathbf{k})(\mathbf{k}+\mathbf{G}_m)_\alpha(\mathbf{k}+\mathbf{G}_m)_\beta. \tag{G.10}$$

In Ch. 15 use is made of the fact that tight-binding and local orbital forms of the matrix elements of the kinetic energy operator can be cast in terms of functions of distances between atoms [418, 617], so that a two-body form like Eq. (G.7) can be used instead of a generic form like (G.10).

Ewald contribution to stress

Using the above forms, many different expressions for stress can be found that may be more or less convenient in various methods. The application to the Ewald term is given in Sec. F.2. Here we reproduce the expression for the stress corresponding to the plane wave formula, Eq. (13.1), for total energy, as given in Eq. (2) of [104]: The strain derivative is

$$\begin{aligned}\frac{\partial \gamma_{\mathrm{Ewald}}}{\partial \epsilon_{\alpha\beta}} = {} & \frac{\pi}{2\Omega\epsilon}\sum_{G\neq 0}\frac{\mathrm{e}^{-G^2/4\epsilon}}{G^2/4\epsilon}\left|\sum_\tau Z_\tau \mathrm{e}^{i\mathbf{G}\cdot\mathbf{x}_\tau}\right|^2\left[\frac{2G_\alpha G_\beta}{G^2}(G^2/4\epsilon+1)-\delta_{\alpha\beta}\right] \\ & + \frac{1}{2}\epsilon^{1/2}{\sum_{\tau\tau'\mathbf{T}}}' Z_\tau Z_{\tau'} H'(\epsilon^{1/2}D)\frac{D_\alpha D_\beta}{D^2} + \frac{\pi}{2\Omega\epsilon}\left[\sum_\tau Z_\tau\right]^2\delta_{\alpha\beta},\end{aligned} \tag{G.11}$$

where $\mathbf{D} = \mathbf{x}_{\tau'} - \mathbf{x}_\tau + \mathbf{T}$ and the sum is only for terms with $D \neq 0$. Note that here ϵ denotes a convergence parameter (*not* the strain $\epsilon_{\alpha\beta}$) which may be chosen for computational performance. Z_τ denotes the atomic core charge of atom τ, $\mathbf{T}$ the lattice translation vectors, and $\mathbf{x}_\tau$ the atomic positions in the unit cell. The function $H'(x)$ is

$$H'(\mathbf{x}) = \partial[\mathrm{erfc}(\mathbf{x})]/\partial\mathbf{x} - \mathbf{x}^{-1}\mathrm{erfc}(\mathbf{x}), \tag{G.12}$$

with $\mathrm{erfc}(\mathbf{x})$ denoting the complementary error function.

G.4 Internal strain

The expressions for stress in the previous sections have been derived assuming a homogeneous scaling of space, including the electron wavefunctions and positions of the nuclei [104, 129]. However, this is not the whole story for the actual measured stress. The proof

that this is a correct expression for the stress hinges upon the requirement that the energy be minimum with respect to all internal degrees of freedom. In addition to the requirement that the electron wavefunction be at the variational minimum, one must add the requirement that each nucleus I be at the minimum energy position, i.e. that the force on each nucleus vanishes, $\mathbf{F}_I = 0$, in the presence of the strain. Only for simple crystal structures and certain symmetry strains are the positions of the nuclei fixed by symmetry. In general, one must find the positions given by condition $\mathbf{F}_I = 0$, and the displacement at which this occurs is defined to be

$$\mathbf{u}_{s,\alpha} = \sum_{\beta} \epsilon_{\alpha\beta} \tau_{s,\beta} + \mathbf{u}^{\text{int}}_{s,\alpha}, \tag{G.13}$$

where the first term represents uniform scaling of the basis and the second, the deviations or "internal strains" (see, e.g., [90] and [104] and references given there). To linear order the internal strains are proportional to the external strain, defining "internal strain parameter" Γ,

$$\mathbf{u}^{\text{int}}_{s,\gamma} = \sum_{\alpha\beta} \Gamma_{s,\gamma\alpha\beta}\ \epsilon_{\alpha\beta}. \tag{G.14}$$

The effect can be understood in simple examples, such as diamond or zinc-blende structures. In the unstrained crystal, planes of atoms perpendicular to the (1 1 1) direction are spaced alternately $1/4$ and $3/4$ times $\sqrt{3}a/4$; for a uniaxial strain in the (1 1 1) direction, the spacing is not determined by symmetry. The problem is equivalent to the one-dimensional chain of molecules described in Exercise G.4.

Internal strains are crucial for understanding and predicting stress–strain relations. However, internal strain parameters have been measured in only a few cases because of the difficulty of experimental measurements of atomic positions in a strained crystal. Thus this is a crucial area where theory adds information to our knowledge of elasticity even in cases where macroscopic elastic constants are well established.

SELECT FURTHER READING

Basic theory of elasticity:

Landau, L. D. and Lifshitz, E. M., *Theory of Elasticity*, Pergamon Press, Oxford, England, 1958.

General Theory:

Nielsen, O. H. and Martin, R. M., "Quantum-mechanical theory of stress and force," *Phys. Rev.* B 32(6):3780–3791, 1985.

Applications in a plane wave basis:

Nielsen, O. H. and Martin, R. M., "Stresses in semiconductors: *ab initio* calculations on Si, Ge, and GaAs," *Phys. Rev.* B 32(6):3792–3805, 1985.

Expressions in localized bases:

Soler, J. M., Artacho, E., Gale, J., Garcia, A., Junquera, J., Ordejon, P. and Sanchez-Portal, D., "The SIESTA method for *ab intio* order-N materials simulations," *J. Phys. : Condens. Matter* 14:2745–2779, 2002.

Feibelman, P. J., "Calculation of surface stress in a linear combination of atomic orbitals representation," *Phys. Rev.* B 50:1908–1911, 1994.

Exercises

G.1 Show that for particles interacting via two-body central potentials the contribution to the stress tensor is given by the generalized virial expression (G.6). Further, transform the expression to the symmetric form (G.7).

G.2 Derive the expression, (G.8), for the Hartree contribution to the stress tensor.

G.3 Using the argument of the scaling of reciprocal space, show that the kinetic contribution to the stress can be written in the form (G.10), which is convenient for plane wave calculations.

G.4 Find the elastic constant $C = \mathrm{d}^2 E/\mathrm{d}L^2$ and the internal strain parameter Γ defined by Eq. (G.14) for a one-dimensional chain of diatomic molecules. The atoms in a molecule are spaced a distance R_1 and are connected by a spring with constant K_1; spacing between the molecules is R_2 and they are connected by a spring with constant K_2. The cell length is $L = R_1 + R_2$. Show that the system has the expected behavior that the molecules are incompressible for $K_1 \gg K_2$.

G.5 Show that in any crystal with one atom per cell the internal strain is zero by symmetry.

G.6 As an example of the condition in the previous problem, show that for the molecular chain in Exercise G.4, internal strain vanishes for $R_1 = R_2$ and $K_1 = K_2$. For a homonuclear case, this means one atom per cell. Note that the internal strain is still zero for a diatomic ionic crystal with two different atoms so long as $R_1 = R_2$ and $K_1 = K_2$.

G.7 Show that it is *impossible* to have a chain with three inequivalent atoms per cell and still have zero internal strain.

Appendix H

Energy and stress densities

Summary

A *density* is a field defined at each position **r**, for example the particle number density $n(\mathbf{r})$, which is a well-defined, experimentally measurable function. It would be desirable to have expressions for other densities, in particular, energy and stress densities. However, energy and stress densities are not unique on a microscopic quantum scale, even though they are the basis of the theory of elasticity on a macroscopic scale. This appendix brings out three points: (1) certain integrals of energy and stress densities are unique and very useful; (2) there are important contributions to the energy or stress density that are completely unique – these include all terms that arise from the fact that electrons are a many-body system of fermions; (3) all other terms that are non-unique can be shown to involve only the single scalar number density – there are different possible choices for these terms, each involving only derivatives of the density $n(\mathbf{r})$ or the classical Coulomb potential $V^{CC}(\mathbf{r})$ which is directly related to $n(\mathbf{r})$. It follows that all the issues of non-uniqueness are exactly the same as in a one-particle problem.

Only one density is widely used in electronic structure – the particle density $n(\mathbf{r})$. It is the fundamental measurable quantity in quantum mechanics and the fundamental density in density functional theory. Theoretical expressions for $n(\mathbf{r})$ are well defined and lead to unique results. Here we emphasize that other densities have the potential to play a useful role in electronic structure theory. In particular, energy and stress densities have the potential to be very useful in electronic structure, beyond their limited use thus far.

The difficulty in formulating energy and stress densities is their inherent non-uniqueness. The problem is that, unlike the particle density $n(\mathbf{r})$ which is defined by Eq. (3.8), there are no operators in quantum mechanics that uniquely define "energy at a point" or "stress at a point." Of course, there are expressions for the total energy and stress, but this is not sufficient to define an energy or stress density. The value at any point is always subject to "gauge transformations" that leave the total invariant.

Is there any sense in which an energy or stress density can be useful? The answer is yes, for two reasons:

1. Many important quantities can be shown to be invariant to the choice of gauge. For example, total surface energy and surface stress are defined by integrals over the surface region. Because the integral extends from the vacuum to the bulk interior of the system, it can be shown [895, 899, 900] that gauge-dependent terms vanish in the integrals. Similarly, the expressions for force in terms of surface integrals of the stress density in App. I are invariant and can be very useful (Sec. H.3). For such quantities, it may be convenient to choose a particular gauge, even though one must not associate any physical meaning to the gauge-dependent integrand.
2. Specific analysis can identify terms in the energy and stress densities that are well defined. As shown below, with appropriate definitions, *unique densities result from all contributions to the energy or stress that arise from the fact that electrons constitute a many-body system of fermions*

All non-unique terms in the energy or stress densities involve *only derivatives of the total density* $n(\mathbf{r})$ *and the classical Coulomb potential* $V^{\mathrm{CC}}(\mathbf{r})$. It follows that *all issues of non-uniqueness are exactly the same as in a one-particle problem.*

H.1 Energy density

The total energy of a system of electrons and nuclei can be written in the general form, Eq. (3.16), or the Kohn–Sham form, Eq. (7.5),

$$E = \langle \hat{T} \rangle + [\langle \hat{V}_{\mathrm{int}} \rangle - E_{\mathrm{Hartree}}] + E^{\mathrm{CC}} = T_s + E^{\mathrm{CC}} + E_{\mathrm{xc}}, \tag{H.1}$$

where T_s is the independent-particle kinetic energy and the Coulomb terms are grouped to ensure they are well defined in an infinite system. An energy density $e(\mathbf{r})$ (denoted by a lower case, italic Roman letter) or a density per particle $\epsilon(\mathbf{r}) \equiv e(\mathbf{r})/n(\mathbf{r})$ (lower case Greek letter) is a function which when integrated over all space yields the total energy E, e.g.

$$E = \int \mathrm{d}\mathbf{r} e(\mathbf{r}), \tag{H.2}$$

with

$$e(\mathbf{r}) = t_{\mathrm{ip}}(\mathbf{r}) + e^{\mathrm{CC}}(\mathbf{r}) + e_{\mathrm{xc}}(\mathbf{r}). \tag{H.3}$$

If we separate out the ion–ion interaction E_{II}, which has no effect on the equations for the electrons except to ensure neutrality, the total energy can be written

$$E = \int \mathrm{d}\mathbf{r} n(\mathbf{r}) \epsilon(\mathbf{r}) + E_{II}, \tag{H.4}$$

with[1]

$$\epsilon(\mathbf{r}) = \tau_{\mathrm{ip}}(\mathbf{r}) + V_{\mathrm{ext}}(\mathbf{r}) + \frac{1}{2} V_{\mathrm{Hartree}}(\mathbf{r}) + \epsilon_{\mathrm{xc}}(\mathbf{r}). \tag{H.5}$$

[1] The factor of $1/2$ in the Hartree term might be thought of as an *ad hoc* assignment of $1/2$ of the energy to each particle; however, it follows from the much deeper fact that electrons are identical. Any other assignment of the energy would violate this symmetry.

Classical Coulomb energy density

The first problem in defining an energy density is the classical Coulomb term. There are two forms for the energy density in electrostatics [448, 790]

$$E^{\mathrm{CC}} = \frac{1}{8\pi}\int \mathrm{d}\mathbf{r}|\mathbf{E}^{\mathrm{CC}}(\mathbf{r})|^2 = \frac{1}{2}\int \mathrm{d}\mathbf{r}V^{\mathrm{CC}}(\mathbf{r})[n(\mathbf{r}) + n^+(\mathbf{r})], \tag{H.6}$$

where $\mathbf{E}^{\mathrm{CC}} = -\nabla V^{\mathrm{CC}}$ is the electric field due to the total charge density of the electrons and nuclei $n(\mathbf{r}) + n^+(\mathbf{r})$. Each of the integrands can be viewed as an energy density $e^{\mathrm{CC}}(\mathbf{r})$ and each has advantages in different situations. The first expression is the Maxwell energy density assigned to the field instead of the particles. The second expression has the form of the interaction of particles with the energy assigned to the position of the particles. Even though this part of the energy density is not unique, it is purely classical and all forms can be expressed in terms of the charge density. (Note the close analogy with the "boson" part of the kinetic energy, Eq. (H.14), below.)

There is an important practical distinction between the two forms in Eq. (H.6). Only in the second case can the energy be written in the form of (H.5), with $V_{\mathrm{ext}}(\mathbf{r})$ the Coulomb the potential due to the nuclei $n^+(\mathbf{r})$ and $V_{\mathrm{Hartree}}(\mathbf{r})$ the classical Coulomb potential due to the electrons $n(\mathbf{r})$.

Exchange–correlation energy density

Chapter 7 discusses the physical reasoning for expressions for $\epsilon_{\mathrm{xc}}(\mathbf{r})$ as a functional of the exchange–correlation hole around an electron at point $\mathbf{r}$. Even though it is not defined by the fact that its integrals must yield the total E_{xc}, $\epsilon_{\mathrm{xc}}(\mathbf{r})$ *is uniquely specified by the definition that it is the additional energy per electron at point* $\mathbf{r}$ *due to exchange and correlation*. This follows from expression (7.17) as a coupling constant integration, and it can be understood by an independent derivation [901]: the potential part of $\epsilon_{\mathrm{xc}}(\mathbf{r})$ is obviously unique because it is given in terms of the pair correlation functions, which are measurable functions. The kinetic energy contribution to $\epsilon_{\mathrm{xc}}(\mathbf{r})$ is only the change in kinetic energy due to correlation; this density $\tau_c(\mathbf{r})$ is also unique [901–903] by extension of the arguments given below for $\tau_x(\mathbf{r})$.

Kinetic energy density for independent-particles

Finally, we consider the first term in the Kohn–Sham energy, the kinetic energy of independent particles T_s. This is treated in some detail because the analysis leads to expressions that are useful in construction of functionals and in analysis of actual electronic structure calculations.

In analogy to the Coulomb energy in Eq. (H.6), the kinetic energy of N independent fermions can be expressed in different forms

$$T_s = -\frac{1}{2}\sum_{i=1}^{N}\int \mathrm{d}\mathbf{r}\psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r}) = \frac{1}{2}\sum_{i=1}^{N}\int \mathrm{d}\mathbf{r}|\nabla\psi_i(\mathbf{r})|^2. \tag{H.7}$$

The equivalence of the two forms follows from integration by parts, where boundary terms vanish for bound states since ψ_i vanish at the boundary or for period functions where boundary terms cancel. Thus either integrand,

$$t^{(1)}(\mathbf{r}) = -\frac{1}{2}\sum_{i=1}^{N} \psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r}) \quad \text{or} \quad t^{(2)}(\mathbf{r}) = \frac{1}{2}\sum_{i=1}^{N} |\nabla\psi_i(\mathbf{r})|^2, \tag{H.8}$$

can be regarded as a "kinetic energy density" $t(\mathbf{r})$ since the integral of either density is the total kinetic energy.

How is possible to find any part of the kinetic energy density that is unique and useful? First divide the problem into parts: the kinetic energy density of independent bosons with density $n(\mathbf{r})$ plus the excess "exchange kinetic energy density" of the fermions can be written

$$t(\mathbf{r}) = t_n(\mathbf{r}) + t_x(\mathbf{r}). \tag{H.9}$$

This can be accomplished[2] by expressing the wavefunctions as

$$\psi_i(\mathbf{r}) = s(\mathbf{r})\phi_i(\mathbf{r}); \quad s(\mathbf{r}) = n(\mathbf{r})^{1/2}. \tag{H.10}$$

Thus $\sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2 = 1$ at each point $\mathbf{r}$, from which it immediately follows that (Exercise H.1)

$$\sum_{i=1}^{N} \boldsymbol{\nabla}|\phi_i(\mathbf{r})|^2 = 0; \quad \sum_{i=1}^{N} \nabla^2|\phi_i(\mathbf{r})|^2 = 0 \tag{H.11}$$

at each point $\mathbf{r}$. From the first equation in (H.11), it follows that cross terms involving $\boldsymbol{\nabla}s(\mathbf{r})$ and $\boldsymbol{\nabla}\phi_i(\mathbf{r})$ vanish in any expression for the kinetic energy density. Using the second equality, it is straightforward to show that $\sum_{i=1}^{N} |\nabla\phi_i(\mathbf{r})|^2 = -\sum_{i=1}^{N} \phi_i(\mathbf{r})\nabla^2\phi_i(\mathbf{r})$, so that

$$t_x(\mathbf{r}) = n(\mathbf{r})\tau_x(\mathbf{r}), \tag{H.12}$$

with

$$\tau_x(\mathbf{r}) = \frac{1}{2}\sum_{i=1}^{N} |\nabla\phi_i(\mathbf{r})|^2 = -\frac{1}{2}\sum_{i=1}^{N} \phi_i(\mathbf{r})\nabla^2\phi_i(\mathbf{r}), \tag{H.13}$$

which is manifestly invariant to the choice of form of the kinetic energy, Eq. (H.8). The "exchange kinetic energy per particle" $\tau_x(\mathbf{r})$ also has a clear physical meaning; it is the curvature of the exchange hole [904–906], which can be shown to be the relative kinetic energy of pairs of electrons [907].[3] The curvature is clear from plots of the exchange hole in the Ne atom in Fig. 7.2, as well as Figs. 5.5 and 7.4, which also include correlation. Therefore, the excess exchange kinetic energy density $t_x(\mathbf{r})$ (and the density per particle $\tau_x(\mathbf{r}) = t_x(\mathbf{r})/n(\mathbf{r})$) is a unique, meaningful density.

[2] The approach taken here was pointed out to the author by E. Stechel.

[3] The excess fermion kinetic energy density itself is the appropriate physically meaningful density for some properties. For example, exchange should depend upon only the fermion part. In fact, exchange functionals [908] have been constructed in terms of τ_x, based upon the fact that the short-range shape of the exchange hole is determined by $t_x(\mathbf{r})$.

The remaining term involves only derivatives of $s(\mathbf{r}) = n(\mathbf{r})^{1/2}$. Its contribution to the total kinetic energy in Eq. (H.7) is the same as that of N non-interacting bosons with density $n(\mathbf{r})$, i.e. with each boson having wavefunction $s(\mathbf{r})$, which can be written

$$T_n = \frac{1}{2}\int \mathrm{d}\mathbf{r}|\nabla s(\mathbf{r})|^2 = -\frac{1}{2}\int \mathrm{d}\mathbf{r} s(\mathbf{r})\nabla^2 s(\mathbf{r}). \tag{H.14}$$

Clearly, either $\frac{1}{2}|\nabla s(\mathbf{r})|^2$ or $-\frac{1}{2}s(\mathbf{r})\nabla^2 s(\mathbf{r})$ are acceptable choices for the density t_n. (The latter is the same as the Weizsacker [319] term in Sec. 6.1.) Thus the issue of non-uniqueness of the kinetic energy density has been reduced to the simplest form involving only the density. Since the density is a scalar function of one coordinate, the non-uniqueness issues are the same as for a one-particle problem.

Like the classical Coulomb terms, only one form of the kinetic energy density has the form of an energy per particle, the second expression in (H.14), for which the density can be written

$$t_n(\mathbf{r}) = n(\mathbf{r})\tau_n(\mathbf{r}), \tag{H.15}$$

with

$$\tau_n(\mathbf{r}) = \frac{1}{2}\left[\frac{\nabla s(\mathbf{r})}{s(\mathbf{r})}\right]^2 = \frac{1}{8}\left[\frac{\nabla n(\mathbf{r})}{n(\mathbf{r})}\right]^2. \tag{H.16}$$

Energy density per particle: convenient expressions

The expressions for the energy density per electron at each point $\mathbf{r}$ have the advantage that they are closely related to the Kohn–Sham equation which is cast in terms of the density of electrons and is derived from variational equations, Eq. (7.8). Combining expression (H.5) with (H.16) leads to (Exercise H.3)

$$\epsilon(\mathbf{r}) = \sum_i \epsilon_i |\psi_i(\mathbf{r})|^2 - \frac{1}{2}V_{\mathrm{Hartree}}(\mathbf{r}) + [\epsilon_{\mathrm{xc}}(\mathbf{r}) - V_{\mathrm{xc}}(\mathbf{r})], \tag{H.17}$$

where the first term is an eigenvalue weighted density and the other terms correct for overcounting.[4] The first term in (H.17) is essentially a projected density of states that can be used to identify local energies and bonding [909].

H.2 Stress density

In an inhomogeneous system, the stress field is not uniform (even for uniform strain). Is it possible to define a unique stress density field $\sigma_{\alpha\beta}(\mathbf{r})$? Forces are well-defined measurable

[4] This expression is the same as that given by Cohen and Burke [896], $\frac{1}{2}V_{\mathrm{ext}}(\mathbf{r})$, except that they also subtracted, i.e. they assigned 1/2 the interaction energy to the external potential (the nuclei). Unlike electron–electron interaction, where the factor of 1/2 follows from particle symmetry, this assignment is arbitrary. The choice made in (H.5) and (H.17) is consistent with the definition of energy in the Kohn–Sham equations.

quantities; however, the stress density related to the force density $f(\mathbf{r})$ acting on particles at point $\mathbf{r}$ is

$$\boldsymbol{\nabla}_\beta \sigma_{\alpha\beta}(\mathbf{r}) = f_\alpha(\mathbf{r}). \tag{H.18}$$

For dimension $d > 1$, this relation does not uniquely determine the stress density [129, 721, 910, 911], since the curl of any vector field can be added to $\sigma_{\alpha\beta}(\mathbf{r})$ with no change in the forces. The stress field can also be defined as the generalization of Eq. (G.4) to an inhomogeneous metric field $g_{\alpha\beta}(\mathbf{r})$,

$$\sigma_{\alpha\beta}(\mathbf{r}) = -\frac{1}{2\Omega}\frac{\partial E_{\text{total}}}{\partial g_{\alpha\beta}(\mathbf{r})}, \tag{H.19}$$

but this still leads to a non-unique expression [898] of the same form as given earlier by Godfrey [910].

An illuminating point that has been clarified in recent work [901] is that, just as in the energy density, *all non-unique terms can be written as simple expressions in terms of derivatives of the charge density* $n(\mathbf{r})$ *and electrostatic potential* $V^{\text{CC}}(\mathbf{r})$. For the case of the Kohn–Sham independent-particle theory within the local density approximation for exchange and correlation ϵ_{xc}, the expressions given by Nielsen and Martin (NM) [129] (see also [911]), Godfrey [910], and Rogers and Rappe [898] can all be written as

$$\begin{aligned}\sigma_{\alpha\beta}(\mathbf{r}) = &-\frac{\hbar^2}{m_e}\left[n\sum_i \boldsymbol{\nabla}_\alpha\phi_i\boldsymbol{\nabla}_\beta\phi_i\right]_{\mathbf{r}} \\ &-\frac{\hbar^2}{4m_e}\left[\frac{\boldsymbol{\nabla}_\alpha n\boldsymbol{\nabla}_\beta n}{n} + \delta_{\alpha\beta}\left[C-1\right]\nabla^2 n - C\nabla_\alpha\nabla_\beta n\right]_{\mathbf{r}} \\ &+\frac{1}{4\pi}\left[\mathbf{E}_\alpha\mathbf{E}_\beta - \frac{1}{2}\delta_{\alpha\beta}\mathbf{E}_\gamma\mathbf{E}_\gamma\right]_{\mathbf{r}} + \delta_{\alpha\beta}n(\mathbf{r})\left[\epsilon_{\text{xc}}^{\text{LDA}}(n) - V_{\text{xc}}^{\text{LDA}}(n)\right]_{\mathbf{r}},\end{aligned} \tag{H.20}$$

where all non-uniqueness in the kinetic terms is subsumed into the parameter C ($= 4\beta$ in the notation of [898]). The other terms involving ϕ_i are unique for the same reasons as in the energy density.

H.3 Applications

The energy density has been used in various ways. For example, the first term in Eq. (H.17) is essentially a projected density of states that can be used to identify local energies and bonding [909].

Integrals of energy densities: surface energies

Certain integrals over the energy density can be shown to be well defined, independent of any "gauge transformations." For example, the form of the energy density involving the Maxwell density $|\mathbf{E}(\mathbf{r})|^2$ and the analogous kinetic density involving $|\nabla n(\mathbf{r})|^2$ has been

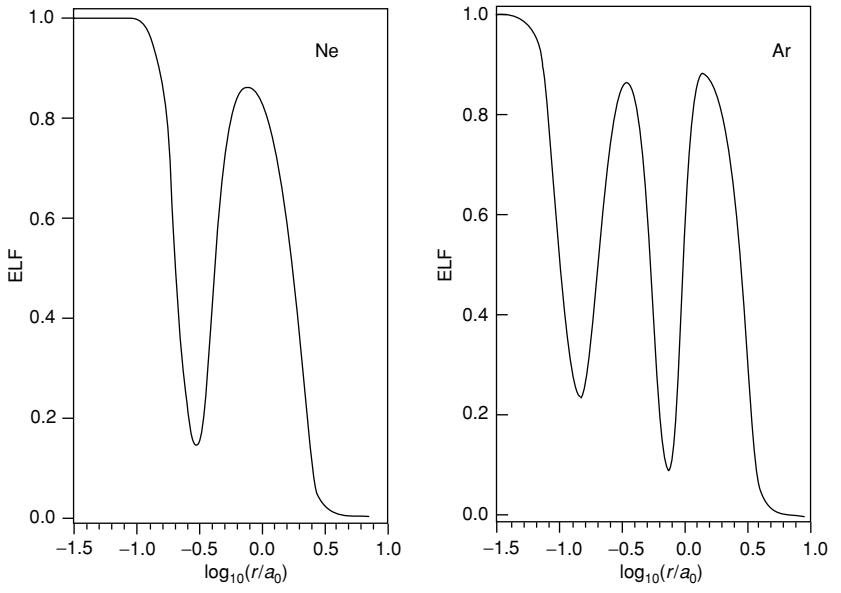


Figure H.1. The "electron localization function" (ELF): Eq. (H.21) versus radius for Ne and Ar [897]. The minima clearly indicate the shell structure separated in space. On the other hand, the density is monotonic and has almost no structure. This indicates that the ELF (or any function of the kinetic energy density) holds the potential for improved functionals. From [897].

used [895, 899, 900] to calculate absolute surface energies of semiconductors that would not be possible by the usual total energy methods.

Electron localization function (ELF)

As emphasized above, the exchange kinetic energy density is well defined and is related to exchange hole curvature. This is the basis for definition of the "electron localization function" (ELF) which is a transformation of $\tau_x(\mathbf{r})$. The form proposed by Becke [897] is defined for each spin σ,

$$\mathrm{ELF}(\mathbf{r}) \equiv [1 + \chi^\sigma(\mathbf{r})]^{-1}, \tag{H.21}$$

where $\chi^\sigma = t_x^\sigma / t_{\mathrm{TF}}^\sigma$, with $t_x^\sigma(\mathbf{r})$ the exchange kinetic energy density given by Eq. (H.13) for spin σ at point $\mathbf{r}$ and $t_{\mathrm{TF}}^\sigma(\mathbf{r})$ the Thomas–Fermi expression, Eq. (6.1), for spin σ in a homogeneous gas at a density equal to $n(\mathbf{r})$, which is a convenient normalization. The definition in (H.21) is chosen so that $0 < \mathrm{ELF} < 1$, with larger values corresponding to larger "localization," i.e. a tendency of an electron of spin σ *not* to have another same spin electron in its vicinity. Among the properties brought out by the ELF function is the shell structure, which is difficult to visualize from the density alone. For example, Fig. H.1 shows the ELF function for Ne and Ar calculated in the Hartree–Fock approximation [897]. The shells are shown by the distinct minima, whereas the density is monotonic and has almost

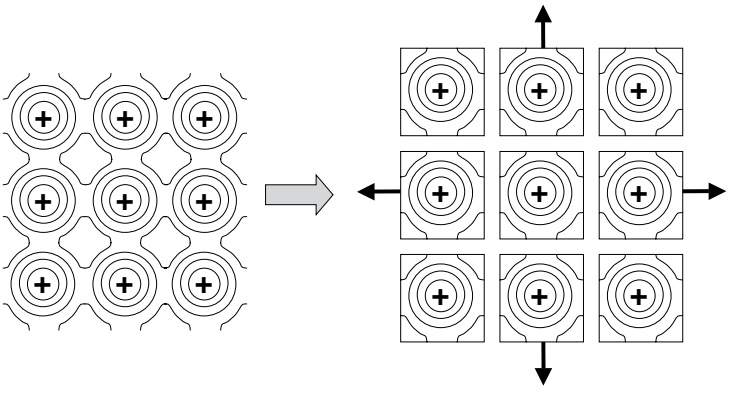


Figure H.2. A different way to expand the crystal from the uniform scaling illustrated in Fig. G.1. The regions around each nucleus are kept rigid and all the variation is the regions between the atoms, i.e. a non-uniform change of coordinates which can be expressed as the original coordinates with a non-uniform metric. Any such variation is a valid form for the generalized force theorem and this approach has great advantages in methods that treat the cores explicitly.

no structure. Also in any one-electron system $\chi = t_x = 0$ so that ELF $= 1$. Similarly, in a many-electron problem, large values of the ELF function indicate a tendency of electrons to be in non-bonded states. This has been used to analyze complex problems, e.g. simulations of water involving proton transfer [912].

Stress density

The stress density provides alternative ways to calculate the macroscopic stress in a solid. The basic idea follows from the definition of stress as a force per unit area [129]. Consider a material that is in equilibrium in the presence of a macroscopic stress, i.e. all internal variables (the electron wavefunction and the positions of the nuclei) are at equilibrium. Then the macroscopic stress is the force per unit area transmitted across any surface that divides the macroscopic solid into two parts. Thus, the stress is the first derivative of the total energy for a displacement of the two half-spaces. A convenient procedure in a crystal is the non-uniform expansion illustrated in Fig. H.2, which shows a crystal with cells pulled apart on the boundaries. The linear change in energy is just the surface integral of the stress field on the boundaries. Since each of the cell boundaries can be considered to divide the crystal into two half-spaces, macroscopic stress is given by the surface integral of the stress density on the cell boundaries. The contributions to the stress calculated in this way are: (1) the Coulomb forces per unit area *transmitted across the boundary* (i.e. the force on one side due to charge on the other side of the boundary), (2) the kinetic stress density at the boundary, which has the same form as a gas of particles that carry momentum across the boundary, and (3) the exchange–correlation terms that are unique but difficult to determine exactly. Finally, the result is unique for any valid form of the stress tensor, since the integral over the surface of a unit cell is invariant to gauge transformations [129].

The calculation of stress from surface integrals closely resembles the calculation of forces by the expressions given in App. I. In particular, the formula for the pressure in the local density and atomic sphere approximations, Eqs. (I.8) or (I.9), is a very useful special case of the more general formulation of the stress field given here. In methods that explicitly deal with core states, this approach has great advantages: the core states remain invariant and only the outer valence states evaluated at the cell boundaries are needed in the calculation of stress.

This idea has been derived in several different ways[5] in the context of the atomic sphere approximation [462–465, 913], where the pressure can be found from an integral over the sphere surface. In a monatomic close-packed solid, the atomic sphere approximation (ASA) is very good and the equations simplify because there are no Coulomb interactions between the neutral spheres, leaving only kinetic and exchange–correlation terms. Two different forms that are convenient for evaluation are given in Sec. I.3. These are very useful in actual calculations, and examples of results for close-packed metals are cited in Sec. 17.7 and App. I.

SELECT FURTHER READING

Classical theory of stress, strain and energy fields:

Landau, L. D. and Lifshitz, E. M., *Theory of Elasticity*, Pergamon Press, Oxford, England, 1958.

Energy density:

Chetty, N. and Martin, R. M., "First-principles energy density and its applications to selected polar surfaces." *Phys. Rev.* B 45: 6074–6088, 1992.

Cohen, M. H., Frydel, D., Burke, K. and Engel, E., "Total energy density as an interperative tool," *J. Chem. Phys.* 113: 2990–2994, 2000.

Electron localization function:

Becke, A. D. and Edgecombe, K. E., "A simple measure of electron localization in atomic and molecular systems," *J. Chem. Phys.* 92: 5397–5403, 1990.

Quantum theory of stress:

Nielsen, O. H. and Martin, R. M., "Quantum-mechanical theory of stress and force," *Phys. Rev.* B 32(6): 3780–3791, 1985.

Rogers, C. and Rappe, A., "Geometric formulation of quantum stress fields," *Phys. Rev.* B 65:224117, 2002.

Exercises

H.1 Show that $\sum_{i=1}^{N} |\nabla\phi_i(\mathbf{r})|^2 = -\sum_{i=1}^{N} \phi_i(\mathbf{r})\nabla^2\phi_i(\mathbf{r})$ follows from the requirement $\sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2 = 1$ at all $(\mathbf{r})$.
Hint: Use Eqs. (H.10) and (H.11).

[5] A good exposition of the relation of the derivations is given by Heine [465].

H.2 Show that the excess fermion kinetic energy density in Eq. (H.13) follows from (H.11). The previous problem may be helpful.

H.3 Show that (H.17) follows from the definitions of the terms in the energy density given before and the Kohn–Sham equation for the eigenvalues.

H.4 Show that the formulas for the stress given by Nielsen and Martin [129] in their Eqs. (33) and (34) can be written in the form of Eq. (H.21), using definition (H.10).

Appendix I

Alternative force expressions

Summary

It is very useful to formulate expressions for forces alternative to the usual force theorem of Sec. 3.3. The basic idea is that since the wavefunction is required to be at a variational minimum, the energy is invariant to *any* change in the wavefunction to linear order. The usual force theorem assumes that the wavefunction remains unchanged when a parameter is changed, but there are an infinite number of other possibilities. An important example involves core electrons; it is much more physical and leads to simpler expressions if the core states are assumed to be rigidly attached to nuclei when a nucleus moves or the crystal is strained. This leads to very useful expressions for forces, stress (pressure), and generalized forces that are energy differences taken to first order for various changes.

The "force theorem" or "Hellmann–Feynman theorem," Eqs. (3.19) or (9.26)

$$\mathbf{F}_I = -\frac{\partial E}{\partial \mathbf{R}_I} = -\int \mathrm{d}^3 r n(\mathbf{r}) \frac{\partial V_{\mathrm{ext}}(\mathbf{r})}{\partial \mathbf{R}_I} - \frac{\partial E_{II}}{\partial \mathbf{R}_I}, \tag{I.1}$$

or the generalized form, Eq. (3.21)

$$\frac{\partial E}{\partial \lambda} = \left\langle \Psi_\lambda \middle| \frac{\partial \hat{H}}{\partial \lambda} \middle| \Psi_\lambda \right\rangle, \tag{I.2}$$

applies to any variation and to non-local potentials as in Eq. (13.3) for pseudopotentials. The same fundamental ideas lead to the "stress theorem," Eq. (3.26), and the practical expression in App. G. These expressions follow from first-order variation of the energy, assuming all the electronic degrees of freedom are at the variational minimum. The expressions correspond to evaluating the force as the derivative of the energy with respect to the parameter λ, keeping the electrons fixed. This is illustrated on the left-hand side of Fig. I.1.

The subject of this appendix is alternative formulas that take advantage of the fact that the electronic degrees of freedom are at a variational minimum. Because the derivative of the energy with respect to any of these variables vanishes, any linear change can be added *with no change in the force*. The resulting degrees of freedom can be used to transform

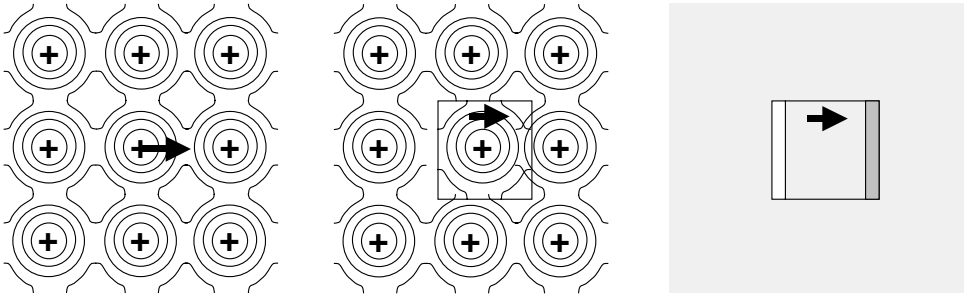


Figure I.1. Illustration of two ways of calculating forces. Left: The usual force theorem, Eq. (I.1), follows if the electron density is held constant to first order as the nucleus moves. Center: A region of charge is moved rigidly with the nucleus. Right: Definition of the region B that is "cut out" and moved rigidly leading to the changed density in region C.

the expressions into different forms that can be more useful in specific cases. An extreme form – that is useful in practice – is shown in the center of Fig. I.1 and explained in Sec. I.1.

There are two general approaches for choosing alternative expressions:

- *Use of the variational principle, involving the effective potential and density* to re-express Eq. (I.1) in forms that involve changes in the density $n(\mathbf{r})$ and/or changes in the total internal potential $V_{\text{eff}}(\mathbf{r})$. The advantage is that the resulting expressions may be easier to evaluate.
- *Geometric relations* that relate the force acting on the nucleus to the force *transferred across a boundary that surrounds the nucleus*. This can be formulated in terms of a stress field which is a force per unit area, establishing relations with the stress density field. Actual expressions can often be shown to be equivalent to specific choices of variations of $n(\mathbf{r})$ and $V_{\text{eff}}(\mathbf{r})$.

I.1 Variational freedom and forces

The usual form of the "force theorem" follows immediately since the first-order change in energy can be considered to arise solely from the term $\int \mathrm{d}\mathbf{r}\, V_{\text{ext}}(\mathbf{r})n(\mathbf{r})$ in Eq. (9.3) with all other changes summing to zero. Alternative formulations for force in density functional theory can be understood using the functionals derived in Sec. 9.2. Different expression can be derived using the most general functional, Eq. (9.13), which is *variational with respect to both the effective potential and the density* for a given external potential [417, 419, 423, 424, 914].

The essential point for our purpose is that one can add *any change in* $V_{\text{eff}}(\mathbf{r})$ *or* $n(\mathbf{r})$ *with no change in the force.* The disadvantage of this approach is that one must calculate the first-order change in the individual terms in Eq. (9.13); the advantage is that difficult problems can be greatly reduced or eliminated. For example, consider the force acting on a nucleus. The usual expression is derived by displacing the nucleus while holding the density

constant, *even the density of core electrons of that nucleus* as illustrated on the left-hand side of Fig. I.1.

An alternative approach is to displace the electron density in a region around the nucleus rigidly with the nucleus; then there are no changes in the large core–nucleus interaction and one arrives at the physically appealing picture that the force is due to the nucleus and core moving together relative to the other atoms. It is important to note that *this is not an approximation*; it is merely a rearrangement of terms. How can this be done? One way that at first appears extremely artificial is to "cut out a region of space" and displace it. This leaves a slice of vacuum on one side and double density on the other, as shown on the right-hand side of Fig. I.1. The effect in the case of moving a nucleus is shown in the center of the figure.

Despite the totally unphysical nature of this change of density, the final consequences are physical and there are advantages in the way the equations can be formulated as first shown by Mackintosh and Andersen [464] and described by Heine [465]. A very simple derivation has been presented by Jacobsen, Norskov, and Puska ([423] App. A) using the properties of the functional, Eq. (9.13), where we can choose any variation of the density and effective potential. Since the densities are frozen, it is straightforward to evaluate all the terms involving the electron density to linear order:

$$\begin{aligned} \delta E^{\mathrm{CC}} &= \delta E^{\mathrm{CC}}_{A\leftrightarrow B} + \int_C n(\mathbf{r}) V^{\mathrm{CC}}_{A+B}(\mathbf{r}), \\ \delta E_{\mathrm{xc}} &= \int_C n(\mathbf{r}) \epsilon_{\mathrm{xc}}[n(\mathbf{r})], \\ \delta T &= \delta \left[\sum_i \varepsilon_i \right] - \int_C n(\mathbf{r}) V_{\mathrm{eff}}(\mathbf{r}), \end{aligned} \tag{I.3}$$

where $\delta E^{\mathrm{CC}}_{A\leftrightarrow B}$ denotes classical Coulomb interactions *between* regions A and B (interactions inside A and B do not change); V^{CC}_{A+B} is the potential due to regions A and B (that due to region C is higher order); δE_{xc} is only considered in the local density approximation; and δT is the change in kinetic energy. Then the total change is (Exercise I.1)

$$\delta E_{\mathrm{total}} = \delta \left[\sum_i \varepsilon_i \right] + \delta E^{\mathrm{CC}}_{A\leftrightarrow B} + \int_C n(\mathbf{r}) \left\{ V^{\mathrm{CC}}_{A+B}(\mathbf{r}) \epsilon_{\mathrm{xc}}[n(\mathbf{r})] - V_{\mathrm{eff}}(\mathbf{r}) \right\}. \tag{I.4}$$

Finally, one has freedom to choose $V_{\mathrm{eff}}(\mathbf{r})$ in region C and a clever choice is to make [423] the last term vanish. This means simply to *define* $V_{\mathrm{eff}}(\mathbf{r})$ to have an added term $\epsilon_{\mathrm{xc}}(n(\mathbf{r}) - V_{\mathrm{xc}}(n(\mathbf{r})$ only in region C. With this definition of the derivative, $\partial V_{\mathrm{eff}}(\mathbf{r})/\partial \mathbf{R}_I$ is a delta function on the boundary on region B, and for this change in V_{eff}, the force is given strictly in terms of the eigenvalues plus the force from electrostatic interactions *that cross the A–B boundary*:

$$-\frac{\partial E_{\mathrm{total}}}{\partial \mathbf{R}_I} = -\frac{\partial \sum_i \varepsilon_i}{\partial \mathbf{R}_I} - \frac{\partial E^{\mathrm{CC}}_{A\leftrightarrow B}}{\partial \mathbf{R}_I}. \tag{I.5}$$

I.2 Energy differences

The expressions for "force" in terms of the eigenvalue sums are actually most useful for calculation of small, but finite, energy differences between cases that involve small changes in the potential. Thus a convenient way of calculating the energy difference due to a small change (adding an external field, change of volume or shape, displacement of an atom, etc.) using standard programs is to calculate the finite energy difference

$$\Delta E_{\text{total}} = \Delta \sum_i \varepsilon_i + \Delta E^{\text{CC}}_{A\leftrightarrow B} \tag{I.6}$$

for potentials as defined above. Perhaps the simplest example in solid state physics – which is also very useful – is the difference in energy between fcc, hcp, and bcc structure metals. In each case the potential is well approximated as neutral and spherical, i.e. the ASA of Sec. 16.6. Then the Coulomb terms vanish and the energy difference is just

$$\Delta E_{\text{total}} \to \Delta \sum_i \varepsilon_i\,, \tag{I.7}$$

where *it is essential not to use the self-consistent potential for each structure: instead the eigenvalues ε_i for each structure are calculated using the same potential!* To linear order of accuracy, the difference, Eq. (I.7), can be evaluated using the potential from any one of the structures. Since the differences are small, this procedure is very useful, taking advantage of the variational freedom to make the calculation more accurate and at the same time easier!

I.3 Pressure

The same ideas apply for any derivative, e.g. stress and pressure, as described in general in Sec. H.2. Figure H.2 illustrates the choice of "pulling apart" rigid units in a crystal, leaving space in between. From appropriate derivatives of the total energy one can calculate the stress on the boundaries. Furthermore, the stress on a boundary that cuts all space into two parts (e.g. a boundary drawn through the spaces in Fig. H.2) can be used to define the *macroscopic stress*: because all forces between the two half-spaces must cross the boundary, the average stress on the boundary is rigorously the macroscopic stress with no non-unique gauge terms [129].

An extremely useful application of the alternative form of the force theorem is the calculation of pressure in an isotropic situation. In a crystal, this means the ASA (Sec. 16.5, especially Fig. 16.10). Also the idea is useful for liquids and matter at high pressure and temperature where the average environment is spherical [462–465]. This is the fortunate situation in which the electrostatic terms vanish because there are no Coulomb fields outside a neutral sphere. The pressure is given simply by the change in sum of eigenvalues for the change in effective potential defined following Eq. (I.4). The resulting expression can be written in terms of the wavefunctions as [462, 465] (see Exercise I.2)

$$3P\Omega = \int \mathrm{d}\mathbf{S}\cdot\left\{\sum_i \left[\boldsymbol{\nabla}\psi_i^*(\mathbf{r}\cdot\boldsymbol{\nabla}\psi_i) - \psi_i^*\boldsymbol{\nabla}(\mathbf{r}\cdot\boldsymbol{\nabla}\psi_i) + \text{c.c.}\right] + \frac{1}{3}n\epsilon_{\text{xc}}\mathbf{r}\right\}. \tag{I.8}$$

Using the fact that ψ_i is a solution of the Kohn–Sham equations in spherical geometry, it was shown by Pettifor [465, 913] that the expression can be rewritten as

$$4\pi S^2 P = \sum_l \int \mathrm{d}E n_l(E)\psi_l^2(S, E) \left\{[E - V_{\mathrm{xc}}(S)]S^2 + (D_l - l)(D_l + l + 1) + \frac{1}{3}\epsilon_{\mathrm{xc}}(S)S^2\right\}. \quad \text{(I.9)}$$

These expressions are particularly convenient for calculation of the equation of state of materials in the ASA approximation because they give the experimentally measurable pressure P directly instead of the total energy (which is very large since it includes all the core electrons). The equilibrium volume Ω is for $P(\Omega) = 0$; the bulk modulus is the slope $B = -\mathrm{d}P/\mathrm{d}\Omega$; the cohesive energy as a function of volume can be found as the integral $\Delta E_{\mathrm{total}} = \int P\mathrm{d}\Omega$; and, finally, the absolute total energy is the cohesive energy plus the total energy of the atom that can be found separately.

Expressions for the pressure in the ASA can also be derived from the stress density (Sec. H.2) as was shown by Nielsen and Martin [129] and as can be seen in Eq. (H.21). There is no ambiguity of the stress field in this case because it is a one-dimensional (radial) problem, and the pressure is the radial stress, i.e. the force per unit area. Also, Eq. (H.21) simplifies because there are no Coulomb terms at the sphere boundary so that the final expressions involve only kinetic and exchange–correlation terms, finally leading to expressions equivalent to those above [129].

I.4 Force and stress

An alternative expression [129, 915] for the total force *on a volume* is given by the well-known relation that a force field is a divergence of a stress field [721]

$$f_\alpha(\mathbf{r}) = \sum_\beta \nabla_\beta \sigma_{\alpha,\beta}(\mathbf{r}). \quad \text{(I.10)}$$

By integrating over a volume containing a nucleus, e.g. region B in Fig. I.1, and using Gauss' theorem, the total force on the region is given in terms of a surface integral of the stress field

$$F_\alpha^{\mathrm{total}} = \sum_\beta \int_S \mathrm{d}S \hat{S}_\beta \sigma_{\alpha,\beta}(\mathbf{r}), \quad \text{(I.11)}$$

where S is the surface of the volume and $\hat{S}$ is the outward normal unit vector. Although there are non-unique terms in the stress field (Sec. H.2), the force is well defined and gauge invariant because such terms vanish in the divergence or in the integral.

Grafenstein and Ziesche [915] have shown that Eq. (I.11) leads to the form of the generalized force expression given in Eq. (I.5) for the particular case of the local density approximation. This provides an additional way of understanding the meaning of the terms, since it does not depend upon the seemingly arbitrary tricks used in the derivation of (I.5).

In addition, the relation to the stress field provides a simple interpretation valid for both independent-particle and many-body problems. First, since the system is assumed to be in equilibrium, the force on the region is the force of constraint, i.e. the external force that is needed to hold the nucleus fixed. This is the same as in the application of the usual force theorem, but here it is essential to add that there are no other constraints on the system in the volume considered. The expression for the stress (e.g., see App. H) is a sum of potential and kinetic terms. The kinetic term is due to particles *crossing the surface* and is present in classical systems at finite temperature and quantum systems at all temperatures. The potential terms are due to interactions *crossing the surface*; interactions within the region, e.g. a nucleus with its own core electrons are not counted. For electrostatic interactions, this is simply the force on the multi-poles inside the region due to fields from outside, which can be conveniently written as volume integrals over the sphere. Finally, the exchange–correlation contribution is the effect of the exchange–correlation hole extending across the surface; for the LDA this is a delta function.

I.5 Force in APW-type methods

An approach to calculation of forces in APW and LAPW methods has been developed by Soler and Williams [673] and by Yu, Singh, and Krakauer [674]. The general idea is quite close to the spirit of the alternative force approaches described above, but the implementation is very different. These authors work directly with the APW or LAPW expressions for the total energy, and calculate a force from the derivative of the total energy with respect to the displacement of an atom *relative to the rest of the lattice*, or equivalently *the displacement of the rest of the lattice relative to the given atom*. There are many choices for the change in the wavefunction with displacement of the nucleus, and the latter interpretation suggests a most convenient one. The sphere and all it contents (nucleus, core electrons, . . .) are held fixed, and the energy changes only because of the change in boundary conditions on the sphere and Coulomb potentials that propagate into the sphere. The change in energy to first order can be found straightforwardly [673] by differentiating each of the terms in the expression for the APW or LAPW total energy with respect to the position of the sphere, evaluated for the unchanged wavefunction. This avoids any need to evaluate the derivative of the large Coulomb energy of interaction of the nucleus with the change density in its sphere; the effect is replaced by forces on the sphere due to its displacement relative to surrounding spheres.

SELECT FURTHER READING

An intuitive discussion of the meaning of the terms in alternative expressions:

Heine, V., in *Solid State Physics*, edited by Ehenreich, H. Seitz, F. and Turnbull, D. Academic Press, New York, 1980, Vol. 35, p. 1.

A simple, compact derivation of the basic ideas using the the variational properties of functions:

Jacobsen, K. W. Norskov, J. K. and Puska, M. J., "Interatomic interactions in the effective-medium theory," *Phys. Rev.* B 35:7423–7442, 1987, App. A.

Exercises

I.1 Show that the expression, (I.4), for an energy difference to first order follows from the form of the energy functional given in Eq. (9.13). Use this result with the special choice for the change in potential to derive the final result, Eq. (I.5).

I.2 Using the fact that ψ_i is a solution of the Kohn–Sham equations in a spherical geometry, show that the potential can be eliminated and the expression for pressure can be written in terms of the wavefunction and its derivatives as in Eq. (I.8). Also show that there is an added term for exchange and correlation that can be written in the form in Eq. (I.8) in the local approximation. Hint: The first part can be done by partial integration and the second is the correction due to the fact that the potential is not fixed as the spherical system is scaled.

Appendix J

Scattering and phase shifts

Summary

Scattering and phase shifts play a central role in many fields of physics and are especially relevant for electronic structure in the properties of pseudopotentials (Ch. 11) and the formulation of augmented and multiple-scattering KKR methods (Chs. 16 and 17). The purpose of this appendix is to collect the formulas together and to make added connections to scattering cross sections and electrical resistivity.

J.1 Scattering and phase shifts for spherical potentials

Scattering plays an essential role in interesting physical properties of electronic systems and in basic electronic structure theory. Scattering due to defects leads to such basic phenomena as resistivity in metals and is the basis for pseudopotential theory (Ch. 11) and all the methods that involve augmentation (Ch. 16). The basic element is the scattering from a single center, which we will consider here only in the spherical approximation, although the formulation can be extended to general symmetries (see [641]). A schematic figure of the scattering of plane waves is shown in Fig. J.1.

Consider the problem of scattering from a potential that is localized. This applies to a neutral atom (and charged ions with appropriate changes) and to the problem of a single muffin-tin potential, where the potential is explicitly set to a constant outside the muffin-tin sphere of radius S. Since the problem is inherently spherical, scattering of plane waves is described by first transforming to spherical functions using the well-known identity [11, 266, 448]

$$e^{i\mathbf{q}\cdot\mathbf{r}} = 4\pi \sum_L i^l \; j_l(qr) \; Y_L^*(\hat{\mathbf{q}}) \; Y_L(\hat{\mathbf{r}}), \tag{J.1}$$

where $j_l(qr)$ are spherical Bessel functions (Sec. K.1) and $Y_L(\hat{r}) \equiv Y_{l,m}(\theta, \phi)$ denotes a spherical harmonic with $\{l, m\} \equiv L$ (Sec. K.2). Since there is no dependence upon the angle around the axis defined by $\hat{r}$, this can also be written as a function of r and θ,

$$e^{i\mathbf{q}\cdot\mathbf{r}} = e^{iqr\cos(\theta)} = \sum_l (2l+1) \; i^l \; j_l(qr) \; P_l[\cos(\theta)], \tag{J.2}$$

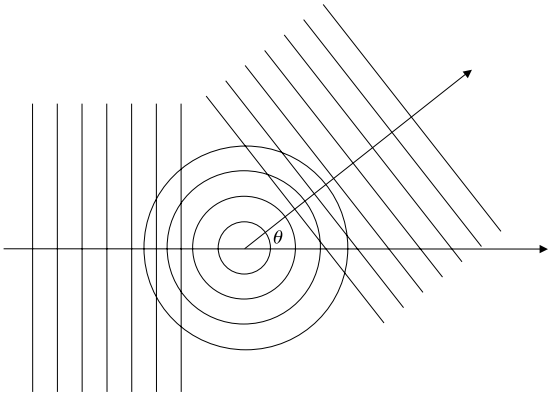


Figure J.1. Schematic illustration of scattering of a plane wave by a spherical potential.

where $P_l(x)$ are the Legendre polynomials (Sec. K.2). Using spherical symmetry, the scattering can be classified in terms of wavefunctions of angular momentum $L \equiv \{l, m\}$, i.e.

$$\psi_L(\mathbf{r}) = i^l \psi_l(r) Y_L(\theta, \phi) = i^l r^{-1} \phi_l(r) Y_L(\theta, \phi), \tag{J.3}$$

as in Eq. (10.1). Inside the region, where the potential is non-zero, radial function $\psi_l(r)$ or $\phi_l(r)$ can be found by numerical integration of the radial Schrödinger equation, (10.12). Outside the region at large r the solution must be a linear combination of regular and irregular solutions, i.e. spherical Bessel and Neumann functions $j_l(\kappa r)$ and $n_l(\kappa r)$,

$$\psi_l^>(\varepsilon, r) \;=\; C_l \;[j_l(\kappa r) \;-\; \tan\eta_l(\varepsilon) \;\; n_l(\kappa r)], \tag{J.4}$$

where $\kappa^2 = \varepsilon$. The energy-dependent phase shifts $\eta_l(\varepsilon)$ are determined by the condition that $\psi_l^>(\varepsilon, S)$ must match the inner solution $\psi_l(\varepsilon, S)$ in value and slope at the chosen radius S. In terms of the dimensionless logarithmic derivative of the inner solution (see Eq. (11.20))

$$D_l(\varepsilon, r) \equiv r\psi_l'(r)/\psi_l(r) = r\frac{\mathrm{d}}{\mathrm{d}r}\ln\psi_l(r), \tag{J.5}$$

this leads to the result

$$\tan\eta_l(\varepsilon) = \frac{S\dfrac{\mathrm{d}}{\mathrm{d}r} j_l(\kappa r)|_S - D_l(\varepsilon)\; j_l(\kappa S)}{S\dfrac{\mathrm{d}}{\mathrm{d}r} n_l(\kappa r)|_S - D_l(\varepsilon)\; n_l(\kappa S)}. \tag{J.6}$$

The scattering cross-section for a single site at positive energies can be expressed in terms of the phase shift. Using asymptotic forms of the Bessel and Neumann functions at positive energies $\varepsilon = \frac{1}{2}k^2$, the wave function, Eq. (J.4), at large radius approaches [96, 266, 704]

$$\psi_l^>(\varepsilon, r) \;\to\; \frac{C_l}{kr}\sin\left[kr + \eta_l(\varepsilon) - \frac{l\pi}{2}\right], \tag{J.7}$$

which shows that each η_l is a phase shift for a partial wave. The full scattered function can

be written

$$\psi_l^{>}(\varepsilon, r) \;\rightarrow\; \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{r}} + i\frac{\mathrm{e}^{\mathrm{i}qr}}{qr}\sum_l (2l+1)\mathrm{e}^{\mathrm{i}\eta_l}\sin(\eta_l)P_l[\cos(\theta)], \tag{J.8}$$

and the scattering cross-section is then given by the scattered flux per unit solid angle (see, e.g., [96, 266, 704])

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{q^2}\left|\sum_l (2l+1)\mathrm{e}^{\mathrm{i}\eta_l}\sin(\eta_l)P_l[\cos(\theta)]\right|^2, \tag{J.9}$$

and the total cross-section by

$$\sigma_{\text{total}} = 2\pi\int \sin(\theta)\mathrm{d}\theta\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{4\pi}{q^2}\sum_l (2l+1)\sin^2(\eta_l). \tag{J.10}$$

For negative energy, κ is imaginary and the Neumann function should be replaced by the Hankel function (Sec. K.1) $h_l^{(1)} = j_l + in_l$, which has the asymptotic form $i^{-l}\mathrm{e}^{-|\kappa|r}/|\kappa|r$. The condition for a bound state is that $tan(\eta_l(\varepsilon)) \rightarrow \infty$, so that the coefficient of the Bessel function vanishes in Eq. (J.4) and the Hankel solution is the solution in all space outside the sphere. The bound state wavefunctions are thus real if one adopts a convention of inclusion of a factor i^l in the wavefunction as in Eq. (16.37), for example.

SELECT FURTHER READING

Basic formulas for phase shifts and scattering:

Shankar, R., *Principles of Quantum Mechanics*, Plenum Publishing, New York, 1980.
Thijssen, J. M., *Computational Physics*, Cambridge University Press, Cambridge, England, 2000.

References on augmented and multiple scattering methods:

Kübler, J., *Theory of Itinerant Electron Magnetism*, Oxford University Press, Oxford, 2001.
Kübler, J. and Eyert, V., in *Electronic and Magnetic Properties of Metals and Ceramics*, edited by K. H. J. Buschow, VCH-Verlag, Weinheim, Germany, 1992, p. 1.
Lloyd, P. and Smith, P. V., "Multiple scattering theory in condensed materials," *Adv. Phys.* 21:29, 1972.

Appendix K

Useful relations and formulas

K.1 Bessel, Neumann, and Hankel functions

Spherical Bessel, Neumann, and Hankel functions are radial solutions of the Helmholtz equation in three dimensions. Spherical Bessel and Neumann functions are related to the half-order functions and can be represented as

$$j_m(x) = \sqrt{\frac{\pi}{2x}} J_{m+\frac{1}{2}}(x) = (-1)^m x^m \left(\frac{\mathrm{d}}{x\mathrm{d}x}\right)^m \frac{\sin(x)}{x}, \tag{K.1}$$

and

$$n_m(x) = \sqrt{\frac{\pi}{2x}} N_{m+\frac{1}{2}}(x) = -(-1)^m x^m \left(\frac{\mathrm{d}}{x\mathrm{d}x}\right)^m \frac{\cos(x)}{x}. \tag{K.2}$$

Examples are

$$\begin{aligned}
j_0(x) &= \frac{\sin(x)}{x}, & n_0(x) &= -\frac{\cos(x)}{x}, \\
j_1(x) &= \frac{\sin(x)}{x^2} - \frac{\cos(x)}{x}, & n_1(x) &= -\frac{\cos(x)}{x^2} - \frac{\sin(x)}{x}, \\
j_2(x) &= \left(\frac{3}{x^3} - \frac{1}{x}\right)\sin(x) - \frac{3}{x^2}\cos(x), & n_2(x) &= \left(-\frac{3}{x^3} + \frac{1}{x}\right)\cos(x) - \frac{3}{x^2}\sin(x).
\end{aligned} \tag{K.3}$$

Hankel functions are defined by $h_l^{(1)} = j_l + in_l$ and $h_l^{(2)} = j_l - in_l$ which are convenient combinations for many problems. In particular, for positive imaginary arguments, $h_l^{(1)}$ has the asymptotic form $i^{-l}\mathrm{e}^{-|\kappa|r}/|\kappa|r$ corresponding to a bound state solution.

K.2 Spherical harmonics and Legendre polynomials

Spherical harmonics are the angular part of the solutions of the Laplace equation in spherical coordinates. They are given by,[1]

[1] The definitions here are the same as given by Condon and Shortley [916], Jackson [448], and in "Numerical Recipes" [854]. However, some authors define $Y_{l,m}$ with a factor $(-1)^m$ and omit the factor $(-1)^m$ in the associated Legendre polynomials, Eq. (K.6). Of course, the final form for $Y_{l,m}$ is the same, but one must be careful to use consistent definitions.

$$Y_{l,m}(\theta,\phi)=\sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m[\cos(\theta)]\mathrm{e}^{\mathrm{i}m\phi}, \tag{K.4}$$

which define an orthonormal representation on a sphere

$$\int_0^{\pi}\mathrm{d}\theta\sin(\theta)\int_0^{2\pi}\mathrm{d}\phi Y_{l,m}^*(\theta,\phi)Y_{l',m'}(\theta,\phi)=\delta_{ll'}\delta_{mm'}. \tag{K.5}$$

The functions $P_l^m(\cos(\theta))$ are associated Legendre polynomials, which are related to the ordinary Legendre polynomials $P_l(x)$ by

$$P_l^m(x)=(-1)^m(1-x^2)^{m/2}\frac{\mathrm{d}^m P_l(x)}{\mathrm{d}x^m},\ m=0,\ldots,l. \tag{K.6}$$

The Legendre polynomials $P_l(x)$ are defined to be orthogonal on the interval $[-1,1]$; a compact expression valid for any order is (Rodrigues formula)

$$P_l(x)=\frac{1}{2^l l!}\frac{\mathrm{d}^l(x^2-1)^l}{\mathrm{d}x^l}. \tag{K.7}$$

Using the Rodrigues formula for $P_l(x)$, a definition for $P_l^m(x)$ can be derived valid for both negative and positive m (see previous footnote regarding the factor $(-1)^m$).

$$P_l^m(x)=\frac{(-1)^m}{2^l l!}(1-x^2)^{m/2}\frac{\mathrm{d}^{l+m}(x^2-1)^l}{\mathrm{d}x^{l+m}}. \tag{K.8}$$

It can be shown that

$$P_l^{-m}(x)=(-1)^m\frac{(l-m)!}{(l+m)!}P_l^m(x). \tag{K.9}$$

It is helpful to give explicit examples for low orders in terms of angles with $P_l^m \equiv P_l^m(\cos(\theta))$:

$$\begin{aligned}
&P_0^0=1,\ P_1^0=\cos(\theta),\quad P_2^0=\tfrac{1}{2}[3\cos^2(\theta)-1],\ P_3^0=\tfrac{1}{2}\cos(\theta)[5\cos^2(\theta)-3],\\
&\qquad P_1^1=-\sin(\theta),\ P_2^1=-3\sin(\theta)\cos(\theta),\ P_3^1=-\tfrac{3}{2}\sin(\theta)[5\cos^2(\theta)-1],\\
&\qquad\qquad P_2^2=3\sin^2(\theta),\qquad\quad P_3^2=15\cos(\theta)\sin^2(\theta),\\
&\qquad\qquad\qquad\qquad\quad P_3^3=-15\sin^3(\theta).
\end{aligned} \tag{K.10}$$

K.3 Real spherical harmonics

It is often convenient to work with real functions instead of $Y_{l,m}(\theta,\phi)$ that are eigenfunctions of angular momentum. The general definition is simply the normalized real and imaginary parts of $Y_{l,m}(\theta,\phi)$, which can be denoted $S_{l,m}(\theta,\phi)$ given by

$$\begin{aligned}
S_{l,m}^{+}(\theta,\phi)&=\frac{1}{\sqrt{2}}[Y_{l,m}(\theta,\phi)+Y_{l,m}^*(\theta,\phi)],\\
S_{l,m}^{-}(\theta,\phi)&=\frac{1}{\sqrt{2}i}[Y_{l,m}(\theta,\phi)-Y_{l,m}^*(\theta,\phi)].
\end{aligned} \tag{K.11}$$

These functions are used, e.g., in Ch. 14.

K.4 Clebsch–Gordon and Gaunt coefficients

The Clebsch–Gordan coefficients are extensively used in the quantum theory of angular momentum and play an important role in the decomposition of reducible representations of a rotation group into irreducible representations. Clebsch–Gordan coefficients are given in terms of Wigner $3jm$ symbols by the expression

$$C^{j_3 m_3}_{j_1 m_1, j_2 m_2} = (-1)^{j_1 - j_2 + m_3} \sqrt{2j_3 + 1} \begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & -m_3 \end{pmatrix}, \tag{K.12}$$

where the Wigner $3jm$ symbol is defined by

$$\begin{aligned} \begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix} &= \delta_{m_1+m_2+m_3,0} (-1)^{j_1 - j_2 - m_3} \\ &\times \left[\frac{(j_3 + j_1 - j_2)!(j_3 - j_1 + j_2)!(j_1 + j_2 - j_3)!(j_3 - m_3)!(j_3 + m_3)!}{(j_1 + j_2 + j_3 + 1)!(j_1 - m_1)!(j_1 + m_1)!(j_2 - m_2)!(j_2 + m_2)!} \right]^{1/2} \\ &\times \sum_k \frac{(-1)^{k + j_2 + m_2} (j_2 + j_3 - m_1 - k)!(j_1 - m_1 + k)!}{k!(j_3 - j_1 + j_2 - k)!(j_3 - m_3 - k)!(k + j_1 - j_2 + m_3)!}. \end{aligned} \tag{K.13}$$

The summation over k is over all integers for which the factorials are non-negative.

The Gaunt coefficients [917] (also given by Condon and Shortley [916], pp. 178–179) are defined as

$$c^{l''}(l\ m, l'\ m') = \sqrt{\frac{2}{2l'' + 1}} \int_0^\pi \mathrm{d}\theta \sin(\theta) \Theta(l'', m - m')\ \Theta(l, m)\ \Theta(l', m'), \tag{K.14}$$

where $\Theta(l, m)$ are given by

$$\Theta(l, m) = \sqrt{\frac{2l + 1}{2} \frac{(l - m)!}{(l + m)!}} P_l^m[\cos(\theta)]. \tag{K.15}$$

Like the Clebsch–Gordan coefficients, the Gaunt coefficients can be expressed in terms of the Wigner $3jm$ symbols

$$\begin{aligned} c^{l''}(l\ m, l'\ m') = (-1)^m &\left[\frac{(2l + 1)(2l' + 1)}{2l'' + 1} \right]^{1/2} \\ &\times \begin{pmatrix} l & l' & l'' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} l & l' & l'' \\ m & -m' & -m + m' \end{pmatrix}. \end{aligned} \tag{K.16}$$

The product of two Wigner $3jm$ symbols is associated with the coupling of two angular momentum vectors. In order to make the connection between the two coefficients more transparent we express the Gaunt coefficients in terms of the Clebsch–Gordan

$$c^{l''}(l\ m, l'\ m') = (-1)^{m'} \frac{\left[(2l + 1)(2l' + 1)\right]^{1/2}}{2l'' + 1} C^{l''0}_{l0,l'0} C^{l''m - m'}_{lm,l'-m'}. \tag{K.17}$$

K.5 Chebyshev polynomials

A Taylor series is a direct expansion in powers of the variable

$$f(x) \to c_0 + c_1 x + c_2 x^2 + \cdots + c_M x^M. \tag{K.18}$$

In operator expansions, such as needed in Eq. (23.20), this has the advantage that each successive term is simply obtained recursively using $x^{n+1} = x x^n$; however, for high powers there can be problems with instabilities and the expansion becomes worse as x increases. On the other hand, Chebyshev polynomials of type I, $T_n(x)$, are defined to be orthogonal on the interval $[-1, +1]$, so that any function on this interval can be expanded as a unique linear combination of $T_n(x)$. Furthermore, the expansion has the property that it fits the function $f(x)$ over the entire interval in a least-squares sense, and the polynomials can be computed recursively. The polynomials can be expressed by defining the first two and all others by the recursion relation [854]

$$T_0(x) = 1;\ T_1(x) = x;\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{K.19}$$

The resulting expansion is

$$f(x) \to \frac{c_0}{2} + \sum_{n=1}^{M_p} c_n T_n(x). \tag{K.20}$$

It is a simple exercise to derive the first few polynomials and to demonstrate the orthogonality.

Appendix L

Numerical methods

Summary

The methods described here are widely used in numerical analysis, selected because of their importance in electronic structure calculations. These methods are used primarily in iterative improvements of the wavefunctions (iterative diagonalization), updates of the charge density in the Kohn–Sham self-consistency loop, and displacements of atoms in structure relaxation. Because the size and nature of the problems are so varied, different methods are more appropriate for different cases.

L.1 Numerical integration and the Numerov method

Equations (10.8) and (10.12) are examples of second derivative equations which play a prominent role in physics, e.g. the Poisson equation which we also need to solve in finding the self-consistent solution of the full Kohn–Sham or Hartree–Fock equations. These equations can be written in the general form

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}u_l(r) + k_l^2(r)u_l(r) = S_l(r), \tag{L.1}$$

where $S = 0$ for Schrödinger-like equations. For the Poisson equation, $u_l(r)$ is the electrostatic potential and $S_l(r) = 4\pi\rho_l(r)$ is the l angular momentum component of the charge density. The equations may be discretized on a grid and integrated using a numerical approximation for the second derivative. (A good description can be found in [444].) An efficient approach is to use the Numerov algorithm [921] to integrate the equations outward from the origin, and inward from infinity to a matching point. The solution is given by requiring that the wavefunction and its derivative match at a chosen radius R_c. Since the amplitude of the wavefunction can be required to match (only the overall amplitude is set by normalization), actually it is required only to match the ratio $x(r) \equiv (\mathrm{d}\phi_l(r)/\mathrm{d}r)/\phi_l(r)$, which is the logarithmic derivative of $\mathrm{d}\phi_l(r)$.

We want to discretize the differential equation (L.1) on a grid with spacing h. The second derivative operator can be expressed as (here we drop the subscript l and denote discrete

points by $r_j \to j$)

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}u(j) = \frac{u(j+1) - 2u(j) + u(j-1)}{h^2} + \frac{h^2}{12}\frac{\mathrm{d}^4}{\mathrm{d}r^4}u(r) + O(h^4). \tag{L.2}$$

Here we have explicitly written out the leading error, which is $O(h^2)$ (see Exercise L.1). Direct application of this discretized derivative allows one to calculate all values of $u(j)$ recursively given two initial values, say $u(1)$ and $u(2)$. The error in calculating the new $u(j)$ at each step is of order h^4. However, with a little extra work we can obtain a method that is of order h^6, a substantial improvement known as the Numerov method.

The leading error in the second derivative formula (L.2) is from the fourth derivative of the function. But this can be found by differentiating the differential equation, (Eq. L.1), twice which leads to the relation $\frac{\mathrm{d}^4}{\mathrm{d}r^4}u(r) = \frac{\mathrm{d}^2}{\mathrm{d}r^2}(S(r) - k^2(r)u(r))$. That is, knowledge about the curvature of the source and potential terms leads to a more accurate integration scheme. Substituting this expression into (L.2), defining $F(r) = S(r) - k^2(r)u(r)$, and using (L.2) to lowest order for $\frac{\mathrm{d}^2}{\mathrm{d}r^2}F(r)$, we find the improved formula (Exercise L.2)

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}u(j) = \frac{u(j+1) - 2u(j) + u(j-1)}{h^2} + \frac{F(j+1) - 2F(j) + F(j-1)}{12} + O(h^4). \tag{L.3}$$

Substituting this into the original equation leads to the final formula (here $a \equiv h^2/12$)

$$\begin{aligned}[1 + ak^2(j+1)]u(j+1) - 2[1 - 5ak^2(j)]u(j) + [1 + ak^2(j-1)]u(j-1) \\ = a[S(j+1) - 2S(j) + S(j-1)] + O(h^6),\end{aligned} \tag{L.4}$$

which can be solved recursively (forward or backward) starting with u at two grid points.

The idea behind the Numerov method can be extended to any dimension, where the pattern of points is given the name "Mehrstellen" (see [209] which cites [537], p. 164)). The key point is the use of the differential equation itself to find an expression for both kinetic and potential terms valid to higher order than the original finite difference expression.

L.2 Steepest descent

Minimization of a function $F(\{x_i\})$ in space of variables x_i, $i = 1, N$ is a widely studied problem in numerical anlysis [854, 920, 922].[1] In the absence of any other information the best choice for a direction of displacement from a point x_i^0 to reach the minimum is the steepest descent (SD) direction

$$g_i^0 = -\frac{\partial F}{\partial x_i}|_{x_i = x_i^0}, \tag{L.5}$$

which is shown by the initial direction from point 0 in Fig. L.1. The lowest energy along this direction can be found by "line minimization" in one-dimensional space, i.e. the minimum of F as a function of α^1, where $x_i^1 = x_i^0 + \alpha^1 g_i^0$. Of course, a series of such

[1] For simplicity we assume that there is only one minimum, which is valid in large classes of problems in electronic structure. For special cases, such as level crossing at transition states, one may need to adopt special measures.

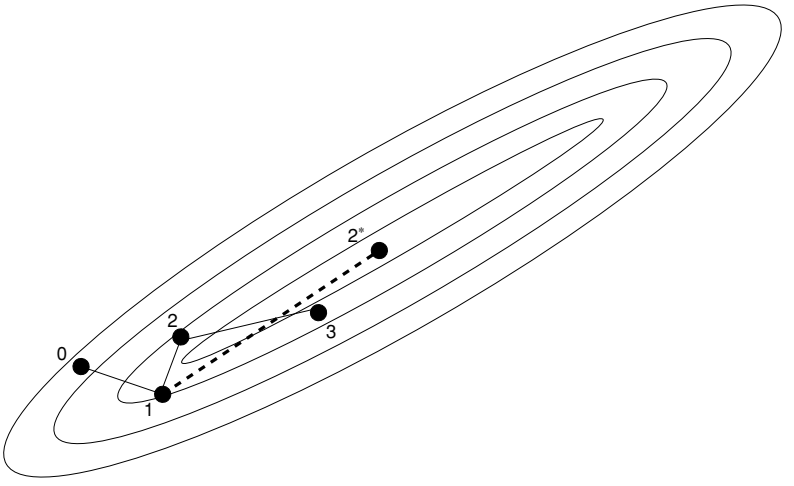


Figure L.1. Schematic illustration of minimization of a function in two dimensions. The steps 1, 2, 3, . . ., denote the steepest descent steps, and the point 2* denotes the conjugate gradient path that reaches the exact solution after two steps if the functional is quadratic.

steps must be taken to approach the absolute minimum, generating the sets of points $x_i^0, x_i^1, x_i^2, \ldots,$. This process is illustrated in Fig. L.1 for a very simple function of two variables $F(x_1, x_2) = A(x_1)^2 + B(x_2)^2$, with $B \gg A$. We see that even though the function F decreases at each step, the steps do not move directly to the minimum. Furthermore, the method suffers from a real version of the "Zeno paradox" and one never reaches the minimum exactly. The SD method is particularly bad if the function F has very different dependence on the different variables so that the region around the minimum forms a long narrow valley.

L.3 Conjugate gradient

Although it may seem surprising, there is a faster way to reach the minimum than to always follow the "downhill" steepest descent direction. After the first step, one not only has the gradient F at the present point, but also the value and gradient at previous points. The additional information can be used to choose a more optimal direction along which the line minimization will lead to a lower energy. In fact, for a quadratic functional in N dimensions, the conjugate gradient (CG) method is guaranteed to reach the minimum in N steps [704, 854, 920, 922]. We will consider this case explicitly to illustrate the power of method. In addition, CG can be applied to more complicated functionals (such as the Kohn–Sham functional) and we expect many advantages still to accrue since the functional is quadratic near the minimum.

Consider the quadratic functional

$$F(\{x_i\}) \equiv F(\mathbf{x}) = \frac{1}{2}\mathbf{x} \cdot \mathbf{H} \cdot \mathbf{x}, \tag{L.6}$$

with gradients

$$\mathbf{g} = -\frac{\partial F}{\partial \mathbf{x}} = -\mathbf{H} \cdot \mathbf{x}. \tag{L.7}$$

The first step is the same as steepest descent, i.e. minimization of F along a line $\mathbf{x}^1 = \mathbf{x}^0 + \alpha^1 \mathbf{d}^0$, where $\mathbf{d}^0 = \mathbf{g}^0$. For this and for all steps, the minimum occurs for

$$\mathbf{d}^n \cdot \mathbf{g}(\mathbf{x}^{n+1}) = 0. \tag{L.8}$$

For the $n+1$ step the best choice is to move in a direction where the gradient along the previous direction $\mathbf{d}^n$ remains zero. Since the change in gradient as we move in the new direction $\mathbf{d}^{n+1}$ is $\Delta\mathbf{g} = \alpha^{n+1}\mathbf{H} \cdot \mathbf{d}^{n+1}$, it follows that the desired condition is satisfied if

$$\mathbf{d}^n \cdot \mathbf{H} \cdot \mathbf{d}^{n+1} = 0. \tag{L.9}$$

This equation defines the "conjugate direction" in the sense of orthogonality in the space with metric $\mathbf{H} = H_{ij}$. If this condition is satisfied at each step, then it can be shown (Exercise L.3) that the conjugate condition is maintained for *all steps*

$$\mathbf{d}^{n'} \cdot \mathbf{H} \cdot \mathbf{d}^{n+1} = 0, \text{ for all } n' \leq n. \tag{L.10}$$

The key point is that (unlike SD) each line minimization *preserves* the minimization done in all previous steps and only adds independent (i.e. conjugate) variations. This is manifested in the fact that (unlike the SD method that never reaches the minimum) for a quadratic functional the conjugate gradient method reaches the minimum *exactly* in N steps, where N is the dimension of the space x_i, $i = 1, N$. This is illustrated in Fig. L.1 where the exact solution is reached in two steps for a problem with two variables.

For actual calculations it is useful to specify the new conjugate gradient direction $\mathbf{d}^{n+1}$ in terms of the quantities at hand, the current gradient and the previous direction,

$$\mathbf{d}^{n+1} = \mathbf{g}^{n+1} + \gamma^{n+1}\mathbf{d}^n. \tag{L.11}$$

Also available is the quantity $\mathbf{y}^n = \mathbf{d}^n \cdot \mathbf{H}$ which is needed for the evaluation of F in the line minimization for direction $\mathbf{d}^n$. Using Eq. (L.8), (L.11) can be written

$$\gamma^{n+1} = -\frac{\langle \mathbf{y}^n | \mathbf{g}^{n+1} \rangle}{\langle \mathbf{y}^n | \mathbf{d}^n \rangle}. \tag{L.12}$$

It is straightforward to show (Exercise L.4) that the directions are also given by

$$\gamma^{n+1} = \frac{\mathbf{g}^{n+1} \cdot \mathbf{g}^{n+1}}{\mathbf{g}^n \cdot \mathbf{g}^n}, \tag{L.13}$$

with the definition $\gamma^1 = 0$. These forms[2] are equivalent for the quadratic case; however, they are different in the applications needed for electronic structure, where there are constraints or non-linearities (App. M).

[2] Equation (L.12) is often called the Hestens–Steifel form; Eq. (L.13) is the Fletcher–Reeves expression; and an alternative Pollak–Ribiere form is particularly useful for non-quadratic functionals [854].

How does one apply the CG method to problems that are not quadratic? The basic idea is to define conjugate directions as above, but to carry out the line minimization for the given non-linear functional. This is essential for the CG algorithm since one must reach the line minimum in order for the new gradient to be perpendicular to the present direction, so that the functional is fully minimized along each direction in turn.

L.4 Quasi-Newton–Raphson methods

Consider the problem of solving the equation

$$\mathbf{F}(\mathbf{x}) = \mathbf{x}, \tag{L.14}$$

where $\mathbf{x}$ denotes a vector in many dimensions. For example, this could be the problem in Sec. 9.3 of finding the solution of the Kohn–Sham equations where the output density $n^{\text{out}}(\mathbf{r})$ (which is a function of the input density $n^{\text{in}}(\mathbf{r})$) is equal to the input density $n^{\text{in}}(\mathbf{r})$. This problem has exactly the form of (L.14) if the density is expanded in a set of M functions $n^{\text{in}}(\mathbf{r}) = \sum_k^M x^k h^k(\mathbf{r})$, with $\mathbf{x} = \{x^k\}$. This becomes a minimization problem for the norm of the residual $|\mathbf{R}[\mathbf{x}]|$, where

$$\mathbf{R}[\mathbf{x}] \equiv \mathbf{F}(\mathbf{x}) - \mathbf{x}. \tag{L.15}$$

In Eqs. (9.21) and (13.7) it was shown how to solve this problem if one is in a region where $\mathbf{R}$ is a linear function of $\mathbf{x}$ and the Jacobian,

$$\mathbf{J} \equiv \frac{\delta \mathbf{R}}{\delta \mathbf{x}}, \tag{L.16}$$

is known. Then one can follow the Quasi-Newton–Raphson approach to minimize the residual. In terms of $\mathbf{x}_i$ at step i, the value that would give $\mathbf{r}_{i+1} = 0$ at the next iteration is

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{J}^{-1}\mathbf{R}_i. \tag{L.17}$$

The problem is that, in general, the Jacobian is not known (or it is hard to invert) and one needs to resort to other methods which iterate to the solution in a space of functions, i.e. a Krylov subspace.

L.5 Pulay DIIS full-subspace method

The idea behind the "discrete inversion in the iterative subspace" (DIIS) method[3] is to minimize the residual at any step i by using the best possible combination of *all* previously generated vectors, i.e. making use of the full Krylov subspace.

$$\mathbf{x}_{i+1} = \sum_{j=0}^{i} a_j \mathbf{x}_j = c_0 \mathbf{x}_0 + \sum_{j=1}^{i} c_j \delta \mathbf{x}_j. \tag{L.18}$$

[3] The present discussion follows [718].

If we assume linearity of the residual near the solution, then

$$\mathbf{R}[\mathbf{x}_{i+1}] = \mathbf{R}[\sum_{j=0}^{i} a_j \mathbf{x}_j] = \sum_{j=0}^{i} a_j \mathbf{R}[\mathbf{x}_j]. \tag{L.19}$$

The condition is that $\mathbf{x}_{i+1}$ be chosen to minimize the square norm of the residual

$$\langle \mathbf{R}[\mathbf{x}_{i+1}] | \mathbf{R}[\mathbf{x}_{i+1}] \rangle = \sum_{j,k} a_j a_k A_{j,k};\ A_{j,k} = \langle \mathbf{R}[\mathbf{x}_j] | \mathbf{R}[\mathbf{x}_k] \rangle, \tag{L.20}$$

subject to any auxiliary conditions. In electronic structure problems, the two most relevant conditions are:

- electronic bands, where one requires orthonormalization of eigenvectors;
- density mixing, where $\sum_{j=0}^{i} a_j = 1$ for charge conservation.

In the latter case the solution is [718]

$$a_i = \sum_j A^{-1}_{j,i} \Big/ \sum_{j,k} a_j a_k A^{-1}_{j,k}. \tag{L.21}$$

Through Eq. (L.18), this provides the optimal new vector at each step i in terms of the results of all previous steps. For extremely large problems, such as many eigenvectors for the Schrödinger equation, it is not feasible to store many sets of vectors. However, for density mixing, especially for only a few troublesome components of the density, one can store several previous densities.

Kresse and Furthmüller [718] have shown that the Pulay DIIS method described above is equivalent to updating a Jacobian that is closely related to the modified Broyden schemes [433, 434]. In addition, van Lenthe and Pulay have shown that it is possible to carry out the Davidson and DIIS algorithms with only three vectors at each step [923].

L.6 Broyden Jacobian update methods

The Broyden method [431] is a way to generate the inverse Jacobian successively in the course of an iterative process.[4] The modified Broyden method [433, 434] given at the end is similar to the result of the DIIS method, except that it explicitly involves only the two states at a time. This method is widely used and it is illuminating to derive the form in steps that show the relevant points.

The method starts with a reasonable guess $\mathbf{J}_0^{-1}$ (e.g. that for linear mixing $\mathbf{J}_0^{-1} = \alpha \mathbf{1}$ [430]). The approximate form may be used for several steps after which the inverse $\mathbf{J}^{-1}$ is improved at subsequent steps. Since the Jacobian is not exact at any step, Eq. (L.17) and the actual calculation at step i provide two quantities: (1) the prediction from (L.17) for step i, $\delta \mathbf{x}_i = \mathbf{x}_i - \mathbf{x}_{i-1} = -\mathbf{J}_{i-1}^{-1} \mathbf{R}_{i-1}$; and (2) the actual result from step i, the change in the residual $\delta \mathbf{R}_i = \mathbf{R}_i - \mathbf{R}_{i-1}$. The new, improved $\mathbf{J}_i^{-1}$ is chosen by requiring that at each

[4] The description here follows that of Pickett in [413].

step i the $\mathbf{J}_i^{-1}$ be able to reproduce the result of the iteration just completed, i.e.

$$0 = \delta\mathbf{x}_i - \mathbf{J}_i^{-1}\delta\mathbf{R}_i. \tag{L.22}$$

This provides M equations for the M^2 components of $\mathbf{J}_i^{-1}$. The other conditions are fixed by requiring the norm of the change in the Jacobian matrix

$$Q = \left|\left|\mathbf{J}_i^{-1} - \mathbf{J}_{i-1}^{-1}\right|\right| \tag{L.23}$$

be minimized. The last may be accomplished by the method of Lagrange multipliers, and is equivalent to the condition that $\mathbf{J}_i^{-1}$ produces the same result as $\mathbf{J}_{i-1}^{-1}$ acting on *all* vectors orthogonal to the current change $\delta\mathbf{R}_i$. The result is [413,430] (see Exercise L.8)

$$\mathbf{J}_i^{-1} = \mathbf{J}_{i-1}^{-1} \frac{(\delta\mathbf{x}_i - \mathbf{J}_i^{-1}\delta\mathbf{R}_i)\delta\mathbf{R}_i}{\langle\delta\mathbf{R}_i|\delta\mathbf{R}_i\rangle}. \tag{L.24}$$

As it stands, Eq. (L.24) can be used if the Jacobian matrix is small, e.g. in plane wave methods where only a few troublesome components of the density need to be treated in this way. However, it is not useful in cases where storage of a full Jacobian matrix is not feasible, e.g. in the update of the charge density on a large grid needed in many calculations. Srivastava [430] introduced a way to avoid storage of the matrices completely by using Eq. (L.24) to write the predicted change $\delta\mathbf{x}_{i+1}$ in terms of a sum over all the previous steps involving only the initial $\mathbf{J}_0^{-1}$ (see also [432]).

A modified Broyden method has been proposed by Vanderbilt and Louie [433] and adapted by Johnson [434] to include the advantages of Srivistava's method [430] that requires less storage. The idea is that the requirement that the immediate step be reproduced exactly is too restrictive, and an improved algorithm can take into account information from previous iterations. Then one finds $\mathbf{J}_i^{-1}$ by minimizing a weighted norm

$$Q^{\text{modified}} = \sum_{j=1}^{i} w_j \left|\delta\mathbf{x}_j - \mathbf{J}_i^{-1}\delta\mathbf{R}_j\right|^2 + w_0\left|\left|\mathbf{J}_i^{-1} - \mathbf{J}_0^{-1}\right|\right|. \tag{L.25}$$

This has the advantage that the weights w_j can be chosen to emphasize the most relevant prior steps and the term w_0 adds stability. Vanderbilt and Louie [433] showed a simple example in which the modified method approached the exact Jacobian rapidly, compared to a slower approach using the original Broyden scheme. Clearly, there are strong resemblances to the Pulay DIIS algorithm of the previous section.

L.7 Moments, maximum entropy, kernel polynomial method, and random vectors

The direct determination of the spectral properties of a Hermitian matrix via conventional Householder tridiagonalization has computational cost scaling as N^3. However, if one is interested only in the density of states of such a matrix (whatever its origin – dynamical matrix, Hamiltonian matrix, etc.), then there are more efficient schemes based on the relative ease of extracting *power moments* of the spectral densities. The utility of moments in physical calculations was recognized before the era of quantum mechanics, when Thirring used moments of the dynamical matrix to estimate thermodynamic quantities [924]. Montroll employed

moments to compute vibrational state densities as referenced in Born and Huang [90], p. 74.

The moments of the eigenvalue spectrum about an energy E_0 are defined as

$$\langle [H - E_0]^n \rangle = \sum_i [\varepsilon_i - E_0]^n = \int \mathrm{d}\varepsilon [\varepsilon - E_0]^n n(\varepsilon), \tag{L.26}$$

where $n(\varepsilon)$ is the density of states (see Sec. 4.7)

$$n(\varepsilon) = \sum_i \delta(\varepsilon - \varepsilon_i). \tag{L.27}$$

The zeroth moment is the total number of states; the first moment, the average eigenvalue; the second, a measure of the spectral width; the third moment, a measure of the spectral asymmetry about E_0; etc. From many moments, one can approximately reconstruct the density of states. Similarly, local information is derived using the local projected density of states, such as the angular momentum projected density around an atomic site in Eq. (16.33) or the basis function projection in Eq. (23.17). Thus the fundamental quantities in electronic structure can be determined from the moments if there are useful ways to compute the moments and there are stable algorithms to reconstruct the spectrum.

The first aspect, finding the moments given the hamiltonian matrix, is beautifully solved by the recursion method [828]. The expressions given in (23.17) relate the moments to the coefficients generated in the Lanczos algorithm, which have the interpretation of creating a "chain" of hops whereby the hamiltonian connects one state to the next. If the hamiltonian matrix is localized in space (short-range hops) this means that information about the local density of states at a site can be efficiently generated in a small number of applications of the hamiltonian because only hops within some local range are needed, as explained in Ch. 23.

If the global (rather than projected) DOS is needed, one can compute approximate moments of the global DOS by repeated matrix-on-vector operations where the vectors needed have random components (a suitable choice is to sample each component independently from the unit normal distribution). For the global DOS for large matrices, very few vectors are needed to provide moments leading to accurate spectra (there is a "self-averaging" which requires *fewer* vectors for larger system sizes [832]. Very accurate determination of partial integrals of the DOS is another matter and requires more careful convergence of the moment data with respect to random vectors. Much of this was grasped earlier with characteristic prescience by Lanczos [925].

The second aspect, reconstructing the spectrum, is a long-standing problem in applied mathematics in the nineteenth and twentieth centuries called the "classical moment problem" [830]:

> Given a finite number of moments over some interval of a non-negative function, find the function from which the moments arose.

Two classes of practical solutions have emerged. Most naturally, one may adopt a polynomial solution using polynomials suitably orthogonal on the interval. With a sufficient number of

moments, it is possible to obtain very accurate reconstructions, e.g. with plausible jagged spikes approximating δ functions. The method is numerically robust [838] and has been extended to non-orthogonal bases [926]. It is routine in these computations to work with several hundred or more moments.

Alternately, one can seek to find a "best" solution given incomplete information (a finite moment sequence). A modern method that has been applied to this problem is the method of maximum entropy, which utilizes a variational principle to maximize the "entropy" $-\int n(\varepsilon)\ln(n(\varepsilon))$ subject to the constraints that the moment conditions are satisfied [927]. The utility of maximum entropy for moments was shown with examples by Mead and Papanicolaou [928]; Skilling [927] used maximum entropy with random vectors to extract state densities of large matrices; Drabold and Sankey [832] applied the method to electronic structure problems and introduced "importance sampling" in selecting vectors to improve the convergence of integrated quantities (like the band energy for determining Fermi level); and Stephan, Drabold, and Martin [855] demonstrated that this scheme was useful in density functional schemes for determining the Fermi level order-N in several thousand atom models. An example of calculation of the phonon density of states from a sparse dynamical matrix is given in Fig. 23.4. Maximum entropy converges much faster than the orthogonal polynomial solution, but is more delicate numerically and it is difficult to use more than a few hundred moments in current maximum entropy schemes.

SELECT FURTHER READING

Golub, G. H. and Van Loan, C. F., *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 1980.

Heath, M. T., *Scientific Computing: An introductory Survey*, McGraw-Hill, New York, 1997.

Koonin, S. E. and Meredith, D. C., *Computational Physics*, Addison Wesley, Menlo Park, CA, 1990.

Parlett, B. N., *The Symmetric Eigenvalue Problem*, Prentice Hall, Engelwood Cliffs, N. J., 1980.

Press, W. H. and Teukolsky, S. A., *Numerical Recipes*, Cambridge University Press, Cambridge, 1992.

Thijssen, J. M., *Computational Physics*, Cambridge University Press, Cambridge, England, 2000.

Exercises

L.1 Derive the leading error in the finite difference approximation to the second derivative that is $O(h^2)$ and is given explicitly in Eq. (L.2).

L.2 Derive the Numerov expressions (L.3) and (L.4) and show the leading error in the solutions are, respectively, $O(h^4)$ and $O(h^6)$.

L.3 Show that the conjugate gradient minimization equations, (L.8) and (L.9), follow from differentiating the functional and assuming it is quadratic. Then derive the key equation, (L.10), that if each direction is made conjugate to the previous one, then it is also conjugate to all previous directions. This can be shown by induction given that each direction is defined to be conjugate to the previous direction and it is a linear combination only of the new steepest descent gradient and the previous direction, as in Eq. (L.11).

L.4 For the quadratic functional (L.6), show that the conjugate directions (L.12) are also given by (L.13).

L.5 Consider a two-dimensional case $F(x, y) = Ax^2 + By^2$, with $B = 10A$. Show that the CG method reaches the exact minimum in two steps, starting from any point (x, y), whereas SD does not. What is the value of F in the SD method after two steps starting from $x = 1;\ y = 1$.

L.6 As the simplest three-dimensional example, consider $F(x, y, z) = Ax^2 + By^2 + Cz^2$, and show that the third direction $\mathbf{d}^3$ is conjugate to the first direction $\mathbf{d}^1 = \mathbf{g}^1$.

L.7 Make a short computer program to do the CG minimization of a function $F = \mathbf{G} \cdot \mathbf{x} + \mathbf{x} \cdot \mathbf{G} \cdot \mathbf{x}$ in any dimension for any $\mathbf{G}$ and $\mathbf{H}$.

L.8 The Broyden method generates a new approximation to the inverse Jacobian $\mathbf{J}_i^{-1}$ at each step i based upon the conditions outlined before Eq. (L.24). Verify that $\mathbf{J}_i^{-1}$, defined by Eq. (L.24), satisfies (L.22) and that $\mathbf{J}_i^{-1} - \mathbf{J}_{i-11}^{-1}$ gives a null result when acting on any residual orthogonal to $\delta\mathbf{R}_i$.

Appendix M

Iterative methods in electronic structure

Summary

This appendix describes technical aspects of advances made in recent years, stimulated by the work of Car and Parrinello in 1985 (Ch. 18), that have brought entire new classes of problems and properties under the umbrella of *ab initio* electronic structure. In fact, the methods belong to general classes of iterative algorithms that have a long history in eigenvalue problems, even though their widespread use in electronic structure in condensed matter followed the work of Car and Parrinello. This chapter is devoted to features particularly relevant to electronic structure, and aspects that are inherent to general numerical algorithms are deferred to App. L. The methods may be classified in many ways: as minimization of the energy versus minimization of a residual; single vector update versus full iterative subspace methods; etc. Nevertheless, they can all be brought into a common framework, in which the key features are to:

- replace matrix diagonalization by iterative equations for the wavefunctions ψ_i in an iterative (Krylov) subspace;
- find new ψ_i^{n+1} using ψ_i^n (and possibly previous $\psi_i^{n'}$, $n' < n$) and the gradient $\mathrm{d}E/\mathrm{d}\psi_i^{n*} = H_{\mathrm{KS}}\psi_i^n$ (the algorithms for this step is where methods differ);
- For plane waves, replace dense matrix multiplications with fast Fourier transforms (FFTs).

M.1 Why use iterative methods?

Electronic structure methods can be grouped into two camps differentiated by the types of basis functions. Methods such as LCAO and LMTO are predicated upon the goal of constructing a *minimal basis* of size N_b; the work goes into constructing the basis, which may be highly optimized for a given class of problems. Except for very large systems (see Ch. 23), the hamiltonian is expressed as a small, dense matrix of size $N_b \times N_b$, for which it is appropriate to employ traditional dense matrix diagonalization techniques, for which the computational effort scales as N_b^3 or as $N_e N_b^2$, where N_e is the number of desired eigenvectors.

On the other hand, methods that use general bases such as plane waves[1] and grids often involve a much larger number of basis functions than the number of desired eigenstates ($N_b \gg N_e$); the hamiltonian is very simple to construct and it can be made sparse, i.e. mainly zero elements so that only the non-zero elements need to be calculated and/or stored. Except for small problems, it is much more efficient to use iterative methods, in which the $N_b \times N_b$ hamiltonian is never explicitly constructed and the computational effort scales as $N_e^2 N_b$ or as $N_e^2 N_b \ln(N_b)$. These approaches have been applied most successfully to plane waves (where they are built upon the pioneering work of Car and Parrinello [156] using fast Fourier transforms and regular grids as described in Sec. M.11), and real-space methods [525]: finite difference [526, 535], finite element [544, 545, 547], multigrid [209, 538]; and wavelets [553, 814]

The iterative methods described in this appendix have much in common with the problem of finding the self-consistent Kohn–Sham potential, which is in general an iterative process as described in Ch. 9, and "order-N" approaches of Ch. 23, which are useful for very large systems with iterative methods employed that take advantage of the sparseness of the hamiltonian. Since many of the methods employed are useful in many contexts, the general forms for the methods are discussed in App. L and their application to calculations of eigenvalues and eigenvectors emphasized in this appendix.

We first consider the problem of solving the Schrödinger equation for a fixed hamiltonian

$$(H - \varepsilon)|\psi\rangle = 0. \tag{M.1}$$

This is the problem in many-body simulations where the hamiltonian never changes, and it is the inner loop in a Kohn–Sham problem where the effective independent-particle hamiltonian may be taken as fixed during the iterations (the solution inside of the loop in Fig. 9.1) to find the eigenvalues and eigenvectors of that effective hamiltonian. Iterative methods also have the advantage that the hamiltonian can be updated simultaneously with improvements to the wavefunctions (e.g. in the Car–Parrinello unified method, Ch. 18) to achieve self-consistency as well as to solve the Schrödinger equations. However, logically it is simpler to first consider the case of a fixed hamiltonian after which the extension is not difficult.

M.2 Simple relaxation algorithms

The algorithm [939] proposed by Jacobi in 1848 is in many ways the grandfather of iterative eigenvalue methods. The basic idea is to iterate a from of the equation

$$(H - \varepsilon^n)|\psi^n\rangle = |R[\psi^n]\rangle, \tag{M.2}$$

where n is the iteration step, $|\psi^n\rangle$ and ε^n are approximate eigenvectors and eigenvalues, and $|R[\psi^n]\rangle$ is a "residual" vector. The iterations continue with a particular choice of the

[1] The APW, LAPW, and PAW methods are in some ways intermediate and it may be possible to take advantage of both types of approaches.

improved eigenvector $|\psi^{n+1}\rangle$ and eigenvalue ε^{n+1} until the eigenvalue is converged or the norm of the residual vanishes to within some tolerance [930].

If the matrix is diagonally dominant (as is the case for the hamiltonian expressed in the bases most commonly chosen in electronic structure calculations) then we can rewrite the eigenvalue problem, Eq. (M.1), as

$$|\psi\rangle = D^{-1}(H - \varepsilon)|\psi\rangle + |\psi\rangle, \tag{M.3}$$

where D is a non-singular matrix. This form suggests many variations and the choice of D can be viewed as a "preconditioning" of the hamiltonian operator, as discussed below. If we define the iteration sequence [930]

$$\begin{aligned} \varepsilon^n &= \frac{\langle\psi^n|H|\psi^n\rangle}{\langle\psi^n|\psi^n\rangle}, \\ \delta\psi^{n+1} &= D^{-1}(H - \varepsilon^n)\psi^n, \\ \psi^{n+1} &= \psi^n + \delta\psi^{n+1}, \end{aligned} \tag{M.4}$$

then the middle equation of (M.4) is just the linear set of equations

$$D\delta\psi^{n+1} = R^n \quad \text{or} \quad \delta\psi^{n+1} = D^{-1}R^n \equiv K R^n, \tag{M.5}$$

where R^n is the residual at step n and $K \equiv D^{-1}$.

The sequence Eq. (M.4) with Eq. (M.5) corresponds to updates of ψ using the residual R multiplied by a "preconditioning" matrix K (see Sec. M.3). For the methods to be efficient, the matrix D must be easier to invert than the original matrix $(H - \varepsilon)$, and yet be chosen so that the change $\delta\psi^{n+1}$ is as close as possible to the improvement needed to bring ψ^n to the correct eigenvector. From perturbation theory we know that if the hamiltonian is diagonally dominant a good choice is D equal to the diagonal part of H (the choice made by Jacobi). If D is the lower (or upper) triangular part of H, then this becomes the Gauss–Seidel relaxation method [854, 940] which is useful in "sweep methods" where the points on one side have already been updated. At each iteration the new vector is updated with only information from the previous step.

M.3 Preconditioning

The basic idea behind "preconditioning" is to modify the functional dependence upon the variables to be more "isotropic," i.e. to make the curvature more similar for the different variables, which is exactly the idea behind the improved convergence in (M.5). For the problems encountered in electronic structure, it often happens that the original formulation is very badly conditioned; but on physical grounds it is simple to see how to improve the conditioning. In general, the choice of formula depends upon the problem and we will be content to list two characteristic examples.

The simplest example is the energy expressed in a plane wave basis, where the functions are expressed as $\psi_{i,\mathbf{k}}(\mathbf{r}) = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})u_{i,\mathbf{k}}(\mathbf{r})$, with u the Bloch function given in (12.12), where $c^n_{i,m}(\mathbf{k})$ are the $m = 1, N_{\mathrm{PW}}$ variables describing the $i = 1, N_e$ eigenvectors at step n.

Because high Fourier components $|\mathbf{k}+\mathbf{G}_m|$ have high kinetic energy, the total energy varies much more rapidly as a function of coefficients $c^n_{i,m}(\mathbf{k})$ with large $|\mathbf{G}_m|$ than for coefficients with small $|\mathbf{G}_m|$. Preconditioning can be used to modify the gradients and cancel this effect; a simple form suggested in [931] is

$$K(x) = \frac{27+18x+12x^2+8x^3}{27+18x+12x^2+8x^3+16x^4}, \tag{M.6}$$

where

$$x^n_i(\mathbf{G}_m) = \frac{1}{2}\frac{|\mathbf{k}+\mathbf{G}_m|^2}{T^n_i}, \tag{M.7}$$

which multiplies each steepest descent vector $g^n_i(\mathbf{G}_m)$. Here $x^n_i(\mathbf{G}_m)$ is the ratio of the kinetic energy of the $|\mathbf{k}+\mathbf{G}_m|$ Fourier component to the kinetic energy T^n_i of the state i at step n. Since $K \propto 1/|\mathbf{G}_m|^2$ for large $|\mathbf{G}_m|$, this cancels the increase in $g^n_i(\mathbf{G}_m)$ which grows as $|\mathbf{G}_m|^2$.

Seitsonen [941] proposed to precondition the steepest descent vector in real-space methods by extending the form of (M.6) to represent a local kinetic energy at point $\mathbf{r}$. The variable x in (M.6) is defined to be

$$x^n_i(\mathbf{r}) = A\frac{|\lambda^m_i - V(\mathbf{r})|}{T^n_i}, \tag{M.8}$$

which is the ratio of the local kinetic energy $|\lambda^n_i - V(r)|$ to the total kinetic energy for state i, $T^n_i = \langle\psi^n_i|\nabla^2|\psi^n_i\rangle$, and A is an adjustable parameter. At each step n the factor $K(x^n_i(\mathbf{r}))$ multiplies the residual for state i at each point $\mathbf{r}$ of the real-space grid as in (M.5).

M.4 Iterative (Krylov) subspaces

Iterative methods are based upon repeated application of some operator A to generate new vectors. Starting from a trial vector ψ^0, a set of vectors $A^n\psi^0$ is generated by recursive application of A. Linear combinations of the vectors can be chosen to construct the set $\{\psi^0, \psi^1, \psi^2, \ldots\}$, which forms a Krylov subspace [919]. In many cases, an accurate solution for desired states can be found in terms of a number of states in this new basis that is much smaller than the number of states in the original basis. The distinction between the various methods is the choice of operator A and the way that new vectors ψ^{n+1} are created at each step using $A\psi^n$ and the previously generated ψ^i, $i = 0, n$.

There are three choices for A that are most directly applicable to problems related to electronic structure: the (shifted) hamiltonian, $A = [H-\varepsilon]$; the shifted inverse hamiltonian operator, $A = [H-\varepsilon]^{-1}$; and the imaginary time propagator, $A = \exp(-\delta\tau(H-\varepsilon))$. Each of these choices has important advantages. The first is closely related to the variational Schrödinger and Kohn–Sham equations, (3.13) and (7.11), which leads to helpful physical interpretations and suggests solution in terms of well-established minimization techniques and subspace matrix diagonalization techniques [919, 920, 942]. The second choice, employing inverse powers, is especially appropriate for finding eigenvectors close to a trial

eigenvalue ε. The inverse is useful for proofs of principle, but in practice one uses approximations with easily invertible operators; these are closely related to perturbation expansions for the wavefunctions and eigenvalues. Imaginary time projection has the advantage that it is closely related to real-time methods for time-dependent phenomena (see Ch. 20) and to statistical mechanics involving thermal expectation values, where $\beta = 1/k_B T \to \delta\tau$.

Methods differ in the extent of the Krylov subspace explicitly treated at each iteration. Simple relaxation methods such as the Jacobi algorithm find the new approximate eigenvector ψ^{n+1} in terms of the previous ψ^n only. This is analogous to steepest descent minimization. Others, such as the Lanczos, Davidson, and the RMM–DIIS (Sec. M.7) methods consider the entire subspace generated up to the given iteration. In general, a great price must be paid to keep the entire Krylov subspace for very large problems; however, the widely used Lanczos and conjugate gradient (CG) minimization methods are full subspace methods, able to generate a new vector orthogonal (or conjugate) to *all* previous vectors even though ψ^{n+1} is found only in terms of the previous two vectors ψ^n and ψ^{n-1}. Thus these methods can be much more powerful than simple relaxation methods, with only a moderate increase in requirements at each step of the iteration. The original Davidson method [935] requires keeping the entire subspace, but it can also be cast in a form requiring only three vectors using a CG approach [923].

M.5 The Lanczos algorithm and recursion

The Lanczos method [943] was one of the first iterative methods used by modern computers to solve eigenvalue problems. It is remarkably simple and amazingly powerful as a tool to bring out physical interpretations and analogies. The algorithm automatically generates an orthogonal basis (a Krylov or iterative subspace) in which the given operator A is tridiagonal. (In electronic structure problems $A = H$, where H is often the hamiltonian.) It is especially powerful for generating a number of the lowest (or highest) eigenvectors of large matrices. The simplest version suffers from the "Lanczos disease" of spurious solutions due to numerical rounding errors as the number of desired eigenvectors increases; however, this can be easily controlled by orthogonalizing after a number of iterative steps. In addition, it can be formulated as a continued fraction which leads to powerful methods for finding moments of the spectral distribution.

The Lanczos algorithm proceeds as follows (good descriptions can be found in [444] and [944]): Starting with a normalized trial vector ψ_1, form a second vector $\psi_2 = C_2 [A\psi_1 - A_{11}\psi_1]$, where $A_{11} = \langle\psi_1|A|\psi_1\rangle$ and C_2 is chosen so that ψ_2 is normalized. It is easy to see that ψ_2 is orthogonal to ψ_1. Subsequent vectors are constructed recursively by

$$\psi_{n+1} = C_{n+1} [A\psi_n - A_{nn}\psi_n - A_{nn-1}\psi_{n-1}] . \tag{M.9}$$

The matrix $A_{nn'}$ is explicitly tridiagonal since Eq. (M.9) shows that A operating on ψ_n yields only terms proportional to ψ_n, ψ_{n-1}, and ψ_{n+1}. Furthermore, each vector ψ_n is orthogonal to *all* the other vectors, as may be shown by induction (see Exercise M.1). Going to step M yields a tridiagonal matrix

$$A = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \alpha_3 & . & & \\ & & . & . & . & \\ & & & . & . & \beta_M \\ & & & & \beta_M & \alpha_M \end{bmatrix}, \tag{M.10}$$

where α and β are given by Eq. (M.9). Standard routines exist to find eigenvalues of an order-M tridiagonal matrix in time order-M. (The ideas are straightforward and can be cast in the form of the roots of a function defined by a recursive set of polynomials [444]. See Exercise M.2.) Eigenvectors can easily be found by inverse iteration if the eigenvalues are known. If the original basis contains N vectors, then N steps are required to generate the full hamiltonian in this tridiagonal form.

The importance of the method is that eigenvectors and eigenvalues near the bottom and top of the spectrum can be determined with great accuracy *without generating the entire matrix*. One can understand why the highest and lowest vectors are generated first by noting that if one starts with a trial vector ψ_1 that is a linear combination of eigenvectors, each step of Eq. M.9 is an operation by H that increases the weight of those eigenvectors that are farthest away in energy from the average energy of the trial state. Since the average energy must be somewhere in the middle of the spectrum the states at the edges are projected out with greatest efficiency. In typical cases involving millions of states, a few steps of the algorithm are sufficient to find the ground state with great precision [944].

There are many variations in the way the Lanczos algorithm can be used. For example, the recursive relation (M.9) can be used to generate a matrix of size $M \times M$, which is diagonalized to find the desired state, e.g., the lowest eigenstate. The state thus found can be used as the starting vector for the next iteration, etc., until the desired accuracy is achieved. This is particularly stable [944] and useful in a block form for finding several states. This approach may even be used for M as small as $M = 2$, which makes a particularly efficient algorithm for the ground state that does not require storing many vectors.

The Lanczos algorithm is widely used for a few extreme eigenvalues in extremely large problems such as many-body problems where it is synonymous with "exact diagonalization." The most complete application in large-scale electronic structure calculations has been presented by Wang and Zunger [945] who have incorporated procedures for avoiding spurious states.

The Arnoldi method [854] is a variation of Lanczos that explicitly orthogonalizes each vector ψ_n to all the previous vectors by a Gram–Schmidt procedure. This eliminates the instability and allows calculation of as many eigenvectors as desired. This approach can be applied to non-hermitian matrices.

An elegant consequence of the Lanczos tridiagonal form is a "continued fraction" representation of the spectrum for any dynamical correlation function, expressed as a function of complex frequency z. This leads to the recursion method [696, 828], for which pertinent aspects are summarized in Sec. 23.3.

M.6 Davidson algorithms

Davidson [923,935–937,946] has devised methods that are now widely applied to electronic structure problems. There are a number of variations that cannot be covered here. A primary point is that the Davidson approach is closely related to the Lanczos algorithm, but adapted to be more efficient for problems in which the operator is diagonally dominant. This is often the case in electronic structure problems, e.g. plane wave algorithms.

The flavor of the Davidson methods can be illustrated by defining the diagonal part of the hamiltonian matrix as $D_{mm'} = H_{mm}\delta_{mm'}$ and rewriting the eigenvalue problem $H\psi = \varepsilon\psi$ as

$$(H - D)\psi = (\varepsilon I - D)\psi, \tag{M.11}$$

or

$$\psi = (\varepsilon I - D)^{-1}(H - D)\psi. \tag{M.12}$$

Here I is the unit matrix, inversion of $I - D$ is trivial, and $H - D$ involves only off-diagonal elements. The latter equation is very similar to perturbation theory and suggests iterative procedures that converge rapidly if the diagonal part of the hamiltonian is dominant. An algorithm has been suggested by Lenthe and Pulay [923] that involves three vectors at each step of the iteration.

M.7 Residual minimization in the subspace – RMM–DIIS

The approaches described up to now (and the minimization methods described below) converge to the lowest state with no problems because the ground state is an absolute minimum. In order to find higher states, they must ensure orthogonality, either implicitly as in the Lanczos methods or by explicit orthogonalization. The residual minimization method (RMM) proposed by Pulay [934] avoids this requirement and converges to the state in the spectrum with eigenvalue closest to the trial eigenvalue ε because it minimizes the norm of a "residual vector" instead of the energy. Since the approach of Pulay minimizes the residual in the full Krylov iterative space generated by previous iterations, the method is known as RMM–DIIS for "residual minimization method by direct inversion in the iterative subspace." The general idea is to replace the last equation in (M.5) with

$$\psi^{n+1} = c_0\psi^0 + \sum_{j=1}^{n+1} c_j \delta\psi^j, \tag{M.13}$$

where the entire set of c_j is chosen to minimize the norm of the residual R^{n+1}. (Preconditioning can also be applied at each step [718] to speed the convergence.) The c_j coefficients can be obtained by diagonalizing the hamiltonian in the iterative subspace $\{\psi^0, \psi^1, \psi^2, \ldots, \psi^n\}$, which is a miniscule operation since the number of vectors is at most 10 or so. The time-consuming step is the operation $H\psi$, which is a matrix operation requiring, in general, $O(N_b^2)$ operations for each eigenvector ψ, where N_b is the size of the basis. However, for sparse operations this reduces to $O(N_b)$ for large bases, and to

$O(N_b \ln(N_b))$ if FFTs are used as described in Sec. M.11. In practice, for large problems, it is prohibitive to store many vectors and only small matrices are actually diagonalized corresponding to only a few steps n before restarting the process.

The application of this approach in electronic structure was initiated by Wood and Zunger [929] and used subsequently by many authors in various modifications [718, 930, 947]. The DIIS method involves construction of a full matrix of the size of the subspace, which can be efficient so long as the matrix is small and all the vectors spanning the space can be stored. This can be achieved in solving the Kohn–Sham equations in two regimes. If the number of eigenstates needed is small, then all states can be generated at once using the RMM-DIIS approach. If the number of eigenstates needed is large, then the problem can be broken up into energy ranges, and a few states with eigenvalues nearest the chosen energy can be generated by solving a small matrix equation. In this case, care must be taken not to miss or to overcount eigenstates [718]. A great advantage of this method compared to the conjugate gradient methods of Sec. M.8 is that any eigenvector can be found even in the middle of the spectrum with no explicit need to require orthogonality to the other vectors.

M.8 Solution by minimization of the energy functional

The energy minimization approach has the virtue that it parallels exactly the physical picture of minimizing the total energy and the analytic variational equations (3.10)–(3.12) and (7.8), which is also given below in (M.15). To accomplish the minimization one can utilize the steepest descent (SD) and conjugate gradient (CG) algorithms, which are general minimization methods widely used in numerical analysis [854, 920] and in electronic structure calculations [425, 440, 931–933]. As in all iterative methods, one starts from trial functions ψ_i^0, for the $i = 1, N$ orbitals, and generates improved functions ψ_i^n by n successive iterations. The basic SD and CG algorithms are described in App. L; however, applications in electronic structure require special choices and modifications due to the *constraint of orthonormality* of the functions ψ_i^n. The explicit equations and the sequence of operations in electronic structure calculations are given in Fig. M.1 which is described in this section.

Minimization algorithm with constraints

The analytic variational equations, (3.10)–(3.12) and (7.8), including the constraint of orthonormality follow from the Lagrange multiplier formulation[2] with

$$\mathcal{L} = E[\psi_i] - \sum_{ij} \Lambda_{ij} \left(\int \mathrm{d}\mathbf{r} \psi_i^*(\mathbf{r}) \psi_j(\mathbf{r}) - \delta_{ij} \right), \tag{M.14}$$

where $E[\psi_i]$ is the usual Kohn–Sham expression for energy, Eq. (7.5). The derivative of the Lagrangian gives the steepest descent direction including the constraint

[2] Note the similarity to the lagrangian in the Car–Parrinello method, Ch. 18, where there the "fictitious electronic mass" is also added.

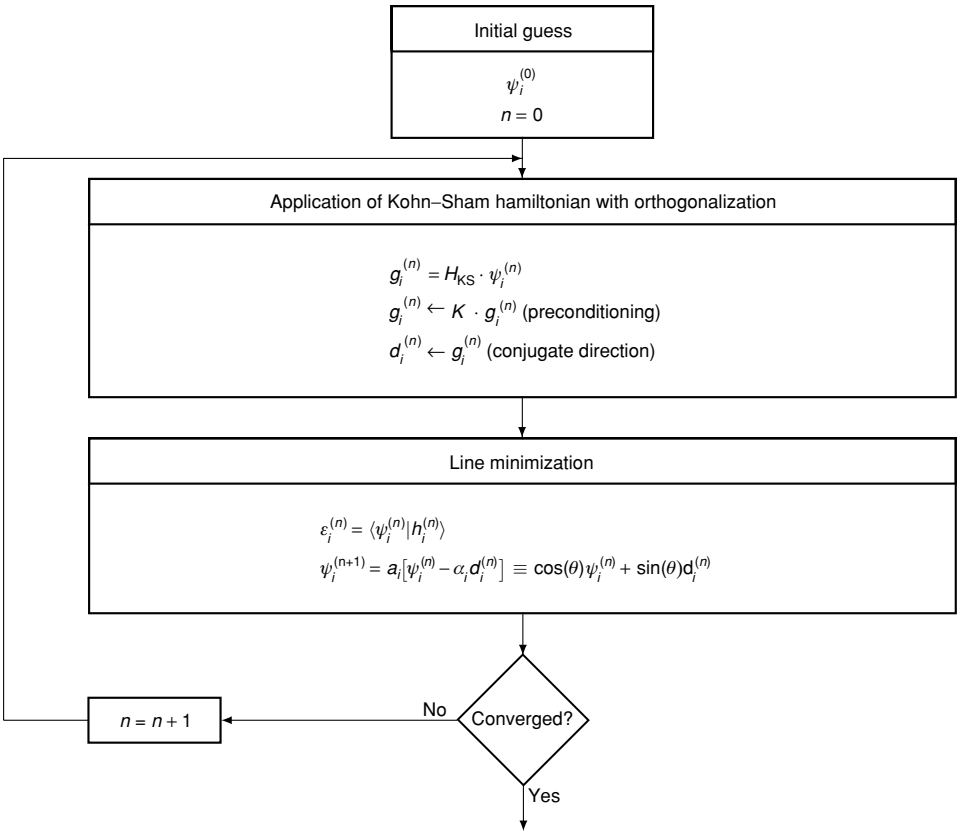


Figure M.1. Iterative loop for solving the non-self-consistent Schrödinger equation in the "band-by-band" conjugate gradient method [440, 931] (or steepest descent where $d_i = g_i$), using the notation of App. L. For unconstrained functionals of non-orthonormal orbitals (Eqs. (M.16) or (23.26)), this is the usual SD or CG method and no other steps are needed. If the constraint of orthonormality is explicitly imposed, then in addition to the steps shown there are orthonormalization operations $g_i^{(n)} \leftarrow g_i^{(n)} \perp \left(\psi_1, \psi_2, \ldots, \psi_{i-1}; \psi_i^{(n)}\right)$; $d_i^{(n)} \leftarrow d_i^{(n)} \perp \left(\psi_1, \psi_2, \ldots, \psi_{i-1}; \psi_i^{(n)}\right)$.

$$\frac{\delta \mathcal{L}}{\delta \psi_i^*} = H_{\mathrm{KS}} \psi_i - \sum_j \Lambda_{ij} \psi_j . \tag{M.15}$$

This is completely sufficient for infinitesimal variations; however, for finite steps in a numerical procedure additional steps must be taken to conserve orthonormality.

In Car–Parrinello MD simulations (Ch. 18), constraints are enforced using a Lagrange multiplier in a way that conserves energy [711]. Minimization methods have the opposite philosophy: the goal is *energy minimization*: to lose energy as efficiently as possible to reach the ground state (which of course must obey the constraints). In SD and CG minimization, the constraint is violated at each step, so that orthonormalization is needed after each of the

intermediate steps. The most common method is the Gram–Schmidt procedure, which produces one of the possible sets of orthonormal vectors.

There are two basic approaches for SD or CG minimization in electronic structure: "band-by-band" [931] and "all-bands" [933]. The former approach diagonalizes the hamiltonian, i.e. all the desired eigenvalues and vectors are found. The latter finds vectors that span the desired subspace, which is sufficient for many purposes; subspace diagonalization can be added if needed. We will describe the steps in the "band-by-band" method; the only change needed for "all bands" is that all desired eigenvectors $i = 1, N_e$ are treated together as a "supervector" of size in each line minimization. The basic strategy is outlined in Fig. M.1 and further details are given in [931] and [440]. The algorithm for finding the direction in Hilbert space for minimization of ψ_i^n, the ith vector at step n, is as follows: calculate the SD gradient $g_i^{(n)} = H_{\rm KS}\psi_i^n$ as in Fig. M.1; orthogonalize $g_i^{(n)}$ to all the previously calculated eigenvectors ψ_j, $j < i$, and to the present vector ψ_i^n; precondition and orthonormalize; find the conjugate direction $d_i^{(n)}$ in the case of CG minimization followed by another orthonormalization.

The next step is line minimization, i.e. $\psi_i^n \to \psi_i^{n+1} = a\psi + b\Delta\psi$ to find the minimum eigenvalue ε_i as a function of a and b. A simple procedure to maintain normalization is to construct the new vector as $\cos(\theta)\psi + \sin(\theta)\Delta\psi$, *and* minimize as a function of θ. Since both ψ and $\Delta\psi$ are orthogonal to the previous vectors by construction, this maintains orthonormalization along the line. Repeating for $i = 1, N_e$, produces the desired set of eigenvectors ψ_i, $i = 1, N_e$.

The CG method is well known to speed convergence greatly in some problems, as discussed in App. L. However, the basic ideas of CG are violated by the constraints or for nonlinear functionals (as is the case for the unconstrained quadratic functionals in Sec. M.8); there is no proof that the new direction is conjugate to all previous directions. Furthermore, the two formulas for the CG direction, Eqs. (L.12) and (L.13), are no longer equivalent and tests must be made to find the most efficient approach in any given problem.[3]

Finally, there is another important choice: when to update the density $n(\mathbf{r})$ and the potential $V_{\rm eff}(\mathbf{r})$ in the Kohn–Sham or any other self-consistent method. The "band-by-band" method has the advantage that one can update during or after the line minimization for each band. If there are many bands, each update of the density is a small perturbation, which can improve convergence (see Sec. 9.3). If the update is done after all bands are completed, then various extrapolation techniques can be used to choose a new $V_{\rm eff}^{n+1}(\mathbf{r})$ given the potential and/or density at previous steps $n, n-1, \ldots$ In addition, there are choices of the way to update the density most effectively during iterations toward self-consistency or when the atoms move [716].

Functionals of non-orthogonal orbitals

One can also construct functionals that do not require orthonormal orbitals so that the SD and CG methods can be used directly. One approach is the CG method of [932], where the

[3] The widely used form in Sec. L.13 does not converge for the simplest case of a fixed hamiltonian; however, it may still be useful if one only needs inaccurate solutions for a given hamiltonian in a self-consistency cycle.

density and energy are defined using well-known expressions in terms of the inverse overlap matrix S_{ij}^{-1}, where $S_{ij}^n = \langle \psi_i^n | \psi_j^n \rangle$ at step n (see also Sec. 23.5). Using the expressions in Sec. 9.2, such as Eq. (9.7) the energy can be written as

$$E_{\rm KS} = \sum_{ij} \langle \psi_i | H_{\rm KS} | \psi_j \rangle S_{ij}^{-1} + G[n]. \tag{M.16}$$

Another approach is described in Ch. 23: to define a modified functional, Eq. (23.26), that can be minimized with no constraint and which equals the Kohn–Sham energy at the minimum. The new functional is closely related to Eq. (M.16), except that the inverse matrix is replaced by $S^{-1} = [\mathbf{1} + (\mathbf{S} - \mathbf{1})]^{-\mathbf{1}} \to [\mathbf{1} - (\mathbf{S} - \mathbf{1})] = \mathbf{2} - \mathbf{S}$. This can be viewed as the first term [844] in the expansion of S^{-1} or as an interpretation [845] of the Lagrange equations, (M.15).

Non-extremal eigenstates

How can one use minimization methods to find states in the middle of a spectrum? The first and simplest approach follows by noting that the eigenfunctions of the "folded" operator $(H - \varepsilon)^2$ are the same as those of H, and the eigenvalues are always positive with absolute minimum for the state with eigenvalue closest to ε. Any minimization method or power method (such as Lanczos) that rapidly converges to extreme states can be used to find the states closest to ε. However, there is a problem with this approach due to poor convergence that is inherent in the use of the "folded" operator $(H - \varepsilon)^2$. The eigenvalue spectrum of the squared operator is compressed $\propto (\varepsilon_i - \varepsilon)^2$ close to the chosen energy ε, making the problem poorly conditioned and leading to difficulties in separating the states in the desired energy range near ε. Nevertheless, there is an important case where the method is effective: the states closest to the gap (the HOMO and LUMO) in a semiconductor or insulator [948] can be found choosing ε in the gap. Since there are no states with very small values of $\varepsilon_i - \varepsilon$, the spectrum has a positive lower bound and there is no essential difficulty. More than one state can be found if orthonormalization is explicitly required or if a "block" method of several states is used.

A much more robust method is the "shift and invert" approach often attributed to Ericsson and Ruhe [949], which is a transformation of the Lanczos method. There is a price to pay for the inversion, but the full inverse is not required – only the operation of the inverse on vectors. The advantage is that the spectrum is spread out near the desired energy ε making it easier to obtain the eigenstates near ε. In fact, if the eigenvalues are separated by $\approx \Delta E$, the separation of the eigenvalues of the shift-invert operator is $\propto \Delta E/(\varepsilon_i - \varepsilon)^2$.

M.9 Comparison/combination of methods: minimization of residual or energy

Perhaps the most extensive comparison to date for different iterative methods has been presented by Kresse and Furthmüller [718] for the CG and RMM–DIIS methods. The "VASP" program created by these authors, which is one of the most widely used pseudopotential codes, uses a combination of CG and RMM–DIIS steps. The methods require very similar

operations except that CG minimization requires explicit orthonormalization of each vector at each step. They report that for large systems, orthonormalization becomes the dominant factor because one vector must be orthonormalized to a large number of other vectors at every single band update. This requires access to memory which then dominates over the cost of the floating point operations. The RMM–DIIS method operates on each vector separately and needs no such orthonormalization. However, for small systems, the cost can be comparable. The main disadvantage of RMM–DIIS is that it always finds the vector closest to the trial vector, so that care must be taken to find all the vectors.

Kresse and Furthmüller [718] have made an algorithm that combines CG and RMM–DIIS methods, applying them sequentially to a set of vectors equal to the number of bands, and performing only a few updates during each iteration. After each iteration the bands are explicitly orthogonalized (only for the RMM–DIIS method where they may be non-orthogonal due to numerical error) and this becomes the input for the next iteration. Many examples of convergence are given in [718] and the method has been applied to many systems.

M.10 Exponential projection in imaginary time

The Schrödinger equation in imaginary time $\tau = it$ is

$$-\frac{\mathrm{d}\psi}{\mathrm{d}\tau} = H\psi, \tag{M.17}$$

which has the formal solution

$$\psi(\tau) = \mathrm{e}^{-H\tau}\psi(0). \tag{M.18}$$

It is straightforward to see that the operation in (M.18) projects out of the ground state as $\tau \to \infty$.

This is a widely used approach in many problems (e.g. many-body quantum Monte Carlo simulations) and it has the conceptual advantage that it is closely related to time-dependent phenomena and to statistical mechanics. It has not been widely applied in solving the Kohn–Sham equations for condensed matter, but has been adapted to calculations on electrons confined to "quantum dot" structures [840].

M.11 Algorithmic complexity: transforms and sparse hamiltonians

All iterative methods replace diagonalization of the hamiltonian matrix by the application of an operator $\hat{A}$, such as the hamiltonian $\hat{H}$ or a function of $\hat{H}$,

$$H_{\mathrm{KS}}\psi_i = \frac{\delta E_{\mathrm{KS}}}{\delta \psi_i^*} \equiv -F_i^e, \tag{M.19}$$

to approximate wavefunctions, where we have omitted spin and space labels. The interpretation as a gradient of the total energy follows from the Kohn–Sham equations, (7.8) and (7.12), which can be considered as a the negative of a generalized "force" on the electrons

$-F_i^e$. The solution at the minimum is that the force be zero, and iterative procedures arrive at this condition in various ways.

In this appendix we will consider plane waves as the primary example for iterative methods. (It is straightforward to translate the arguments and algorithms for other bases, e.g. real-space grids treated in Sec. 12.8.) The explicit form of (M.19) needed for plane waves is given by (using Eq. (12.9))

$$-F_i(\mathbf{G}_m) = \sum_{m'} H_{m,m'}(\mathbf{k}) c_{i,m'}(\mathbf{k}), \tag{M.20}$$

where the variables in the wavefunctions are the $c_{n,m}(\mathbf{k})$ coefficients in the Bloch functions (see Eq. (12.12)),

$$u_{i\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega_{\text{cell}}}} \sum_m c_{i,m}(\mathbf{k}) \exp(i\mathbf{G}_m \cdot \mathbf{r}). \tag{M.21}$$

Here $i = 1, N_e$, where N_e is the number of desired eigenvectors (often the number of filled bands) and $m = 1, N_{\text{PW}}$, where $N_{\text{PW}} = N_b$ is the number of plane waves included in the basis. Applied straightforwardly, however, this does *not* lead to an efficient algorithm for plane waves. The reason is that the matrix operator form for $H_{m,m'}(\mathbf{k})$ in plane waves $\mathbf{G}_m$, $\mathbf{G}_{m'}$ given in Eq. (12.10) is a dense matrix due to the fact that the potential part $V_{\text{eff}}(\mathbf{G}_m - \mathbf{G}_{m'})$ is, in general, non-zero for all $\mathbf{G}_m$, $\mathbf{G}_{m'}$. Multiplication by a full square matrix on each of the N_e eigenvectors requires $N_e N_{\text{PW}}^2$ operations. In addition, there are other operations such as construction of the charge density in real space that require convolutions in Fourier space that involve $O(N_e N_{\text{PW}}^2)$ operations if done by the direct sums over $\mathbf{G}$ vectors as Eq. (12.29).

How can an efficient *sparse* algorithm be created for plane waves? The idea has already been used in Sec. 12.7 to calculate the density from the wavefunctions using fast Fourier transforms (FFTs) and the fact that the density is easily expressed in real space as $n(\mathbf{r}_j) = \sum_{i,\mathbf{k}} |u_{i,\mathbf{k}}(\mathbf{r}_j)|^2$. If each wavefunction is expanded in N_{PW} plane waves, the density requires a larger number of Fourier components $\bar{N}_{\text{PW}}$ (see explanation below). In order to calculate the density each wavefunction is represented by $c_{i,m}(\mathbf{k})$ with $m = 1, N_{\text{PW}}$ non-zero components and the other $\bar{N}_{\text{PW}} - N_{\text{PW}}$ components set equal to zero. This expanded $c_{i,m}(\mathbf{k})$ is then transformed using an FFT to a grid in real space, leading to the Bloch function (M.21) on a grid of $\bar{N}_{\text{grid}} = \bar{N}_{\text{PW}}$ regularly spaced points $\mathbf{r}_j$. The density is then simply the sum of squares of the wavefunctions at each point, as shown in Fig. 12.4.

Now consider the operation $H\psi_i$ needed in Eq. (M.20) (and the corresponding equations (18.14), (M.2), or (M.15)). Multiplication of the kinetic energy term is very simple since the kinetic energy part of $H_{m,m'}(\mathbf{k})$ given by (12.10) is a diagonal matrix in Fourier space. On the other hand, multiplication by V is simple in real space, where V is diagonal. The operations are done by FFTs as shown in the sequence of steps in Fig. M.2 very much like the operations for the charge density. The FFTs are done on an expanded grid with $\bar{N}_{\text{PW}}$ points and the new wavefunction is truncated to the original size N_{PW} when the results are collected in Fourier space. This procedure can be used in any of the iterative plane wave methods described here as well as in the Car–Parrinello method of Ch. 18.

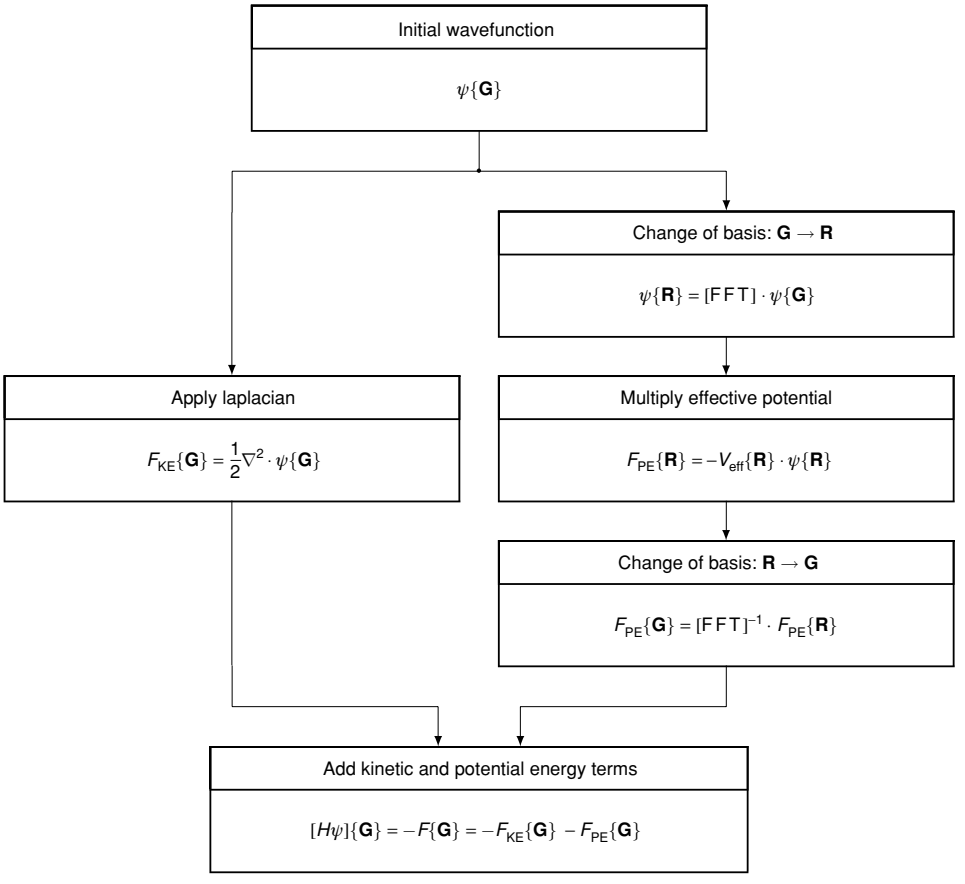


Figure M.2. Schematic representation describing the application of the hamiltonian using Fourier transforms (FFTs). The "force" in Eq. (M.19) is denoted by F_i. The operations are diagonal in each space respectively as long as the potential is local; non-local pseudopotentials require generalization to a non-local expression on the grid points in real space.

Clearly, this approach can be applied to any operator A involving the hamiltonian. In general, additional applications of the FFT will be needed; however, this need not be a major increase in complexity. In particular, powers of H may be treated by repeated application of the FFT.

Aliasing and the FFT transforms

When is the FFT operation exact? Clearly, the Fourier analysis is a mathematical identity if done with an infinite number of plane waves. But the question is: what is required for the FFT operations to give the exact answers for a given finite basis of plane waves? One of the great advantages of the plane wave method is that it truly is a basis, i.e. it is variational and the energy always decreases as more plane waves are added. We do not want to add some

uncontrolled approximation that would destroy this property. See Exercise M.3 for further discussion.

For the density, it is easy to see the required conditions. If the wavefunction is limited to Fourier components with $|\mathbf{G}| < |\mathbf{G}_{\max}|$, then the density can have components up to $|2\mathbf{G}_{\max}|$. If the box for the FFT is defined to be *greater than twice as large as* $|\mathbf{G}_{\max}|$ *in all directions*, then every Fourier component will be calculated exactly. Note that in three dimensions, this means a box of size *at least* as large as $2^3 = 8$ times larger than the smallest box that contains the sphere of $\mathbf{G}$ vectors, i.e. N^*_{PW} is larger than N_{PW} by at least a factor of $8\pi/3 = 8.4$, *roughly an order of magnitude*. Despite this fact, it is still much more efficient to carry out the operations using the FFT for all but the smallest problems.

Note that the estimate for the size of the FFT box depends upon the assumption that the problem is roughly isotropic so that the $\mathbf{G}$ vectors are defined in a cube. If one chooses non-orthorhombic primitive vectors of the reciprocal lattice, then the number of $\mathbf{G}$ vectors will be larger than the above estimate in order to circumscribe a sphere of radius $|2\mathbf{G}_{\max}|$. Fortunately, for large systems, where the methods are most useful, the cell can usually be chosen so that the FFT operations are efficient.

The condition for multiplication of the potential times the wavefunction does not appear so obvious at first sight. There is no reason to suppose that $V(\mathbf{G})$ has a limited number of Fourier components; the ionic potential has a $1/G^2$ form which is reduced by screening but not to zero. The Hartree potential has exactly the same range as the density due to the Poisson equation; however, there is no such limitation on the exchange–correlation potential (more on this below). Thus $V\psi$ extends to all $\mathbf{G}$ vectors even if ψ is limited. *Nevertheless, the range up to* $|2\mathbf{G}_{\max}|$ *is sufficient for an exact calculation.* The reason is that only the Fourier components of $V\psi$ with $|\mathbf{G}| < |\mathbf{G}_{\max}|$ are relevant for the Schrödinger equation. This is easily seen from the definition of the matrix elements of the potential, Eq. (12.8), which involves only components of V up to $|2\mathbf{G}_{\max}|$ if the wavefunctions extend up to $|\mathbf{G}_{\max}|$. In an iterative approach, the potential enters by explicit multiplication of V times a trial vector; even though multiplication would give Fourier components with $|\mathbf{G}| > |\mathbf{G}_{\max}|$, *only those with* $|\mathbf{G}| < |\mathbf{G}_{\max}|$ *are relevant.* Even if the higher Fourier components are calculated, the contribution to the wavefunction is explicitly omitted. *In fact, it is essential that such components be omitted; otherwise one violates the original statement of the problem: the solution of the Schrödinger equation with wavefunctions expanded in a fixed finite basis set.*

The algorithm shown in Fig. M.2 denotes the operations on a wavefunction $\psi(\mathbf{G})$ defined on a set of N_{PW} Fourier components. The algorithm, in fact, generates the product $V\psi$ on a large grid of size N^*_{PW} and the product is explicitly truncated to produce the "force" $F(\mathbf{G})$ defined on the small set of N_{PW} Fourier components. This force is then used to update the wavefunction in any of the iterative methods described in this chapter.

A few words are in order regarding the exchange–correlation energy and potential. There is no simple relation of the reciprocal space and real-space formulations since $\epsilon_{\mathrm{xc}}(n)$ is a non-linear function of n. For example, the fact that exchange involves $n^{1/3}$ means that a single Fourier component of $n(\mathbf{G})$ gives rise to an infinite set of components of $\epsilon_{\mathrm{xc}}(\mathbf{G})$ and $V_{\mathrm{xc}}(\mathbf{G})$. Thus the problem is in $\mathbf{G}$ space formulation: direct sums in $\mathbf{G}$ space can

never give exact $\epsilon_{xc}(\mathbf{G})$ and $V_{xc}(\mathbf{G})$ in terms of $n(\mathbf{G})$. However, FFT formulation allows the exchange–correlation terms to be treated in real space with no problem. So long as one includes all components up to $|2\mathbf{G}_{max}|$, the resulting $\epsilon_{xc}(\mathbf{G})$ and $V_{xc}(\mathbf{G})$ can be used to *define* those terms in a way that is sufficiently accurate for the solution of Kohn–Sham equations (Exercise M.3).

SELECT FURTHER READING

Bylander, D. M., Kleinman, L. and Lee, S., "Self-consistent calculations of the energy bands and bonding properties of $B_{12}C_3$," *Phys. Rev. B* 42:1394–1403, 1990.

Davidson, E. R., "The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices." *J. Comp. Phys.* 17:87, 1975.

Davidson, E. R., "Monster matrices: their eigenvalues and eigenvectors," *Computers in Phys.* 7:519, 1993.

Gillan, M. J., "Calculation of the vacancy formation energy in aluminum," *J. Phys.: Condens. Matter* 1:689, 1989.

Kresse, G. and Furthmüller, J., "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," *Phys. Rev. B* 54:11169–11186, 1996.

Martins, J. L. and Cohen, M. L., "Diagonalization of large matrices in pseudopotential band-structure calculations: Dual-space formalism," *Phys. Rev. B* 37:6134–6138, 1988.

Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A. and Joannopoulos, J. D., "Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients," *Rev. Mod. Phys.* 64:1045–1097, 1992.

Pulay, P., "Convergence acceleration of iterative sequences, the case of SCF iteration," *Chem. Phys. Lett.* 73:393–397, 1980.

Stich, I., Car, R., Parrinello, M. and Baroni, S., "Conjugate gradient minimization of the energy functional: A new method for electronic structure calculation," *Phys. Rev. B* 39:4997, 1989.

Teter, M. P., Payne, M. C. and Allan, D. C., "Solution of Schrödinger's equation for large systems," *Phys. Rev. B* 40:12255–12263, 1989.

Wood, D. M. and Zunger, A., "A new method for diagonalizing large matrices," *J. Phys. A* 18:1343–1359, 1985.

Computational physics books:

Koonin, S. E. and Meredith, D. C., *Computational Physics*, Addison Wesley, Menlo Park, CA, 1990.

Thijssen, J. M., *Computational Physics*, Cambridge University Press, Cambridge, England, 2000.

Numerical analysis:

Booten, A. and van der Vorst, H., "Cracking large scale eigenvalue problems, part I: Algorithms," *Comp. in Phys.* 10:239–242, 1996.

Booten, A. and van der Vorst, H., "Cracking large scale eigenvalue problems, part II: Implementations," *Comp. in Phys.* 10:331–334, 1996. [941]

Golub, G. H. and Van Loan, C. F., *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 1980.

Heath, M. T., *Scientific Computing: An Introductory Survey*, McGraw-Hill, New York, 1997.

Parlett, B. N., *The Symmetric Eigenvalue Problem*, Prentice Hall, Engelwood Cliffs, N. J., 1980.

Exercises

M.1 Show by induction that each vector ψ_n generated by the Lanczos algorithm is orthogonal to *all* the other vectors, and that the hamiltonian has tridiagonal form, Eq. (M.10). Regarding the problem that orthogonality is guaranteed only for infinite numerical precision, show how errors in each step can accumulate in the deviations from orthogonality.

M.2 The solution for the eigenvalues of the tridiagonal matrix H in Eq. (M.10) is given by $|H_{ij} - \lambda\delta_{ij}| = 0$, which is polynomial $P_M(\lambda)$ of degree M. This may be solved in a recursive manner starting with the subdeterminant with $M = 1$. The first two polynomials are $P_1(\lambda) = \alpha_1 - \lambda$ and $P_2(\lambda) = (\alpha_2 - \lambda)P_1(\lambda) - \beta_2^2$. Show that the general relation for higher polynomials is

$$P_n(\lambda) = (\alpha_n - \lambda)P_{n-1}(\lambda) - \beta_n^2[P_{n-2}(\lambda)], \tag{M.22}$$

and thus that the solution can be found by root tracing (varying λ successively to reach condition $P_M(\lambda) = 0$ in computer time proportional to M for each eigenvalue.

M.3 Consider a plane wave calculation with the wavefunction limited to Fourier components with $|\mathbf{G}| < |\mathbf{G}_{\max}|$. Show that all Fourier components of the external potential and the Hartree potential are given exactly (with no "aliasing") by the FFT algorithm, so long as the FFT extends to $|2\mathbf{G}_{\max}|$. For the non-linear exchange–correlation potential, show that there is no exact expression and that the expressions are "reasonable."

Appendix N

Code for empirical pseudopotential and tight-binding

The "TBPW" code is a modular code for Slater–Koster orthogonal tight-binding (TB, Sec. 14.4) and plane wave empirical pseudopotential (EPM, Sec. 12.6) calculations of electron energies in finite systems or bands in crystals. This appendix is a schematic description. A much more complete description, sample files, and the full source codes are available on-line at the site given in Ch. 24.

TBPW is meant to be an informative, instructional code that can bring out much of the physics of electronic structure. It has many of the main features of full density functional codes, but is much simpler. A main characteristic is that the codes are modular and are organized to separate the features common to all electronic structure calculations from the aspects that are specialized to a given method.

A sample input file is given below in Sec. N.4. The input file uses keywords that are recognized by the input routines, so that the same file can be used for either TB or PW calculations.

N.1 Calculations of eigenstates: modules common to all methods

The functional features are listed below with the corresponding inputs in square brackets [ ].

- Crystal structure:
  - Space: number of space dimensions [dimension = 1, 2, 3, . . .]
  - Lattice: real-space cell [translation vectors]
  - Atomic basis: [positions, types of atoms]
- kpoints: points in BZ for calculation [list of points, specification of lines in BZ, or "special points" (Sec. 4.6)]
- Hamiltonian: generated by specialized modules for TB or PW
- Diagonalization: calls to standard dense matrix routines to diagonalize hamiltonians generated by either PW or TB methods (for PW there is an option for the conjugate gradient method of App. M) [choice of method]
- Plotting resulting bands: [information for plots, if desired]

N.2 Plane wave empirical pseudopotential method (EPM)

The EPM calculation requires information on the potential to set up the hamiltonian matrix. Local empirical potentials for Si, Ga, and As [532] are included. Options are given for a

user to create new potentials. (See exercises 12.11 and 12.12 for examples of problems that illustrate the calculations and the results.)

Features of the code special to the EPM calculations are:

- Basis: list of plane waves included [cutoff in energy for plane waves]
- Hamiltonian: Fourier components of potential [list of values or choice available in subroutines]
- Diagonalization method: standard dense matrix routines or conjugate gradient method (App. M) [choice of method]
- Charge density in real space [choice of real-space points, lines, or planes]

N.3 Slater–Koster tight-binding (TB) method

The tight-binding (TB) calculations require procedures to find all the relevant neighbors and hamiltonian matrix elements. Basis states can have arbitrary angular momentum using the rotation operator methods described in Sec. N.5. The Slater–Koster parameters can be read in from a file or calculated using Harrison's "universal" parameters [344,590]. The user can specifiy arbitrary interactions (not necessarily two-center), which requires changing one subroutine.

Features of the code special to the TB calculations are:

- Basis: orbitals included for each atom [atom type, angular momentum]
- Neighbor list: general program to find neighbors in any cluster or crystal [cutoff radius for neighbor distance]
- Hamiltonian: choice of Slater–Koster parameters [read in from a file or specify choice available]

N.4 Sample input file for TBPW

Input is specified by keywords defined in the on-line manual. (The verbose keywords are meant to be self-explanatory.) The lines can be in any order and comments can be added simply by inserting lines. Some comments are marked with *** for visibility, but this is not essential.

```
*** Information for PW or TB calculation of Si

*** Information for lattice
NumberOfDimensions 3
LatticeConstant 10.2612170
LatticeVectors
0.0  0.5  0.5
0.5  0.0  0.5
0.5  0.5  0.0
```

```
*** Information for atomic positions
NumberOfAtoms 2
NumberOfSpecies 1
ChemicalSpeciesLabel
1  14  Si
(Chemical Species label (first number) assumed to be in
sequential order)

AtomicCoordinatesFormat ScaledByLatticeVectors
AtomicCoordinatesAndAtomicSpecies
-0.125  -0.125  -0.125  1
 0.125   0.125   0.125  1

*** Information for k points and number of bands to plot
(This example is a set of k points along lines in the fcc BZ
for plotting band structure. This is used for the free electron
bands and Si and Ni bands in Chapters 12 and 14.)
NumberOfBands 8
NumberOfLines 5
NumberOfDivisions 15
KPointsScale ReciprocalLatticeVectors
KPointsAndLabels
0.0    0.0    0.0      Ga
0.375  0.375  0.75     K
0.5    0.5    0.5      L
0.0    0.0    0.0      Ga
0.0    0.5    0.5      X
0.25   0.625  0.625    U

*** Information for Plane Wave Calculation (atomic units
assumed)
EnergyCutOff 6.0

*** Information for Tight Binding Calculation
MaximumDistance 5.5
EnergiesInEV
TightBindingModelType 1 (Harrison Model)
OrbitsAndEnergies

4
0  0  -13.55
1  1  -6.52
1  2  -6.52
1  3  -6.52
```

N.5 Two-center matrix elements: expressions for arbitrary angular momentum *l*

Two-center matrix elements for any particular angular momenta can be worked out [950], with increasing effort for increasing L. Is it possible to make an algorithm that works for

any angular momenta? By using rotation operator algebra, the rotations (analogous to those shown explicitly for p states in Fig. 14.2) can be generated to define the quantization axis for the orbitals along the line between the atoms.[1] The general formulation is most easily cast in terms of the complex orbitals that are eigenfunctions of L_z, where the z-axis is the same for all orbitals. (It is straightforward at the end to convert back to real orbitals.) Thus the two orbitals involved in any matrix element are l, m and l', m'. These orbitals must be written in a representation quantized along the z'-axis, which is parallel to the direction $\hat{\mathbf{R}}$. The transformation is applied to the set of orbitals $-l \leq m \leq l$ for a given l, since the transformation preserves the angular momentum l but the different m components are mixed. Let the set of $2l + 1$ states for a given l be denoted by $|l\{m\}\rangle$. The rotation is a unitary transformation given by [951]

$$|l\{m'\}\rangle = \mathrm{e}^{-\mathrm{i}\theta \hat{L}_y} \mathrm{e}^{-\mathrm{i}\phi \hat{L}_z} |l\{m\}\rangle, \tag{N.1}$$

where the rotation angles θ and ϕ are defined by

$$\hat{\mathbf{R}} = \sin\theta(\hat{\mathbf{x}}\cos\phi + \hat{\mathbf{y}}\sin\phi) + \hat{\mathbf{z}}\cos\theta. \tag{N.2}$$

The two exponential operators in (N.1) rotate the quantization axis first about the z-axis, and then about the new y'-axis to define the quantization axis along z'. Then the matrix elements for the $(2l + 1) \times (2l' + 1)$ block of the K matrix corresponding to l and l' can be written

$$K_{l\{m\},l'\{m'\}} = \langle l\{m\}|\mathrm{e}^{\mathrm{i}\phi \hat{L}_z}\, \mathrm{e}^{\mathrm{i}\theta \hat{L}_y}\, \hat{K}\, \mathrm{e}^{-\mathrm{i}\theta \hat{L}_y} \mathrm{e}^{-\mathrm{i}\phi \hat{L}_z} |l'\{m'\}\rangle, \tag{N.3}$$

where the right-hand side is expressed in terms of the operator $\hat{K}$ (e.g. the overlap) which is diagonal in the azimuthal quantum number m defined about the z'-axis.

The operations involving $\hat{L}_z$ are straightforward; the states are eigenfunctions of $\hat{L}_z$ so that $\mathrm{e}^{-\mathrm{i}\phi \hat{L}_z}|l, m\rangle = \mathrm{e}^{-\mathrm{i}m\phi}|l, m\rangle$, which is diagonal in the set $\{m\}$. However, $\hat{L}_y$ is more difficult. The matrix elements of $\hat{L}_y$ are well known [951]

$$\begin{aligned}\langle l, m|\hat{L}_y|l', m'\rangle &= \frac{1}{2i}\delta_{ll'} \\ &\times \left[\sqrt{l(l+1) - m'(m'+1)}\,\delta_{m,m'+1} - \sqrt{l(l+1) - m'(m'-1)}\,\delta_{m,m'-1}\right],\end{aligned} \tag{N.4}$$

but there is still a difficulty since this non-diagonal operator appears in an exponential. This can be solved by diagonalizing the matrix, Eq. (N.4), for the $\hat{L}_y$ operator in the basis of eigenfunctions $|l, m\rangle$ of $\hat{L}_z$ for each l. Standard numerical routines can be used to find the eigenvalues and eigenvectors so that the $\hat{L}_y$ operator can be written as

$$\hat{L}_y = M_y L_z M_y^{\dagger}, \tag{N.5}$$

where M_y is a matrix whose columns are the eigenstates of the $\hat{L}_y$ operator written in $\hat{L}_z$ basis. Using the identity $\mathrm{e}^{VAV^{\dagger}} = V\mathrm{e}^{A}V^{\dagger}$, where V is unitary, the resulting expression for

[1] This formulation is due to N. Romero and T. Arias.

the matrix elements takes the form

$$K_{l\{m\},l'\{m'\}} = \langle l\{m\}|\mathrm{e}^{\mathrm{i}\phi\hat{L}_z} M_y\, \mathrm{e}^{\mathrm{i}\theta\hat{L}_z} M_y^\dagger\, \hat{K}\, M_y \mathrm{e}^{-\mathrm{i}\theta\hat{L}_y} M_y^\dagger \mathrm{e}^{-\mathrm{i}\phi\hat{L}_z}|l\{m'\}\rangle. \tag{N.6}$$

Finally, it is a small step to transform $K_{l\{m\},l'\{m'\}}$ to a representation with real orbitals $S^{\pm}_{l,m}$ that are combinations of $\pm m$ and $\pm m'$ give in (K.11). A more detailed description of the method and computer codes for tight-binding calculations that use this algorithm are available online at the site in Ch. 24. The codes can treat more than one orbital per angular momentum channel and are used for the calculation of bands shown in Figs. 14.6 and 14.7.

Appendix 0

Units and conversion factors

Quantity	Symbol	Hartree atomic units	Conventional units
Electron mass	m_e	1	$9.109{,}381{,}88(72) \times 10^{-31}$ kg
Electron charge	e	1	$1.602{,}176{,}462(63) \times 10^{-19}$ C
Planck constant/(2π)	$\hbar$	1	$1.054{,}571{,}596(82) \times 10^{-34}$ J s
Speed of light	c	137.036,000	299,792,458 m/s
Bohr radius	$a_0 = \frac{\hbar^2}{m_e e^2}$	1	$0.529{,}2083(19) \times 10^{-10}$ m
Hartree	$Ha = e^2/a_0$	1	27.211,3834(11) eV
Rydberg	$\text{Ryd} = \frac{1}{2}e^2/a_0$	0.5	13.605,6917(6) eV
Electron volt	eV	0.036,749,3260	$1.602{,}176{,}462(63) \times 10^{-19}$ J
Proton–electron mass ratio	m_p/m_e	1,836.152,6675(39)	

Other conversion factors			
1 eV =	23.06 Kcal/mol	8.0685×10^3 cm^{-1}	1.1604×10^4 K
1 GPa =	10 kbar	6.241 meV/Å^3	3.399×10^{-5} Hartree/a_0^3
1 Hartree/a_0^3 =	2.9418×10^4 GPa	294.18 Mbar	

Source: P. J. Mohr and B. N. Taylor, "CODATA recommended values of the fundamental physical constants: 1998," *Rev. Mod. Phys.* 72 (2000), 351.
The latest CODATA values available can be found at http://physics.nist.gov/Constants.

References

[1] F. Seitz, *The Modern Theory of Solids*, McGraw-Hill Book Company, New York, 1940, reprinted in paperback by Dover Press, New York, 1987.

[2] J. C. Slater, *Quantum Theory of Atomic Structure, Vol. 1–4*, McGraw-Hill, New York, 1960–1972.

[3] H. A. Lorentz, *Theory of Electrons [Reprint of volume of lectures given at Columbia University in 1906]*, Dover, New York, 1952.

[4] P. Zeeman, "The effect of magnetisation on the nature of light emitted by a substance (Translated by Arthur Stanton from the Proceedings of the Physical Society of Berlin.)," *Nature* 55:347, 1897.

[5] J. J. Thomson, "Cathode rays," *Phil. Mag., Series 5* 44:310–312, 1897.

[6] J. J. Thomson, "Cathode rays," *The Electrician: a weekly illustrated journal of Electrical Engineering, Industry and Science* 39, 1897.

[7] E. Rutherford, "The scattering of α and β particles by matter and the structure of the atom," *Phil. Mag., Series 6* 21:669–688, 1911.

[8] N. Bohr, "On the constitution of atoms and molecules," *Phil. Mag., Series 6* 26:1–25, 1913.

[9] M. Jammer, *The Conceptual Development of Quantum Mechanics,* McGraw-Hill, New York, 1966.

[10] *Sources of Quantum Mechanics*, edited by B. L. van de Waerden, North Holland, Amsterdam, 1967.

[11] A. Messiah, *Quantum Mechanics, Vol. I*, Wiley, New York, 1964.

[12] L. Hoddeson and G. Baym, "The development of the quantum-mechanical electron theory of metals: 1900–1928," *Proc. Roy. Soc. A* 371:8, 1987.

[13] L. Hoddeson and G. Baym, "The development of the quantum-mechanical electron theory of metals: 1928–1933," *Rev. Mod. Phys.* 59:287, 1987.

[14] L. Hoddeson, E. Braun, J. Teichmann, and S. Weart, *Out of the Crystal Maze [Chapters for the History of Solid State Physics]*, Oxford University Press, New York, Oxford, 1992.

[15] O. Stern, "Ein Weg zur experimentellen Prüfung der Richtungsquantelung im Magnetfeld (Experiment to test the applicability of the quantum theory to the magnetic field)," *Z. Physik* 7:249–253, 1921.

[16] W. Gerlach and O. Stern, "Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld (Experimental test of the applicability of the quantum theory to the magnetic field)," *Z. Physik* 9:349–352, 1922.

[17] A. H. Compton, "Possible magnetic polarity of free electrons: Estimate of the field strength of the electron," *Z. Phys.* 35:618–625, 1926.

[18] S. A. Goudschmidt and G. H. Uhlenbeck, "Die Kopplungsmöglichkeiten der Quantenvektoren im Atom," *Z. Phys.* 35:618–625, 1926.
[19] W. Pauli, "Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplex Struktur der Spektren," *Z. Phys.* 31:765, 1925.
[20] E. C. Stoner, "The distribution of electrons among atomic levels," *Phil. Mag.* 48:719, 1924.
[21] E. Fermi, "Zur Quantelung des Idealen Einatomigen Gases," *Z. Phys.* 36:902, 1926.
[22] S. N. Bose, "Plancks Gesetz und Lichtquanten-hypothese," *Z. Phys.* 26:178, 1924.
[23] A. Einstein, "Quantheorie des Idealen Einatomigen Gases," *Sber. preuss Akad. Wiss.* p. 261, 1924.
[24] W. Heisenberg, "Mehrkörperproblem und Resonanz in der Quantenmechanik," *Z. Phys.* 38:411, 1926.
[25] P. A. M. Dirac, "On the theory of quantum mechanics," *Proc. Roy. Soc. London* Ser. A 112:661, 1926.
[26] J. C. Slater, "The theory of complex spectra," *Phys. Rev.* 34:1293, 1929.
[27] P. A. M. Dirac, "The quantum theory of the electron," *Proc. Roy. Soc. London* Ser. A 117:610, 1928.
[28] G. N. Lewis, "The atom and the molecule," *J. Am. Chem. Soc.* 38:762–786, 1916.
[29] W. Heitler and F. London, "Wechselwirkung neutraler Atome und homopolare Bindung nach der Quantenmechanik," *Z. Phys.* 44:455, 1927.
[30] W. Pauli, "Über Gasentartung und Paramagnetismus," *Z. Phys.* 41:91, 1927.
[31] A. Sommerfeld, "Zur Elektronen Theorie der Metalle auf Grund der Fermischen Statistik," *Z. Phys.* 47:43, 1928.
[32] P. Drude, "Bestimmung optischer Konstanten der Metalle," *Wied. Ann.* 39:481–554, 1897.
[33] P. Drude, *Lehrbuch der Optik (Textbook on Optics),* S. Hirzel, Leipzig, 1906.
[34] G. E. Kimball, "The electronic structure of diamond," *J. Chem. Phys.* 3:560, 1935.
[35] H. Bethe, "Theorie der Beugung von Elektronen in Kristallen," *Ann. Phys. (Leipzig)* 87:55, 1928.
[36] F. Bloch, "Über die Quantenmechanik der Elektronen in Kristallgittern," *Z. Phys.* 52:555, 1928.
[37] R. E. Peierls, "Zur Theorie der galvanomagnetischen Effekte," *Z. Phys.* 53:255, 1929.
[38] R. E. Peierls, "Zur Theorie der electrischen und thermischen Leitfähigkeit von Metallen," *Ann. Phys. (Leipzig)* 4:121, 1930.
[39] A. H. Wilson, "The theory of electronic semiconductors," *Proc. Roy. Soc. London* Ser. A 133:458, 1931.
[40] A. H. Wilson, "The theory of electronic semiconductors – II," *Proc. Roy. Soc. London* Ser. A 134:277, 1931.
[41] J. C. Slater, *Solid-State and Molecular Theory: A Scientific Biography,* John Wiley and Sons, New York, 1975.
[42] D. R. Hartree, *The Calculation of Atomic Structures,* John Wiley and Sons, New York, 1957.
[43] D. R. Hartree, "The wave mechanics of an atom with non-coulombic central field: parts I, II, III," *Proc. Cambridge Phil. Soc.* 24:89,111,426, 1928.
[44] E. Hylleraas, "Neue Berechtnung der Energie des Heeliums im Grundzustande, sowie tiefsten Terms von Ortho-Helium," *Z. Phys.* 54:347, 1929.
[45] E. A. Hylleraas, "Über den Grundterm der Zweielektronenprobleme von H^-, He, Li^+, Be^+ usw," *Z. Phys.* 65:209, 1930.
[46] V. Fock, "Näherungsmethode zur Losung des quanten-mechanischen Mehrkörperprobleme," *Z. Phys.* 61:126, 1930.

[47] A. Sommerfeld and H. Bethe, "Elektronentheorie der Metalle," *Handbuch der Physik* 24/2:333, 1933.
[48] J. C. Slater, "The electronic structure of metals," *Rev. Mod. Phys.* 6:209–280, 1934.
[49] E. P. Wigner and F. Seitz, "On the constitution of metallic sodium," *Phys. Rev.* 43:804, 1933.
[50] E. P. Wigner and F. Seitz, "On the constitution of metallic sodium II," *Phys. Rev.* 46:509, 1934.
[51] J. C. Slater, "Electronic energy bands in metals," *Phys. Rev.* 45:794–801, 1934.
[52] H. M. Krutter, "Energy bands in copper," *Phys. Rev.* 48:664, 1935.
[53] W. Shockley, "Electronic energy bands in sodium chloride," *Phys. Rev.* 50:754–759, 1936.
[54] J. C. Slater, "Wavefunction in a periodic potential," *Phys. Rev.* 51:846–851, 1937.
[55] J. C. Slater, "An augmented plane wave method for the periodic potential problem," *Phys. Rev.* 92:603–608, 1953.
[56] M. M. Saffren and J. C. Slater, "An augmented plane wave method for the periodic potential problem. II," *Phys. Rev.* 92:1126, 1953.
[57] W. C. Herring, "A new method for calculating wave functions in crystals," *Phys. Rev.* 57:1169, 1940.
[58] E. Fermi, "Displacement by pressure of the high lines of the spectral series," *Nuovo Cimento* 11:157, 1934.
[59] H. Hellmann, "A new approximation method in the problem of many electrons," *J. Chem. Phys.* 3:61, 1935.
[60] H. Hellmann, "Metallic binding according to the combined approximation procedure," *J. Chem. Phys.* 4:324, 1936.
[61] F. Herman and J. Callaway, "Electronic structure of the germanium crystal," *Phys. Rev.* 89:518–519, 1953.
[62] F. Herman, "Theoretical investigation of the electronic energy band structure of solids," *Rev. Mod. Phys.* 30:102, 1958.
[63] F. Herman, "Elephants and mahouts – early days in semiconductor physics," *Phys. Today* June, 1984:56, 1984.
[64] W. Heisenberg, "Zur Theorie des Ferromagnetismus," *Z. Phys.* 49:619, 1928.
[65] P. A. M. Dirac, "Quantum mechanics of many-electron systems," *Proc. Roy. Soc. London* Ser. A 123:714–733, 1929.
[66] N. Bohr, Studier over Metallernes Elektrontheori (thesis), 1911.
[67] H. J. van Leeuwen, Vraagstukken uit de Electrontheorie van het Magnetisme (thesis), 1911.
[68] H. J. van Leeuwen, "Problemes de la Theorie Electronique du Magnetisme," *J. Phys. Radium* 6:361, 1921.
[69] L. Pauling, *The Nature of the Chemical Bond, Third Edition*, Cornell University Press, Ithaca, N. Y., 1960.
[70] E. P. Wigner, "On the interaction of electrons in metals," *Phys. Rev.* 46:1002–1011, 1934.
[71] N. F. Mott and R. Peierls, "Discussion of the paper by De Boer and Verwey," *Proc. Phys. Soc. London*, Ser. A 49:72, 1937.
[72] N. F. Mott, "The basis of the theory of electron metals, with special reference to the transition metals," *Proc. Phys. Soc. London*, Ser. A 62:416, 1949.
[73] N. F. Mott, *Metal-Insulator Transitions*, Taylor and Francis, London/Philadelphia, 1990.
[74] F. Hund, "Zur Deutung verwickelter Spektren, insbesondere der Elemente Scandium bis Nickel," *Z. Physik* 33:345, 1925.
[75] F. Hund, "Zur Deutung verwickelter Spektren.II," *Z. Physik* 34:296, 1925.
[76] F. Hund, *Linienspektren und periodisches System der Elemente*, Springer-Verlag, Berlin, 1927.

[77] P. W. Anderson, *Basic Notions of Condensed Matter Physics*, Addison-Wesley, Reading, Mass., 1984.
[78] P. W. Anderson, "More is different," *Science* 177:393–396, 1972.
[79] *More is Different: Fifty Years of Condensed Matter Physics*, edited by N.-P. Ong and R. Bhatt, Princeton University Press, Princeton, N. J., 2001.
[80] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, "Microscopic theory of superconductivity," *Phys. Rev.* 106:162–164, 1957.
[81] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, "Quantum Monte Carlo simulations of solids," *Rev. Mod. Phys.* 73:33–83, 2001.
[82] W. G. Aulbur, L. Jonsson, and J. W. Wilkins, "Quasiparticle calculations in solids," *Solid State Physics*, 54:1–218, 2000.
[83] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, "Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions," *Rev. Mod. Phys.* 68:13–125, 1996.
[84] N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, W.B. Saunders Company, Philadelphia, 1976.
[85] H. Ibach and H. Luth, *Solid State Physics An Introduction to Theory and Experiment*, Springer-Verlag, Berlin, 1991.
[86] C. Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, New York, 1996.
[87] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*, Cambridge University Press, Cambridge, U. K., 1995.
[88] M. Marder, *Condensed Matter Physics*, John Wiley and Sons, New York, 2000.
[89] M. Born and J. R. Oppenheimer, "Zur Quantentheorie der Molekeln," *Ann. Physik* 84:457, 1927.
[90] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Oxford University Press, Oxford, 1954.
[91] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.* 136:B864–871, 1964.
[92] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.* 140:A1133–1138, 1965.
[93] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989.
[94] M. H. Kalos and Paula Whitlock, *Monte Carlo Methods*, John Wiley and Sons, New York, 1986.
[95] B. L. Hammond, W. A. Lester, Jr., and P. J. Reynolds, *Monte Carlo Methods in ab initio Quantum Chemistry*, World Scientific, Singapore, 1994.
[96] G. D. Mahan, *Many-Particle Physics, 3rd Ed.*, Kluwer Academic/Plenum Publishers, New York, 2000.
[97] J. M. Zuo, J. C. H. Spence, and M. O'Keeffe, "Bonding in GaAs," *Phys. Rev. Lett.* 61:353–356, 1988.
[98] F. Franks, *Water: A Comprehensive Treatise, Vol. 1*, Plenum, New York, 1972.
[99] P. Ball, *H_2O: A Biography of Water*, Weidenfeld and Nicholson, London, 1999.
[100] J. M. Zuo, P. Blaha, and K. Schwarz, "The theoretical charge density of silicon: Experimental testing of exchange and correlation potentials," *J. Phys. Condens Matter* 9:7541–7561, 1997.
[101] P. Coppens, *X-ray Charge Densities and Chemical Bonding*, Oxford University Press, Oxford, 1997.

[102] Z. W. Lu, A. Zunger, and M. Deutsch, "Electronic charge distribution in crystalline diamond, silicon, and germanium," *Phys. Rev. B* 47:9385–9410, 1993.
[103] M. T. Yin and M. L. Cohen, "Theory of static structural properties, crystal stability, and phase transformations: Application to Si and Ge," *Phys. Rev. B* 26:5668–5687, 1982.
[104] O. H. Nielsen and R. M. Martin, "Stresses in semiconductors: *Ab initio* calculations on Si, Ge, and GaAs," *Phys. Rev. B* 32(6):3792–3805, 1985.
[105] J. R. Chelikowsky and M. L. Cohen, "Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors," *Phys. Rev. B* 14:556–582, 1976.
[106] V. L. Moruzzi, A. R. Williams, and J. F. Janak, "Local density theory of metallic cohesion," *Phys. Rev. B* 15:2854–2857, 1977.
[107] V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals*, Pergamon Press, New York, 1978.
[108] F. D. Murnaghan, "The compressibility of media under extreme pressures," *Proc. Nat. Acad. Sci. USA* 50:244–247, 1944.
[109] L. P. Howland, "Band structure and cohesive energy of potassium chloride," *Phys. Rev.* 109:1927, 1958.
[110] P. D. DeCicco, "Self-consistent energy bands and cohesive energy of potassium chloride," *Phys. Rev.* 153:931, 1967.
[111] W. E. Rudge, "Variation of lattice constant in augmented-plane-wave energy-band calculation for lithium," *Phys. Rev.* 181:1033, 1969.
[112] M. Ross and K. W. Johnson, "Augmented-plane-wave calculation of the total energy, bulk modulus, and band structure of compressed aluminum," *Phys. Rev. B* 2:4709, 1970.
[113] E. C. Snow, "Total energy as a function of lattice parameter for copper via the self-consistent augmented-plane-wave method," *Phys. Rev. B* 8:5391, 1973.
[114] J. F. Janak, V. L. Moruzzi, and A. R. Williams, "Ground-state thermomechanical proerties of some cubic elements in the local-density formalism," *Phys. Rev. B* 12:1257–1261, 1975.
[115] W. B. Holzapfel, "Physics of solids under strong compression," *Rep. Prog. Phys.* 59:29, 1996.
[116] *High-pressure techniques in chemistry and physics*, edited by W. B. Holzapfel and N. S. Issacs, Oxford University Press, Oxford/New York/Tokyo, 1997.
[117] R. Biswas, R. M. Martin, R. J. Needs, and O. H. Nielsen, "Complex tetrahedral structures of silicon and carbon under pressure," *Phys. Rev. B* 30(6):3210–3213, 1984.
[118] M. T. Yin, "Si-III (BC-8) crystal phase of Si and C: Structural properties, phase stabilities, and phase transitions," *Phys. Rev. B* 30:1773–1776, 1984.
[119] G. J. Ackland, "High-pressure phases of group IV and III-V semiconductors," *Rep. Prog. Phys.* 64:483–516, 2001.
[120] A. Mujica, A. Rubio, A. Munoz, and R. J. Needs, "High-pressure phases of group IVa, IIIa-Va and IIb-VIa compounds," *Rev. Mod. Phys.* 75:863–912 (2003).
[121] C. Mailhiot, L. H. Yang, and A. K. McMahan, "Polymeric nitrogen," *Phys. Rev. B* 46:14419–14435, 1992.
[122] J. S. Kasper and Jr. R. H. Wentorf, "The crystal structures of new forms of silicon and germanium," *Acta Cryst.* 17:752, 1964.
[123] R. J. Needs and A. Mujica, "Theoretical description of high-pressure phases of semiconductors," *High Pressure Research* 22:421, 2002.
[124] H. Olijnyk, S. K. Sikka, and W. B. Holzapfel, "Structural phase transitions in Si and Ge under pressures up to 50 GPa," *Phys. Lett.* 103A:137, 1984.

[125] J. Z. Hu and I. L. Spain, "Phases of silicon at high pressure," *Solid State Commun.* 51:263, 1984.
[126] A. K. McMahan, "Interstitial-sphere linear muffin-tin orbital structural calculations for C and Si," *Phys. Rev. B* 30:5835–5841, 1984.
[127] N. Moll, M. Bockstedte, M. Fuchs, E. Pehlke, and M. Scheffler, "Application of generalized gradient approximations: The diamond-beta-tin phase transition in Si and Ge," *Phys. Rev. B* 52:2550–2556, 1995.
[128] O. H. Nielsen and R. M. Martin, "First-principles calculation of stress," *Phys. Rev. Lett.* 50(9):697–700, 1983.
[129] O. H. Nielsen and R. M. Martin, "Quantum-mechanical theory of stress and force," *Phys. Rev. B* 32(6):3780–3791, 1985.
[130] O. H. Nielsen, "Optical phonons and elasticity of diamond at megabar stresses," *Phys. Rev. B* 34:5808–5819, 1986.
[131] C. Herring, in *Magnetism IV*, edited by G. Rado and H. Suhl, Academic Press, New York, 1966.
[132] J. Kübler, *Theory of Itinerant Electron Magnetism*, Oxford University Press, Oxford, 2001.
[133] E. C. Stoner, "Collective electron ferromagnetism II. Energy and specific heat," *Proc. Roy. Soc. London* Ser. A 169:339, 1939.
[134] J. Kübler and V. Eyert, in *Electronic and Magnetic Properties of Metals and Ceramics*, edited by K. H. J. Buschow, VCH-Verlag, Weinheim, Germany, 1992, p. 1.
[135] Q. Niu and L. Kleinman, "Spin-wave dynamics in real crystals," *Phys. Rev. Lett.* 80:2205–2208, 1998.
[136] R. Gebauer and S. Baroni, "Magnons in real materials from density-functional theory," *Phys. Rev. B* 61:R6459–R6462, 2000.
[137] *Lattice Dynamics*, edited by R. F. Wallis, Pergamon Press, London, 1965.
[138] *Dynamical Properties of Solids, Vol. III*, edited by G. K. Horton and A. A. Maradudin, North-Holland, Amsterdam, 1979.
[139] K. Kunc and R. M. Martin, "Density-functional calculation of static and dynamic properties of GaAs," *Phys. Rev. B* 24(4):2311–2314, 1981.
[140] K. Kunc, I. Loa, K. Syassen, R. K. Kramer, and A. K. Ahn, "MgB_2 under pressure: phonon calculations, Raman spectroscopy, and optical reflectance," *J. Phys. Condensed Matter* 13:9945–9962, 2001.
[141] P. Ordejon, E. Artacho, R. Cachau, J. Gale, A. Garcia, J. Junquera, J. Kohanoff, M. Machado, D. Sanchez-Portal, J. M. Soler, and R. Weht, "Linear scaling DFT calculations with numerical atomic orbitals," *Mat. Res. Soc. Symp. Proc.* 677, 2001.
[142] R. E. Cohen and H. Krakauer, "Electronic structure studies of the differences in ferroelectric behavior of $BaTiO_3$ and $PbTiO_3$," *Ferroelectrics* 136:65, 1992.
[143] D. J. Chadi and R. M. Martin, "Calculation of lattice dynamical properties from electronic energies: application to C, Si and Ge," *Solid State Commun.* 19(7):643–646, 1976.
[144] H. Wendel and R. M. Martin, "Theory of structural properties of covalent semiconductors," *Phys. Rev. B* 19(10):5251–5264, 1979.
[145] J. Nagamatsu, N. Nakagawa, T. Muranaka, Y. Zenitani, and J. Akimitsu, "Superconductivity at 39 K in magnesium diboride," *Nature* 410:63–64, 2001.
[146] U. V. Waghmare and K. M. Rabe, "*Ab initio* statistical mechanics of the ferroelectric phase transition in $PbTiO_3$," *Phys. Rev. B* 55:6161–6173, 1997.

[147] R. D. King-Smith and D. Vanderbilt, "Theory of polarization of crystalline solids," *Phys. Rev. B* 47:1651–1654, 1993.
[148] R. Resta, "Macroscopic polarization in crystalline dielectrics: the geometric phase approach," *Rev. Mod. Phys.* 66:899–915, 1994.
[149] M. V. Berry, "Quantal phase factors accompanying adiabatic changes," *Proc. Roy. Soc. London* A 392:45, 1984.
[150] P. D. De Cicco and F. A. Johnson, "The quantum theory of lattice dynamics. IV," *Proc. R. Soc. London, Ser. A* 310:111–119, 1969.
[151] L. J. Sham, "Electronic contribution to lattice dynamics in insulating crystals," *Phys. Rev.* 188:1431–1439, 1969.
[152] R. Pick, M. H. Cohen, and R. M. Martin, "Microscopic theory of force constants in the adiabatic approximation," *Phys. Rev. B* 1:910–920, 1970.
[153] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, "Phonons and related crystal properties from density-functional perturbation theory," *Rev. Mod. Phys.* 73:515–562, 2001.
[154] P. Giannozzi, S. de Gironcoli, P. Pavoni, and S. Baroni, "*Ab initio* calculation of phonon dispersion in semiconductors," *Phys. Rev. B* 43:7231, 1991.
[155] Y. Kong, O. V. Dolgov, O. Jepsen, and O. K. Andersen, "Electron-phonon interaction in the normal and superconducting states of MgB_2," *Phys. Rev. B* 64:020501, 2001.
[156] R. Car and M. Parrinello, "Unified approach for molecular dynamics and density functional theory," *Phys. Rev. Lett.* 55:2471–2474, 1985.
[157] M. P. Grumbach and R. M. Martin, "Phase diagram of carbon at high pressures and temperatures," *Phys. Rev. B* 54:15730–15741, 1996.
[158] F. P. Bundy, W. A. Bassettand, M. S. Weathers, R. J. Hemley, H. K. Mao, and A. F. Goncharov, "The pressure-temperature phase and transformation diagram for carbon; updated through 1994," *Carbon* 34:141–153, 1996.
[159] G. Galli, R. M. Martin, R. Car, and M. Parrinello, "*Ab initio* calculation of properties of carbon in the amorphous and liquid states." *Phys. Rev. B* 42:7470, 1990.
[160] G. Galli, R. M. Martin, R. Car, and M. Parrinello, "Melting of diamond at high pressure." *Science* 250:1547, 1990.
[161] A. C. Mitchell, J. W. Shaner, and R. N. Keller, "The use of electrical-conductivity experiments to study the phase diagram of carbon," *Physica* 139:386, 1986.
[162] C. H. Xu, C. Z. Wang, C. T. Chan, and K. M. Ho, "A transferable tight-binding potential for carbon," *J. Phys.: Condens. Matter* 4:6047, 1992.
[163] D. Alfe and M. J. Gillan, "First-principles simulations of liquid Fe-S under earth's core conditions," *Phys. Rev. B* 58:8248–56, 1998.
[164] D. Alfe, G. Kresse, and M. J. Gillan, "Structure and dynamics of liquid iron under earth's core conditions," *Phys. Rev. B* 61:132–142, 2000.
[165] M. Sprik, J. Hutter, and M. Parrinello, "*Ab initio* molecular dynamics simulation of liquid water: Comparison of three gradient-corrected density functionals," *J. Chem. Phys.* 105:1142, 1996.
[166] E. Schwegler, G. Galli, F. Gygi, and R. Q. Hood, "Dissociation of water under pressure," *Phys. Rev. Lett.* 87:265501, 2001.
[167] D. R. Hamann, "H_2O hydrogen bonding in density-functional theory," *Phys. Rev. B* 55:R10 157, 1997.
[168] P. L. Geissler, C. Dellago, D. Chandler, J. Hutter, and M. Parrinello, "Autoionization in liquid water," *Science* 291:2121, 2001.

[169] C. Cavazzoni, G. L. Chiarotti, S. Scandolo, E. Tosatti, M. Bernasconi, and M. Parrinello, "Superionic and metallic states of water and ammonia at giant planet conditions," *Science* 283:44, 1999.

[170] M. Boero, K. Terakura, T. Ikeshoji, C. C. Liew, and M. Parrinello, "Hydrogen bonding and dipole moment of water at supercritical conditions: A first-principles molecular dynamics study," *Phys. Rev. Lett.* 85:3245–3248, 2000.

[171] I. Bako, J. Hutter, and G. Palinkas, "Car-Parrinello molecular dynamics simulation of the hydrated calcium ion," *J. Chem. Phys.* 117:9838, 2002.

[172] M. Boero, M. Parrinello, and K. Terakura, "First principles molecular dynamics study of Ziegler-Natta heterogeneous catalysis," *J. Am. Chem. Soc.* 120:746–2752, 1998.

[173] M. Boero, M. Parrinello, S. Huffer, and H. Weiss, "First principles study of propene polymerization in Ziegler-Natta heterogeneous catalysis," *J. Am. Chem. Soc.* 122:501–509, 2000.

[174] A. R. Smith, V. Ramachandran, R. M. Feenstra, D. W. Greve, M.-S. Shin, M. Skowronski, J. Neugebauer, and J. E. Northrup, "Wurtzite GaN surface structures studied by scanning tunneling microscopy and reflection high energy electron diffraction," *J. Vac. Sci. Tech. A* 16:1641, 1998.

[175] W. A. Harrison, "Theory of polar semiconductor surfaces," *J. Vac. Sci. Tech.* 16:1492–1496, 1979.

[176] R. M. Martin, "Atomic reconstruction at polar interfaces of semiconductors," *J. Vac. Sci. Tech.* 17(5):978–981, 1980.

[177] A. A. Wilson, *Thermodynamics and Statistical Mechanics*, Cambridge University Press, Cambridge, England, 1957.

[178] A. A. Wilson, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, New York, 1965.

[179] G. X. Qian, R. M. Martin, and D. J. Chadi, "First-principles calculations of atomic and electronic structure of the GaAs (110) surface." *Phys. Rev. B* 37:1303, 1988.

[180] A. Garcia and J. E. Northrup, "First-principles study of Zn- and Se-stabilized ZnSe(100) surface reconstructions," *J. Vac. Sci. Tech. B* 12:2678–2683, 1994.

[181] J. E. Northrup and S. Froyen, "Structure of GaAs(001) surfaces: the role of electrostatic interactions," *Phys. Rev. B* 50:2015, 1994.

[182] A. Franciosi and C. G. Van de Walle, "Heterojunction band offset engineering," *Surf. Sci. Rep.* 25:1, 1996.

[183] C. G. Van de Walle and R. M. Martin, "Theoretical study of band offsets at semiconductor interfaces," *Phys. Rev. B* 35:8154–8165, 1987.

[184] C. G. Van de Walle and R. M. Martin, "'Absolute' deformation potentials: Formulation and *ab initio* calculations for semiconductors." *Phys. Rev. Lett.* 62:2028–2031, 1989.

[185] D. J. Chadi and K. J. Chang, "Energetics of DX-center formation in GaAs and $Al_xGa_{1-x}As$ alloys," *Phys. Rev. B* 39:10063–10074, 1989.

[186] D. J. Chadi and K. J. Chang, "Theory of the atomic and electronic structure of DX centers in GaAs and $Al_xGa_{1-x}As$ alloys," *Phys. Rev. Lett.* 61:873–876, 1988.

[187] C. Herring, N. M. Johnson, and C. G. Van de Walle, "Energy levels of isolated interstitial hydrogen in silicon," *Phys. Rev. B* 64:125209, 2001.

[188] W. D. Knight, K. Clemenger, W. A. de Heer, W. A. Saunders, M. Y. Chou, and M. L. Cohen, "Electronic shell structure and abundances of sodium clusters," *Phys. Rev. Lett.* 52:2141, 1984.

[189] M. Brack, "The physics of simple metal clusters: self-consistent jellium model and semiclassical approaches," *Rev. Mod. Phys.* 65:677–732, 1993.

[190] U. Rothlisberger, W. Andreoni, and P. Giannozzi, "Thirteen-atom clusters: equilibrium geometries, structural transformations, and trends in Na, Mg, Al, and Si," *J. Chem. Phys.* 92:1248, 1992.

[191] J. C. Phillips, "Electron-correlation energies and the structure of Si_{13}," *Phys. Rev. B* 47:14132, 1993.

[192] J. C. Grossman and L. Mitas, "Quantum Monte Carlo determination of electronic and structural properties of Si_n clusters," *Phys. Rev. Lett.* 74:1323–1325, 1995.

[193] J. C. Grossman and L. Mitas, "Family of low-energy elongated Si_n ($n \leq 50$) clusters," *Phys. Rev. B* 52:16735–16738, 1995.

[194] N. Troullier and J. L. Martins, "Structural and electronic properties of C_{60}," *Phys. Rev. B* 46:1754–1765, 1992.

[195] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, "C_{60}: Buckminsterfullerene," *Nature* 318:162, 1985.

[196] S. Iijima, "Helical microtubules of graphitic carbon," *Nature* 354:56, 1991.

[197] W. Kratschmer, L.D. Lamb, K. Fostiropoulos, and D.R. Huffman, "Solid C_{60}: a new form of carbon," *Nature* 347:354, 1990.

[198] R. C. Haddon et al., "Conducting films of C_{60} and C_{70} by alkali-metal doping," *Nature* 350:320, 1991.

[199] O. E. Gunnarsson, "Superconductivity in fullerides," *Rev. Mod. Rev.* 69:575–606, 1997.

[200] J. I. Pascual and et al., "Seeing molecular orbitals," *Chem. Phys. Lett.* 321:78–82, 2000.

[201] J. Tersoff and D. R. Hamann, "Theory of the scanning tunneling microscope," *Phys. Rev. B* 31(2):805–813, 1985.

[202] V. Meunier, C. Roland, J. Bernholc, and M. Buongiorno Nardelli, "Electronic and field emission properties of boron nitride carbon nanotube superlattices," *Appl. Phys. Lett.* 81:46, 2002.

[203] N. Hamada, S. Sawada, and A. Oshiyama, "New one-dimensional conductors: Graphitic microtubules," *Phys. Rev. Lett* 68:1579–1581, 1992.

[204] R. Saito, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, "Electronic structure of graphene tubules based on C_{60}," *Phys. Rev. B* 46:1804–1811, 1992.

[205] R. Saito, G. Dresselhaus, and M. S. Dresselhaus, *Physical Properties of Carbon Nanotubes*, Imperial College Press, London, 1998.

[206] X. Blase, L. X. Benedict, E. L. Shirley, and S. G. Louie, "Are fullerene tubules metallic?" *Phys. Rev. Lett* 72:1878–1881, 1994.

[207] A. Rubio, J. L. Corkill, and M. L. Cohen, "Theory of graphitic boron nitride nanotubes," *Phys. Rev. B* 49:5081–5084, 1994.

[208] N. G. Chopra, R. J. Luyken, K. Cherrey, V. H. Crespi, M. L. Cohen, S. G. Louie, and A. Zettl, "Boron nitride nanotubes," *Science* 269:966, 1995.

[209] E. L. Briggs, D. J. Sullivan, and J. Bernholc, "Real-space multigrid-based approach to large-scale electronic structure calculations," *Phys. Rev. B* 54:14362–14375, 1996.

[210] *Photoelectron Spectroscopy, 2nd Ed.*, edited by S. Hüffner, Springer, Berlin, 1995.

[211] T. Miller, E. D. Hansen, W. E. McMahon, and T.-C. Chiang, "Direct transitions, indirect transitions, and surface photoemission in the prototypical system Ag(111)," *Surf. Sci.* 376:32, 1997.

[212] P. Thiry, D. Chandesris, J. Lecante, C. Guillot, R. Pinchaux, and Y. Petroff, "E vs k and inverse lifetime of Cu(110)," *Phys. Rev. Lett.* 43:82–85, 1979.

[213] G. A. Burdick, "Energy band structure of copper," *Phys. Rev.* 129:138–150, 1963.

[214] T. C. Chiang, J. A. Knapp, M. Aono, and D. E. Eastman, "Angle-resolved photoemission, valence-band dispersions e(k), and electron and hole lifetimes for GaAs," *Phys. Rev. B* 21:3513–3522, 1980.

[215] K. C. Pandey and J. C. Phillips, "Nonlocal pseudopotentials for Ge and GaAs," *Phys. Rev. B* 9:1552–1559, 1974.

[216] M. Imada, A. Fujimori, and Y. Tokura, "Metal-insulator transitions," *Rev. Mod. Phys.* 70:1039–1263, 1998.

[217] M. S. Hybertsen and S. G. Louie, "Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies," *Phys. Rev. B* 34:5390–5413, 1986.

[218] R. W. Godby, M. Schlüter, and L. J. Sham, "Quasiparticle energies in GaAs and AlAs," *Phys. Rev. B* 35:4170–4171, 1987.

[219] M. Rohlfing, P. Krüger, and J. Pollmann, "Quasiparticle band-structure calculations for C, Si, Ge, GaAs, and SiC using gaussian-orbital basis sets," *Phys. Rev. B* 48:17791–17805, 1993.

[220] L. Hedin and S. Lundquist, in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1969, Vol. 23, p. 1.

[221] A. L. Wachs, T. Miller, T. C. Hsieh, A. P. Shapiro, and T. C. Chiang, "Angle-resolved photoemission studies of Ge(111)-c(2 × 8), Ge(111)-(1 × 1)H, Si(111)-(7 × 7), and Si(100)-(2 × 1)," *Phys. Rev. B* 32:2326–2333, 1985.

[222] J. E. Ortega and F. J. Himpsel, "Inverse-photoemission study of Ge(100), Si(100), and GaAs(100): Bulk bands and surface states," *Phys. Rev. B* 47:2130–2137, 1993.

[223] M. Staedele, M. Moukara, J. A. Majewski, P. Vogl, and A. Gorling, "Exact exchange Kohn-Sham formalism applied to semiconductors," *Phys. Rev. B* 59:10031–10043, 1999.

[224] W. Koch and M. C. Holthausen, *A Chemists' Guide to Density Functional Thoery*, Wiley-VCH, Weinheim, 2001.

[225] D. Pines, *Elementary Excitations in Solids*, Wiley, New York, 1964.

[226] D. Pines and P. Nozières, *The Theory of Quantum Liquids, Vol. I*, Addison-Wesley Inc., Redwood City, 1989.

[227] M. Rohlfing and S. G. Louie, "Electron-hole excitations and optical spectra from first principles," *Phys. Rev. B* 62:4927, 2000.

[228] L. X. Benedict and E. L. Shirley, "*Ab initio* calculation of $\epsilon_2(\omega)$ including the electron-hole interaction: Application to GaN and CaF_2," *Phys. Rev. B* 59:5441–5451, 1999.

[229] A. Zangwill and P. Soven, "Density-functional approach to local-field effects in finite systems: Photoabsorption in the rare gases," *Phys. Rev. A* 21:1561, 1980.

[230] E. Runge and E. K. U. Gross, "Density-functional theory for time-dependent systems," *Phys. Rev. Lett.* 52:997–1000, 1984.

[231] M. E. Casida, in *Recent Developments and Applications of Density Functional Theory*, edited by J. M. Seminario, Elsevier, Amsterdam, 1996, p. 391.

[232] A. Gorling, "Exact exchange-correlation kernel for dynamic response properties and excitation energies in density-functional theory," *Phys. Rev. A* 57:3433–3436, 1998.

[233] K. Yabana and G. F. Bertsch, "Time-dependent local-density approximation in real time," *Phys. Rev. B* 54:4484–4487, 1996.

[234] I. Vasiliev, S. Ogut, and J. R. Chelikowsky, "*Ab initio* excitation spectra and collective electronic response in atoms and clusters," *Phys. Rev. Lett.* 82:1919–1922, 1999.

[235] F. Kootstra, P. L. de Boeij, and J. G. Snijders, "Application of time-dependent density-functional theory to the dielectric function of various nonmetallic crystals," *Phys. Rev. B* 62:7071–7083, 2000.

[236] J. B. Staunton, J. Poulter, B. Ginatempo, E. Bruno, and D. D. Johnson, "Incommensurate and commensurate antiferromagnetic spin fluctuations in Cr and Cr alloys from *ab initio* dynamical spin susceptibility calculations," *Phys. Rev. Lett.* 82:3340–3343, 1999.
[237] R. van Leeuwen, "Key concepts of time-dependent density-functional theory," *Int. J. Mod. Phys. B* 15:1969–2023, 2001.
[238] G. Onida, L. Reining, and A. Rubio, "Electronic excitations: density-functional versus many-body Green's-function approaches," *Rev. Mod. Phys.* 74:601, 2002.
[239] H. Uchiyama, K. M. Shen, S. Lee, A. Damascelli, D. H. Lu, D. L. Feng, Z.-X. Shen, and S. Tajima, "Electronic structure of MgB_2 from angle-resolved photoemission spectroscopy," *Phys. Rev. Lett.* 88:157002, 2002.
[240] J. M. An and W. E. Pickett, "Superconductivity of MgB_2: Covalent bonds driven metallic," *Phys. Rev. Lett.* 86:4366–4369, 2001.
[241] J. Kortus, I. I. Mazin, K. D. Belashchenko, V. P. Antropov, and L. L. Boyer, "Superconductivity of metallic boron in MgB_2," *Phys. Rev. Lett.* 86:4656–4659, 2001.
[242] P. B. Allen and B. Mikovic, in *Solid State Phys., Vol. 37*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1982, p. 1.
[243] D. Rainer, *Prog. Low Temp. Phys.*, North-Holland, Amsterdam, 1986, Vol. 10, pp. 371–424.
[244] H. J. Choi, D. Roundy, H. Sun, M. L. Cohen, and S. G. Louie, "First-principles calculation of the superconducting transition in MgB_2 within the anisotropic eliashberg formalism," *Phys. Rev. B* 66:020513, 2002.
[245] G. R. Stewart, "Heavy-fermion systems," *Rev. Mod. Phys.* 56:755–787, 1984.
[246] W. Jones and N. H. March, *Theoretical Solid State Physics, Vol. I*, John Wiley and Sons, New York, 1976.
[247] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory (Unabridged reprinting of 1989 version)*, Dover, Mineola, New York, 1996.
[248] L. Mihaly and M. C. Martin, *Solid State Physics: Problems and Solutions*, John Wiley and Sons, New York, 1996.
[249] J. W. Strutt (Lord Rayleigh), *Theory of Sound, Vol. 1, Sec 88*, Reprint: Dover Publications, New York, 1945.
[250] W. Ritz, "Über eine neue Methode zur Losung Gewisser Variationsprobleme der mathematischen Physik," *Reine Angew. Math.* 135:1, 1908.
[251] P. Ehrenfest, "Bemurkung über die angenäherte Gültigkeit der klassischen Mechanik innerhalb der Quantenmechanik," *Z. Phys.* 45:455, 1927.
[252] M. Born and V. Fock, "Beweis des Adiabatensatzes," *Z. Phys.* 51:165, 1928.
[253] P. Güttiger, "Das Verhalten von Atomen im magnetischen Drefeld," *Z. Phys.* 73:169, 1931.
[254] W. Pauli, *Handbuch der Physik*, Springer, Berlin, 1933, pages 83–272 relates to force and stress.
[255] H. Hellmann, *Einfuhrung in die Quantumchemie*, Franz Duetsche, Leipzig, 1937.
[256] R. P. Feynman, "Forces in molecules," *Phys. Rev.* 56:340, 1939.
[257] M. Born, W. Heisenberg, and P. Jordan, "Zur Quantenmechanik, II," *Z. Phys.* 35:557, 1926.
[258] B. Finkelstein, "Über den Virialsatz in der Wellenmechanik," *Z. Phys.* 50:293, 1928.
[259] V. Fock, "Naherungsmethode zur Losung des quanten-mechanischen Mehrkorperprobleme," *Z. Phys.* 63:855, 1930.
[260] J. C. Slater, "The virial and molecular structure," *J. Chem. Phys.* 1:687, 1933.
[261] R. D. McWeeny and B. T. Sutcliffe, *Methods of Molecular Quantum Mechanics, second edition*, Academic Press, New York, 1976.

[262] Y. H. Shao, C. A. White, and M. Head-Gordon, "Efficient evaluation of the Coulomb force in density-functional theory calculations," *J. Chem. Phys.* 114:6572–6577, 2001.
[263] E. M. Landau and L. P. Pitaevskii, *Statistical Physics: Part 1*, Pergamon Press, Oxford, England, 1980.
[264] J. K. L. MacDonald, "Successive approximations by the Rayleigh-Ritz variation method," *Phys. Rev.* 43:830, 1933.
[265] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: non-relativistic theory*, Pergamon Press, Oxford, England, 1977.
[266] R. Shankar, *Principles of Quantum Mechanics*, Plenum Publishing, New York, 1980.
[267] X. Gonze and J. P. Vigneron, "Density functional approach to non-linear response coefficients in solids," *Phys. Rev. B* 39:13120, 1989.
[268] X. Gonze, "Perturbation expansion of variational principles at arbirary order," *Phys. Rev. A* 52:1086–1095, 1995.
[269] J. C. Slater, *Symmetry and Energy Bands in Crystals (Corrected and reprinted version of 1965 Quantum Theory of Molecules and Solids, Vol. 2)*, Dover, New York, 1972.
[270] V. Heine, *Group Theory*, Pergamon Press, New York, 1960.
[271] M. Tinkham, *Group Theory and Quantum Mechanics*, McGraw-HIll, New York, 1964.
[272] M. J. Lax, *Symmetry Principles in Solid State and Molecular Physics*, Wiley, New York, 1974.
[273] H. J. Monkhorst and J. D. Pack, "Special points for Brillouin-zone integrations," *Phys. Rev. B* 13:5188–5192, 1976.
[274] A. H. MacDonald, "Comment on special points for Brillouin-zone integrations," *Phys. Rev. B* 18:5897–5899, 1978.
[275] A. Baldereschi, "Mean-value point in the Brillouin zone," *Phys. Rev. B* 7:5212–5215, 1973.
[276] D. J. Chadi and M. L. Cohen, "Electronic structure of $Hg_{1-x}Cd_xTe$ alloys and charge-density calculations using representative k points," *Phys. Rev. B* 7:692–699, 1973.
[277] J. Moreno and J. M. Soler, "Optimal meshes for integrals in real- and reciprocal-space unit cells," *Phys. Rev. B* 45:13891–13898, 1992.
[278] J. F. Janak, in *Computational Methods in Band Theory*, edited by P. M. Marcus, J. F. Janak, and A. R. Williams, Plenum, New York, 1971, pp. 323–339.
[279] G. Gilat, "Analysis of methods for calculating spectral properties in solids," *J. Comput. Phys.* 10:432–65, 1972.
[280] G. Gilat, "Methods of Brillouin zone integration," *Methods Comput. Phys.* 15:317–70, 1976.
[281] A. H. MacDonald, S. H. Vosko, and P. T. Coleridge, "Extensions of the tetrahedron method for evaluating spectral properties of solids," *J. Phys. C: Solid State Phys.* 12:2991–3002, 1979.
[282] P. E. Blöchl, O. Jepsen, and O. K. Andersen, "Improved tetrahedron method for Brillouin-zone integrations," *Phys. Rev. B* 49:16223–16233, 1994.
[283] L. Van Hove, "The occurrence of singularities in the elastic frequency distribution of a crystal," *Phys. Rev.* 89:1189–1193, 1953.
[284] W. Jones and N. H. March, *Theoretical Solid State Physics, Vol. II*, John Wiley and Sons, New York, 1976.
[285] J. M. Luttinger, "Fermi surface and some simple equilibrium properties of a system of interacting fermions," *Phys. Rev.* 119:1153, 1960.
[286] R. M. Martin, "Fermi-surface sum rule and its consequences for periodic Kondo and mixed-valence systems," *Phys. Rev. Lett.* 48(5):362–365, 1982.

[287] S. Goedecker, "Decay properties of the finite-temperature density matrix in metals," *Phys. Rev. B* 58:3501–3502, 1998.
[288] J. W. Gibbs, "Fourier series," *Nature (Letter to the Editor)* 59:200, 1898.
[289] S. Ismail-Beigi and T. A. Arias, "Locality of the density matrix in metals, semiconductors and insulators," *Phys. Rev. Lett.* 82:2127–2130, 1999.
[290] J. Bardeen, "Theory of the work function. II. the surface double layer," *Phys. Rev.* 49:653, 1936.
[291] U. von Barth and L. Hedin, "A local exchange-correlation potential for the spin polarized case: I," *J. Phys. C* 5:1629, 1972.
[292] E. P. Wigner, "Effects of the electron interaction on the energy levels of electrons in metals," *Trans. Faraday Soc.* 34:678, 1938.
[293] M. Gell-Mann and K. A. Brueckner, "Correlation energy of an electron gas at high-density," *Phys. Rev.* 106:364, 1957.
[294] W. J. Carr and A. A. Maradudin, "Ground state energy of a high-density electron gas," *Phys. Rev.* 133:371, 1964.
[295] W. J. Carr, "Energy, specific heat, and magnetic properties of the low-density electron gas," *Phys. Rev.* 122:1437, 1961.
[296] B. Holm, "Total energies from GW calculations," *Phys. Rev. Lett.* 83:788–791, 1999.
[297] D. M. Ceperley and B. J. Alder, "Ground state of the electron gas by a stochastic method," *Phys. Rev. Lett.* 45:566–569, 1980.
[298] G. Ortiz and P. Ballone, "Correlation energy, structure factor, radial distribution function and momentum distribution of the spin-polarized uniform electron gas," *Phys. Rev. B* 50:1391–1405, 1994.
[299] Y. Kwon, D. M. Ceperley, and R. M. Martin, "Effects of backflow correlation in the three-dimensional electron gas: Quantum Monte Carlo study," *Phys. Rev. B* 58:6800–6806, 1998.
[300] J. P. Perdew and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems," *Phys. Rev. B* 23:5048, 1981.
[301] S. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid corelation energies for local spin density calculations: a critical analysis," *Can. J. Phys.* 58:1200, 1983.
[302] G. Ortiz, M. Harris, and P. Ballone, "Correlation energy, structure factor, radila density distribution function, and momentum distribution of the spin-polarized electron gas," *Phys. Rev. B* 50:1391–1405, 1994.
[303] P. Gori-Giorgi, F. Sacchetti, and G. B. Bachelet, "Analytic structure factors and pair correlation functions for the unpolarized electron gas," *Phys. Rev. B* 61:7353–7363, 2000.
[304] J. C. Slater, "Cohesion in monovalent metals," *Phys. Rev.* 35:509, 1930.
[305] I.-W. Lyo and E. W. Plummer, "Quasiparticle band structure of Na and simple metals," *Phys. Rev. Lett.* 60:1558–1561, 1988.
[306] E. Jensen and E. W. Plummer, "Experimental band structure of Na," *Phys. Rev. Lett.* 55:1912, 1985.
[307] J. Lindhard, "On the properties of a gas of charged particles," *Kgl. Danske Videnskab. Selskab, Mat.-fys. Medd.* 28:8, 1954.
[308] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.* 136:B864–871, 1964.
[309] N. David Mermin, "Thermal properties of the inhomogeneous electron gas," *Phys. Rev.* 137:A1441–1443, 1965.
[310] *Theory of the Inhomogeneous Electron Gas*, edited by S. Lundqvist and N. H. March, Plenum, New York, 1983.

[311] *Density Functional Methods in Physics*, edited by R. M. Dreizler and J. da Providencia, Plenum, New York, 1985.
[312] R. M. Dreizler and E. K. U. Gross, *Density Functional Theory: An Approach to the Quantum Many-body Problem*, Springer, Berlin, 1990.
[313] *Density Functional Theory*, edited by E. K. U. Gross and R. M. Dreizler, Plenum, New York, 1995.
[314] R. O. Jones and O. Gunnarsson, "The density functional formalism, its applications and prospects," *Rev. Mod. Phys.* 61:689–746, 1989.
[315] W. Kohn, "Nobel lecture: Electronic structure of matter – wave functions and density functionals," *Rev. Mod. Phys.* 71:1253–1266, 1999.
[316] L. H. Thomas, "The calculation of atomic fields," *Proc. Cambridge Phil. Roy. Soc.* 23:542–548, 1927.
[317] E. Fermi, "Un metodo statistico per la determinazione di alcune priorieta dell'atome," *Rend. Accad. Naz. Lincei* 6:602–607, 1927.
[318] P. A. M. Dirac, "Note on exchange phenomena in the Thomas-Fermi atom," *Proc. Cambridge Phil. Roy. Soc.* 26:376–385, 1930.
[319] C. F. von Weizsacker, "Zur Theorie der Kernmassen," *Z. Phys.* 96:431, 1935.
[320] D. A. Kirzhnits, "Quantum corrections to the Thomas-Fermi equation," *Soviet Phys. – JETP* 5:64, 1957.
[321] R. P. Feynman, N. Metropolis, and E. Teller, "Equations of state of elements based on the generalized Fermi-Thomas theory," *Phys. Rev.* 75:1561–1573, 1949.
[322] E. Teller, "On the stability of molecules in the Thomas-Fermi theory," *Rev. Mod. Phys.* 34:627–630, 1962.
[323] W. Kohn, in *Highlights in Condensed Matter Theory*, edited by F. Bassani, F. Fumi, and M. P. Tosi, North Holland, Amdterdam, 1985, p. 1.
[324] M. Levy, "Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the n-representability problem." *Proc. Nat. Acad. Sci. USA* 76:6062, 1979.
[325] M. Levy, "Electron densities in search of hamiltonians," *Phys. Rev. A* 26:1200, 1982.
[326] M. Levy and J. P. Perdew, in *Density Functional Methods in Physics*, edited by R. M. Dreizler and J. da Providencia, Plenum, New York, 1985, p. 11.
[327] E. Lieb, in *Physics as Natural Philosophy*, edited by A. Shimony and H. Feshbach, MIT Press, Cambridge, 1982, p. 111.
[328] E. Lieb, "Density functionals for coulomb systems," *Int. J. Quant. Chem.* 24:243, 1983.
[329] E. Lieb, in *Density Functional Methods in Physics*, edited by R. M. Dreizler and J. da Providencia, Plenum, New York, 1985, p. 31.
[330] T. L. Gilbert, "Hohenberg-Kohn theorem for nonlocal external potentials," *Phys. Rev. B* 12:2111, 1975.
[331] O. Gunnarsson, B. I. Lundqvist, and J. W. Wilkins, "Contribution to the cohesive energy of simple metals: Spin-dependent effect," *Phys. Rev. B* 10:1319–1327, 1974.
[332] G. Vignale and M. Rasolt, "Current- and spin-density-functional theory for inhomogeneous electronic systems in strong magnetic fields," *Phys. Rev. B* 37:10685–10696, 1988.
[333] G. Vignale and W. Kohn, "Current-dependent exchange-correlation potential for dynamical linear response theory," *Phys. Rev. Lett.* 77:2037–2040, 1996.
[334] K. Capelle and E. K. U. Gross, "Spin-density functionals from current-density functional theory and vice versa: A road towards new approximations," *Phys. Rev. Lett.* 78:1872–1875, 1997.

[335] R. van Leeuwen, "Causality and symmetry in time-dependent density-functional theory," *Phys. Rev. Lett.* 80:1280–1283, 1998.
[336] J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, Jr. "Density-functional theory for fractional particle number: Derivative discontinuities of the energy," *Phys. Rev. Lett.* 49:1691–1694, 1982.
[337] N. T. Maitra, I. Souza, and K. Burke, "Current-density functional theory of the response of solids," *Phys. Rev. B* 68:045109, 2003.
[338] G. Wannier, "Dynamics of band electrons in electric and magnetic fields," *Rev. Mod. Phys.* 34:645, 1962.
[339] G. Nenciu, "Dynamics of band electrons in electric and magnetic fields: rigorous justification of the effective hamiltonians," *Rev. Mod. Phys.* 63:91, 1991.
[340] X. Gonze, Ph. Ghosez, and R. W. Godby, "Density-polarization functional theory of the response of a periodic insulating solid to an electric field," *Phys. Rev. Lett.* 74:4035–4038, 1995.
[341] R. M. Martin and G. Ortiz, "Functional theory of extended coulomb systems," *Phys. Rev. B* 56:1124–1140, 1997.
[342] R. M. Martin and G. Ortiz, "Recent developments in the theory of polarization in solids," *Solid State Commun.* 102:121–126, 1997.
[343] J. E. Harriman, "Orthonormal orbitals for the representation of an arbitrary density," *Phys. Rev. A* 24:680–682, 1981.
[344] W. A. Harrison, *Electronic Structure and the Properties of Solids*, Dover, New York, 1989.
[345] J. Harris, "Adiabatic-connection approach to Kohn-Sham theory," *Phys. Rev. A* 29:1648, 1984.
[346] O. Gunnarsson and B. I. Lundqvist, "Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism," *Phys. Rev. B* 13:4274, 1976.
[347] M. Levy and J. P. Perdew, "Hellmann–Feynman, virial, and scaling requisites for the exact universal density functionals, shape of the correlation potential and diamagnetic susceptibility for atoms," *Phys. Rev. A* 32:2010–2021, 1985.
[348] O. Gunnarsson, M. Jonson, and B. I. Lundqvist, "Descriptions of exchange and correlation effects in inhomogeneous electron systems," *Phys. Rev. B* 20:3136, 1979.
[349] R. Q. Hood, M. Y. Chou, A. J. Williamson, G. Rajagopal, R. J. Needs, and W. M. C. Foulkes, "Exchange and correlation in silicon." *Phys. Rev. B* 57:8972–8982, 1998.
[350] O. Gritsenko, R. van Leeuwen, and E. J. Baerends, "Analysis of electron interaction and atomic shell structure in terms of local potentials," *J. Chem. Phys.* 101:8455, 1994.
[351] J. P. Perdew and M. Levy, "Physical content of the exact Kohn–Sham orbital energies: Band gaps and derivative discontinuities," *Phys. Rev. Lett.* 51:1884–1887, 1983.
[352] L. J. Sham and M. Schlüter, "Density-functional theory of the energy gap," *Phys. Rev. Lett.* 51:1888–1891, 1983.
[353] C. Almbladh and U. von Barth, "Exact results for the charge and spin densities, exchange-correlation potentials, and density-functional eigenvalues," *Phys. Rev. B* 31:3231, 1985.
[354] M. Levy, J. P. Perdew, and V. Sahni, "Exact differential equation for the density and ionization energy of a many-particle system," *Phys. Rev. A* 30:2745, 1984.
[355] A. Gorling, "Density-functional theory for excited states," *Phys. Rev. A* 54:3912–3915, 1996.
[356] J. F. Janak, "Proof that $\partial e/\partial n_i = \epsilon_i$ in density-functional theory," *Phys. Rev. B* 18:7165, 1978.
[357] D. Mearns, "Inequivalence of physical and Kohn-Sham Fermi surfaces," *Phys. Rev. B* 38:5906, 1988.
[358] E. K. U. Gross, C. A. Ullrich, and U. J. Gossmann, in *Density Functional Theory*, edited by E. K. U. Gross and R. M. Dreizler, Plenum Press, New York, 1995, p. 149.

[359] I. Vasiliev, S. Ogut, and J. R. Chelikowsky, "First-principles density-functional calculations for optical spectra of clusters and nanocrystals," *Phys. Rev. B* 65:115416, 2002.
[360] D. J. Thouless and J. G. Valatin, "Time-dependent Hartree-Fock equations and rotational states of nuclei," *Nucl. Phys.* 31:211, 1962.
[361] T. Ando, "Density-functional calculation of sub-band structure in accumulation and inversion layers," *Phys. Rev. B* 13:3468–3477, 1976.
[362] G. Vignale and M. Rasolt, "Density-functional theory in strong magnetic field," *Phys. Rev. Lett.* 59:2360–2363, 1987.
[363] H. J. F. Jansen, "Many-body properties calculated from the Kohn-Sham equations in density-functional theory," *Phys. Rev. B* 43:12025, 1991.
[364] Y.-H. Kim and A. Gorling, "Excitonic optical spectrum of semiconductors obtained by time-dependent density-functional theory with the exact-exchange kernel," *Phys. Rev. Lett.* 89:096402, 2002.
[365] T. Grabo, T. Kreibich, S. Kurth, and E. K. U. Gross, in *Strong Coulomb Correlations in Electronic Structure: Beyond the Local Density Approximation*, edited by V. I. Anisimov, Gordon & Breach, Tokyo, 1998.
[366] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, "First principles calculations of the electronic structure and spectra of strongly correlated systems: The LDA + U method," *J. Phys.: Condensed Matter* 9:767–808, 1997.
[367] J. P. Perdew and K. Burke, "Comparison shopping for a gradient-corrected density functional," *Int. J. Quant. Chem.* 57:309–319, 1996.
[368] M. D. Towler, A. Zupan, and M. Causa, "Density functional theory in periodic systems using local gaussian basis sets," *Computer Physics Commun.* 98:181–205, 1996.
[369] F. Herman, J. P. Van Dyke, and I. P. Ortenburger, "Improved statistical exchange approximation for inhomogeneous many-electron systems," *Phys. Rev. Lett.* 22:807, 1969.
[370] P. S. Svendsen and U. von Barth, "Gradient expansion of the exchange energy from second-order density response theory," *Phys. Rev. B* 54:17402–17413, 1996.
[371] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," *Phys. Rev. A* 38:3098–3100, 1988.
[372] J. P. Perdew and Y. Wang, "Accurate and simple analytic representation of the electron-gas correlation energy," *Phys. Rev. B* 45:13244–13249, 1992.
[373] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.* 77:3865–3868, 1996.
[374] B. Hammer, L. B. Hansen, and J. K. Norskov, "Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals," *Phys. Rev. B* 59:7413–7421, 1999.
[375] S.-K. Ma and K. A. Brueckner, "Correlation energy of an electron gas with slowly varying high density," *Phys. Rev.* 165:18–31, 1968.
[376] C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," *Phys. Rev. B* 37:785–789, 1988.
[377] R. Colle and O. Salvetti, "Approximate calculation of the correlation energy for the closed and open shells," *Theo. Chim. Acta* 53:59–63, 1979.
[378] J. B. Krieger, Y. Chen, G. J. Iafrate, and A. Savin, "Construction of an accurate SIC-corrected correlation energy functional based on an electron gas with a gap," *preprint*, 2000.
[379] J. Rey and A. Savin, "Virtual space level shifting and correlation energies." *Int. J. Quant. Chem.* 69:581–587, 1998.

[380] D. R. Hamann, "Generalized gradient theory for silica phase transitions," *Phys. Rev. Lett.* 76(4):660–663, 1996.

[381] J. A. White and D. M. Bird, "Implementation of gradient-corrected exchange-correlation potentials in Car-Parrinello total-energy calculations," *Phys. Rev. B* 50:4954–4957, 1994.

[382] V. P. Antropov, M. I. Katsnelson, M. van Schilfgaarde, and B. N. Harmon, "*Ab initio* spin dynamics in magnets," *Phys. Rev. Lett.* 75:729–732, 1995.

[383] M. Uhl and J. Kübler, "Exchange-coupled spin-fluctuation theory: Application to Fe, Co, and Ni," *Phys. Rev. Lett.* 77:334–337, 1996.

[384] T. Oda, A. Pasquarello, and R. Car, "Fully unconstrained approach to noncollinear magnetism: Application to small fe clusters," *Phys. Rev. Lett.* 80:3622–3625, 1998.

[385] D. M. Bylander, Q. Niu, and L. Kleinman, "Fe magnon dispersion curve calculated with the frozen spin-wave method," *Phys. Rev. B* 61:R11875–R11878, 2000.

[386] Y.-H. Kim, I.-H. Lee S. Nagaraja, J. P. Leburton, R. Q. Hood, and R. M. Martin, "Two-dimensional limit of exchange-correlation energy functional approximations," *Phys. Rev. B* 61:5202–5211, 2000.

[387] A. Svane and O. Gunnarsson, "Localization in the self-interaction-corrected density-functional formalism," *Phys. Rev. B* 37:9919, 1988.

[388] A. Svane and O. Gunnarsson, "Transition-metal oxides in the self-interaction-corrected density functional formalism," *Phys. Rev. Lett.* 65:1148–1151, 1990.

[389] W. M. Temmerman, Z. Szotek, and H. Winter, "Self-interaction corrected electronic structure of La_2CuO_4," *Phys. Rev. B* 47, 11533–11536, 1993.

[390] A. Svane, Z. Szotek, W. M. Temmerman, J. Lægsgaard, and H. Winter, "Electronic structure of cerium pnictides under pressure," *J. Phys.: Condens. Matter*.

[391] V. I. Anisimov, J. Zaanen, and O. K. Andersen, "Band theory and Mott insulators: Hubbard U instead of Stoner I," *Phys. Rev. B* 44:943, 1991.

[392] J. Hubbard, "Electron correlations in narrow energy bands IV the atomic representation," *Proc. Roy. Soc. London, series A* 285:542, 1965.

[393] D. Baeriswyl, D. K. Campbell, J. M. P. Carmelo, and F. Guinea, *The Hubbard Model*, Plenum Press, New York, 1995.

[394] M. R. Norman, "Band theory and the insulating gap in CoO," *Phys. Rev. B* 40:10632–10634, 1989.

[395] L. J. Sham and M. Schlüter, "Density functional theory of the band gap," *Phys. Rev. B* 32:3883, 1985.

[396] R. T. Sharp and G. K. Horton, "A variational approach to the unipotential many-electron problem," *Phys. Rev.* 90:317, 1953.

[397] D. M. Bylander and L. Kleinman, "The optimized effective potential for atoms and semiconductors," *Int. J. Mod. Phys.* 10:399–425, 1996.

[398] J. B. Krieger, Y. Li, and G. J. Iafrate, "Exact relations in the optimized effective potential method employing an arbitrary $E_{xc}[\{\psi_{i\sigma}\}]$," *Phys. Lett. A* 148:470–473, 1990.

[399] J. B. Krieger, Y. Li, and G. J. Iafrate, "Construction and application of an accurate local spin-polarized Kohn-Sham potential with integer discontinuity: Exchange-only theory," *Phys. Rev. A* 45:101, 1992.

[400] J. B. Krieger, Y. Li, and G. J. Iafrate, in *Density Functional Theory*, edited by E. K. U. Gross and R. M. Dreizler, Plenum Press, New York, 1995, p. 191.

[401] J. C. Slater, "A simplification of the Hartree-Fock method," *Phys. Rev.* 81:385–390, 1951.

[402] M. Ernzerhof and G. E. Scuseria, "Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional," *J. Chem. Phys.* 98:5029–5036, 1999.

[403] A. D. Becke, "A new mixing of Hartree-Fock and local density-functional theories," *J. Chem. Phys.* 98:1372–1377, 1993.

[404] A. D. Becke, "Density functional thermochemistry III. The role of exact exchange," *J. Chem. Phys.* 98:5648–5652, 1993.

[405] J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations," *J. Chem. Phys.* 105:9982–9985, 1996.

[406] C. Filippi, C. J. Umrigar, and X. Gonze, "Excitation energies from density functional perturbation theory," *J. Chem. Phys.* 107:9994–10002, 1997.

[407] S. Kurth, J. P. Perdew, and P. Blaha, "Molecular and solid-state tests of density functional approximations: LSD, GGAs, and meta-GGAs," *Int. J. Quantum Chem.* 75:889, 1999.

[408] W. Kolos and L. Wolniewicz, "Potential-energy curves for the X $^1\sigma_g^+$, b $^3\sigma_u^+$, and C $^1\pi_u$ states of the hydrogen molecule," *J. Chem. Phys.* 43:2429, 1965.

[409] C. O. Almbladh and A. C. Pedroza, "Density-functional exchange-correlation potentials and orbital eigenvalues for light atoms," *Phys. Rev. A* 29:2322–2330, 1984.

[410] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, "Assessment of Gaussian-2 and density functional methods for the computation of enthalpies of formation," *J. Chem. Phys.* 106:1063, 1997.

[411] D. C. Patton, D. V. Porezag, and M. R. Pederson, "Simplified generalized-gradient approximation and anharmonicity: Benchmark calculations on molecules," *Phys. Rev. B* 55:7454–7459, 1997.

[412] A. Zupan, P. Blaha, K. Schwarz, and J. P. Perdew, "Pressure-induced phase transitions in solid Si, SiO_2, and Fe: Performance of local-spin-density and generalized-gradient-approximation density functionals," *Phys. Rev. B* 58:11266, 1998.

[413] W. E. Pickett, "Pseudopotential methods in condensed matter applications," *Computer Physics Reports* 9:115, 1989.

[414] D. J. Singh, *Planewaves, Pseudopotentials, and the APW Method*, Kluwer Academic Publishers, Boston, 1994, and references therein.

[415] J. Harris, "Simplified method for calculating the energy of weakly interacting fragments," *Phys. Rev. B* 31:1770–1779, 1985.

[416] M. Weinert, R. E. Watson, and J. W. Davenport, "Total-energy differences and eigenvalue sums," *Phys. Rev. B* 32:2115–2119, 1985.

[417] W. M. C. Foulkes and R. Haydock, "Tight-binding models and density-functional theory," *Phys. Rev. B* 39:12520–12536, 1989.

[418] Otto F. Sankey and David J. Niklewski, "*Ab initio* multicenter tight-binding model for molecular dynamics simulations and other applications in covalent systems," *Phys. Rev. B* 40:3979–3995, 1989.

[419] M. Methfessel, "Independent variation of the density and potential in density functional methods," *Phys. Rev. B* 52:8074, 1995.

[420] A. J. Read and R. J. Needs, "Tests of the Harris energy functional," *J. Phys. Cond. Matter* 1:7565, 1989.

[421] E. Zaremba, "Extremal properties of the Harris energy functional," *J. Phys. Cond. Matter* 2:2479, 1990.

[422] I. J. Robertson and B. Farid, "Does the Harris energy functional possess a local maximum at the ground-state density?" *Phys. Rev. Lett.* 66:3265–3268, 1991.

[423] K. W. Jacobsen, J. K. Norskov, and M. J. Puska, "Interatomic interactions in the effective-medium theory," *Phys. Rev. B* 35:7423–7442, 1987.

[424] D. M. C. Nicholson, G. M. Stocks, Y. Wang, W. A. Shelton, Z. Szotek, and W. M. Temmerman, "Stationary nature of the density-functional free energy: Application to accelerated multiple-scattering calculations," *Phys. Rev. B* 50:14686–14689, 1994.

[425] M. J. Gillan, "Calculation of the vacancy formation energy in aluminum," *J. Phys.: Condens. Matter* 1:689, 1989.

[426] N. Marzari, D. Vanderbilt, and M. C. Payne, "Ensemble density-functional theory for *ab initio* molecular dynamics of metals and finite-temperature insulators," *Phys. Rev. Lett.* 79:1337–1340, 1997.

[427] P. H. Dederichs and R. Zeller, "Self-consistency iterations in electronic-structure calculations," *Phys. Rev. B* 28:5462, 1983.

[428] K.-M. Ho, J. Ihm, and J. D. Joannopoulos, "Dielectric matrix scheme for fast convergence in self-consistent electronic-structure calculations," *Phys. Rev. B* 25:4260–4262, 1982.

[429] P. Bendt and A. Zunger, "New approach for solving the density-functional self-consistent-field problem," *Phys. Rev. B* 26:3114–3137, 1982.

[430] G. P. Srivastava, "Broyden's method for self-consistent field convergence acceleration," *J. Phys. A* 17:L317, 1984.

[431] C. G. Broyden, "A class of methods for solving nonlinear simulataneous equations," *Math. Comput.* 19:577–593, 1965.

[432] D. Singh, H. Krakauer, and C. S. Wang, "Accelerating the convergence of self-consistent linearized augmented-plane-wave calculations," *Phys. Rev. B* 34:8391–8393, 1986.

[433] D. Vanderbilt and S. G. Louie, "Total energies of diamond (111) surface reconstructions by a linear combination of atomic orbitals method," *Phys. Rev. B* 30:6118, 1984.

[434] D. D. Johnson, "Modified Broyden's method for accelerating convergence in self-consistent calculations," *Phys. Rev. B* 38:12807–12813, 1988.

[435] D. G. Anderson, "Iterative procedures for non-linear integral equations," *Assoc. Comput. Mach.* 12:547, 1965.

[436] P. Pulay, "Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules. I. theory," *Mol. Phys.* 17:197–204, 1969.

[437] M. Allen and D. Tildesley, *Computer simulation of liquids*, Oxford University Press, New York, Oxford, 1989.

[438] M. Parrinello and A. Rahman, "Crystal structure and pair potentials: A molecular-dynamics study," *Phys. Rev. Lett.* 45:1196–1199, 1980.

[439] I. Souza and J. L. Martins, "Metric tensor as the dynamical variable for variable-cell-shape molecular dynamics," *Phys. Rev. B* 55:8733–8742, 1997.

[440] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, "Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients," *Rev. Mod. Phys.* 64:1045–1097, 1992.

[441] C. F. Fischer, *The Hartree-Fock Method for Atoms: A Numerical Approach*, John Wiley and Sons, New York, 1977.

[442] J. C. Slater, *Quantum Theory of Atomic Structure, Vol. 1*, McGraw-Hill, New York, 1960.

[443] J. C. Slater, *Quantum Theory of Atomic Structure, Vol. 2*, McGraw-Hill, New York, 1960.

[444] S. E. Koonin and D. C. Meredith, *Computational Physics*, Addison Wesley, Menlo Park, CA, 1990.

[445] F. Herman and S. Skillman, *Atomic Structure Calculations*, Prentice-Hall, Engelwood Cliffs, N. J., 1963.

[446] D. D. Koelling and B. N. Harmon, "A technique for relativistic spin-polarized calculations," *J. Phys. C* 10:3107–3114, 1977.

[447] A. H. MacDonald, W. E. Pickett, and D. Koelling, "A linearised relativistic augmented-plane-wave method utilising approximate pure spin basis functions," *J. Phys. C: Solid State Phys.* 13:2675–2683, 1980.

[448] J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1962.

[449] M. S. Hybertsen and S. G. Louie, "Spin-orbit splitting in semiconductors and insulators from the ab initio pseudopotential," *Phys. Rev. B* 34:2920, 1986.

[450] G. Theurich and N. A. Hill, "Self-consistent treatment of spin-orbit coupling in solids using relativistic fully separable ab initio pseudopotentials," *Phys. Rev. B* 64:073106, 2001.

[451] P. A. M. Dirac, "The quantum theory of the electron, Part II," *Proc. Roy. Soc. London* Ser. A 118:351, 1928.

[452] A. Messiah, *Quantum Mechanics, Vol. II*, Wiley, New York, 1964.

[453] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics*, McGraw-Hill, New York, 1964.

[454] F. R. Vukajlovic, E. L. Shirley, and R. M. Martin, "Single-body methods in 3d transition-metal atoms." *Phys. Rev. B* 43:3994, 1991.

[455] J. C. Slater, *The Self-Consistent Field Theory for Molecules and Solids: Quantum Theory of Molecules and Solids, Vol. 4*, McGraw-Hill, New York, 1974.

[456] A. K. McMahan, R. M. Martin, and S. Satpathy, "Calculated effective hamiltonian for La_2CuO_4 and solution in the impurity Anderson approximation," *Phys. Rev. B* 38:6650, 1988.

[457] J. F. Herbst, R. E. Watson, and J. W. Wilkins, "Relativistic calculations of 4f excitation energies in the rare-earth metals: Further results," *Phys. Rev. B* 17:3089–3098, 1978.

[458] M. S. Hybertsen and M. Schlüter and N. E. Christensen, "Calculation of coulomb interaction parameters for La_2CuO_4 using a constrained-density-functional approach," *Phys. Rev. B* 39:9028, 1989.

[459] O. K. Andersen and O. Jepsen, "Explicit, first-principles tight-binding theory," *Physica* 91B:317, 1977.

[460] G. K. Straub and Walter A. Harrison, "Analytic methods for the calculation of the electronic structure of solids," *Phys. Rev. B* 31:7668–7679, 1985.

[461] O. K. Andersen, "Simple approach to the band structure problem," *Solid State Commun.* 13:133–136, 1973.

[462] D. A. Liberman, "Virial theorem in self-consistent-field calculations," *Phys. Rev. B* 3:2081–2082, 1971.

[463] J. F. Janak, "Simplification of total-energy and pressure calculations in solids," *Phys. Rev. B* 9:3985–3988, 1974.

[464] A. R. Mackintosh and O. K. Andersen, in *Electrons at the Fermi Surface*, edited by M. Springford, Cambridge Press, Cambridge, 1975, p. 149.

[465] V. Heine, in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1980, Vol. 35, p. 1.

[466] J. M. Ziman, *Principles of the Theory of Solids*, Cambridge University Press, Cambridge, 1989.

[467] V. Heine, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1970, p. 1.

[468] W. A. Harrison, *Pseudopotentials in the Theory of Metals*, Benjamin, New York, 1966.

[469] M. L. Cohen and V. Heine, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1970, p. 37.
[470] M. L. Cohen and J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors*, 2nd ed., Springer-Verlag, Berlin, 1988.
[471] D. R. Hamann, M. Schlüter, and C. Chiang, "Norm-conserving pseudopotentials," *Phys. Rev. Lett.* 43:1494–1497, 1979.
[472] L. Kleinman and D. M. Bylander, "Efficacious form for model pseudopotentials," *Phys. Rev. Lett.* 48:1425–1428, 1982.
[473] P. E. Blöchl, "Generalized separable potentials for electronic-structure calculations," *Phys. Rev. B* 41:5414–5416, 1990.
[474] D. Vanderbilt, "Soft self-consistent pseudopotentials in a generalized eigenvalue formalism," *Phys. Rev. B* 41:7892, 1990.
[475] P. E. Blöchl, "Projector augmented-wave method," *Phys. Rev. B* 50:17953–17979, 1994.
[476] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," *Phys. Rev. B* 59:1758–1775, 1999.
[477] E. Amaldi, O. D'Agostino, E. Fermi, B. Pontecorvo, F. Rasetti, and E. Segre, "Artificial radioactivity induced by neutron bombardment – II," *Proc. Ry. Soc. (London) Series A* 149:522–558, 1935.
[478] J. Callaway, "Electron energy bands in sodium," *Phys. Rev.* 112:322, 1958.
[479] E. Antoncik, "A new formulation of the method of nearly free electrons," *Czech. J. Phys.* 4:439, 1954.
[480] E. Antoncik, "Approximate formulation of the orthogonalized plane-wave method," *J. Phys. Chem. Solids* 10:314, 1959.
[481] J. C. Phillips and L. Kleinman, "New method for calculating wave functions in crystals and molecules," *Phys. Rev.* 116:287, 1959.
[482] W. C. Herring and A. G. Hill, "The theoretical constitution of metallic beryllium," *Phys. Rev.* 58:132, 1940.
[483] F. Herman, "Calculation of the energy band structures of the diamond and germanium crystals by the method of orthogonalized plane waves," *Phys. Rev.* 93:1214, 1954.
[484] T. O. Woodruff, "Solution of the Hartree-Fock-Slater equations for silicon crystal by the method of orthogonalized plane waves," *Phys. Rev.* 98:1741, 1955.
[485] F. Herman, "Speculations on the energy band structure of Ge-Si alloys," *Phys. Rev.* 95:847, 1954.
[486] F. Bassani, "Energy band structure in silicon crystals by the orthogonalized plane-wave method," *Phys. Rev.* 108:263–264, 1957.
[487] B. Lax, "Experimental investigations of the electronic band structure of solids," *Rev. Mod. Phys.* 30:122, 1958.
[488] M. H. Cohen and V. Heine, "Cancellation of kinetic and potential energy in atoms, molecules, and solids," *Phys. Rev.* 122:1821, 1961.
[489] N. W. Ashcroft, "Electron-ion pseudopotentials in metals," *Phys. Lett.* 23:48–53, 1966.
[490] I. V. Abarenkov and V. Heine, "The model potential for positive ions," *Phil. Mag.* 12:529, 1965.
[491] A. O. E. Animalu, "Non-local dielectric screening in metals," *Phil. Mag.* 11:379, 1965.
[492] A. O. E. Animalu and V. Heine, "The screened model potential for 25 elements," *Phil. Mag.* 12:1249, 1965.
[493] P. A. Christiansen, Y. S. Lee, and K. S. Pitzer, "Improved *ab initio* effective core potentials for molecular calculations," *J. Chem. Phys.* 71:4445–4450, 1979.

[494] M. Krauss and W. J. Stevens, "Effective potentials in molecular quantum chemistry," *Ann. Rev. Phys. Chem* 35:357, 1984.

[495] W. C. Topp and J. J. Hopfield, "Chemically motivated pseudopotential for sodium," *Phys. Rev. B* 7:1295–1303, 1973.

[496] E. Engel, A. Hock, R. N. Schmid, R. M.Dreizler, and N. Chetty, "Role of the core-valence interaction for pseudopotential calculations with exact exchange," *Phys. Rev. B* 64:125111–125122, 2001.

[497] E. L. Shirley, D. C. Allan, R. M. Martin, and J. D. Joannopoulos, "Extended norm-conserving pseudopotentials." *Phys. Rev. B* 40:3652, 1989.

[498] G. Lüders, "Zum zusammenhang zwischen S-Matrix und Normierungsintegrassen in der Quantenmechanik," *Z. Naturforsch.* 10a:581, 1955.

[499] G. B. Bachelet, D. R. Hamann, and M. Schlüter, "Pseudopotentials that work: From H to Pu," *Phys. Rev. B* 26:4199, 1982.

[500] D. Vanderbilt, "Optimally smooth norm-conserving pseudopotentials," *Phys. Rev. B* 32:8412, 1985.

[501] G. P. Kerker, "Non-singular atomic pseudopotentials for solid state applications," *J. Phys. C* 13:L189, 1980.

[502] N. Troullier and J. L. Martins, "Efficient pseudopotentials for plane-wave calculations," *Phys. Rev. B* 43:1993–2006, 1991.

[503] A. M. Rappe, K. M. Rabe, E. Kaxiras, and J. D. Joannopoulos, "Optimized pseudopotentials," *Phys. Rev. B* 41:1227, 1990.

[504] G. Kresse, J. Hafner, and R. J. Needs, "Optimized norm-conserving pseudopotentials," *J. Phys.: Condens. Matter* 4:7451, 1992.

[505] C. W. Greeff and W. A. Lester, Jr., "A soft Hartree-Fock pseudopotential for carbon with application to quantum Monte Carlo," *J. Chem. Phys.* 109:1607–1612, 1998.

[506] I. Ovcharenko, A. Aspuru-Guzik, and W. A. Lester, Jr., "Soft pseudopotentials for efficient quantum Monte Carlo calculations: From Be to Ne and Al to Ar," *J. Chem. Phys.* 114:7790–7794, 2001.

[507] S. G. Louie, S. Froyen, and M. L. Cohen, "Nonlinear ionic pseudopotentials in spin-density-functional calculations," *Phys. Rev. B* 26:1738–1742, 1982.

[508] S. Goedecker and K. Maschke, "Transferability of pseudopotentials," *Phys. Rev. A* 45:88–93, 1992.

[509] M. Teter, "Additional condition for transferability in pseudopotentials," *Phys. Rev. B* 48:5031–5041, 1993.

[510] A. Filippetti, David Vanderbilt, W. Zhong, Yong Cai, and G. B. Bachelet, "Chemical hardness, linear response, and pseudopotential transferability," *Phys. Rev. B* 52:11793–11804, 1995.

[511] X. Gonze, R. Stumpf, and M. Scheffler, "Analysis of separable potentials," *Phys. Rev. B* 44:8503, 1991.

[512] N. A. W. Holzwarth, G. E. Matthews, A. R. Tackett, and R. B. Dunning, "Comparison of the projector augmented-wave, pseudopotential, and linearized augmented-plane-wave formalisms for density-functional calculations of solids," *Phys. Rev. B* 55:2005–2017, 1997.

[513] P. E. Blöchl, 'The Projector Augmented Wave method: Algortithm and Results', Conference of the Asian Consortium for Computational Materials Science, Bangalore, India, 2001.

[514] S. Baroni and R. Resta, "*Ab initio* calculation of the macroscopic dielectric constant in silicon," *Phys. Rev. B* 33:7017, 1986.

[515] M. S. Hybertsen and S. G. Louie, "*Ab initio* static dielectric matrices from the density-functional approach. I. formulation and application to semiconductors and insulators," *Phys. Rev. B* 35:5585, 1987.

[516] C. P. Slichter, *Principles of Magnetic Resonance, Third Ed.*, Springer Verlag, Berlin, 1996.

[517] F. Mauri, B. G. Pfrommer, and S. G. Louie, "*Ab initio* theory of NMR chemical shifts in solids and liquids," *Phys. Rev. Lett.* 77:5300–5303, 1996.

[518] T. Gregor, F. Mauri, and R. Car, "A comparison of methods for the calculation of NMR chemical shifts," *J. Chem. Phys.* 111:1815–1822, 1999.

[519] G. B. Bachelet, D. M. Ceperley, and M. G. B. Chiocchetti, "Novel pseudo-hamiltonian for quantum Monte Carlo simulations," *Phys. Rev. Lett.* 62:2088–2091, 1989.

[520] M. W. C. Foulkes and M. Schlüter, "Pseudopotentials with position-dependent electron masses," *Phys. Rev. B* 42:11505–11529, 1990.

[521] A. Bosin, V. Fiorentini, A. Lastri, and G. B. Bachelet, "Local norm-conserving pseudo-hamiltonians," *Phys. Rev. A* 52:236, 1995.

[522] E. L. Shirley and R. M. Martin, "*GW* quasiparticle calculations in atoms," *Phys. Rev. B* 47:15404–15412, 1993.

[523] E. L. Shirley and R. M. Martin, "Many-body core-valence partitioning," *Phys. Rev. B* 47:15413–15427, 1993.

[524] M. Dolg, U. Wedig, H. Stoll, and H. Preuss, "Energy-adjusted *ab initio* pseudopotentials for the first row transition elements," *J. Chem. Phys.* 86:866–872, 1987.

[525] T. L. Beck, "Real-space mesh techniques in density-functional theory," *Rev. Mod. Phys.* 72:1041–1080, 2000.

[526] J. R. Chelikowsky, N. Troullier, and Y. Saad, "Finite-difference-pseudopotential method: Electronic structure calculations without a basis," *Phys. Rev. Lett.* 72:1240–1243, 1994.

[527] B. Segall, "Energy bands of aluminum," *Phys. Rev.* 124:1797–1806, 1961.

[528] V. Heine, "The band structure of aluminum III. A self-consistent calculation," *Proc. Roy. Soc. (London)* A240:361, 1957.

[529] V. Heine and D. Weaire, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1970, p. 249.

[530] J. Ihm, A. Zunger, and M. L. Cohen, "Momentum-space formalism for the total energy of solids," *J. Phys. C* 12:4409, 1979.

[531] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*, Springer-Verlag, Berlin, 1996.

[532] S. B. Zhang, C.-Y. Yeh, and A. Zunger, "Electronic structure of semiconductor quantum films," *Phys. Rev. B* 48:11204–11219, 1993.

[533] L. W. Wang and A. Zunger, "Solving Schrödinger's equation around a desired energy: Application to silicon quantum dots," *J. Chem. Phys.* 48:2394–2397, 1994.

[534] L. W. Wang, J. Kim, and A. Zunger, "Electronic structures of [110]-faceted self-assembled pyramidal InAs/GaAs quantum dots," *Phys. Rev. B* 59:5678–5687, 1999.

[535] J. R. Chelikowsky, N. Troullier, Y. Saad, and K. Wu, "Higher-order finite-difference pseudopotential method: An application to diatomic molecules," *Phys. Rev. B* 50:11355–11364, 1994.

[536] B. Fornberg and D. Sloan, in *Acta Numerica 94*, edited by A. Iserles, Cambridge Press, Cambridge, 1994, pp. 203–267.

[537] L. Collatz, *The Numerical Treatment of Differential Equations, 3rd ed.*, Springer-Verlag, Berlin, 1960.

[538] E. L. Briggs, D. J. Sullivan, and J. Bernholc, "Large-scale electronic-structure calculations with multigrid acceleration," *Phys. Rev. B* 52:R5471–R5474, 1995.

[539] F. Gygi and G. Galli, "Real-space adaptive-coordinate electronic-structure calculations," *Phys. Rev. B* 52:R2229–R2232, 1995.

[540] N. A. Modine, Gil Zumbach, and E. Kaxiras, "Adaptive-coordinate real-space electronic-structure calculations for atoms, molecules, and solids," *Phys. Rev. B* 55:10289–10301, 1997.

[541] Y.-H. Kim, M. Staedele, and R. M. Martin, "Density-functional study of small molecules within the Krieger-Li-Iafrate approximation," *Phys. Rev. A* 60:3633–3640, 1999.

[542] I.-H. Lee, Y.-H. Kim, and R. M. Martin, "One-way multigrid method in electronic-structure calculations," *Phys. Rev. B* 61:4397–4400, 2000.

[543] S. C. Brenner and L. R. Scott, *A Multigrid Tutorial*, Springer, New York, 1994.

[544] J. E. Pask, B. M. Klein, C. Y. Fong, and P. A. Sterne, "Real-space local polynomial basis for solid-state electronic-structure calculations: A finite-element approach," *Phys. Rev. B* 59:12352–12358, 1999.

[545] J. E. Pask, B. M. Klein, P. A. Sterne, and C. Y. Fong, "Finite-element methods in electronic-structure theory," *Comput. Phys. Commun.* 135:1, 2001.

[546] E. Hernandez, M. J. Gillan, and C. M. Goringe, "Basis functions for linear-scaling first-principles calculations," *Phys. Rev. B* 55:13485–13493, 1997.

[547] E. Tsuchida and M. Tsukada, "Electronic-structure calculations based on the finite-element method," *Phys. Rev. B* 52:5573–5578, 1995.

[548] W. L. Briggs, *A Multigrid Tutorial*, SIAM, Philadelphia, 1987.

[549] A. Brandt, "Multiscale scientific computation: Six year research summary," *Available at www.wisdom.weizmann.ac.il/ achi.*, 1999.

[550] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[551] C. K. Chui, *Wavelets: Theory, Algorithms, and Applications*, Academic, San Diego, 1994.

[552] M. Heiskanen, T. Torsti, M. J. Puska, and R. M. Nieminen, "Multigrid method for electronic structure calculations," *Phys. Rev. B* 63:245106, 2001.

[553] T. A. Arias, "Multiresolution analysis of electronic structure: semicardinal and wavelet bases," *Rev. Mod. Phys.* 71:267–311, 1999.

[554] F. Gygi, "Electronic-structure calculations in adaptive coordinates," *Phys. Rev. B* 48:11692–11700, 1993.

[555] D. R. Hamann, "Application of adaptive curvilinear coordinates to the electronic structure of solids," *Phys. Rev. B* 51(11):7337–7340, 1995.

[556] D. R. Hamann, "Comparison of global and local adaptive coordinates for density-functional calculations," *Phys. Rev. B* 63:075107, 2001.

[557] P. J. H. Denteneer and W. van Haaringen, "The pseudopotential-density-functional method in momentum space: details and test cases," *J. Phys. C* 18:4127, 1985.

[558] G. P. Srivastava and D. Weaire, "The theory of the cohesive energy of solids," *Advances in Physics* 36:463–517, 1987.

[559] N. A. W. Holzwarth, A. R. Tackett, and G. E. Matthews, "A projector augmented wave (PAW) code for electronic structure calculations, part I: atompaw for generating atom-centered functions," *Comp. Phys. Commun.* 135:329–347, 2001.

[560] N. A. W. Holzwarth, A. R. Tackett, and G. E. Matthews, "A projector augmented wave (PAW) code for electronic structure calculations, part II: pwpaw for periodic solids in a plane wave basis," *Comp. Phys. Commun.* 135:348–376, 2001.

[561] M. T. Yin and M. L. Cohen, "Theory of *ab initio* pseudopotential calculations," *Phys. Rev. B* 25:7403–7412, 1982.
[562] W. C. Herring and M. H. Nichols, "Thermionic emission," *Rev. Mod. Phys.* 21:185–270, 1949.
[563] K. Kunc and R. M. Martin, "Atomic structure and properties of polar Ge-GaAs(100) interfaces," *Phys. Rev. B* 24(6):3445–3455, 1981.
[564] E. Wimmer, H. Krakauer, M. Weinert, and A. J. Freeman, "Full-potential self-consistent linearized-augmented-plane-wave method for calculating the electronic structure of molecules and surfaces: O_2 molecule," *Phys. Rev. B* 24:864–875, 1981.
[565] K. Laasonen, R. Car, C. Lee, and D. Vanderbilt, "Implementation of ultrasoft pseudopotentials in *ab initio* molecular dynamics," *Phys. Rev. B* 43:6796, 1991.
[566] L. Stixrude, R. E. Cohen, and D. J. Singh, "Iron at high pressure: Linearized-augmented-plane-wave computations in the generalized-gradient approximation," *Phys. Rev. B* 50:6442–6445, 1994.
[567] A. F. Goncharov, E. Gregoryanz, H.-K. Mao amd Z. Liu, and R. J. Hemley, "Optical evidence for a nonmolecular phase of nitrogen above 150 GPa," *Phys. Rev. Lett.* 85:1262–1265, 2000.
[568] J. Cho and M. Scheffler, "*Ab initio* pseudopotential study of Fe, Co, and Ni employing the spin-polarized LAPW approach," *Phys. Rev. B* 53:10685–10689, 1996.
[569] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation," *Phys. Rev. B* 46(11):6671–6687, 1992.
[570] F. Gygi, "Adaptive riemannian metric for plane-wave electronic-structure calculations," *Europhys. Lett.* 19:617–620, 1992.
[571] G. B. Bachelet and N. E. Christensen, "Relativistic and core-relaxation effects on the energy bands of gallium arsenide and germanium," *Phys. Rev. B* 31:879–887, 1985.
[572] Ph. Ghosez, J.-P. Michenaud, and X. Gonze, "Dynamical atomic charges: The case of ABO_3 compounds," *Phys. Rev. B* 58:6224–6240, 1998.
[573] W. Zhong, R. D. King-Smith, and D. Vanderbilt, "Giant LO-TO splittings in perovskite ferroelectrics," *Phys. Rev. Lett.* 72:3618–3621, 1994.
[574] K. Kunc and R. M. Martin, "*Ab initio* force constants in GaAs: a new approach to calculation of phonons and dielectric properties," *Phys. Rev. Lett.* 48(6):406–409, 1982.
[575] M. T. Yin and M. L. Cohen, "*Ab initio* calculation of the phonon dispersion relation: Application to Si," *Phys. Rev. B* 25:4317–4320, 1982.
[576] S. Wei and M. Y. Chou, "*Ab initio* calculation of force constants and full phonon dispersions," *Phys. Rev. Lett.* 69:2799–2802, 1992.
[577] N. Marzari and D. J. Singh, "Dielectric response of oxides in the weighted density approximation," *Phys. Rev. B* 62:12724–12729, 2000.
[578] C. G. Van de Walle and R. M. Martin, "Theoretical study of Si/Ge interfaces," *J. Vac. Sci. Tech. B* 3(4):1256–1259, 1985.
[579] A. Baldereschi, S. Baroni, and R. Resta, "Band offsets in lattice-matched heterojunctions: A model and first-principles calculations for GaAs/AlAs," *Phys. Rev. Lett.* 61:734–737, 1988.
[580] N. E. Christensen, "Dipole effects and band offsets at semiconductor interfaces," *Phys. Rev. B* 37:4528, 1988.
[581] W. R. L. Lambrecht, B. Segall, and O. K. Andersen, "Self-consistent dipole theory of heterojunction band offsets," *Phys. Rev. B* 41:2813, 1990.
[582] W. A. Harrison, E. A. Kraut, J. R. Waldrop, and R. W. Grant, "Polar heterojunction interfaces," *Phys. Rev. B* 18:4402–4410, 1978.

[583] G. Bratina, L. Vanzetti, L. Sorba, G. Biasiol, A. Franciosi, M. Peressi, and S. Baroni, "Lack of band-offset transitivity for semiconductor heterojunctions with polar orientation: ZnSe-Ge(001); Ge-GaAs(001); and ZnSe-GaA(001)," *Phys. Rev. B* 50:11723–11729, 1994.

[584] A. A. Stekolnikov, J. Furthmüller, and F. Bechstedt, "Absolute surface energies of group-IV semiconductors: Dependence on orientation and reconstruction," *Phys. Rev. B* 65:115318, 2002.

[585] E. Penev, P. Kratzer, and M. Scheffler, "Effect of the cluster size in modeling the H2 desorption and dissociative adsorption on Si(001)," *J. Chem. Phys.* 110:3986–3994, 1999.

[586] S. B. Healy, C. Filippi, P. Kratzer, E. Penev, and M. Scheffler, "Role of electronic correlation in the Si(100) reconstruction: A quantum Monte Carlo study," *Phys. Rev. Lett.* 87:016105, 2001.

[587] M. Rohlfing, P. Krüger, and J. Pollmann, "Quasiparticle band structures of clean, hydrogen- and sulfur-terminated Ge(001) surfaces," *Phys. Rev. B* 54:13759–13766, 1996.

[588] M. Machon, S. Reich, C. Thomsen, D. Sanchez-Portal, and P. Ordejon, "*Ab initio* calculations of the optical properties of 4-a-diameter single-walled nanotubes," *Phys. Rev. B* 66:155410, 2002.

[589] J. C. Slater and G. F. Koster, "Simplified LCAO method for the periodic potential problem," *Phys. Rev.* 94:1498–1524, 1954.

[590] W. A. Harrison, *Elementary Electronic Structure*, World Publishing, Singapore, 1999.

[591] D. A. Papaconstantopoulos, *Handbook of Electronic Structure of Elemental Solids*, Plenum, New York, 1986.

[592] *Tight-Binding Approach to Computational Materials Science*, edited by P. E. A. Turchi, A. Gonis, and L. Columbo, Materials Research Society, Warrendale PA, 1998.

[593] C. M. Goringe, D. R. Bowler, and E. Hernandez, "Tight-binding modelling of materials," *Rep. Prog. Phys.* 60:1447–1512, 1997.

[594] H. Jones, N. Mott, and Skinner, "A theory of the form of the x-ray emission bands of metals," *Phys. Rev.* 45:379, 1934.

[595] M. D. Stiles, "Generalized Slater–Koster method for fitting band structures," *Phys. Rev. B* 55:4168–4173, 1997.

[596] P. Vogl, H. P. Hjalmarson, and J. D. Dow, "A semi-empirical tight-binding theory of the electronic structure of semiconductors," *Europhys. Lett.* 44:365, 1983.

[597] N. Bernstein, M. J. Mehl, D. A. Papaconstantopoulos, N. I. Papanicolaou, M. Z. Bazant, and E. Kaxiras, "Energetic, vibrational, and electronic properties of silicon using a nonorthogonal tight-binding model," *Phys. Rev. B* 62:4477–4487, 2000.

[598] S. G. Louie, in *Carbon Nanotubes, Topics Appl. Phys.*, edited by M. S. Dresselhaus, G. Dresselhaus, and Ph. Avouris, Springer-Verlag, Berlin, 2001, Vol. 80, pp. 113–145.

[599] D. A. Papaconstantopoulos, M. J. Mehl, J. C. Erwin, and M. R. Pederson, in *Tight-Binding Approach to Computational Materials Science*, edited by P. E. A. Turchi, A. Gonis, and L. Columbo, Materials Research Society, Warrendale PA, 1998.

[600] E. J. Mele and P. Kral, "Electric polarization of heteropolar nanotubes as a geometric phase," *Phys. Rev. Lett.* 88:056803, 2002.

[601] O. F. Sankey and D. J. Niklewski, "*Ab initio* multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems," *Phys. Rev. B* 40:3979, 1989.

[602] R. E. Cohen, M. J. Mehl, and D. A. Papaconstantopoulos, "Tight-binding total-energy method for transition and noble metals," *Phys. Rev. B* 50(19):14694–14697, 1994.

[603] D. Porezag, Th. Frauenheim, Th. Köhler, G. Seifert, and R. Kaschner, "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon," *Phys. Rev. B* 51:12947–12957, 1995.

[604] L. Goodwin, A. J. Skinner, and D. G. Pettifor, "Generating transferable tight-binding parameters: application to silicon," *Europhys. Lett.* 9:701, 1989.

[605] I. Kwon, R. Biswas, C. Z. Wang, K. M. Ho, and C. M. Soukoulis, "Transferable tight-binding models for silicon," *Phys. Rev. B* 49:7242, 1994.

[606] T. J. Lenosky, J. D. Kress, I. Kwon, A. F. Voter, B. Edwards, D. F. Richards, S. Yang, and J. B. Adams, "Highly optimized tight-binding model of silicon," *Phys. Rev. B* 55:1528–1544, 1997.

[607] C. Z. Wang, B. C. Pan, and K. M. Ho, "An environment-dependent tight-binding potential for Si," *J. Phys.: Condens. Matter* 11:2043–2049, 1999.

[608] J. Kim, J. W. Wilkins, F. S. Khan, and A. Canning, "Extended Si —P[311—P] defects," *Phys. Rev. B* 55:16186, 1997.

[609] J. Kim, F. Kirchhoff, J. W. Wilkins, and F. S. Khan, "Stability of Si-interstitial defects: From point to extended defects," *Phys. Rev. Lett.* 84:503, 2000.

[610] N. C. Bacalis, D. A. Papaconstantopoulos, M. J. Mehl, and M. Lachhab, "Transferable tight-binding parameters for ferromagnetic and paramagnetic iron," *Physica B* 296(1–3):125–129, 2001.

[611] F. Jensen, *An Introduction to Computational Chemistry*, John Wiley and Sons, New York, 1998.

[612] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, Wiley, New York, 2002.

[613] H. Eschrig, *Optimized LCAO Methods*, Springer, Berlin, 1987.

[614] R. Orlando, R. Dovesi, C. Roetti, and V. R. Saunders, "*Ab initio* Hartree-Fock calculations for periodic compounds: application to semiconductors," *J. Phys. Condens. Matter* 2:7769, 1990.

[615] V. R. Saunders, R. Dovesi, C. Roetti, M. Causa, N. M. Harrison, R. Orlando, and C. M. Zicovich-Wilson, CRYSTAL 98 User's Manual (University of Torino, Torino). See http://www.theochem.unito.it/, , 2003.

[616] B. Delley, "From molecules to solids with the DMol3 approach," *J. Chem. Phys.* 113:7756–7764, 2000.

[617] J. M. Soler, E. Artacho, J. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal, "The SIESTA method for *ab intio* order-N materials simulations," *J. Phys. : Condens. Matter* 14:2745–2779, 2002.

[618] S. F. Boys, "Electron wave functions I. A general method for calculation for the stationary states of any molecular system," *Proc. Roy. Soc. London, series A* 200:542, 1950.

[619] G. E. Scuseria, "Linear scaling density functional calculations with gaussian orbitals," *J. Phys. Chem. A* 103:4782–4790, 1999.

[620] J. K. Perry, J. Tahir-Kheli, and W. A. Goddard, "Antiferromagnetic band structure of La_2CuO_4: Becke-3-Lee-Yang-Parr calculations," *Phys. Rev. B* 63:144510, 2001.

[621] K. N. Kudin, G. E. Scuseria, and R. L. Martin, "Hybrid density-functional theory and the insulating gap of UO_2," *Phys. Rev. Lett.* 89:266402, 2002.

[622] J. Muscat, A. Wander, and N. M. Harrison, "On the predivction of band gaps from hybrid density-functional theory," *Chem. Phys. Lett.* 342:397, 2001.

[623] P. R. C. Kent, R. Q. Hood, M. D. Towler, R. J. Needs, and G. Rajagopal, "Quantum Monte Carlo calculations of the one-body density matrix and excitation energies of silicon," *Phys. Rev. B* 57:15293, 1998.

[624] B. Delley, "An all-electron numerical method for solving the local density functional for polyatomic molecules," *J. Chem. Phys.* 92:508–517, 1990.

[625] K. Koepernik and H. Eschrig, "Full-potential nonorthogonal local-orbital minimum-basis band-structure scheme," *Phys. Rev. B* 59:1743–1757, 2000.

[626] J. Junquera, O. Paz, D. Sanchez-Portal, and E. Artacho, "Numerical atomic orbitals for linear-scaling calculations," *Phys. Rev. B* 64:235111, 2001.
[627] M. R. Pederson and K. A. Jackson, "Variational mesh for quantum-mechanical simulations," *Phys. Rev. B* 41:7453–7461, 1990.
[628] A. D. Becke, "A multicenter numerical integration scheme for polyatomic molecules," *J. Chem. Phys.* 88:2547–2553, 1988.
[629] P. Ordejon, E. Artacho, and J. M. Soler, "Self-consistent Order-N density functional calculations for very large systems," *Phys. Rev. B* 53:R10441–R10444, 1996.
[630] D. Sanchez-Portal, E. Artacho, J. I. Pascual J. Gomez-Herrero, R. M. Martin, and J. M. Soler, "First principles study of the adsorption of C-60 on Si(111)," *Surface Sci.* 482:39–43, 2001.
[631] G. A. Baraff and M. Schluter, "Self-consistent Green's-function calculation of the ideal Si vacancy," *Phys. Rev. Lett.* 41:892, 1978.
[632] J. Bernholc, N. O. Lipari, and S. T. Pantelides, "Self-consistent method for point defects in semiconductors: Application to the vacancy in silicon," *Phys. Rev. Lett.* 41:895, 1978.
[633] P. J. Feibelman, "First-principles total-energy calculation for a single adatom on a crystal," *Phys. Rev. Lett.* 54:2627–2630, 1985.
[634] P. J. Feibelman, "Force and total-energy calculations for a spatially compact adsorbate on an extended, metallic crystal surface," *Phys. Rev. B* 35:2626–2646, 1987.
[635] S. G. Louie, K.-M. Ho, and M. L. Cohen, "Self-consistent mixed-basis approach to the electronic structure of solids," *Phys. Rev. B* 19:1774–1782, 1979.
[636] G. Li and Y. Chang, "Planar-basis pseudopotential calculations of the Si(001)2 $\times$ 1 surface with and without hydrogen passivation," *Phys. Rev. B* 48:12032–12036, 1993.
[637] J. Ziman, in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1971, Vol. 26, pp. 1–101.
[638] *Computational Methods in Band Theory*, edited by P. M. Marcus, J. F. Janak, and A. R. Williams, Plenum, New York, 1971.
[639] T. Loucks, *The Augmented Plane Wave Method*, Benjamin, New York, 1967.
[640] J. O. Dimmock, in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1971, Vol. 26, pp. 104–274.
[641] P. Lloyd and P. V. Smith, "Multiple scattering theory in condensed materials," *Adv. Phys.* 21:29, 1972.
[642] W. H. Butler, P. H. Dederichs, A. Gonis, and R. L. Weaver, *Applications of Multiple Scattering Theory to Material Science*, Materials Reserach Society, Pittsburg, Penn., 1992.
[643] H. Skriver, *The LMTO Method*, Springer, New York, 1984.
[644] O. K. Andersen, "Linear methods in band theory," *Phys. Rev. B* 12:3060–3083, 1975.
[645] M. I. Chodorow, "Energy band structure of copper," *Phys. Rev.* 55:675, 1939.
[646] L. F. Mattheiss, "Energy bands for the iron transition series," *Phys. Rev.* 134:A970–A973, 1964.
[647] V. Heine, "s-d interaction in transition metals," *Phys. Rev.* 153:673–682, 1967.
[648] J. W. D. Connolly, "Energy bands in ferromagnetic nickel," *Phys. Rev.* 159:415, 1967.
[649] J. Korringa, "On the calculation of the energy of a Bloch wave in a metal," *Physica* 13:392, 1947.
[650] W. Kohn and N. Rostocker, "Solution of the Schrodinger equation in periodic lattices with an application to metallic lithium," *Phys. Rev.* 94:1111, 1954.
[651] J. W. Strutt [Lord Rayleigh], "On the influence of obstacles arranged in rectangular order upon the properties of the medium," *Phil. Mag.* 23:481–502, 1892.

[652] R. Zeller, P. H. Dederichs, B. Ujfalussy, L. Szunyog, and P. Weinberger, "Theory and convergence properties of the screened Korringa-Kohn-Rostoker method," *Phys. Rev. B* 52:8807–8812, 1995.

[653] T. Huhne, C. Zecha, H. Ebert, P. H. Dederichs, and R. Zeller, "Full-potential spin-polarized relativistic Korringa-Kohn-Rostoker method implemented and applied to bcc Fe, fcc Co, and fcc Ni," *Phys. Rev. B* 58:10236, 1998.

[654] E. N. Economou, *Green's Functions in Quantum Physics, 2nd Ed.*, Springer-Verlag, Berlin, 1992.

[655] B. L. Gyorffy, in *Applications of Multiple Scattering Theory to Material Science*, edited by W. H. Butler, P. H. Dederichs, A. Gonis, and R. L. Weaver, Materials Research Society, Pittsburg, Penn., 1992, pp. 5–25.

[656] P. Lloyd, "Wave propagtation through an assembly of spheres II: The density of single particle eigenstates," *Proc. Phys. Soc, London* 90:207–216, 1967.

[657] S. Müller and A. Zunger, "Structure of ordered and disordered alpha-brass," *Phys. Rev. B* 63:094204, 2001.

[658] P. Soven, "Coherent-potential model of substitutional disordered alloys," *Phys. Rev.* 156:809–813, 1967.

[659] B. Velicky, S. Kirkpatrick, and H. Ehrenreich, "Single-site approximations in the electronic theory of simple binary alloys," *Phys. Rev.* 175:747–766, 1968.

[660] M. Lax, "Multiple scattering of waves," *Rev. Mod. Phys.* 23:287–310, 1951.

[661] J. L. Beeby, "Electronic structure of alloys," *Phys. Rev.* 135:A130, 1964.

[662] G. M. Stocks, W. M. Temmerman, and B. L. Gyorffy, "Complete solution of the Korringa-Kohn-Rostocker coherent-potential-approximation equations: Cu-Ni alloys," *Phys. Rev. Lett.* 41:339–343, 1978.

[663] J. S. Faulkner and G. M. Stocks, "Calculating properties with the coherent-potential approximation," *Phys. Rev. B* 21:3222–3244, 1980.

[664] W. H. Butler, "Theory of electronic transport in random alloys: Korringa–Kohn–Rostoker coherent-potential approximation," *Phys. Rev. B* 31:3260, 1985.

[665] A. F. Tatarchenko, V. S. Stepanyuk, W. Hergert, P. Rennert, R. Zeller, and P. H. Dederichs, "Total energy and magnetic moments in disordered Fe_xCu_{1-x} alloys," *Phys. Rev. B* 57:5213–5219, 1998.

[666] D. D. Johnson, D. M. Nicholson, F. J. Pinski, B. L. Gyorffy, and G. M. Stocks, "Total-energy and pressure calculations for random substitutional alloys," *Phys. Rev. B* 41:9701–9716, 1990.

[667] O. K. Andersen, in *Computational Methods in Band Theory*, edited by P. M. Marcus, J. F. Janak, and A. R. Williams, Plenum, New York, 1971, p. 178.

[668] O. K. Andersen and O. Jepsen, "Explicit, first-principles tight-binding theory," *Phys. Rev. Lett.* 53:2571–2574, 1984.

[669] J. Keller, "Modified muffin tin potentials for the band structure of semiconductors," *J. Phys. C: Solid State Phys.* 13:L85–L87, 1980.

[670] D. Glötzel, B. Segall, and O. K. Andersen, "Self-consistent electronic structure of Si, Ge and diamond LMTO-ASA method," *Solid State Commun.* 36:403, 1980.

[671] Y. Wang, G. M. Stocks, W. A. Shelton, and D. M. C. Nicholson, "Order-N multiple scattering approach to electronic structure calculations," *Phys. Rev. Lett.* 75:2867–2870, 1995.

[672] O. K. Andersen, Z. Pawlowska, and O. Jepsen, "Illustration of the LMTO tight-binding representation: Compact orbitals and charge density in Si," *Phys. Rev. B* 34:5253–5269, 1986.

[673] J. M. Soler and A. R. Williams, "Augmented-plane-wave forces," *Phys. Rev. B* 42:9728–9731, 1990.

[674] R. Yu, D. Singh, and H. Krakauer, "All-electron and pseudopotential force calculations using the linearized-augmented-plane-wave method," *Phys. Rev. B* 93:6411–6422, 1991.

[675] S. Mishra and S. Satpathy, "Kronig-penny model with the tail-cancellation method," *Am. J. Phys.* 69:512–513, 2001.

[676] P. M. Marcus, "Variational methods in the computation of enegy bands," *Int. J. Quant. Chem.* 1S:567–588, 1967.

[677] P. Blaha, K. Schwarz, P. Sorantin, and S.B. Trickey, "Full-potential, linearized augmented plane wave programs for crystalline systems," *Computer Phys. Commun.* 59(2):399, 1990.

[678] A. R. Williams, J. Kübler, and Jr. C. D. Gelatt, "Cohesive properties of metallic compounds: Augmented-spherical-wave calculations," *Phys. Rev. B* 19:6094–6118, 1979.

[679] D. D. Koelling and G. O. Arbman, "Use of energy derivative of the radial solution in an augmented plane wave method: application to copper," *J. Phys. F: Met. Phys.* 5:2041–2054, 1975.

[680] H. Krakauer, M. Posternak, and A. J. Freeman, "Linearized augmented plane-wave method for the electronic band structure of thin films," *Phys. Rev. B* 19:1706–1719, 1979.

[681] M. Weinert, E. Wimmer, and A. J. Freeman, "Total-energy all-electron density functional method for bulk solids and surfaces," *Phys. Rev. B* 26:4571–4578, 1982.

[682] L. F. Mattheiss and D. R. Hamann, "Linear augmented-plane-wave calculation of the structural properties of bulk cr, mo, and w," *Phys. Rev. B* 33:823–840, 1986.

[683] W. E. Pickett, "Electronic structure of the high-temperature oxide superconductors," *Rev. Mod. Phys.* 61:433, 1989.

[684] H. J. F. Jansen and A. J. Freeman, "Total-energy full-potential linearized augmented-plane-wave method for bulk solids: Electronic and structural properties of tungsten," *Phys. Rev. B* 30:561–569, 1984.

[685] R. E. Cohen, W. E. Pickett, and H. Krakauer, "Theoretical determination of strong electron-phonon coupling in $YBa_2Cu_3O_7$," *Phys. Rev. Lett.* 64:2575–2578, 1990.

[686] H. Krakauer, W. E. Pickett, and R. Cohen, "Analysis of electronic structure and charge density of the high-temperature superconductor $YBa_2Cu_3O_7$," *J. Superconductivity* 1:111, 1988.

[687] M. Methfessel, "Elastic constants and phonon frequencies of Si calculated by a fast full-potential linear-muffin-tin-orbital method," *Phys. Rev. B* 38:1537, 1988.

[688] M. Methfessel, C. O. Rodriguez, and O. K. Andersen, "Fast full-potential calculations with a converged basis of atom-centered linear muffin-tin orbitals: Structural and dynamic properties of silicon," *Phys. Rev. B* 40:2009, 1989.

[689] M. Methfessel and M. van Schilfgaarde, in *Electronic Strcuture and Physical Properties of Solids: The uses of the LMTO method*, edited by H. Dreysse, Springer, Heidelberg, 1999, pp. 114–147.

[690] T. Fujiwara, "Electronic structure calculations for amorphous alloys," *J. Non-Crystalline Solids* 61–62:1039–48, 1984.

[691] H. J. Nowak, O. K. Andersen, T. Fujiwara, O. Jepsen, and P. Vargas, "Electronic-structure calculations for amorphous solids using the recursion method and linear muffin-tin orbitals: Application to fe80b20," *Phys. Rev. B* 44:3577–3598, 1991.

[692] S. K. Bose, O. Jepsen, and O. K. Andersen, "Real-space calculation of the electrical resistivity of liquid 3d transition metals using tight-binding linear muffin-tin orbitals," *Phys. Rev. B* 48:4265–4275, 1993.

[693] O. Jepsen, O. K. Andersen, and A. R. Mackintosh, "Electronic structure of hcp transition metals," *Phys. Rev. B* 12:3084–3103, 1977.

[694] S. Satpathy and Z. Pawlowska, "Construction of bond-centered wannier functions for silicon bands," *Phys. Stat. Sol. (b)* 145:555–565, 1988.

[695] J. C. Duthi and D. G. Pettifor, "Correlation between d-band occupancy and crystal structure in the rare earths," *Phys. Rev. Lett.* 38:564–567, 1977.

[696] R. Haydock, in *Recursion Method and Its Applications*, edited by D. G. Pettifor and D. L. Weaire, Springer-Verlag, Berlin, 1985.

[697] J. Friedel, "Electronic structure of primary solid solutions in metals," *Adv. Phys.* 3:446, 1954.

[698] O. K. Andersen, T. Saha-Dasgupta, R. Tank, C. Arcangeli, O. Jepsen, and G. Krier, in *Electronic Structure and Physical Properties of Solids*, edited by H. Dreysse, Springer, Berlin, 1998, pp. 3–84.

[699] O. K. Andersen and T. Saha-Dasgupta, "Muffion-tin orbitals of arbitrary order," *Phys. Rev. B* 62:R16219–R16222, 2000.

[700] K. H. Weyrich, "Full-potential linear muffin-tin-orbital method," *Phys. Rev. B* 37:10269–10282, 1988.

[701] R. Car and M. Parrinello, in *Simple Molecular Systems at Very High Density*, edited by A. Polian, P. Loubeyre, and N. Boccara, Plenum, New York, 1989, p. 455.

[702] D. K. Remler and P. A. Madden, "Molecular dynamics without effective potentials via the Car-Parrinello approach," *Molecular Physics* 70:921, 1990.

[703] G. Pastore, E. Smargiassi, and F. Buda, "Theory of *ab initio* molecular-dynamics calculations," *Phys. Rev. A* 44:6334, 1991.

[704] J. M. Thijssen, *Computational Physics*, Cambridge University Press, Cambridge, England, 2000.

[705] G. Galli and M. Parrinello, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis, Kluwer, Dordrecht, 1991, pp. 283–304.

[706] M. E. Tuckerman and M. Parrinello, "Integrating the Car-Parrinello equations. I. Basic integration techniques," *J. Chem. Phys.* 101:1302, 1994.

[707] M. E. Tuckerman and M. Parrinello, "Integrating the Car-Parrinello equations. II. multiple time scale techniques," *J. Chem. Phys.* 101:1316, 1994.

[708] Special issue, "Techiniques for simulations," *Computational Materials Science* 12, 1998.

[709] L. Colombo, "A source code for tight-binding molecular simulations," *Comp. Mat. Sci.* 12:278–287, 1998.

[710] M. C. Payne, J. D. Joannopoulos, D. C. Allan, M. P. Teter, and D. M. Vanderbilt, "Molecular dynamics and *ab initio* total energy calculations," *Phys. Rev. Lett.* 56:2656, 1986.

[711] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes," *J. Comput. Phys.* 23:327, 1977.

[712] M. C. Payne, "Error cancellation in the molecular dynamics method for total energy calculations," *J. Phys.: Condens. Matter* 1:2199–2210, 1989.

[713] R. Car, M. Parrinello, and M. Payne, "Comment on 'error cancellation in the molecular dynamics method for total energy calculations'," *J. Phys.: Condens. Matter* 3:9539–9543, 1991.

[714] M. P. Grumbach and R. M. Martin, "Phase diagram of carbon at high pressure: Analogy to silicon," *Solid State Communications* 100:61, 1996.

[715] O. F. Sankey and R. E. Allen, "Atomic forces from electronic energies via the Hellmann–Feynman theorem, with application to semiconductor (110) surface relaxation," *Phys. Rev. B* 33:7164–7171, 1986.

[716] T. A. Arias, M. C. Payne, and J. D. Joannopoulos, "*Ab initio* molecular dynamics: Analytically continued energy functionals and insights into iterative solutions," *Phys. Rev. Lett.* 69:1077–1080, 1992.

[717] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, Th. Frauenheim, S. Suhai, and G. Seifert, "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties," *Phys. Rev. B* 58:7260–7268, 1998.

[718] G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," *Phys. Rev. B* 54:11169–11186, 1996.

[719] X. Gonze, "Adiabatic density-functional perturbation theory," *Phys. Rev. A* 52:1096–1114, 1995.

[720] S. Y. Savrasov, "Linear response calculations of spin fluctuations," *Phys. Rev. Lett.* 81:2570–2573, 1998.

[721] L. D. Landau and E. M. Lifshitz, *Theory of Elasticity*, Pergamon Press, Oxford, England, 1958.

[722] J. F. Nye, *Physical Properties of Crystals*, Oxford University Press, Oxford, England, 1957.

[723] H. Wendel and R. M. Martin, "Charge density and structural properties of covalent semiconductors," *Phys. Rev. Lett.* 40(14):950–953, 1978.

[724] K.-M. Ho, C.-L. Fu, B. N. Harmon, W. Weber, and D. R. Hamann, "Vibrational frequencies and structural properties of transition metals via total-energy calculations," *Phys. Rev. Lett.* 49:673–676, 1982.

[725] K.-M. Ho, C.-L. Fu, and B. N. Harmon, "Vibrational frequencies via total-energy calculations, applications to transition metals," *Phys. Rev. B* 29:1575–1587, 1984.

[726] V. Heine and J. H. Samson, "Magnetic, chemical and structural ordering in transition metals," *J. Phys. F* 13:2155, 1983.

[727] S. Tinte, J. Iniguez, K. M. Rabe, and D. Vanderbilt, "Quantitative analysis of the first-principles effective hamiltonian approach to ferroelectric perovskites," *Phys. Rev. B* 67:064106, 2003.

[728] S. Baroni, P. Giannozzi, and A. Testa, "Greens function approach to linear response in solids," *Phys. Rev. Lett.* 58:1861–1864, 1987.

[729] A. A. Quong and B. M. Klein, "Self-consistent-screening calculation of interatomic force constants and phonon dispersion curves from first principles," *Phys. Rev. B* 46:10734–10737, 1992.

[730] S. Y. Savrasov and D. Y. Savrasov, "Linear-response theory and lattice dynamics: A muffin-tin-orbital approach," *Phys. Rev. B* 54:16470–16486, 1996.

[731] R. M. Sternheimer, "Electronic polarizabilities of ions from the Hartree-Fock wave functions," *Phys. Rev.* 96:951, 1954.

[732] S. de Gironcoli, "Lattice dynamics of metals from density-functional perturbation theory," *Phys. Rev. B* 51:6773, 1995.

[733] S. Y. Savrasov and D. Y. Savrasov, "Electron-phonon interactions and related physical properties of metals from linear-response theory," *Phys. Rev. B* 54:16487–16501, 1996.

[734] R. Resta and K. Kunc, "Self-consistent theory of electronic states and dielectric response in semiconductors," *Phys. Rev. B* 34:7146–7157, 1986.

[735] P. B. Littlewood, "On the calculation of the macroscopic polarisation induced by an optic phonon," *J. Phys. C* 13:4893, 1980.

[736] R. Resta, M. Posternak, and A. Baldereschi, "Towards a quantum theory of polarization in ferroelectrics: The case of $KNbO_3$," *Phys. Rev. Lett.* 70:1010–1013, 1993.

[737] Ph. Ghosez, X. Gonze, Ph. Lambin, and J.-P. Michenaud, "Born effective charges of barium titanate: Band-by-band decomposition and sensitivity to structural features," *Phys. Rev. B* 51:6765–6768, 1995.

[738] G. M. Eliashberg, "Interactions between electrons and lattice vibrations in a superconductor [Translation: Sov. Phys. JETP 11, 696 (1960)]," *Zh. Eksp. Teor. Fiz.* 38:966, 1960.

[739] G. D. Gaspari and B. L. Gyorffy, "Electron-phonon interactions, d resonances, and superconductivity in transition metals," *Phys. Rev. Lett.* 28:801–805, 1972.

[740] J. J. Hopfield, "Angular momentum and transition-metal superconductivity," *Phys. Rev.* 186:443–451, 1969.

[741] M. M. Dacorogna, M. L. Cohen, and P. K. Lam, "Self-consistent calculation of the q dependence of the electron-phonon coupling in aluminum," *Phys. Rev. Lett.* 55:837–840, 1985.

[742] J. F. Cooke, "Neutron scattering from itinerant-electron ferromagnets," *Phys. Rev. B* 7:1108–1116, 1973.

[743] C. Yannouleas and U. Landman, "Molecular dynamics in shape space and femtosecond vibrational spectroscopy of metal clusters," *J. Phys. Chem. A* 102:2505–2508, 1998.

[744] A. Rubio, J. A. Alonso, X. Blase, L. C. Balbas, and S. G. Louie, "*Ab initio* photoabsorption spectra and structures of small semiconductor and metal clusters," *Phys. Rev. Lett.* 77:247–250, 1996.

[745] A. D. Yoffe, "Semiconductor quantum dots and related systems: electronic, optical, luminescence and related properties of low dimensional systems," *Adv. Phys.* 50:1–208, 2001.

[746] G. Belomoin, A Smith, S. Rao, R. Twesten, J. Therrien, M. Nayfeh, L. Wagner, L. Mitas, and S. Chaieb, "Observation of a magic discrete family of ultrabright Si nanoparticles," *Appl. Phys. Lett* 80:841–843, 2002.

[747] A. Pasquarello and A. Quattropani, "Application of variational techniques to time-dependent perturbation theory," *Phys. Rev. B* 48:5090–5094, 1993.

[748] A. Dal Corso, F. Mauri, and A. Rubio, "Density-functional theory of the nonlinear optical susceptibility: Application to cubic semiconductors," *Phys. Rev. B* 53:15638–15642, 1996.

[749] C. A. Ullrich, U. J. Gossmann, and E. K. U. Gross, "Density-functional approach to atoms in strong laser-pulses," *Ber. Bunsen Phys. Chem.* 99:488–497, 1995.

[750] M. Protopapas, C. H. Keitel, and P. L. Knight, "Atomic physics with super-high intensity lasers," *Rep. Prog. Phys.* 60:389, 1997.

[751] H. Flocard, S. Koonin, and M. Weiss, "Three-dimensional time-dependent Hartree-Fock calculations: Application to $^{16}O + ^{16}O$ collisions," *Phys. Rev. C* 17:1682–1699, 1978.

[752] O. Sugino and Y. Miyamoto, "Density-functional approach to electron dynamics: Stable simulation under a self-consistent field," *Phys. Rev. B* 59:2579–2586, 1999.

[753] H. Talezer and R. Kosloff, "An accurate and efficient scheme for propagating the time-dependent Schrödinger equation," *J. Chem. Phys.* 81:3967–3971, 1984.

[754] A. Tsolakidis, D. Sanchez-Portal, and R. M. Martin, "Calculation of the optical response of c60 and na8 using time-dependent density functional theory and local orbitals," *Phys. Rev. B* 66:235416, 2002.

[755] M. A. L. Marques, A. Castro, and A. Rubio, "Assessment of exchange-correlation functionals for the calculation of dynamical properties of small clusters in time-dependent density functional theory," *J. Chem. Phys.* 115:3006–3014, 2001.

[756] R. van Leeuwen and E. J. Baerends, "Exchange-correlation potential with correct asymptotic behavior," *Phys. Rev. A* 49:2421, 1994.

[757] G. Weinreich, *Solids: Elementary Theory for Advanced Students*, John Wiley and Sons, New York, 1965.
[758] W. Kohn, "Construction of Wannier functions and applications to energy bands," *Phys. Rev. B* 7:4388–4398, 1973.
[759] G. Blount, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1962, p. 305.
[760] S. F. Boys, "Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another," *Rev. Mod. Phys.* 32:296–299, 1960.
[761] C. Edmiston and K. Ruedenberg, "Localized atomic and molecular orbitals," *Rev. Mod. Phys.* 35:457–464, 1963.
[762] N. Marzari and D. Vanderbilt, "Maximally localized generalized Wannier functions for composite energy bands," *Phys. Rev. B* 56:12847–12865, 1997.
[763] G. H. Wannier, "The structure of electronic excitations in the insulating crystals," *Phys. Rev.* 52:191–197, 1937.
[764] G. F. Koster, "Localized functions in molecules and crystals," *Phys. Rev.* 89:67, 1953.
[765] D. W. Bullett, "A chemical pseudopotential approach to covalent bonding. I," *J. Phys. C: Solid State Phys.* 8:2695–2706, 1975.
[766] P. W. Anderson, "Self-consistent pseudopotentials and ultralocalized functions for energy bands," *Phys. Rev. Lett.* 21:13, 1968.
[767] W. Kohn, "Analytic properties of Bloch waves and Wannier functions," *Phys. Rev. B* 115:809–821, 1959.
[768] L. He and D. Vanderbilt, "Exponential decay properties of Wannier functions and related quantities," *Phys. Rev. Lett.* 86:5341–5344, 2001.
[769] S. N. Taraskin, D. A. Drabold, and S. R. Elliott, "Spatial decay of the single-particle density matrix in insulators: Analytic results in two and three dimensions," *Phys. Rev. Lett.* 88:196405, 2002.
[770] A. K. McMahan, J. F. Annett, and R. M. Martin, "Cuprate parameters from numerical Wannier functions." *Phys. Rev. B* 42:6268, 1990.
[771] I. Schnell, G. Czycholl, and R. C. Albers, "Hubbard-u calculations for Cu from first-principle Wannier functions," *Phys. Rev. B* 65:075103, 2002.
[772] S. F. Boys, in *Quantum Theory of Atoms, Molecules and the Solid State*, edited by P.-O. Löwdin, Academic Press, New York, 1982, p. 253.
[773] W. Hierse and E. Stechel, "Robust localized-orbital transferability using the Harris functional," *Phys. Rev. B* 54:16515–16522, 1996.
[774] S. Liu, J. M. Perez-Jorda, and W. Yang, "Nonorthogonal localized molecular orbitals in electronic structure theory," *J. Chem. Phys.* 112:1634, 2000.
[775] I. Souza, T. J. Wilkens, and R. M. Martin, "Polarization and localization in insulators: generating function approach," *Phys. Rev. B* 62:1666–1683, 2000.
[776] U. Stephan and D. A. Drabold, "Order-N projection method for first-principles computations of electronic quantities and Wannier functions," *Phys. Rev. B* 57:6391–6407, 1998.
[777] U. Stephan, R. M. Martin, and D. A. Drabold, "Extended-range computation of Wannier-like functions in amorphous semiconductors," *Phys. Rev. B* 62:6885–6888, 2000.
[778] P. L. Silvestrelli, N. Marzari, D. Vanderbilt, and M. Parrinello, "Maximally-localized Wannier functions for disordered systems: application to amorphous silicon," *Solid State Commun.* 107:7–11, 1998.

[779] G. Berghold, C. J. Mundy, A. H. Romero, J. Hutter, and M. Parrinello, "General and efficient algorithms for obtaining maximally localized Wannier functions," *Phys. Rev. B* 61:10040–10048, 2000.

[780] W. T. Yang, "Absolute-energy-minimum principles for linear-scaling electronic-structure calculations," *Phys. Rev. B* 56:9294–9297, 1997.

[781] I. Souza, N. Marzari, and D. Vanderbilt, "Maximally localized Wannier functions for entangled energy bands," *Phys. Rev. B* 65:035109, 2002.

[782] O. K. Andersen, A. I. Liechtenstein, O. Jepsen, and F. Paulsen, "LDA energy bands, low-energy hamiltonians, t', t", t (k), and J(perpendicular)," *J. Phys. Chem. Solids* 56:1573, 1995.

[783] W. Kohn, "Theory of the insulating state," *Phys. Rev.* 133:A171–181, 1964.

[784] M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, Clarendon Press, Oxford, 1977.

[785] R. M. Martin, "Comment on: Piezoelectricity under hydrostatic pressure," *Phys. Rev. B* 6:4874, 1972.

[786] R. M. Martin, "Comment on: Calculation of electric polarization in crystals," *Phys. Rev. B* 9:1998, 1974.

[787] A. K. Tagantsev, "Review: Electric polarization in crystals and its response to thermal and elastic perturbations," *Phase Transitions* 35:119, 1991.

[788] R. Resta, "Why are insulators insulating and metals conducting?" *J. Phys.: Condens. Matte* 14:R625–R656, 2002.

[789] R. Resta, "The quantum mechanical position operator in extended systems," *Phys. Rev. Lett.* 80:1800–1803, 1998.

[790] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon Press, Oxford, England, 1960.

[791] R. P. Feynman, R. B. Leighton, and M. Sands, *Lectures on Physics, Vol. 2*, Addison Wesley Publishing Company, Reading, Massachusetts, 1982.

[792] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, "Quantized Hall conductance in a two-dimensional periodic potential," *Phys. Rev. Lett.* 49:405–408, 1982.

[793] D. J. Thouless, "Quantization of particle transport," *Phys. Rev. B* 27:6083–6087, 1983.

[794] Q. Niu and D. J. Thouless, "Quantised adiabatic charge transport in the presence of substrate disorder and many-body interaction," *J. Phys. A* 17:2453, 1984.

[795] G. Ortiz and R. M. Martin, "Macroscopic polarization as a geometric quantum phase: Many-body formulation," *Phys. Rev. B* 49:14202–14210, 1994.

[796] G. Ortiz, I. Souza, and R. M. Martin, "The exchange-correlation hole in polarized dielectrics: Implications for the microscopic functional theory of dielectrics," *Phys. Rev. Lett.* 80:353–356, 1998.

[797] R. Resta and S. Sorella, "Electron localization in the insulating state," *Phys. Rev. Lett.* 82:370–373, 1999.

[798] C. Aebischer, D. Baeriswyl, and R. M. Noack, "Dielectric catastrophe at the Mott transition," *Phys. Rev. Lett.* 86:468–471, 2001.

[799] C. Kallin and B. I. Halperin, "Surface-induced charge disturbances and piezoelectricity in insulating crystals," *Phys. Rev. B* 29:2175–2189, 1984.

[800] R. Resta, "Theory of the electric polarization in crystals," *Ferroelectrics* 136:51, 1992.

[801] R. Resta, "Towards a quantum theory of polarization in ferroelectrics: the case of $KNbO_3$," *Europhys. Lett.* 22:133–138, 1993.

[802] G. Arlt and P. Quadflieg, "Piezoelectricity in III-V compounds with a phenomenological analysis of the piezoelectric effect," *Phys. Status Solidi* 25:323, 1968.

[803] D. Vanderbilt, "Berry-phase theory of proper piezoelectric response," *J. Phys. Chem. Solids* 61:147–151, 2000.

[804] X. Gonze, D. C. Allan, and M. P. Teter, "Dielectric tensor, effective charges, and phonons in α-quartz by variational density-functional perturbation theory," *Phys. Rev. Lett.* 68:3603–3606, 1992.

[805] A. Dal Corso and F. Mauri, "Wannier and Bloch orbital computation of the nonlinear susceptibility," *Phys. Rev. B* 50:5756–5759, 1994.

[806] W. Kleemann, F. J. Schäfer, and M. D. Fontana, "Crystal optical studies of spontaneous and precursor polarization in $KNbO_3$," *Phys. Rev. B* 30:1148–1154, 1984.

[807] E. K. Kudinov, "Difference between insulating and conducting states," *Sov. Phys. Solid State* 33:1299–1304, 1991, [Fiz. Tverd. Tela 33, 2306 (1991)].

[808] H. B. Callen and T. A. Welton, "Irreversibility and generalized noise," *Phys. Rev.* 83:34, 1951.

[809] H. B. Callen, M. L. Barasc, and J. L. Jackson, "Statistical mechanics of irreversibility," *Phys. Rev.* 88:1382, 1952.

[810] R. Kubo, "A general expression for the conductivity tensor," *Canad. Journ. Phys.* 34:1274, 1956.

[811] P. C. Martin, *Measurement and Correlation Functions*, Gordon and Breach, New York, 1968.

[812] D. R. Penn, "Wave-number-dependent dielectric function of semiconductors," *Phys. Rev.* 128:2093–2097, 1962.

[813] C. Sgiarovello, M. Peressi, and R. Resta, "Electron localization in the insulating state: Application to crystalline semiconductors," *Phys. Rev. B* 64:115202, 2001.

[814] S. Goedecker, "Linear scaling electronic structure methods," *Rev. Mod. Phys.* 71:1085–1123, 1999.

[815] G. Galli, "Large scale electronic structure calculations using linear scaling methods," *Phys. Stat. Sol.* 217:231–249, 2000.

[816] P. Ordejon, "Linear scaling *ab intio* calculations in nanoscale materials with SIESTA," *Phys. Stat. Sol.* 217:335–356, 2000.

[817] D. R. Bowler and M. J. Gillan, "Recent progress in first principles O(N) methods," *Molecular Simulations* 25:239–255, 2000.

[818] S. Y. wu and C. S. Jayanthi, "Order-N methodologies and their applications," *Phys. rep.* 358:1–74, 2002.

[819] P. Fulde, *Electron Correlation in Molecules and Solids, 2nd Ed.*, Springer-Verlag, Berlin, 1993.

[820] L. Greengard, "Fast algorithms for classical physics," *Science* 265:909–914, 1994.

[821] A. J. Williamson, R. Q. Hood, and J. C. Grossman, "Linear-scaling quantum Monte Carlo calculations," *Phys. Rev. Lett.* 87:246406–246409, 2001.

[822] W. Kohn, "Density functional and density matrix method scaling linearly with the number of atoms," *Phys. Rev. Lett.* 76:3168–3171, 1996.

[823] W. Hierse and E. Stechel, "Order-N methods in self-consistent density-functional calculations," *Phys. Rev. B* 50:17811–17819, 1994.

[824] A. L. Ankudinov, C. E. Bouldin, J. J. Rehr, J. Sims, and H. Hung, "Parallel calculation of electron multiple scattering using Lanczos algorithms," *Phys. Rev. B* 65:104107, 2002.

[825] D. G. Pettifor, "New many-body potential for the bond order," *Phys. Rev. Lett.* 63:2480–2483, 1989.

[826] M. Aoki, "Rapidly convergent bond order expansion for atomistic simulations," *Phys. Rev. Lett.* 71:3842, 1993.
[827] A. P. Horsfield, "A comparison of linear scaling tight-binding methods," *Mater. Sci. Eng.* 5:199, 1996.
[828] R. Haydock, in *Solid State Physics*, edited by H. Ehenreich, F. Seitz, and D. Turnbull, Academic Press, New York, 1980, Vol. 35, p. 1.
[829] R. Haydock, V. Heine, and M. J. Kelly, "Electronic structure based on the local atomic environment for tight-binding bands: II," *J. Phys. C* 8:2591–2605, 1975.
[830] N. I. Akhiezer, *The Classical Moment Problem*, Oliver and Boyd, Edinburgh, 1965.
[831] D. A. Drabold, P. Ordejon, J. J. Dong, and R. M. Martin, "Spectral properties of large fullerenes: from cluster to crystal." *Solid State Commun.* 96:833, 1995.
[832] D. A. Drabold and O. F. Sankey, "Maximum entropy approach for linear scaling in the electronic structure problem," *Phys. Rev. Lett.* 70:3631–3634, 1993.
[833] P. Ordejón, D. A. Drabold, R. M. Martin, and S. Itoh, "Linear scaling method for phonon calculations from electronic structure," *Phys. Rev. Lett.* 75:1324–1327, 1995.
[834] W. T. Yang, "Direct calculation of electron density in density functional theory," *Phys. Rev. Lett.* 66:1438–1441, 1991.
[835] K. Kitaura, S. I. Sugiki, T. Nakano, Y. Komeiji, and M. Uebayasi, "Fragment molecular orbital method: analytical energy gradients," *Chem. Phys. Lett.* 336:163–170, 2001.
[836] S. Goedecker and L. Colombo, "Efficient linear scaling algorithm for tight-binding molecular dynamics," *Phys. Rev. Lett.* 73:122–125, 1994.
[837] A. F. Voter, J. D. Kress, and R. N. Silver, "Linear-scaling tight binding from a truncated-moment approach," *Phys. Rev. B* 53:12733–12741, 1996.
[838] R. N. Silver and H. Roder, "Calculation of densities of states and spectral functions by Chebyshev recursion and maximum entropy," *Phys. Rev. E* 56:4822–4829, 1997.
[839] S. Goedecker, "Low complexity algorithms for electronic structure calculations," *J. Comp. Phys.* 118, 1995.
[840] D. Jovanovic and J. P. Leburton, "Self-consistent analysis of single-electron charging effects in quantum-dot nanostructures," *Phys. Rev. B* 49:7474, 1994.
[841] A. Alavi, Parrinello, and D. Frenkel, "*Ab initio* calculation of the sound velocity of dense hydrogen: implications for models of Jupiter," *Science* 269:1252–4, 1995.
[842] J. L. Corkill and K. M. Ho, "Electronic occupation functions for density-matrix tight-binding methods," *Phys. Rev. B* 54:5340–5345, 1996.
[843] X. P. Li, R. W. Nunes, and D. Vanderbilt, "Density-matrix electronic-structure method with linear system-size scaling," *Phys. Rev. B* 47:10891–10894, 1993.
[844] F. Mauri, G. Galli, and R. Car, "Orbital formulation for electronic structure calculation with linear system-size scaling," *Phys. Rev. B* 47:9973–9976, 1993.
[845] P. Ordejón, D. A. Drabold, M. P. Grumbach, and R. M. Martin, "Unconstrained minimization approach for electronic computations that scales linearly with system size," *Phys. Rev. B* 48:14646–14649, 1993.
[846] J. Kim, F. Mauri, and G. Galli, "Total-energy global optimizations using nonorthogonal localized orbitals," *Phys. Rev. B* 52:1640–1648, 1995.
[847] R. W. Nunes and D. Vanderbilt, "Generalization of the density-matrix method to a nonorthogonal basis," *Phys. Rev. B* 50:17611–17614, 1994.
[848] C. H. Xu and G. Scuseria, "An O(N) tight-binding study of carbon clusters up to C_{8640}: the geometrical shape of the giant icosahedral fullerenes," *Chem. Phys. Lett.* 262:219, 1996.

[849] S. Itoh, P. Ordejón, D. Drabold, and R. M. Martin, "Structure and energetics of giant fullerenes: an order-n molecular dynamics study," *submitted to Phys. Rev. B*, 1995.

[850] S. Y. Qiu, C. Z. Wang, K. M. Ho, and C. T. Chan, "Tight-binding molecular dynamics with linear system-size scaling," *J. Phys.: Condens. Matter* 6:9153, 1994.

[851] R. W. Nunes, J. Bennetto, and D. Vanderbilt, "Atomic structure of dislocation kinks in silicon," *Phys. Rev. B* 57:10388–10397, 1998.

[852] R. W. Nunes, J. Bennetto, and D. Vanderbilt, "Core reconstruction of the 90 degree partial dislocation in nonpolar semiconductors," *Phys. Rev. B* 58:12563–12566, 1998.

[853] E. B. Stechel, A. R. Williams, and P. J. Feibelman, "N-scaling algorithm for density-functional calculations of metals and insulators," *Phys. Rev. B* 49:10088–10101, 1994.

[854] W. H. Press and S. A. Teukolsky, *Numerical Recipes*, Cambridge University Press, Cambridge, 1992.

[855] U. Stephan, D. A. Drabold, and R. M. Martin, "Improved accuracy and acceleration of variational order-n electronic-structure computations by projection techniques," *Phys. Rev. B* 58:13472–13481, 1998.

[856] P. J. de Pablo, F. Moreno-Herrero, J. Colchero, J. G. Herrero, P. Herrero, A. M. Baro, P. Ordejon, J. M. Soler, and E. Artacho, "Absence of dc-conductivity in lambda-DNA," *Phys. Rev. Lett.* 85:4992–4995, 2000.

[857] E. Hernandez and M. J. Gillan, "Self-consistent first-principles technique with linear scaling," *Phys. Rev. B* 51:10157–10160, 1995.

[858] J.-L. Fattebert and J. Bernholc, "Towards grid-based O(N) density-functional theory methods: Optimized nonorthogonal orbitals and multigrid acceleration," *Phys. Rev. B* 62:1713–1722, 2000.

[859] P.D. Haynes and M. C. Payne, "Localised spherical-wave basis set for O(N) total-energy pseudopotential calculations", *Comput. Phys. Commun*. 102, pages 17–27 (1997).

[860] R. Baer and M. Head-Gordon, "Sparsity of the density matrix in Kohn-Sham density functional theory and an assessment of linear system-size scaling methods," *Phys. Rev. Lett.* 79:3962–3965, 1997.

[861] G. C. Evans, *Functionals and Their Applications*, Dover, New York, 1964.

[862] J. Matthews and R. L. Walker, *Mathmatical Methods of Physics*, W. A. Benjamin, Inc., New York, 1964.

[863] S. Doniach and E. H. Sondheimer, *Green's Functions for Solid State Physicists (Reprinted in "'Frontiers in Physics Series, No. 44)*, W. A. Benjamin, Reading, Mass., 1974.

[864] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-particle Systems*, McGraw-Hill, New York, 1971.

[865] C. Kittel, *Quantum Theory of Solids, 2nd Revised Edition*, John Wiley and Sons, New York, 1964.

[866] H. Mori, "A continued-fraction representation of the time-correlation functions," *Prog. Theor. Phys.* 34:399, 1965.

[867] R. Kubo, "Statistical-mechanical theory of irreversible processes. I," *J. Phys. Soc. Japan* 12:570, 1957.

[868] D. A. Greenwood, "The Boltzmann equation in the theory of electrical conduction in metals," *Proc. Phys. Soc. (London)* 71:585, 1958.

[869] P. Nozières and D. Pines, "Electron interaction in solids, collective approach to the dielectric constant," *Phys. Rev.* 109:762–777, 1959.

[870] H. Ehrenreich and M. H. Cohen, "Self-consistent field approach to the many-electron problem," *Phys. Rev.* 115:786–790, 1959.
[871] N. Wiser, "Dielectric constant with local field effects included," *Phys. Rev.* 129:62–69, 1963.
[872] E. P. Wigner, "Über eine Verschärfung des Summensatzes," *Phys. Z.* 32:450, 1931.
[873] H. Kramers, C. C. Jonker, and T. Koopmans, "Wigners Erweiterung des Thomas-Kuhnschen Summensatzes für ein Elektron in einem Zentralfeld," *Z. Phys.* 80:178, 1932.
[874] W. Cochran and R. A. Cowley, "Dielectric constants and lattice vibrations," *J. Phys. Chem. Solids* 23:447, 1962.
[875] R. Zallen, "Symmetry and reststrahlen in elemental crystals," *Phys. Rev.* 173:824–832, 1968.
[876] R. Zallen, R. M. Martin, and V. Natoli, "Infrared activity in elemental crystals," *Phys. Rev. B* 49:7032–7035, 1994.
[877] W. F. Cady, *Piezoelectricity*, McGraw-Hill, New York, 1946.
[878] R. M. Martin, "Piezoelectricity," *Phys. Rev. B* 5(4):1607–1613, 1972.
[879] R. A. Coldwell-Horsfall and A. A. Maradudin, "Zero-point energy of an electron lattice," *J. Math. Phys.* 1:395, 1960.
[880] P. A. Schultz, "Local electrostatic moments and periodic boundary conditions," *Phys. Rev. B* 60:1551–1554, 1999.
[881] G. Makov and M. C. Payne, "Periodic boundary conditions in *ab initio* calculations," *Phys. Rev. B* 51:4014–4022, 1995.
[882] P. P. Ewald, "Die Berechnung optischer und electrostatischer Gitterpotentiale," *Ann. der Physik* 64:253, 1921.
[883] H. Kornfeld, "Die Berechnung electrostatischer Potentiale und der Energie von Dipole- und Quadrupolgittern," *Z. Phys.* 22:27, 1924.
[884] K. Fuchs, "A quantum mechanical investigation of the cohesive forces of metallic copper," *Proc. Roy. Soc.* 151:585, 1935.
[885] M. P. Tosi, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Academic, New York, 1964.
[886] L. M. Fraser, W. M. C. Foulkes, G. Rajagopal, R. J. Needs, S. D. Kenny, and A. J. Williamson, "Finite-size effects and coulomb interactions in quantum Monte Carlo calculations for homogeneous systems with periodic boundary conditions," *Phys. Rev. B* 53:1814, 1996.
[887] J. E. Lennard-Jones and B. M. Dent, "Cohesion at a crystal surface," *Trans. Faraday Soc.* 24:92–108, 1928.
[888] L. N. Kantorovich, "Elimination of the long-range dipole interaction in calculations with periodic boundary conditions," *Phys. Rev. B* 60:15476, 1999.
[889] P. J. Feibelman, "Calculation of surface stress in a linear combination of atomic orbitals representation," *Phys. Rev. B* 50:1908–1911, 1994.
[890] A. Sommerfeld, *Mechanics of Deformable Bodies*, Academic Press, New York, 1950.
[891] E. Schrödinger, "The energy-impulse hypothesis of material waves," *Ann. Phys. (Leipzig)* 82:265, 1927.
[892] R. P. Feynman, Undergraduate thesis, unpublished, massachusetts institute of technology, 1939.
[893] P. C. Martin and J. Schwinger, "Theory of many-particle systems. I," *Phys. Rev.* 115:1342–1373, 1959.
[894] C. Rogers and A. Rappe, "Unique quantum stress fields," *Proceedings for Fundamental Physics of Ferroelectrics,* pp. 91–96, 2001.
[895] N. Chetty and R. M. Martin, "First-principles energy density and its applications to selected polar surfaces." *Phys. Rev. B* 45:6074–6088, 1992.

[896] M. H. Cohen, D. Frydel, K. Burke, and E. Engel, "Total energy density as an interperative tool," *J. Chem. Phys.* 113:2990–2994, 2000.
[897] A. D. Becke and K. E. Edgecombe, "A simple measure of electron localization in atomic and molecular systems," *J. Chem. Phys.* 92:5397–5403, 1990.
[898] C. Rogers and A. Rappe, "Geometric formulation of quantum stress fields," *Phys. Rev. B*, 65:224117, 2002.
[899] N. Chetty and R. M. Martin, "$GaAs$ (111) and (-1-1-1) surfaces and the $GaAs/AlAs$ (111) heterojunction studied using a local energy density." *Phys. Rev. B* 45:6089–6100, 1992.
[900] K. Rapcewicz, B. Chen, B. Yakobson, and J. Bernholc, "Consistent methodology for calculating surface and interface energies," *Phys. Rev. B* 57:7281–7291, 1998.
[901] R. M. Martin, unpublished, 2002.
[902] A. Savin, "Expression of the exact electron-correlation-energy density functional in terms of first-order density matrices," *Phys. Rev. A* 52:R1805–R1807, 1995.
[903] M. Levy and A. Gorling, "Correlation-energy density-functional formulas from correlating first-order density matrices," *Phys. Rev. A* 52:R1808–R1810, 1995.
[904] H. Stoll, E. Golka, and H. Preuss, "Correlation energies in the spin-density functional formalism. II. applications and empirical corrections," *Theor. Chim. Acta* 55:29, 1980.
[905] A. D. Becke, "Hartree-Fock exchange energy of an inhomogeneous electron gas," *Int. J. Quantum Chem.* 23:1915, 1983.
[906] W. L. Luken and J. C. Culbertson, "Localized orbitals based on the fermi hole," *Theor. Chim. Acta* 66:279, 1984.
[907] J. F. Dobson, "Interpretation of the Fermi hole curvature," *J. Chem. Phys.* 94:4328–4333, 1991.
[908] A. D. Becke and M. R. Roussel, "Exchange holes in inhomogeneous systems: A coordinate-space model," *Phys. Rev. A* 39:3761–3767, 1989.
[909] B. Hammer and M. Scheffler, "Local chemical reactivity of a metal alloy surface," *Phys. Rev. Lett.* 74:3487–3490, 1995.
[910] M. J. Godfrey, "Stress field in quantum systems," *Phys. Rev. B* 37:10176–10183, 1988.
[911] A. Filippetti and V. Fiorentini, "Theory and applications of the stress density," *Phys. Rev. B* 61:8433–8442, 2000.
[912] B. L. Trout and M. Parrinello, "Autoionization in liquid water," *J. Phys. Chem.* 103:7340, 1999.
[913] D. G. Pettifor, "Pressure-cell boundary relation and application to transition-metal equation of state," *Commun. Phys.* 1:141, 1976.
[914] M. Methfessel and M. van Schilfgaarde, "Derivation of force theorems in density-functional theory: Application to the full-potential LMTO method," *Phys. Rev. B* 48:4937–4940, 1993.
[915] J. Grafenstein and P. Ziesche, "Andersen's force theorem and the local stress field," *Phys. Rev. B* 53:7143–7146, 1996.
[916] E. U. Condon and G. H. Shortley, *Theory of Atomic Spectra*, Cambridge University Press, New York, 1935.
[917] J. A. Gaunt, "Triplets of helium," *Phil. Trans. Roy. Soc. (London)* 228:151–196, 1929.
[918] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 1980.
[919] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Engelwood Cliffs, N. J., 1980.
[920] M. T. Heath, *Scientific Computing: An introductory Survey*, McGraw-Hill, New York, 1997.
[921] B. Numerov, "Note on the numerical integration of $d^2x/dt^2 = f(x,t)$," *Astronomical Nachr.* 230:359–364, 1927.

[922] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.
[923] J. H. van Lenthe and P. Pulay, "A space-saving modification of Davidson's eigenvector algorithm," *J. Comp. Chem.* 11:1164–1168, 1990.
[924] H. Thirring, "Space lattices and specific heat," *Phys. Zeit.* 14:867, 1913.
[925] C. Lanczos, *Applied Analysis*, Printice Hall, New York, 1956.
[926] H. Roder, R. N. Silver, J. J. Dong, and D. A. Drabold, "Kernel polynomial method for a nonorthogonal electronic structure calculation of amorphous diamond," *Phys. Rev. B* 55:15382, 1997.
[927] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, Kluwer, Dordrecht, 1989, p. 455.
[928] L. R. Mead, "Approximate solution of Fredholm integral equations by the maximum-entropy method," *J Math Phys.* 27:2903, 1986.
[929] D. M. Wood and A. Zunger, "A new method for diagonalizing large matrices," *J. Phys. A* 18:1343–1359, 1985.
[930] J. L. Martins and M. L. Cohen, "Diagonalization of large matrices in pseudopotential band-structure calculations: Dual-space formalism," *Phys. Rev. B* 37:6134–6138, 1988.
[931] M. P. Teter, M. C. Payne, and D. C. Allan, "Solution of Schrödinger's equation for large systems," *Phys. Rev. B* 40:12255–12263, 1989.
[932] I. Stich, R. Car, M. Parrinello, and S. Baroni, "Conjugate gradient minimization of the energy functional: A new method for electronic structure calculation," *Phys. Rev. B* 39:4997, 1989.
[933] D. M. Bylander, L. Kleinman, and S. Lee, "Self-consistent calculations of the energy bands and bonding properties of $B_{12}C_3$," *Phys. Rev. B* 42:1394–1403, 1990.
[934] P. Pulay, "Convergence acceleration of iterative sequences. the case of SCF iteration," *Chem. Phys. Lett.* 73:393–397, 1980.
[935] E. R. Davidson, "The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices." *J. Comp. Phys.* 17:87, 1975.
[936] E. R. Davidson, "Monster matrices: their eigenvalues and eigenvectors," *Computers in Phys.* 7:519, 1993.
[937] A. Booten and H. van der Vorst, "Cracking large scale eigenvalue problems, part i: Algorithms," *Comp. in Phys.* 10:239–242, 1996.
[938] A. Booten and H. van der Vorst, "Cracking large scale eigenvalue problems, part II: Implementations," *Comp. in Phys.* 10:331–334, 1996.
[939] C. G. J. Jacobi, "Über ein leichtes Verfahren die in der Theorie der Säculärstörrungen vorkommenden Gleichungen numerisch aufzulösen," *Crelle's J.* 30:51–94, 1846.
[940] C. M. M. Nex, "A new splitting to solve large hermitian problems," *Comp. Phys. Comm.* 53:141, 1989.
[941] A. P. Seitsonen, M. J. Puska, and R. M. Nieminen, "Real-space electronic-structure calculations: Combination of the finite-difference and conjugate-gradient methods," *Phys. Rev. B* 51:14057–14061, 1995.
[942] Y. Saad, *Iterative Methods for Sparse Linear Systems, 2nd Ed.*, SIAM, Philadelphia, 2003.
[943] C. Lanczos, "An iteration method for the solution of the eigenvalue porblem of linear differential and intergral operators," *J. Res. Nat. Bur. Standards* 45:255, 1950.
[944] H. Q. Lin and J. E. Gubernatis, "Exact diagonalization methods for quantum systems," *Computers in Phys.* 7:400–407, 1993.

[945] L.-W. Wang and A. Zunger, "Large scale electronic strucuture calculations using the lanczos method," *Comp. Mat. Sci.* 2:326–340, 1994.

[946] E. R. Davidson, in *Methods in Computational Molecular Physics*, edited by H. F. Diercksen and S. Wilson, D. Reidel Publishing Co., Dordrecht, 1983, pp. 95–113.

[947] H. Kim, B. D. Yu, and J. Ihm, "Modification of the DIIS method for diagonalizing large matrices," *J. Phys. A* 27:1199–1204, 1994.

[948] A. Canning, L. W. Wang, A. Williamson, and A. Zunger, "Parallel empirical pseudopotential electronic structure calculations for million atom systems," *J. Comp. Phys.* 160:29, 2000.

[949] T. Ericsson and A. Ruhe, "The spectral transformation lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems," *Math. Comp.* 35:1251–1268, 1980.

[950] R. R. Sharma, "General expressions for reducing the Slater-Koster linear combination of atomic orbitals integrals to the two-center approximation," *Phys. Rev. B* 19:2813–2823, 1979.

[951] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Wiley-Interscience, Paris, 1977.

Index