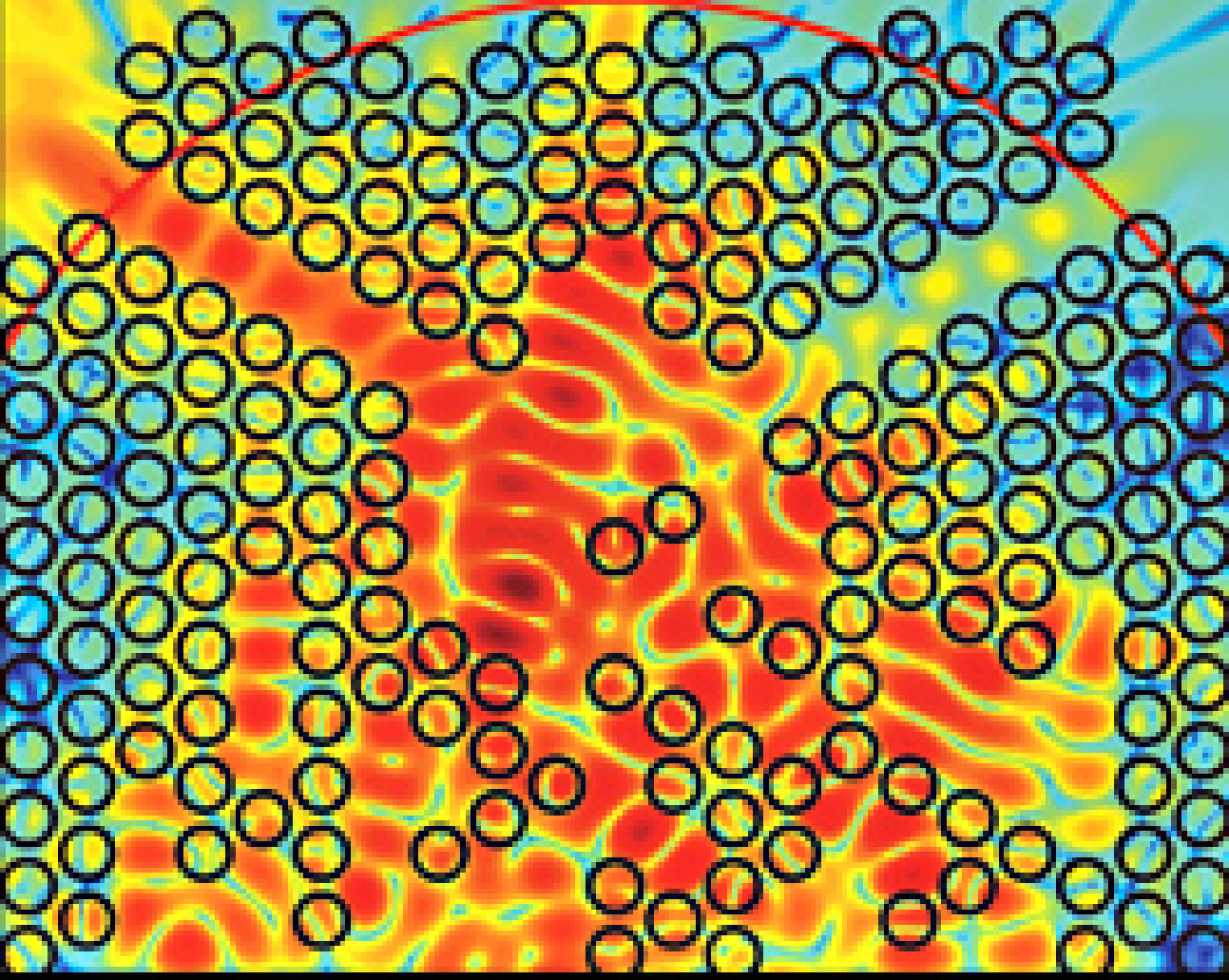

Optimal Device Design

EDITED BY

A. F. J. Levi and Stephan Haas

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521116602

This page intentionally left blank

Optimal Device Design

Explore the frontier of device engineering by applying optimization to nanoscience and device design. This cutting-edge work shows how robust, manufacturable designs that meet previously unobtainable system specifications can be created using a combination of modern computer power, adaptive algorithms, and realistic multi-physics models. Applying this method to nanoscience is a path to creating new devices with new functionality, and it could be the key design element contributing to transitioning nanoscience to a practical technology. Basic introductory examples along with MATLAB code are included, through to more formal and sophisticated approaches, and specific applications and designs are examined. Essential reading for researchers and engineers in electronic devices, nanoscience, materials science, applied mathematics, and applied physics.

A.F.J. Levi is Professor of Electrical Engineering and of Physics and Astronomy at the University of Southern California. He joined USC after working for 10 years at AT&T Bell Laboratories, New Jersey. Professor Levi is the author of the book *Applied Quantum Mechanics*, Second Edition (Cambridge University Press, 2006).

Stephan Haas is Professor of Theoretical Condensed Matter Physics at the University of Southern California.

Optimal Device Design

Edited by

A.F.J. LEVI and S. HAAS

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/9780521116602

First published in print format 2009

ISBN-13 978-0-511-69143-0 eBook (NetLibrary)

ISBN-13 978-0-521-11660-2 Hardback

Contents

Preface

Dramatic advances in the control of physical systems at the atomic scale have provided many new ways to manufacture devices. An important question is how best to design these ultra-small complex systems. Access to vast amounts of inexpensive computing power makes it possible to accurately simulate their physical properties. Furthermore, high-performance computers allow us to explore the large number of degrees of freedom with which to construct new device configurations. This book aims to lay the groundwork for a methodology to exploit these emerging capabilities using optimal device design. By combining applied mathematics, smart computation, physical modeling, and twenty-first-century engineering and fabrication tools it is possible to find atomic and nanoscale configurations that result in components with performance characteristics that have not been achieved using other methods.

Imagine you want to design and build a novel nanoscale device. How would you go about it? A conventional starting point is to look at a macroscopic component with similar functionality, and consider ways to make it smaller. This approach has several potential pitfalls. For one, with continued reduction in size, device behavior will become quantum in character where classical concepts and models cease to be applicable. Moreover, it is limited by ad hoc designs, typically rooted in our unwillingness to consider aperiodic configurations, unless absolutely mandated by physical constraints. Most importantly this conventional approach misses the enormous opportunity of exploring the full landscape of possible system responses, offered by breaking all conceivable symmetries.

Computational resources, realistic physical models, and advanced optimization algorithms now make it possible to efficiently explore the properties of many more configurations than could be tested in a typical laboratory. This is the new paradigm: explore the most improbable, most nonintuitive, system configurations and you will likely be rewarded not only with unprecedented optimized device performance but also with new physical insights into how these small complex systems work.

This book is for those not satisfied with incremental scientific progress but instead willing to explore a vibrant and exciting new direction that acknowledges the richness inherent to small complex systems. We are particularly eager to reach and inspire an emerging generation of computer-savvy individuals who recognize the shortcomings of conventional disciplinary thinking and ad hoc engineering. As incomplete as this book may be in many respects, we wish to show a possible direction out of the science-as-usual mentality. The approach leverages computational resources and advances in controlling and manipulating nanoscale objects. It develops an analysis non-convex via optimization that we believe changes the way one thinks about physical problems.

Chapter 1 offers a statement of the set of problems considered, surveys the mathematical and computational approaches used, and discusses specific applications and designs. The following chapters explore these issues one by one in greater depth, touching on topics at the forefront of our current understanding of nanoscale devices. Of course, as we sincerely hope that this will inspire your scientific thinking and approach towards research, the final word, dear reader, is for you to write.

California S. H. and A. F. J. L.

Acknowledgements

Stephan Haas and Tony Levi are particularly grateful to Dennis Healy for unflinching encouragement and inspiration in many aspects of the research endeavor that made this book possible. They would also like to thank Alexander Balatsky for his insights into how correlated many-body systems work, Ilya Grigorenko for providing powerful computational implementation of adaptive design concepts, and Peter Littlewood for enlightening discussions of complex phenomena.

Dimitris Bertsimas and Omid Nohadani would like to thank J. Birge, F. Kärtner, and K.M. Teo for fruitful discussions and collaboration on some of the projects that are discussed in Chapter 6.

1 Frontiers in device engineering

Philip Seliger and A.F.J. Levi

1.1 Introduction

Today, nanoscience promises to provide an overwhelmingly large number of experimentally accessible ways to configure the spatial position of atoms, molecules, and other nanoscale components to form devices. The central challenge of nano-*technology* is to find the best, most practical, configuration that yields a useful device function. In the presence of what will typically be an enormous non-convex search space, it is reasonable to assume that traditional ad hoc design methods will miss many possible solutions. One approach to solving this difficult problem is to employ machine-based searches of configuration space that discover user-defined objective functions. Such an optimal design methodology aims to identify the best broken-symmetry spatial configuration of metal, semiconductor, and dielectric that produces a desired response. Hence, by harnessing a combination of modern computer power, adaptive algorithms, and realistic physical models, it should be possible to seek robust, manufacturable designs that meet previously unobtainable system specifications. Ultimately one can envision a design process that simultaneously is capable of basic scientific discovery and engineering for technological applications.

This is the frontier of device engineering we wish to explore.

1.1.1 The past success of ad hoc design

For many years an ad hoc approach to device design has successfully contributed to the development of technology. For example, after identifying the cause of poor device performance one typically tries to create a solution by modifying a process or fabrication step. The result is usually a series of

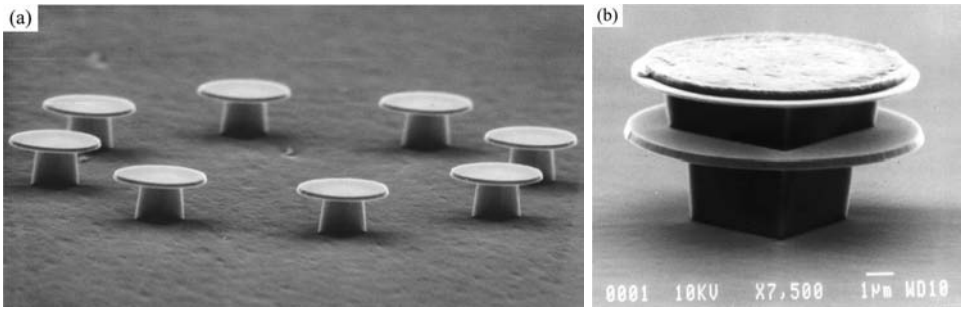

Fig. 1.1. Symmetry, typical of ad hoc device design, is illustrated by SEM micrographs of (a) optically pumped and (b) electrically driven microdisk lasers. The image on the left shows a field of InGaAsP quantum well microdisk lasers supported by InP posts on an InP substrate. Lasing emission is at $\lambda_0 = 1550$ nm wavelength. Disks are 2 μm in diameter and 0.1 μm thick. Nearest-neighbor spacing is 4.7 μm and the outside diameter of the ring of disks is 14 μm. Room temperature threshold power is sub-mW for pump radiation at 850 nm wavelength. The image on the right is an electrically driven 6-quantum well InGaAsP microdisk laser diode that is 10 μm in diameter and 0.3 μm thick. Electrical current is injected from the top metal contact, room temperature threshold current is 2 mA, and lasing emission is at $\lambda_0 = 1550$ nm wavelength.

innovations heavily weighted towards incremental, and hence small, changes in previous practice. The scaling of Complementary Metal Oxide Semiconductor (CMOS) transistors to minimum features sizes of a few nm is a good example of the extraordinary power of such an approach [1, 2].

In addition to incremental improvements there are, of course, new device designs and device concepts that emerge from the research community. Typically, these are also ad hoc in origin and, significantly, tend to have a geometric structure that is highly symmetric. The development of radiation-pressure-driven opto-mechanical resonators is a recent example in which the phenomenon was first explored using highly symmetric toroidal structures [3]. The creation of ultra-small semiconductor lasers, such as the microdisk lasers illustrated in Fig. 1.1, is another [4–6]. The choice of symmetric geometries seems to be a human bias, often driven by ease of analysis. Put simply, symmetric structures are easier to think about. It is symmetric geometries, along with their implicit limited functionality, that are foremost in ad hoc design and are, in general, preferred by the research community.

1.1.2 Looking beyond ad hoc design

Rather than speculate on the reasons for the past success of an ad hoc design methodology, it is more interesting to explore the possibility of an alternative path to design and discovery using new and emerging capabilities such as

nanoscience and access to large computing resources. Part of the motivation comes from the fact that while nanoscience has successfully developed a large number of degrees of freedom with which to create new structures, much of what has been proposed (with the notable exception of quantum computing) is focused on replacing existing electronic and photonic devices such as transistors and lasers with their nearest nanotechnology equivalent. The shortcoming in such an approach is a failure to discover new functions, devices, and systems specific to and only achievable using nanoscience. It is hoped that unbiased machine-based searches for functionality will reveal original, nonintuitive, designs characterized by broken-symmetry geometries.

The rapid and successful development of nanoscale fabrication methods has exposed a critical gap in understanding that appears to represent a very important barrier to fully exploiting nanoscience. Absent from largely experimentally-driven nanoscience research is a methodology or procedure to create new functionalities, new devices, and new system architectures. What is needed is a systematic experimental and theoretical approach that results in the efficient discovery of atom, molecular, and macro-molecular based configurations that exhibit the desired, user specified, functionality. It is the ability to provide made-to-order functions, devices, and systems that will enable the true potential of nanoscience and ensure its adoption in practical systems.

Two key elements of this approach to design are efficient adaptive search algorithms and realistic physical models. Combined they form the basis for the development of optimal design software for small quantum systems. Implemented, such algorithms are capable of discovering initially nonintuitive designs for a given functionality. Typically these designs are highly non-symmetric and usually difficult to interpret. However, as will be illustrated in Section 1.2, sometimes it is possible to analyze the machine-generated solutions and gain new insight into the underlying physical mechanisms driving the system to a given optimal configuration. This potential for learning is another motivation to explore the possibilities of optimal design in nanoscience.

1.2 Example: Optimal design of atomic clusters

The density of electronic states in a solid is a basic attribute that plays a key role in determining material properties. For example, a singular behavior in the density of quasi-particle states can result in enhanced optical activity at a specific photon energy. In fact, a periodic array of atoms in a crystal gives rise to just such peaks in the density of states. This may be illustrated by

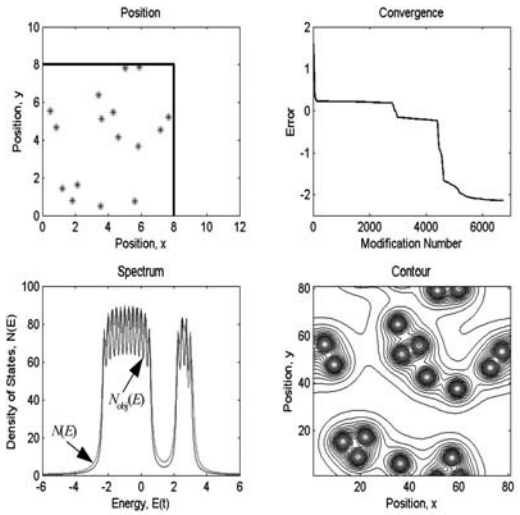

Fig. 1.2. Optimized position of 16 atoms in a square of dimension 8×8 (upper left) giving asymmetric density of states $N(E)$ very close to the objective spectrum $N_{\rm obj}(E)$ [7]. Energy scale is in units of t. Contour plot of interaction potential for atoms in optimized positions is shown in lower right panel. In the calculations $\alpha = 3$, periodic boundary conditions are used, and $\Gamma = 0.2828 \times t$. Convergence as a function of modification number is shown in upper right panel.

considering a two-dimensional square lattice in the nearest-neighbor tight-binding approximation with s-orbitals and eigenenergies E_i. Crystal symmetry and interaction mechanism determine the density of states spectrum. In this case, there is a peak in the density of states at the center of the band.

What we would like to do is find configurations of atoms that are not constrained by crystal symmetry. A key idea is that breaking the spatial symmetry of atom positions creates a truly vast number of possibilities, making it feasible to find configurations of atoms with essentially any desired density of states. The ability to control the response of a material in a user-defined way is a powerful concept which has the potential to change the way one views materials, devices, and systems.

As a first step, consider an algorithm that seeks spatial configurations of atoms characterized by a user-specified or objective density of electronic states $N_{\rm obj}(E)$. The essential physics underpinning the approach is illustrated by considering a *long-range* version of the atomic tight-binding model with Hamiltonian

$$\hat{H} = -\sum_{i,j} t_{i,j} \left(\hat{c}_i^\dagger \hat{c}_j + \hat{c}_i \hat{c}_j^\dagger \right), \tag{1.1}$$

where $c_i^\dagger$ and c_i are creation and annihilation operators respectively at the

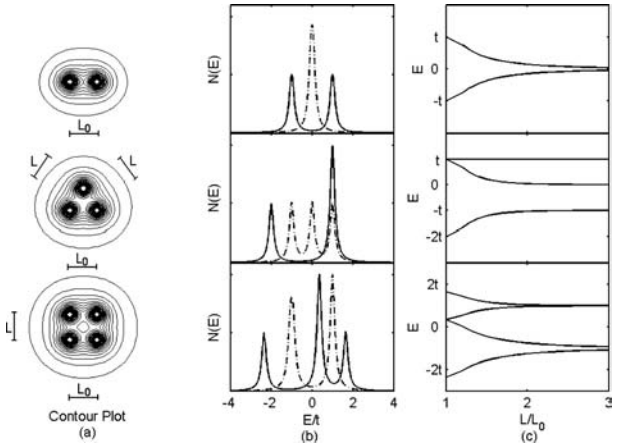

Fig. 1.3. Calculated density of states $N(E)$ for dimer, trimer, and quadrumer with $\alpha = 3$ and showing a hierarchy of non-symmetric contributions as a function of atom separation L normalized to L_0. Energy scale is in units of t. An isolated pair of atoms (dashed line) is symmetrically split (solid line) by a dimer. A trimer forming an equilateral triangle ($L/L_0 = 1$) has an asymmetric density of states (solid line), becoming essentially symmetric when $L/L_0 = 3$ (dashed line). Peak positions for the quadrumer are also controlled by atom separation in the range $L/L_0 = 1$ (solid line) to $L/L_0 = 3$ (dashed line) [7].

atom site $\mathbf{r}_i$. The overlap integrals t_{ij} between an atom at position $\mathbf{r}_i$ and an atom at position $\mathbf{r}_j$ are parameterized by a power law $t_{ij} = t/|\mathbf{r}_i - \mathbf{r}_j|^\alpha$, where t sets the energy scale. The choice of exponent α depends on details of the experimental situation. Here, the Hamiltonian matrix in the basis of single particle states is non-sparse because interaction with all atoms is included. For simplicity only s-orbitals are considered, so it is not necessary to include directionality of atomic electron wave functions. The density of states is

$$N(E) = \sum_i \frac{|\Gamma|/\pi}{(E - E_i)^2 + (\Gamma/2)^2}, \tag{1.2}$$

and Γ is the characteristic energy broadening of each eigenenergy, E_i.

To demonstrate the power of optimal design, consider the non-symmetric objective density of states spectrum in two dimensions, $N_{\rm obj}(E)$, indicated in the lower left of Fig. 1.2. The optimization algorithm finds a spatial configuration of 16 atoms in an 8×8 area with periodic boundary conditions that has a density of states, $N(E)$, essentially identical to the desired or objective spectrum [7]. The implication is both apparent and dramatic: a user who requires new material with a specific quasi-particle density of states can use optimal design software to discover configurations of atoms with the desired behavior. The objective functionality is obtained by broken symmetry so, in

this sense, broken symmetry *is* function.

It is clear from the atom positions indicated in Fig. 1.2 that one could not have guessed the result. However, the output of the computer program can be used to gain new insight into configurations that result in the desired spectrum $N_{\rm obj}(E)$. In this particular case, and as illustrated in Fig. 1.3, one learns that a hierarchy of primitive configurations exists that form the building blocks for any objective density of states. Dimers can be used for symmetric $N(E)$, trimers and larger molecular configurations provide asymmetry to $N(E)$. While, in a strict sense, these heuristics only apply to the dilute limit in which the normalized average spacing between atoms is much greater than unity, it is apparent one may appeal to this insight to explain the more complicated structures that occur in the dense limit.

Remarkably, some aspects of the model have been confirmed experimentally using scanning tunneling microscopy (STM) to precisely position gold atoms on the surface of a nickel-aluminum crystal. STM measurements [8] show that the splitting in the value of eigenenergies E_i for Au dimers on NiAl depends inversely on Au atom separation corresponding to $\alpha = 1$ in the expression $t_{ij} = t/|\mathbf{r}_i - \mathbf{r}_j|^{\alpha}$.

Chapter 2 discusses optimal design of atomic clusters in more detail.

1.3 Design in the age of quantum technology

One of the greatest achievements of semiconductor technology has been the continuous reduction in transistor minimum feature size over the past 35 years. Often described as Moore's Law [1] or scaling, Fig. 1.4 illustrates the historical exponential reduction in CMOS gate length with time. Of course, at some point physical and other limitations will force such geometric scaling to end. Today there seems to be a consensus that a manufacturable technology with minimum feature sizes below 10 nm is achievable [2]. This confidence is based partly on improvements in lithography tools and partly on experience overcoming previously declared limits to scaling [9]. Nevertheless, sometime after 2020 Moore's Law will come to an end and new paths to system innovation will have to be found. Our concern is not how to achieve minimum feature size below 10 nm but rather the approach to design when such capability is available because it is on these nm length scales that the best opportunities to exploit quantum effects will occur.

When simple physical scaling of device geometry no longer provides a path to increased system functionality, improved device performance and function might be achieved by manipulating new quantum degrees of freedom. Examples might include controlling the single electron states of atomic

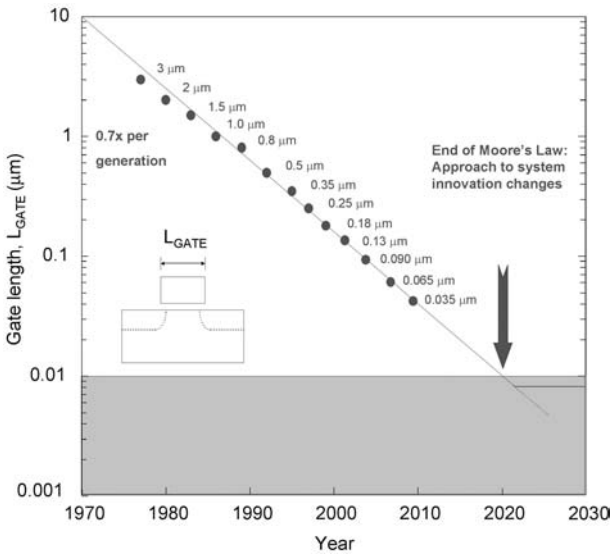

Fig. 1.4. Illustrating reduction in CMOS gate length with time. Gate length has been decreasing, or scaling, consistently for the past 35 years. However, physical and other limitations to continued scaling will impact by about the year 2020 and geometric scaling will come to an end. As this end-point nears, fundamental changes in the approach to system innovation are required.

and nm-sized particles via geometry [10, 11], using interacting electrons in the presence of the coulomb interaction to exploit collective excitations such as plasmons [12], hybridization to control bonding and chemical specificity, using electron and orbital spin to control magnetic response [13], strong light-matter interaction in nm-sized geometries [14, 15], and non-equilibrium processes on fs time scales [16].

Because of the large number of variables, it seems reasonable to consider avoiding ad hoc design and trying to apply a systematic approach to problem solving and analysis. Such an approach is more likely to reap dividends as we move away from devices that behave semi-classically and into a less familiar quantum regime where our intuition might fail. However, to date, much of what has been explored in nanoscience is the result of curiosity-driven research with little direct connection to practical technology development. This approach to discovery may not only be inefficient but may also be susceptible to replacement by more effective methods.

1.3.1 High performance heterostructure bipolar transistors

An example of a high-performance electronic device designed in an ad hoc fashion but with nm control of material composition in one dimension is the heterostructure bipolar transistor (HBT).

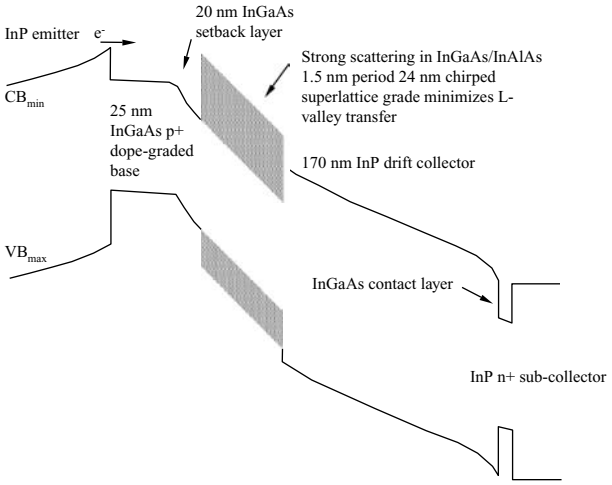

Fig. 1.5. A high-performance heterostructure bipolar transistor (HBT) band profile. The InP emitter injects electrons into the heavily doped p-type base. The collector contains a complex chirped InGaAs/InAlAs superlattice. Designs similar to this have achieved extrapolated characteristic frequency response in excess of 500 GHz at room temperature [17].

The semiconductor band profile of just such an HBT is illustrated in Fig. 1.5. As may be seen, the band profile is quite complex. It is, in fact, the result of many years of iterative improvement in design culminating in the demonstration of an extrapolated characteristic frequency response in excess of 500 GHz at room temperature [17].

Despite this remarkable success, the emitter, base, and complex collector structure has never been systematically optimized. In part this is due to the fact that no reliable, efficient, physically realistic model of electron transport has been developed. Today, no one knows if the designs that have been implemented are optimal, or even close to optimal.

The next section illustrates how one might approach an optimal design strategy for one aspect of nanoscale electronic device design.

1.3.2 Control of electron transmission through a tunnel barrier

As a prototype system, consider the semiconductor $Al_xGa_{1-x}As$, which has a lattice constant 0.5653 nm and atomic layer separation 0.2827 nm. Atomically precise layer-by-layer crystal growth is possible using molecular beam epitaxy (MBE) and the $Al_xGa_{1-x}As$ alloy can be used to form heterojunctions with controlled conduction and valence band off-sets.

As a specific example [18], consider electron transport through a rectangular $Al_xGa_{1-x}As$ barrier of energy $V_0 = 0.3\,\text{eV}$ and width $L = 4\,\text{nm}$, as shown

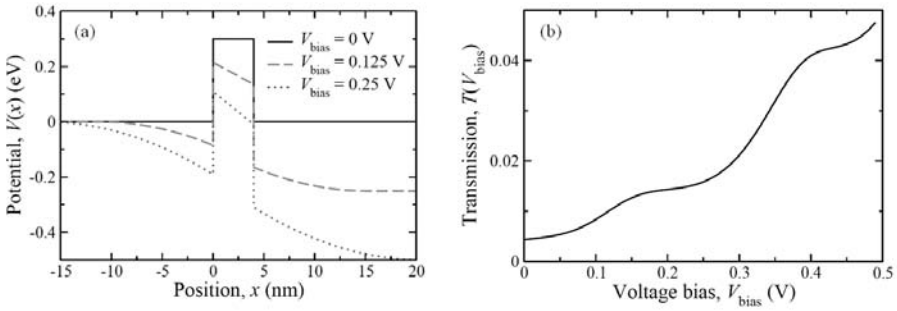

Fig. 1.6. A rectangular potential barrier of energy $V_0 = 0.3\,\text{eV}$ and width $L = 4\,\text{nm}$ gives rise to rapid increase in electron transmission with increasing voltage bias, V_{bias} and resonances. Effective electron mass is $m = 0.07 \times m_0$, where m_0 is the bare electron mass. (a) Conduction band profile of the rectangular potential barrier for the indicated values of V_{bias}. (b) Transmission probability as a function of V_{bias} for an electron of energy $E = 26\,\text{meV}$ incident from the left.

in Fig. 1.6. The barrier is sandwiched between n-type GaAs electrodes with carrier concentration $n = 10^{18}\,\text{cm}^{-3}$. Applying a bias voltage, V_{bias}, results in a depletion region on the right side and an accumulation region on the left side of the barrier. The form of the conduction band profile $V(x)$ in these regions is calculated by solving the Poisson equation. Net electron motion is in the x direction, normal to the barrier plane and there is no confinement in the y and z directions, thereby avoiding possible detrimental consequences of quantized conductance [19–21]. A numerical solution to the Schrödinger equation is obtained piecewise by discretizing the potential profile into 4,000 steps, matching boundary conditions at each interface, and implementing the propagation matrix method [22–24]. An electron of energy $E = 26\,\text{meV}$ incident from the left is partially reflected and partially transmitted, as determined by the wave function boundary conditions $\psi_j = \psi_{j+1}$ and $\partial\psi_j/\partial x = \partial\psi_{j+1}/\partial x$ at each interface. Here ψ_j is a solution of Schrödinger's equation in region j with wave vector $k_j = \sqrt{2m(E - V_j)}$, where V_j is the local potential in the conduction band and m is the effective electron mass.

Exponential increase in electron transmission with bias voltage is a generic feature of the simplest barrier profiles. Potential wells, on the other hand, are known to produce bound-state resonances, leading to sharp transmission peaks. Hence, design of structures with linear and other power-law transmission-voltage characteristics likely involves broken-symmetry potential barrier profiles. As an initial challenge in our exploration of this possibility we use an adaptive quantum design approach to find a potential profile with a transmission function $T(V_{\text{bias}})$ that increases linearly with bias voltage in the window $0\,\text{V} < V_{\text{bias}} < 0.25\,\text{V}$.

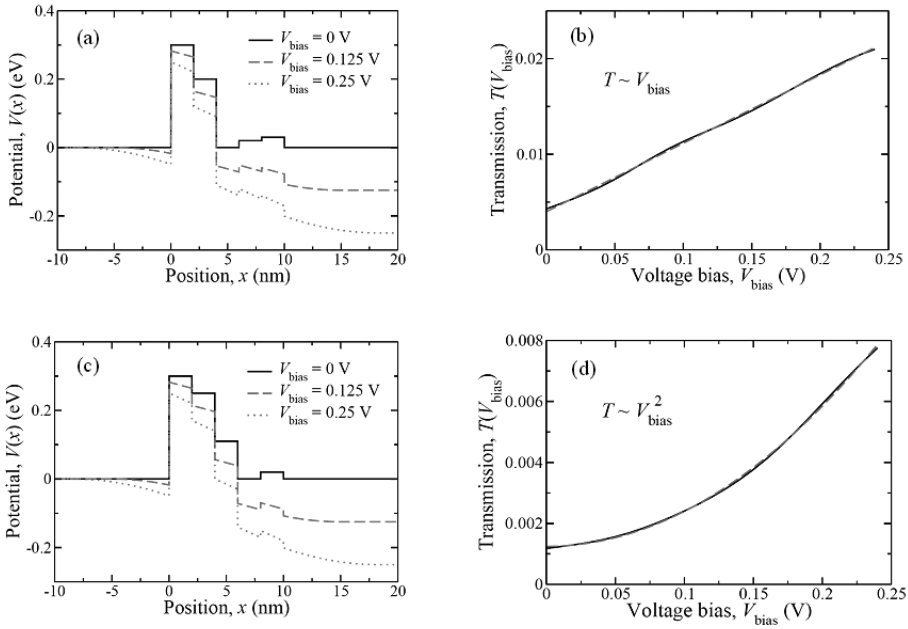

Fig. 1.7. (a) and (c) are solutions from exhaustive numerical searches for conduction band profiles $V(x)$ that yield linear and square dependences of electron transmission as a function of bias voltage, V_{bias}. $V(x)$ is constrained to a region that is 10 nm wide and the maximum local potential is 0.3 eV. The resulting $T(V_{\mathrm{bias}})$ for an electron of energy $E = 26$ meV incident from the left are shown as solid lines in (b) and (d). Broken line is the objective response.

The conduction band potential energy profile is defined on a grid with $\Delta x = 2$ nm (about 8 monolayers in GaAs) spatial increments and $\Delta V = 0.01$ eV energy increments. The numerical search for optimal broken-symmetry barrier profile is constrained to take into account physical as well as computational limitations. Physically, varying the composition of an $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ alloy controls the conduction band potential profile. Each alloy plane normal to the growth direction has an average local potential, $V(x)$. Fabrication inaccuracies of 1–2 monolayers may occur in the epitaxial growth processes, and hence the targeted transmission functionality needs to remain stable against such variations. Moreover, the Al concentration can only be controlled to within a few percent. Computationally, the dimensionality of the search space needs to be constrained in order to match the available computer hardware capabilities. In this example, to keep the search space finite, we focus on nanoscale barrier structures of total width $L = 10$ nm with a maximum on-site potential of 0.3 eV measured from the GaAs conduction band minimum.

Figure 1.7 shows solutions from exhaustive numerical searches for conduction band profiles that give linear and square-law $T(V_{\text{bias}})$ characteristics. For the discrete grid discussed above, the size of the search space is $30^5 = 2.4 \times 10^7$. The resulting broken-symmetry barrier solutions are sequences of rectangular steps. For the case of a linear objective function (Fig. 1.7(a) and (b)), the quadratic deviation of the obtained solution from the objective is $\chi^2 = 5.1 \times 10^{-7}$. When the search is restricted to monotonically decreasing potentials the reduced size of the search space allows one to consider a finer grid in space with $\Delta x = 1$ nm (~ 4 monolayers in GaAs). The size of the search space is now $40!/(10!30!) \sim 8.5 \times 10^8$. In this case the quadratic deviation of the obtained solution from the linear objective is $\chi^2 = 4.5 \times 10^{-6}$. In Fig. 1.7(c) and (d) the solution of an exhaustive search for a barrier profile with a quadratic transmission-voltage characteristic is shown. In this case, the square deviation between solution and objective is $\chi^2 = 5.6 \times 10^{-8}$.

These results show it is possible to construct semiconductor nanoscale structures with desired linear and power-law electron transmission-voltage characteristics. In such devices, elastic scattering limits ballistic electron current flow and dissipative relaxation processes occur in the electrodes. However, there is a hierarchy of objective functionalities, some are more accessible than others using the available building blocks. For example a square-root transmission-voltage objective response poses a much more challenging problem and the best solution identified by the exhaustive numerical search for this functionality only has $\chi^2 = 6.6 \times 10^{-5}$.

It seems remarkable that a linear response in a nanostructure with ballistic electron transport may be achieved by solely utilizing the physical ingredient of elastic scattering and tunneling at potential steps. To better understand the physics enabling power-law transmission as a function of V_{bias}, consider the progressive evolution of solutions for the linear objective from a simple square barrier to the multi-barrier profile of Fig. 1.7(a). As illustrated in Fig. 1.8(a) and (b), the dominant transmission features of the simple square well, i.e. the exponential behavior and resonances, are altered by the addition of steps in the potential barrier profile. It is observed that the superposition of broad resonances due to the presence of different potential steps helps linearize the transmission-voltage curve. It is also this superposition of broad scattering resonances which renders the solution stable against small perturbations.

Robustness of a solution to monolayer changes at each interface in the barrier array may be explored using sensitivity analysis. Results presented in Fig. 1.8(c) show that the slope of $T(V_{\text{bias}})$ is determined by the initial highest

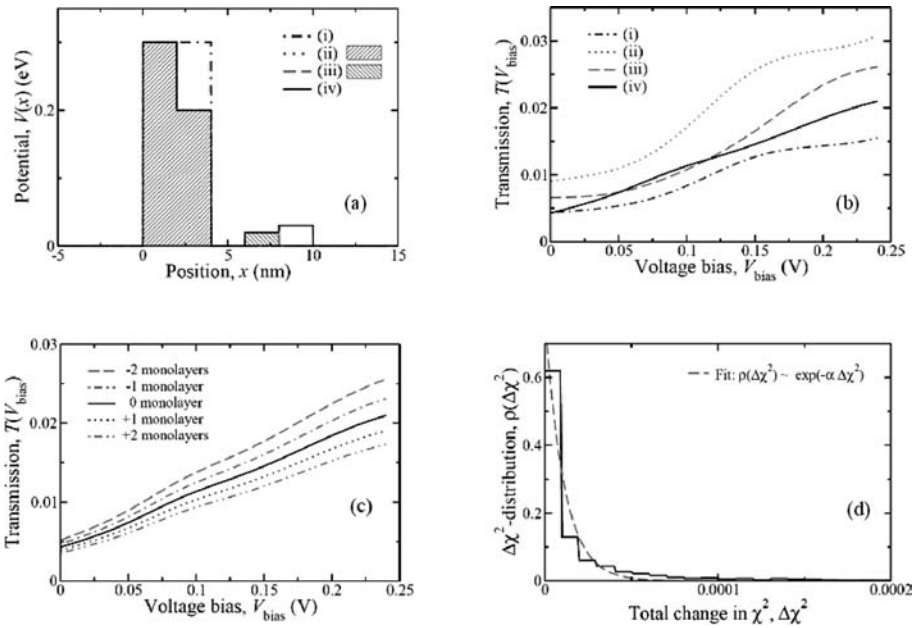

Fig. 1.8. (a) Evolution from a single potential barrier to an array of barriers. (b) $T(V_{\rm bias})$ for the potentials in (a). The superposition of broad resonances enables a linear transmission-voltage response. (i) Single rectangular barrier of 4 nm, (ii) 4 nm wide barrier with a step, (iii) same as (ii) plus 1 nm wide small barrier, (iv) optimized potential profile. (c) Changes of 1–2 monolayers at the initial highest barrier mainly alter the slope of the transmission curve, i.e. the resistance. (d) Randomly selected 0, 1, and 2 monolayer changes at all interfaces only lead to small deviations in χ^2. The characteristic variation in χ^2 is $1/\alpha = 1.1 \times 10^{-5}$.

barrier energy, whereas its smoothness is governed by the low barrier energy tail of the array that controls the fast spatial modulations of the electron wave in the structure. Therefore, depending on the position of the deviations from the original structure, different components of the response function are affected. The effect of smoothing the edges in the conduction band profile has also been explored. Error function rounding of interfaces on the scale of two monolayers changes the linear transmission-voltage response of Fig. 1.7(b) only slightly, yielding a quadratic deviation of $\chi^2 = 5.7 \times 10^{-7}$.

In Fig. 1.8(d) results are shown from a study of randomly selected 0, 1, and 2 monolayer changes at all interfaces in the barrier structure. The effect is relatively small, yielding an average change of $1/\alpha \sim 10^{-5}$ in χ^2, where α is the parameter of the exponential fit in Fig. 1.8(d). Moreover, the sensitivity of χ^2 to changes in the potential energy due to random variations on the order of 1% in the Al concentration of $Al_xGa_{1-x}As$ is found to be similar. Hence, the transmission characteristics enabled by the conduction band

potential profile discussed here are stable against small random variations.

In this section we have demonstrated that adaptive design can be applied to the synthesis of nanoscale devices with power-law transmission-voltage characteristics. Using a constrained exhaustive numerical search, broken-symmetry conduction band profiles of semiconductor heterostructures have been identified which enable desired quantum transport functionality. This custom-designed superposition of broad scattering resonances due to the presence of an optimal potential makes it possible to emulate an Ohm's law transfer function within a window of bias voltages in a system that is dominated by ballistic electron transport. In particular, one can synthesize a nanoscale two-terminal linear resistive element that is robust against monolayer perturbations.

It may be shown numerically that the solution space is non-convex in the design parameters. A global search reveals that there are multiple local minima, but there is only one globally optimal design.

Nanoelectronic device design is discussed further in Chapter 3 and some aspects of the mathematics associated with its optimization are given in Chapter 7.

1.3.3 The need for improved physical models

The elementary barrier transmission problem discussed in the previous section turned out to be quite rich in physical effects. However, it is obvious that the ability to discover useful new functions or improved performance will depend strongly on the realism of the physical model being used. Such models must be capable of providing sufficient variation in types of solution that there is opportunity to explore nonintuitive designs that might lead to new functionality and new understanding.

There is then, a need for the device physics community to create improved models of electron transport as well as of the interaction of light with matter. Both present challenges that relate to many-body aspects of the problem. For example, recently a theory that describes the non-local linear dielectric response of nano-metal structures to an externally applied electric field has been developed [25]. Unlike the conventional phenomenological classical theory [26] for light–metal interaction, the new model is able to describe the transition from classical to quantum response as well as the coexistence of classical and quantum response in structures of arbitrary geometry via the non-local response function $\epsilon(\mathbf{r}, \mathbf{r}', \omega)$. This is of some practical importance because metallic nanoclusters can now be made sufficiently small such that non-local effects due to finite system size and cluster shape dominate the spectral response. In particular, when the ratio between the smallest

characteristic length scale and the Fermi wavelength is comparable to unity, these systems can fail to fully screen external driving fields. Also, in the quantum limit one needs to take into account discreteness of the excitation spectrum as well as the intrinsically strong damping of collective modes. The ability of a model to capture these single and many particle quantum effects is a first step towards accessing new regions of design space with their promise of new device functionality.

As another example, consider a heterostructure bipolar transistor (HBT) in which extreme non-equilibrium electron transport occurs. Modeling the transfer characteristics of such a transistor without including the possibility of inelastic scattering is unrealistic because the existence of base current requires such processes. However, elastic quantum mechanical reflections from an abrupt change in potential at a heterostructure interface strongly influence inelastic scattering rates [27] and so the wave nature of the electron must be included in any model of electron transport. In fact, it has been known for some time that under these conditions base and collector regions may no longer be treated separately [28] and one should model the structure as a single inhomogeneous anisotropic scatterer. Within linear response this requires evaluation of the non-local dielectric response $\epsilon(\mathbf{r}, \mathbf{r}', \omega)$ for the entire transistor structure. Unfortunately, not only can elastic quantum mechanical reflections strongly influence inelastic scattering rates, dissipative processes can in turn modify quantum reflection and transmission rates. This type of feedback can be driven by unitarity [29, 30] and is beyond the self-consistent perturbation theory used so far to calculate non-local dielectric response. An additional difficulty is creating a theory capable of self-consistently including the contribution of non-equilibrium electron distributions to scattering rates. Ultimately, what is needed is a new approach to describing quantum electron transport in modern devices to replace the old semi-classical methods that are no longer either adequate or valid.

Even if a physically realistic model exists that exhibits a suitably rich functionality in its solution space, it is essential that the model be computationally efficient. This is because optimal design will typically require many evaluations of the forward problem.

1.4 Exploring nonintuitive design space

By now it should be clear that there are three categories of design space that can be explored. These are illustrated in Fig. 1.9. Intuitive design space typically consists of elements with spatial symmetry arranged symmetrically and with a low amount of physical connectivity between elements. The

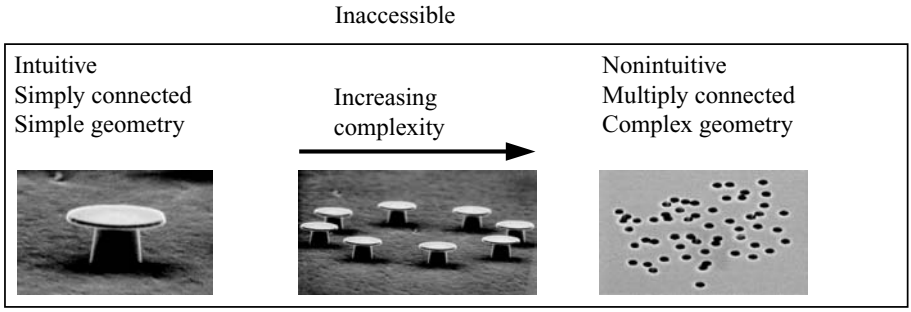

Fig. 1.9. Intuitive and nonintuitive design space. Intuitive design is characterized by elements with simple geometry arranged symmetrically and with low connectivity. This illustration shows a highly symmetric microdisk laser. Nonintuitive design exhibits high spatial complexity and high connectivity. The illustration shows a nano-photonic device with a nonintuitive design that consists of 300 nm diameter cylindrical holes placed in a silicon dielectric slab waveguide. The position of the holes is aperiodic and the device function cannot easily be guessed by looking at the geometry. At the boundary between intuitive and nonintuitive design there is overlap where, for example, a simply connected geometry might exhibit nonintuitive complex collective behavior. The illustration also shows that there are regions of device response that are inaccessible.

illustration shows a microdisk laser that satisfies this criterion. On the other hand, nonintuitive design exhibits high spatial complexity and high connectivity. The illustration shows a nano-photonic device with nonintuitive design that consists of 300 nm diameter cylindrical holes placed in a silicon dielectric slab waveguide. The position of the holes is aperiodic and the device function cannot easily be guessed by looking at the geometry. It is an example of nonintuitive design.

There is an intersection between intuitive and nonintuitive design. This occurs when simply connected symmetric geometries exhibit complex behavior. In the illustration, this might occur when coupling between adjacent microdisks exhibits collective behavior.

There are also regions of device response that are inaccessible. For example a designer might want to have a laser with performance that is impossible to achieve using the physical model and design parameters available.

To systematically explore a design space it is necessary to formalize the problem. In the next section we address this by starting to develop a mathematical formulation of the optimal device design problem.

1.5 Mathematical formulation of optimal device design

As a starting point, we assume that device performance can be predicted using a physical model and that the design parameters of this physical model

are p. The objective is to choose design parameters p that control a measurable quantity s of device performance in a desired manner. To quantify the difference between the desired device performance and the predicted behavior for a given set of design parameter values we define a cost functional J. A better performing design is associated with a lower cost so that design optimization is formulated as a minimization problem. A typical cost function is the least squares performance measure given by

$$J = \sum_{i=1}^{M} \left| s_{\mathrm{obj},i} - s_{\mathrm{sim},i} \right|^2 . \tag{1.3}$$

Often s is a real measurable scalar quantity. However, it can also be a vector. The choice of a least squares cost function has the advantage that it guarantees parabolic, convex, local minima. The least squares performance measure calculates the cumulative difference between the desired objective device response $s_{\mathrm{obj},i}$ and the simulated device response $s_{\mathrm{sim},i}$ at M points. The predicted $s_{\mathrm{sim},i}$ is rarely given as an explicit function of p and the evaluation of $s_{\mathrm{sim},i}$ often involves a computationally costly forward solve of the physical model. During the forward solve $s_{\mathrm{sim},i}$ is computed as the solution of a partial differential equation (PDE) that models the dynamics of the device. An appropriate numerical method such as finite difference, finite element, or a more specialized scheme is chosen to numerically approximate the solution of the given initial-value or boundary-value problem. Usually, the greatest computational burden arises during the forward solve. Typically the computational cost also increases with the accuracy of the numerical approximation of the physical system. Because optimization requires that many forward solves be evaluated, an accurate but efficient forward solver must be used. Most numerical schemes result in one or multiple linear systems that must be solved to evaluate $s_{\mathrm{sim},i}$. Mathematically, the physical system is modeled by solving the linear system

$$\mathbf{L}(p) \cdot x = b, \tag{1.4}$$

where $\mathbf{L}(p)$ contains the device dynamics, the vector x is a physical unknown, and b contains the boundary conditions or other terms driving the system. The matrix $\mathbf{L}(p)$ is an explicit function of the design parameters. Generally, the quantity of interest is derived from the physical unknown x

$$s_{\mathrm{sim},i} = W_i(x), \tag{1.5}$$

and the row-vector W can be a simple operator or even the identity operator in which case $s_{\mathrm{sim},i} = x_j$ for some element j of the vector x. The optimal design problem can now be formulated as a mathematical minimization

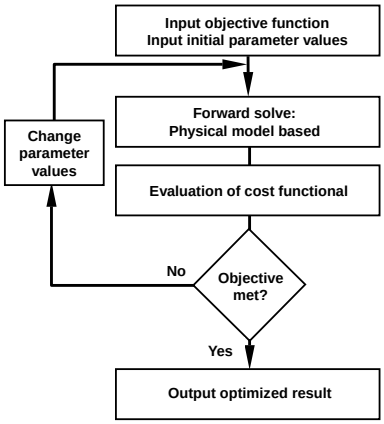

Fig. 1.10. Conventional optimization flowchart.

problem:

$$\begin{aligned} &\min_p J, \\ &\text{subject to} \\ &p \in \Omega \\ &\mathbf{L}(p) \cdot x = b \\ &s_{\mathrm{sim},i} = W_i(x), \end{aligned} \tag{1.6}$$

and Ω represents the feasible (accessible) design parameter space. After each forward solve the cost functional J is evaluated and it can then be determined if the design objectives are met. This occurs if the cost function is sufficiently small and if any additional auxiliary optimality conditions are met. Optimal parameter settings are therefore given by a local minimum of the cost function. The sufficient conditions for locally optimal parameter settings p^* are given by the first and second order conditions

$$\nabla_p(J)\big|_{p^*} = 0, \tag{1.7}$$

and

$$H_p(J)\big|_{p^*} > 0, \tag{1.8}$$

if no constraints are active and H denotes the Hessian operator. Appropriate variations of these conditions are satisfied if there are active constraints at local minimum p^*.

To find locally optimal design parameters an iterative optimization scheme is used as is illustrated in the flowchart shown in Fig. 1.10. The sequence of steps for finding a locally optimal design starts with an initial parameter setting. The device response is calculated using a forward solve. Once $s_{\mathrm{sim},i}$

is computed the cost function J can be evaluated and finally the termination conditions are checked. If the termination conditions are satisfied the optimal design is returned to the user, otherwise the current design parameters are changed and the iterative process continues with the next forward solve.

In many engineering applications the cost function J is a continuously differentiable function of the design parameters p. Efficient local optimization hinges on the choice of a local perturbation of p_l at iteration l and the choice of p_{l+1}. The gradient $\nabla_p(J)$ is the direction that locally causes the greatest decrease of J when p_l is varied. Even though J is an implicit function of p the gradient can be evaluated with extreme efficiency using the adjoint method.

1.6 Local optimization using the adjoint method

To evaluate the gradient of the cost functional J with minimal additional computational cost the adjoint method is used. The N-dimensional gradient $\nabla_p J = (\partial_{p_1} J, \ldots, \partial_{p_N} J)$ can be computed directly by first applying the standard chain rule to (1.3) yielding

$$\partial_{p_k} J = -\sum_{i=1}^{M} 2\left(s_{\mathrm{obj},i} - s_{\mathrm{sim},i}\right)\left(\partial_x W_i\right)\left(\partial_{p_k} x\right), \tag{1.9}$$

where the derivative $\partial_x W_i$ is usually simple and explicit. Using the identity $-\mathbf{L}\cdot\partial_{p_k} x = (\partial_{p_k}\mathbf{L})\cdot x - \partial_{p_k} b$ and the solution h to the adjoint equation

$$\mathbf{L}^T\cdot h = \sum_{i=1}^{M} 2\left(s_{\mathrm{obj},i} - s_{\mathrm{sim},i}\right)\left(\partial_x W_i^T\right), \tag{1.10}$$

the derivative is given directly by

$$\partial_{p_k} J = -h^T\cdot\mathbf{L}\cdot\left(\partial_{p_k} x\right) = h^T\cdot\left(\left(\partial_{p_k}\mathbf{L}\right)\cdot x - \partial_{p_k} b\right). \tag{1.11}$$

Here, N is the dimensionality of the design parameter space and $\mathbf{L}^T$ and W_i^T are the transpose of $\mathbf{L}$ and W_i, respectively. The adjoint variable h must be computed by solving the linear system Eq. (1.10) at the computational cost of one additional forward solve. This is in contrast to N additional forward solves that would be required to approximate the gradient by first-order finite differences. It is remarkable that the efficiency of this method of evaluating the gradient remains practically constant with increasing dimensionality of parameter space Ω, as long as the computational cost is dominated by the forward solve.

It is worth noting that for a compact design parameter space Ω and continuous cost function the Weierstrass theorem guarantees the existence of a global minimum [31]. However, the uniqueness or even cardinality of the set of globally optimal points is generally not known.

As discussed in Chapter 7, local optimization methods, in particular methods using the gradient information, have been well studied and are very efficient even in high dimensions. Finding a global optimum on the other hand is the real challenge of optimal device design. The topology of the device performance over Ω is in general highly nonlinear and non-convex. An example of a smooth, non-convex, two-dimensional function is illustrated in Fig. 1.11(a). The image shows the function values of a sum of multiple exponential functions given by $\sum_{k=1}^{M} m_k \exp\left(-(x-x_k)^2-(y-y_k)^2\right)$ for uniformly distributed seed-points (x_k, y_k) and normally distributed peaks m_k. The gradient information, which is essential to efficient local optimization, is shown in Fig. 1.11(b). Figure 1.11(b) illustrates how most of the features of the smooth cost function in (a) can be reconstructed from the function values and gradients evaluated at only a few points.

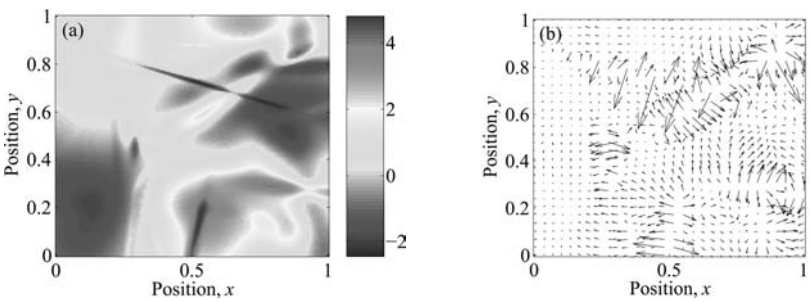

Fig. 1.11. (a) A highly non-convex two-dimensional function illustrated using linear gray scale. (b) The gradient of the function displayed in (a) at uniform grid points. The direction of steepest descent is the most efficient direction for local perturbations of the design parameters x and y while seeking a local minimum.

An example of a highly nonlinear cost function in RF design concerns the scattering of an electromagnetic wave by dielectric cylinders. The design objective in this particular case is to scatter an incident electromagnetic beam by a fixed angle. The cost function measures the difference between the achieved and objective power profiles along a predefined path. Figure 1.12 shows a two-dimensional cut through the four-dimensional cost function space. The image is computed by fixing the location of one scattering cylinder while moving the second scattering cylinder on a uniform grid through the search space. In Fig. 1.12 it is observed that the nonlinearity of the objective function is in this case dominated by the interference patterns

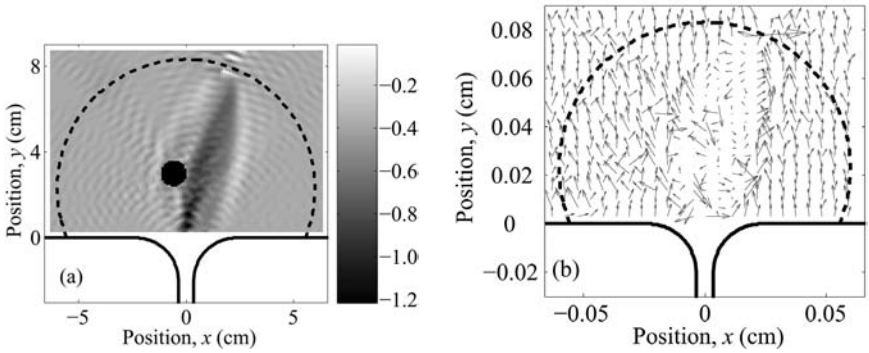

Fig. 1.12. (a) Non-convex cost function for a system similar to that described in [32]. Two dielectric cylinders each are located in the scattering domain to achieve a design objective along the dashed semicircular line. Each cylinder has diameter 6 mm and relative permittivity 3.7. Electromagnetic radiation at frequency 37.5 GHz scatters from the dielectric cylinders. The dark circle located near the center of the parameter space is the location of one dielectric cylinder with fixed position. The grey scale represents the cost function at the position of the second cylinder on a logarithmic ($\log_{10}$) scale. (b) The gradient of the cost function displayed in (a) at uniform grid points. At the given sampling density the information gained by evaluating the gradient is not very helpful in reconstructing the interference pattern shown in (a) nor for locating the global optimum.

associated with the two scattering cylinders. Nevertheless, it is apparent that this cost function has many local minima of similar performance with only a few local minima that perform markedly better than all others. This type of multiple local minima global optimization problem is computationally intensive, a fact that becomes increasingly problematic when the design involves a large number of scattering cylinders [32].

Despite an enormous amount of effort that has been put into solving continuous global optimization problems, it remains an unresolved issue. In the worst case the difficulty of approximating a global minimum increases exponentially with the dimensionality of Ω. An additional challenge of this type of distributed parameter optimization is that the capability of the design is unknown. Unlike inverse problems where parameters for a measured scattering profile need to be determined, the objective for optimal design problems in engineering is often not attainable. Even though no algorithm can solve this problem in finite time there are theoretical results about the existence of a global optimum. The Weierstrass theorem states that any continuous function attains its global minimum on a compact set where the cost function is defined. Because most search spaces in engineering can be approximated as closely as desired by a compact set interior to the search space, the existence of the global optimum is guaranteed. We note that nothing is said about the number of global optima. Many algorithms have been devised to locate a

global optimum and it can be said that no single algorithm performs exceedingly well on all continuous global optimization problems. Depending on the nature of the problem, one algorithm or a hybrid of multiple algorithms and techniques can perform better than others. In the following section some aspects of global optimization are discussed.

1.7 Global optimization

The literature on optimization algorithms is varied and broad. See Appendix A. For any reasonably large search space Ω the consensus is that stochastic algorithms outperform deterministic algorithms unless additional information about the cost function and its domain is given. To see why, consider the pure random search which generates random points over the feasible set and stores the lowest achieved cost function value up to that point. The expected number of function evaluations required by such an algorithm to locate a global optimum grows exponentially with the number of dimensions in the problem. A deterministic, exhaustive search of this size will quickly become impossible to complete with increased dimensionality and conventional computers. Heuristics and additional knowledge are needed to achieve improvement from this baseline performance.

Global optimization algorithms can be classified in terms of their degree of rigor. The classifications are rigorous, complete, asymptotically complete, and incomplete. Rigorous methods reach a global optimum with certainty but give no limit on their run-time. Complete methods are rigorous and provide an approximate global optimum after finite time. Asymptotically complete methods locate the global optimum with probability one after infinite time but provide no approximation after finite time. Incomplete algorithms are search heuristics that work in practice but provide few guarantees about the quality of their solution. Because of time constraints typically imposed on engineering design studies most algorithms are implemented in a manner that renders them incomplete.

1.7.1 Example: Genetic algorithm

The genetic algorithm (GA) is a popular global optimization method. It is extremely flexible with respect to the types of problem it can be applied to. The original GA was designed for binary optimization problems but variations have been applied to continuous as well as mixed global optimization problems and the form that exists today was popularized by Holland [33]. Continuous parameter values can be represented in binary form and a

discrete GA applied or the binary GA can be adapted to handle continuous variables directly.

GA is a population-based search algorithm which uses a current set of sample points called a generation for the creation of the next generation of sample points. The most common form of GA has a constant population size for all generations. Various operators mimicking the evolution of biological organisms serve to select new sample points in the hope of improving the design performance and locating a global optimum. The GA operators are *selection*, *recombination*, and *mutation*. While selection is required to introduce competition between sample points for limited computational resources, recombination can be considered an approximation of the gradient descent method, and the mutation phase serves the global exploration of the parameter space. The initial selection phase is followed by a recombination phase and a mutation phase to form the next generation at which point the process is repeated on the new generation of sample points. In the global optimization literature GAs are mostly heuristic algorithms and are considered slow to converge. Even though evolution in biological systems can be viewed as a form of optimization, the actual criteria under which evolution operates in nature might not be as clear and focused as it is for narrowly outlined engineering problems. Nevertheless, GAs enjoy enduring popularity by practitioners and benefit from the growth and availability of large-scale computing clusters. GAs are considered asymptotically complete due to their stochastic components, mainly mutation. The three operators selection, recombination, and mutation are briefly described in the following, starting with selection.

Selection describes the manner in which members of the current generation are selected for the creation of the next generation. After the cost function of the current set of trial points is evaluated, the selection phase selects a subset of the current generation for the recombination phase. The selection process ensures that relatively good designs are selected but also that sufficient variation of sample designs is present in the selected subset. Certain trial points that are deemed unfit are excluded from the selected subset while the best performers are assigned higher probability of selection. The methods of selection are based on the cost function value J itself or some function of J, for example rank within the current set of cost function values. It is common to assign a probability of survival to a trial point x_i given by $p_i = \frac{1}{J(x_i)} / \sum_j^M \frac{1}{J(x_j)}$ for positive and finite J. Favoring good performers aggressively is also referred to as *elitism*. Sometimes a super-point with extremely low objective function values dominates the selection probabilities. This limits variation in the selected sample of candidate designs and

causes premature convergence of the optimization algorithm. To prevent a super-point from dominating selection, the rank of cost function values instead of the cost function values themselves can be used, or the number of times a candidate can be selected is limited explicitly. More detail about selection probabilities p_i and alternative selection schemes is given in [34]. Intuitively, high performing members of the population contain design parameter elements of good designs. It also assumes that the mixing of good designs can yield further improved designs, an idea that works for convex objective functions in particular. Once the selection is complete the new generation is generated in the recombination phase.

Recombination and, in particular, various forms of crossover are the main evolutionary component of GA. After the selection process, M parents are recombined and modified to form the next generation of sample points. For binary vectors the most popular recombination methods are single-point and multi-point crossover. The crossover uses two parent trial points and generates two offspring for the next generation so that the population size remains constant. Simple crossover is the generation of a random cut between the index 2 and the length of the binary vector. This cut is used as the crossover point. Both parents are severed at the same crossover point to maintain the length of the binary vectors. After making the cut, the head of one parent is combined with the tail of the other parent and vice versa as shown in the following illustration with parents on the left and offspring on the right:

$$\begin{array}{l}(0\,1\,0\,0\,1\,|\,1\,1\,0\,1)\\(0\,0\,1\,1\,1\,|\,0\,1\,0\,0)\end{array}\rightarrow\begin{array}{l}(0\,1\,0\,0\,1\,|\,0\,1\,0\;\;0)\\(0\,0\,1\,1\,1\,|\,1\,1\,0\,1).\end{array}$$

Adoption of this crossover method for continuous search parameters is not obvious. This is troublesome because the recombination step is crucial to the performance of GA. One simple possibility is to cut a vector in R^k in the same manner and recombine them as above by switching "tails" of the parents

$$\begin{array}{l}(x_1\;x_2\;\ldots\,x_l\mid x_{l+1}\;\ldots\;\;x_k)\\(y_1\;y_2\;\ldots\,y_l\mid y_{l+1}\;\ldots\;\;y_k)\end{array}\rightarrow\begin{array}{l}(x_1\;x_2\;\ldots\,x_l\mid y_{l+1}\;\ldots\;\;y_k)\\(y_1\;y_2\;\ldots\,y_l\mid x_{l+1}\;\ldots\;\;x_k).\end{array}$$

The recombination operators must be tailored to a specific problem. One extension to the simple crossover is the multi-point crossover. For multi-point crossover, multiple and possibly a random number of cuts are generated and the parents are recombined in a similar fashion as shown above. There are other algorithms that combine more than two parents as well as other heuristic recombination methods, too many to be listed here. Recombination, followed by a purely stochastic element called mutation, can be used

to guarantee sufficient variation in the population.

Mutation adds a stochastic aspect to GA that is independent of the current generation and their objective function values. With a small probability the parameter vector is perturbed in some manner. For binary vectors the mutation corresponds to flipping a bit with a small probability. The mutation can be adapted for real-valued vectors but again the choice of frequency and scale of the perturbation must be made by the user. It is difficult to determine a priori what kind of perturbations increase efficiency of the algorithm. We note that mutation adds an entirely random element to GA. It is *mutation* that ensures that eventually the entire search space is surveyed.

Most theoretical work on GA has been done for binary vectors. Analysis is concentrated around the concept of schema or hyper-planes [33, 35]. A schema for binary vectors is of the type 101#0## where the wild card # represents either 0 or 1 and the other values are fixed. In this example there are eight vectors that agree with the schema on the fixed elements and so belong to the same schema. Optimization is therefore a search for a schema with zero wild card symbols. In his original work Holland [33] was able to show that a population with μ individuals is able to effectively process $O(\mu^3)$ schemata. Much of the GA's usefulness is attributed to this property. This result is called the *Schema Theorem* and more details are given in [36]. Similar results are lacking for continuous optimizations.

There are many versions of selection, recombination, and mutation as well as various mixes. With each option usually comes an associated parameter that must be selected by the user. This parameter is usually obscured by the biological analog for GA and hence nonintuitive. Recent versions of GA remedy this drawback by automatically adjusting parameters to avoid overwhelming the user.

1.7.2 Constraints

Constraints usually present a significant challenge in optimization problems. In linear as well as quadratic programming for instance it is usually the case that the optimal solution is on the boundary of the feasible set. For continuous global optimization problems it is usually not possible to determine a priori which constraints will be active at the global optimum. Nevertheless it is crucial to handle active constraints correctly during the local as well as the global search. For example, in the electron tunneling problem discussed in Section 1.3.2 the upper bounds on the barrier values were determined by approximate physical constraints. It is likely that these constraints describing the boundary of Ω become active as the device performance demands are increased.

It is noteworthy that the ratio of the *boundary* to the interior of Ω increases with the dimensionality n of the search space. To see why, consider a hypercube $[0,1]^n$ of dimension n. Numerically, constraint i is active if $\min(x_i - 0, 1 - x_i) < \epsilon$ for a given $\epsilon > 0$. The volume of the hypercube is then $1^n = 1$ while the measure of the numerical interior of a hypercube is given by $(1-2\epsilon)^n$. The ratio of interior of Ω to the entire search space Ω is given by $(1-2\epsilon)^n \to 0$ as $n \to \infty$. The space near the boundary grows relative to the entire search space with increasing n and it becomes increasingly important to carefully treat and explore the boundary. For the local search a gradient projection method such as LBFGS-B [37] does account for active constraints. Therefore LBFGS-B or similar methods are preferred to an unconstrained local optimizer despite the additional computational cost.

1.7.2.1 Interior point method

Interior point methods are a natural way to incorporate inequality constraints into the objective function and transform a constrained optimization problem into an unconstrained one. The idea is to add *barrier functions* to the original objective function as in the Lagrange multiplier method

$$\tilde{J}(x, \lambda) = J(x) + \lambda_i \Theta(x), \tag{1.12}$$

where Θ is the barrier function for a constraint. For example, consider simple box constraints on x, $l_i \leq x \leq u_i$. The monotone barrier functions have the property that $\Theta(x) \to \infty$ as $x_i \to l_i$ or $x_i \to u_i$. The method is called an interior point method because for each $\lambda_i > 0$ the search will remain strictly in the interior of the feasible set with respect to the box constraints. This is important for constraints where it is not possible to evaluate the objective function when the constraints are violated. Examples of cost functions where constraint violation has to be suppressed are parameters where values must remain non-negative. For negative parameter values the physical model and the corresponding mathematical formulation may not be valid. The interior point algorithm consists of a sequence of optimization problems with a decreasing sequence of $\lambda_i \to 0$ as $i \to \infty$. For each optimization instance i the optimization problem is only solved approximately before λ_i is decreased and the iterative current state of the optimization remains near an *interior path*, away from the constraints addressed by the barrier functions. Popular barrier functions are logarithms because they diverge to ∞ slowly and therefore are unlikely to impact the original objective function such that local minima are removed. The barrier function adds little computational effort to the optimization algorithm because it is known explicitly including all of its derivatives. It can be shown that under mild assumptions

the minimizers of the modified interior point problems converge to the minimizers of the original minimization problem as $\lambda_i \to 0$ [34]. One problem is that the constraints are always represented in the objective function via global barrier functions, even when they are not violated. The decreasing sequence $\{\lambda_i\}$ nevertheless has the nice property that it makes the weighting between original objective function and added barrier term negligible in the long run. It is usually best to let the sequence decrease slowly during the implementation.

1.7.2.2 Exterior point method

Exterior point methods allow the constraints of the problem to be violated at intermittent points during the optimization. As for interior point methods, constraints are moved into the objective function formulation resulting in an unconstrained minimization problem. For exterior point methods *penalty functions* are added to the cost function instead of barrier functions

$$\tilde{J}(x, \lambda) = J(x) + \lambda_i \Lambda(x), \tag{1.13}$$

where Λ is now a penalty function. This type of algorithm may be used for inequality as well as equality constraints. In general Λ may be chosen in a manner such that $\tilde{J}(x) = J(x), x \in \Omega$ so that for $x \in \Omega$ the minimization problem remains unchanged if no constraints are violated. Depending on the distance from the feasible set, Λ increases the value of $\tilde{J}$ in order to increasingly penalize the constraint violation. The greater the distance, the bigger the penalty. Note that $\Lambda(d)$ should be monotone increasing on R^+ and identically zero for $d < 0$. Janka suggests penalty functions of the form

- $\Lambda(d) = \frac{1}{p} d^p, p > 1$,
- $\Lambda(d) = \cosh(d) - 1$,

where d is some measure of the distance from the feasible set ([36], p. 19). Penalty terms should be continuously differentiable and convex, $\Lambda(0) = 0$, and $\Lambda(d) \to \infty$ monotonically as $d \to \infty$. The exterior point method is again a sequence of minimization problems where λ_i is now a monotonically increasing sequence $\in \mathrm{R}$. As λ_i increases, constraint violations are increasingly penalized, forcing the solution onto the feasible set Ω. An advantage of exterior point methods is that no initially feasible point must be chosen. A disadvantage is that for many constraints such as in PDE-constrained optimization the final solution might not be feasible. This can happen if J decreases much faster outside of Ω than the chosen penalty function Λ increases. In general, interior and exterior point methods can be mixed in a problem with different types of constraint.

1.7.3 Advanced optimization

The core task of optimal design is the numerical search of configurations resulting in global minima (or maxima) in solution space with respect to a user-defined objective function. Typically, a non-convex solution space might consist of a shallow landscape with many local minima. Local measures of curvature around the minima can give an indication of robustness of a given solution with respect to small variations in control parameters. Applying these basic ideas to synthesize a device design, one might program according to the flowchart shown on the left in Fig. 1.13.

The input to the flowchart in Fig. 1.13 is the objective function and a set of initial parameter values. One might think that it is easy to specify the objective function, however, the fact of the matter is this is usually neither the case nor desirable. For example, one does not want to specify a particular transistor transfer characteristic that is inaccessible, rather one would prefer to be presented with a family of characteristics that are both accessible and insensitive to small variations in design parameters. As indicated on the right-hand side of Fig. 1.13, advanced optimization specifies a set-valued objective instead of a single objective. In addition, one will usually be concerned with the robustness of the solutions. The subject of robust optimization in a non-convex solution space is today a focus of much research activity [38].

Measures of distance can have a dramatic influence on the convexity of the cost function. For example, it makes a difference if the cost function is evaluated using the magnitude of differences or the quadratic deviation. Typically a forward solve of the physical model is relatively computationally expensive while evaluation of the cost function is not. Hence, calculating and comparing the performance of, and choosing between, different cost functions can be an effective strategy to improve efficiency.

Because, at least initially, the user probably does not know the best objective function to request, a feature of advanced optimization should also be its adaptivity. As indicated on the right-hand side of Fig. 1.13, inclusion of on-the-fly modification of the cost function, model, and related decision-making tools has the potential to create an adaptive optimization paradigm in which at the end of the calculation both the problem and the algorithm have changed [39]!

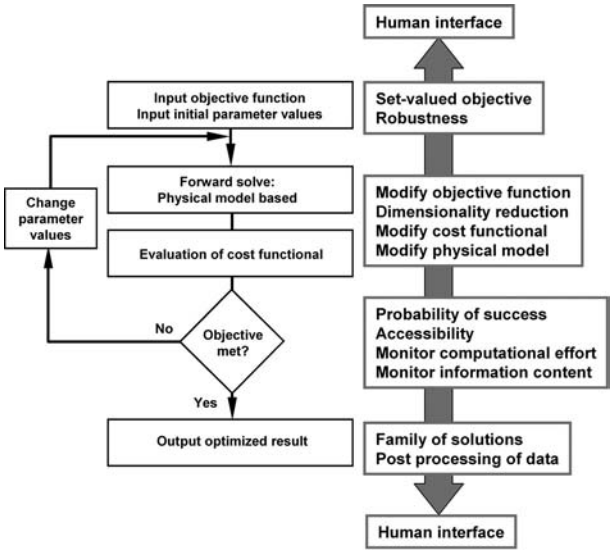

Fig. 1.13. Conventional optimization flowchart (left) and areas where additional adaptivity might assist in the exploration of solution space (right). On-the-fly modification of the cost function, model, and related decision-making tools has the potential to create an adaptive optimization paradigm in which at the end of the calculation both the problem and the algorithm have changed.

1.8 Summary

Control of material composition at the nanometer and atomic scale has potential to dramatically increase electronic and photonic device design space, in part due to quantum effects that provide vast new degrees of design freedom. As a step towards quantum engineering, realistic physical models and synthesis tools based on optimal design should be developed to efficiently explore the nonintuitive parts of this space to learn about and discover the best possible device designs for electronic and photonic systems. Ultimately one can envision a methodology that is capable of basic scientific discovery while simultaneously engineering technological applications. The merging of these two traditionally distinct activities into a unified effort is an opportunity for increased efficiency in twenty-first-century science and technology.

The remaining chapters of this book explore elements of these ideas that represent a new frontier for device engineering.

1.9 References

1. G.E. Moore, *Cramming more components onto integrated circuits*, Electronics **38**, 114–117 (1965). Also reprinted in Proceedings of the IEEE **86**, 82–85 (1998).
2. W. Haensch, E.J. Nowak, R.H. Dennard, *et al.*, *Silicon CMOS devices beyond scaling*, IBM Journal of Research and Development **50**, 339–361 (2006).
3. M. Hossein-Zadeh, H. Rokhsari, A. Hajimiri, and K.J. Vahala, *Characterization of radiation-pressure-driven micromechanical oscillator*, Physical Review A **74**, 023813 1–15 (2006).
4. A.F.J. Levi, S.L. McCall, S.J. Pearton, and R.A. Logan, *Room temperature operation of submicrometre radius disc laser*, Electronics Letters **29**, 1666–1667 (1993).
5. A.F.J. Levi, R.E. Slusher, S.L. McCall, *et al.*, *Room temperature operation of microdisc lasers with submilliamp threshold current*, Electronics Letters **28**, 1010–1012 (1992).
6. N.C. Frateschi, A.P. Kanjamala, A.F.J. Levi, and T. Tanbun-Ek, *Polarization of lasing emission in microdisk laser diodes*, Applied Physics Letters **66**, 1859–1861 (1995).
7. J. Thalken, Y. Chen, A.F.J. Levi, and S. Haas, *Adaptive quantum design of atomic clusters*, Physical Review B **69**, 195410 1–8 (2004).
8. N. Nilius, T.M. Wallis, M. Persson, and W. Ho, *Distance dependence of the interaction between single atoms: gold dimers on NiAl(110)*, Physical Review Letters **90**, 196103 1–4 (2003).
9. J.H. Stathis and D.J. DiMaria, *Reliability projection for ultra-thin oxides at low voltages*, IEDM Technical Digest (Cat. No. 98CH36217) pp. 167–170 (1998).
10. For example, T.M. Wallis, N. Nilius, and W. Ho, *Electronic density oscillations in gold atomic chains assembled atom by atom*, Physical Review Letters **89**, 236802 1–4 (2002).
11. For example, P. Bhattacharya, S. Ghosh, and A.D. Stiff-Roberts, *Quantum dot optoelectronic devices*, Annual Review of Materials Research **34**, 1–40 (2004).
12. For example, W.L. Barns, A. Dereux, and T.W. Ebbesen, *Surface plasmon subwavelength optics*, Nature **424**, 824–830 (2003).
13. For example, S.A. Wolf, A.Y. Chtchelkanova, and D.M. Treger, *Spintronics A retrospective and perspective*, IBM Journal of Research and Development **50**, 101–110 (2006).
14. For example, E. Peter, P. Senellart, D. Martrou, *et al.*, *Exciton photon strong-coupling regime for a single quantum dot in a microcavity*, Physical Review Letters **95**, 067401 1–4 (2005).
15. For example, T. Aoki, B. Dayan, E. Wilcut, *et al.*, *Observation of strong coupling between one atom and a monolithic microresonator*, Nature **443**, 671–674 (2006).
16. For example, D.S. Chemla, *Ultrafast transient nonlinear optical processes in semiconductors*, Semiconductors and Semimetals **58**, pp. 175–256, Academic Press, New York,

New York, 1999.

17. M. Rodwell, M. Le, and B. Brar, *InP bipolar ICs: Scaling roadmaps, frequency limits, manufacturable technologies*, Proceedings of the IEEE **96**, 271–286 (2008).

18. P. Schmidt, S. Haas, and A.F.J. Levi, *Synthesis of electron transmission in nanoscale semiconductor devices*, Applied Physics Letters **88**, 013502 1–3 (2006).

19. R. Landauer, *Spatial variation of currents and fields due to localized scatterers in metallic conduction*, IBM Journal of Research and Development **1**, 223–231 (1957).

20. R. Landauer, *Electrical resistance of disordered one-dimensional lattices*, Philosophical Magazine **21**, 863–867 (1970).

21. M. Buttiker, Y. Imry, R. Landauer, and S. Pinhas, *Generalized many-channel conductance formula with application to small rings*, Physical Review B **31**, 6207–6215 (1985).

22. E.O. Kane, *Basic concepts of tunneling*, in Tunneling Phenomena in Solids, ed. E. Burstein and S. Lundqvist, pp. 1–11, Plenum Press, New York, 1969.

23. G. Bastard, *Superlattice band structure in the envelope-function approximation*, Physical Review B **24**, 5693–5697 (1981).

24. A.F.J. Levi, *Applied Quantum Mechanics*, pp. 171-217, Cambridge University Press, Cambridge, United Kingdom, 2006.

25. I. Grigorenko, S. Haas, and A.F.J. Levi, *Electromagnetic response of broken-symmetry nano-scale clusters*, Physical Review Letters **97**, 036806 1–4 (2006).

26. G. Mie, *Beiträge zur Optik trüber Medien, speziell kolloidaler Metallsungen*, Annalen der Physik **330**, 377–445 (1908).

27. J.F. Müller, A.F.J. Levi, and S. Schmitt-Rink, *Quantum reflections and inelastic scattering of electrons in semiconductor heterojunctions*, Physical Review B **38**, 9843–9849 (1988).

28. A.F.J. Levi, *Nonequilibrium electron transport in heterojunction bipolar transistors*, in InP HBTs: Growth, Processing and Applications, ed. B. Jalali and S.J. Pearton, pp. 89–131, Artech House, Norwood, Massachusetts, 1995.

29. L.D. Landau and E.M. Lifshitz, Quantum Mechanics, Pergamon, Oxford, United Kingdom, 1977.

30. B.Y. Gelfand, S. Schmitt-Rink, and A.F.J. Levi, *Tunneling in the presence of phonons: a solvable model*, Physical Review Letters **62**, 1683–1686 (1989).

31. Rangarajan K. Sundaram, *A First Course in Optimization Theory*, pp. 90-97, Cambridge University Press, Cambridge, United Kingdom, 1996.

32. P. Seliger, M. Mahvash, C. Wang, and A.F.J. Levi, *Optimization of aperiodic dielectric structures*, Journal of Applied Physics **100**, 034310 1–6 (2006).

33. J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, Ann Arbor, Michigan, 1975.

34. E. Janka, *Vergleich stochastischer verfahren zur globalen optimierung*, Diplomarbeit

zur Erlangung des akademischen Grades Magister der Naturwissenschaften, University of Vienna, Vienna, Austria, 1999.

35. L. Altenberg, *The schema theorem and prices theorem*, Foundations of Genetic Algorithms **3**, 23–49 (1995).

36. P.M. Pardalos and H.E. Romeijn, *Handbook of Global Optimization*, volume 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

37. R.H. Byrd, P. Lu, and J. Nocedal, *A limited memory algorithm for bound constrained optimization*, SIAM Journal on Scientific and Statistical Computing **16**, 1190–1208 (1995).

38. For example, A. Ben-Tal, S. Boyd, and A. Nemirovski, *Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems*, Mathematical Programing **107**, 63–89 (2006).

39. A.F.J. Levi, *Towards quantum engineering*, Proceedings of the IEEE **96**, 335–342 (2008).

2 Atoms-up design

Stephan Haas

2.1 Manmade nanostructures

In recent years, it has become possible to detect and control the spatial positions of individual atoms and molecules within nanoclusters, using experimental techniques such as scanning tunneling microscopy. Consider for example the structures shown in Fig. 2.1, made of silver atoms, deposited on a NiAl(110) surface. By bringing a scanning tunneling microscope (STM) sufficiently close to the surface of the substrate, Ag atoms can be moved by following the trajectory of its tip. This capability makes it possible to build one- and two-dimensional atomic and molecular clusters of arbitrary shape on the Ni atom sublattice defined by the NiAl(110) surface.

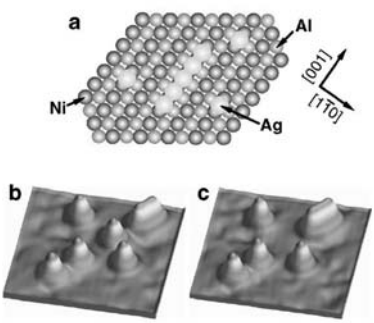

Fig. 2.1. (a) Structure model of a Ag chain (pentamer) on the NiAl(110) surface. Individual Ag atoms are also shown. (b), (c) Three-dimensional representations of STM topography images, taken during the assembly of the Ag chain. Ag atoms and chains appear as round and elongated protrusions correspondingly. The scan size is 8.3 nm × 8.3 nm, $V_{\rm bias} = 3$ V, $I = 1$ nA. (b) An assembled Ag tetramer with five single atoms around it. (c) The atom closest to the tetramer was manipulated along the Ni trough to join it with the tetramer and create a pentamer. Figure from [1].

To complement such emerging capabilities it is clear that a new set of theoretical tools should be developed to assist in the exploration of a potentially vast number of atom configurations and a corresponding enormous range of physical properties. In contrast to classical systems, atomic scale devices exhibit quantum fluctuations and collective quantum phenomena caused by particle interactions. Besides offering an excellent testing ground for models of correlated electrons, they also force us to reconsider conventional paradigms of condensed matter physics, such as crystal symmetries that are imposed by nature. In some instances, such symmetries need to be explicitly broken in order to enable or optimize a desired system response. Consider for example the quasiparticle density of states in tight-binding systems, which is the subject of this chapter. In translationally invariant structures, i.e. crystals, it is well known that the spectral response function exhibits van-Hove singularities at positions of low dispersion, such as the band edges in a one-dimensional chain or the band center in a two-dimensional square lattice. These enhancements of the density of states can be very useful in amplifying system responses such as optical conductivity at specific incident energies. To create new quantum devices, it is therefore important to be able to control the positions and shapes of such features, i.e. by using adaptive design techniques applied to models which capture the essential degrees of freedom of interacting atomic clusters.

Traditional ad hoc methods for the design of nanoscale devices will likely miss many possible configurations. At the same time, it seems unreasonable to expect individuals to manually explore the vast phase space of possibilities for a particular device function. The proposed solution to this difficult design problem is to employ computers to search configuration spaces that enable user-defined target functions. Adaptive quantum design solves an optimal design problem by numerically identifying the best broken-symmetry spatial configuration of atoms and molecules that produces a desired target function response.

The two major ingredients of adaptive quantum design are the physical model, which in the examples given in this chapter evaluates the electronic density of states for a particular spatial arrangement of atoms, and the search algorithm that finds the global minimum in the parameter space of all possible configurations. This problem is typically highly underdetermined and non-convex in the sense that there can be several atomic configurations that yield a system response very close to the desired target function. Often, the associated landscape of solutions is shallow and has many nearly degenerate local minima.

The particular example of a long-range tight-binding model is chosen here because it captures essential features of correlated quantum mechanical systems, and yet permits fast numerical diagonalization of relatively large clusters with broken translational symmetry. In a typical optimization run, these "function calls" occur 100–10,000 times. This model is therefore suitable for developing and testing adaptive design algorithms, and should be viewed as an initial step towards the design of atomic clusters. The goal is to test the applicability and limits of adaptive design techniques on a simple but non-trivial quantum system.

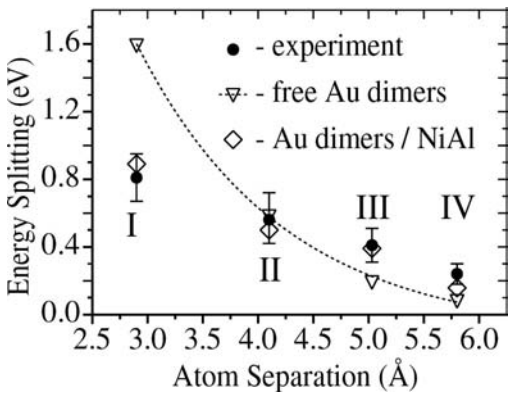

Fig. 2.2. Experimental energy splitting (•) of resonances in Au dimers IIV on NiAl(110) as a function of inter-atomic distance. Error bars reflect the distribution in energy splitting obtained from 200 dimers. Calculated splittings for free (▽) and NiAl-supported Au dimers (◇) are shown for comparison. The dashed line is an exponential fit through the calculated splittings in free dimers. Figure from Ref. [2].

The deposition and manipulation of atoms on metallic substrates using scanning tunneling microscopes [1, 2, 3, 4, 5] is a motivation for this model. The dependence of the tight-binding parameter on the inter-atomic separation is reflected in the electronic densities of states measured in these experiments. More specifically, in a recent set of measurements [1] Nazin *et al.* deposited two gold atoms on a NiAl substrate and monitored the energy split between the bonding and anti-bonding peak in the conductance as a function of how the atoms were positioned with respect to each other. As shown in Fig. 2.2, these measurements clearly indicate a power-law dependence of the effective tight-binding hopping matrix elements on the inter-atomic separation [1, 2]. Thus, within this model description the effects of the substrate are simply absorbed in the matrix elements.

2.2 Long-range tight-binding model

The tight-binding approach is known as an effective tool to describe the band structure of electronic systems. In bulk solids, it is commonly used to model the relevant bands close to the Fermi level, obtained from complex density functional theory calculations. Since in this chapter symmetry-breaking, non-periodic configurations are considered, a long-range variant of the tight-binding model has to be used, in which the overlap integrals depend explicitly on the variable inter-atomic distance. The Hamiltonian then takes the form

$$\hat{H} = -\sum_{i,j} t_{i,j} \left(\hat{c}_i^\dagger \hat{c}_j + \hat{c}_i \hat{c}_j^\dagger \right), \tag{2.1}$$

where $c_i^\dagger$ and c_i are electron creation and annihilation operators at a site $\mathbf{r}_i$, and the spatial decay of the overlap integral $t_{i,j}$ is given by a power-law,

$$t_{i,j} = \frac{t}{|\mathbf{r}_i - \mathbf{r}_j|^\alpha}. \tag{2.2}$$

In the following, the exponent α is taken to be 3.0 unless mentioned otherwise [6]. This parameterization reflects an algebraic variation of the overlap integral with inter-atomic separation, consistent with recent experiments [1, 2]. The choice of sign in the Hamiltonian follows the convention for s-orbitals. However, this simple implementation of the tight-binding model does not account for the orbital directionality of realistic Au, Ag, or Pd atomic wave functions. More sophisticated and numerically expensive techniques, such as the local density approximation, would be needed to make quantitative predictions for these systems.

The Hamiltonian matrix of the long-range tight-binding model in the basis of single-particle states is non-sparse, and only its diagonal matrix elements vanish. In order to obtain the spectrum, the matrix is diagonalized numerically for finite clusters. In Fig. 2.3, the resulting densities of states of translationally invariant chains and square lattices are shown for the nearest-neighbor tight-binding model with $t_{i,j} = t\delta_{i,j}$ in Fig. 2.3(a) and (c), and for the case of long-range overlap integrals in Fig. 2.3(b) and (d). Characteristic van-Hove singularities are observed, in one dimension at the band edges, and in two dimensions at the band center. For the long-range model, the particle-hole symmetry is broken because of frustration introduced by the competing overlaps, leading to asymmetries in the density of states. While the system sizes in Fig. 2.3 are chosen to be rather large in order to make contact with the familiar thermodynamic limit, there are still some visible

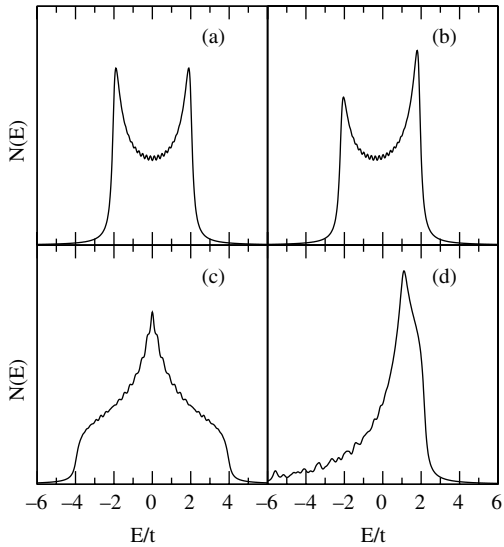

Fig. 2.3. Density of states in spatially invariant tight-binding systems with periodic boundary conditions. (a) Nearest-neighbor chain (no long-range overlap integral) with 30 atoms. (b) Same as (a), but with long-range overlaps according to Eq. (2.2). (c) Nearest-neighbor square lattice with 400 atoms. (d) Same as (c), but with long-range overlaps.

finite-size remnants, i.e. a faint pole structure due to the discreteness of the system. These features become much more pronounced for the few-atom clusters that are studied in the next sections.

2.3 Target functions and convergence criterion

While "nature" gives us densities of states that are constrained by the dimensionality and symmetry of the underlying lattice, our objective here is to engineer specific spectral responses that are useful in designing nanoscale devices. For example, we may wish to produce a quasi-two-dimensional spectrum within a one-dimensional system or to concentrate spectral weight in particular energy windows. These goals are achieved by placing the atomic constituents into optimized symmetry-breaking configurations which are determined by numerical searches.

The specific target functions we wish to consider here are shown in Fig. 2.4. They are (a) a flat top hat density of states centered at $E = 0$,

$$N(E)_{\text{target}} = \theta(E - E_c)\theta(E + E_c), \tag{2.3}$$

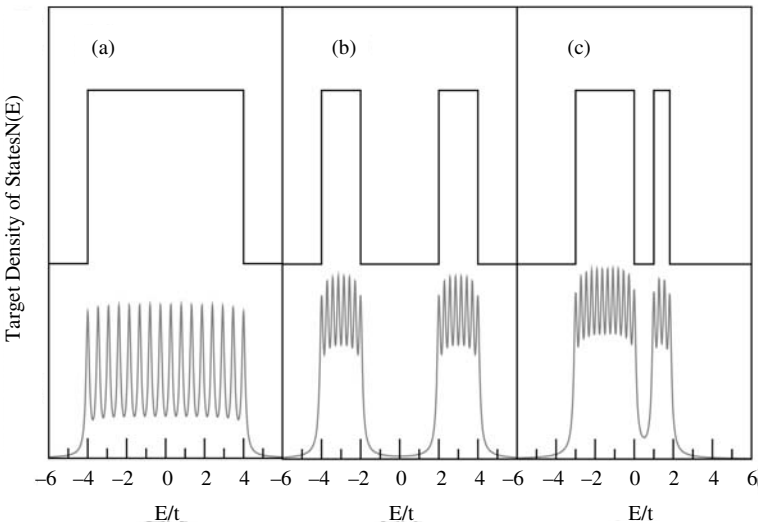

Fig. 2.4. Target densities of states used in this work: (a) particle–hole symmetric top hat function, centered at E = 0, (b) particle-hole symmetric two-peak function, and (c) asymmetric two-peak function. For systems with finite numbers of particles these shapes are approximated by quasiparticle peaks.

where $\theta(x)$ is the Heavyside function, and E_{c} is an energy cut-off, (b) a symmetric two-peak function, centered at $E = 0$,

$$N(E)_{\mathrm{target}} = \theta(E - E_{\mathrm{c}2})\theta(E + E_{\mathrm{c}2}) + \theta(E - E_{\mathrm{c}1})\theta(E + E_{\mathrm{c}2}), \tag{2.4}$$

and (c) a particle–hole-symmetry-breaking function with two unequal peaks, i.e. more spectral weight on the quasi-hole side ($E < 0$) than on the quasi-electron side ($E > 0$) of the spectrum,

$$N(E)_{\mathrm{target}} = \theta(E - E_{\mathrm{c}2})\theta(E + E_{\mathrm{c}1}) + \theta(E - E_{\mathrm{c}4})\theta(E + E_{\mathrm{c}3}). \tag{2.5}$$

In systems with finite numbers of tight-binding atoms, these continuous shapes are approximated by equally spaced poles within the energy windows where $N(E)_{\mathrm{target}}$ is a non-vanishing constant. Here, the delta-functions are given a finite width of $0.02 \times t$. As more atoms are added to the system, these peaks merge together, approaching the bulk result. For other non-flat target functions the quasiparticle peak spacing can be varied, e.g. following a gaussian or Lorentzian shape. Naturally, not all targets can be achieved equally well. Factors that influence the achievable distance to a target include the number of available atoms and, as will be shown, a continuous or discrete number of accessible spatial positions. The dimensionality of

the system poses additional constraints. In particular, there are fewer available configurations in lower dimensions. In the following, we focus on three prototype spectral responses which are targeted by numerically optimizing configurations of clusters with up to 48 atoms in two spatial dimensions.

The optimization algorithms seek to minimize the deviation from a given target density of states, defined by the error function

$$\Delta = \int_{-\infty}^{\infty} dE[N(E) - N(E)_{\text{target}}]^2, \tag{2.6}$$

which is the least-square difference between the system response for a given configuration and the target response function. We have explored a number of numerical techniques, including the Newton–Raphson steepest descent method, simple downhill random walk, simulated and triggered annealing, and genetic algorithms. The advantages and disadvantages of these techniques are briefly discussed in the last section of this chapter. In general, it is found that hybrids of these methods tend to work best. In the next section, we focus on adaptive design of atomic clusters in continuous configuration space without any restrictions to underlying discrete lattices. In this case, the search space is infinite which generally allows better convergence to a given target response than in finite configuration space. However, some experimental realizations of such structures require deposition of atoms on substrates with discrete lattice structures. Therefore, this case is addressed separately in the subsequent section.

2.4 Atoms-up design of tight-binding clusters in continuous configuration space

The first target density of states we would like to study in detail is the particle–hole symmetric top hat function, centered at energy $E = 0$, shown in Fig. 2.4(a). This function represents a constant density of states for a bulk solid between the energy cut-offs $\pm E_c$, giving a bandwidth $W = 2E_c$. Here, we choose $E_c = 3 \times t$. However, it should be noted that with the adaptive quantum design approach we are not restricted to this choice, and target densities of states with quite different bandwidths can be matched, although often to a lesser degree of accuracy.

In Fig. 2.5, the solution for a system with 16 atoms confined to a box with periodic boundary conditions is shown. The guided random walk method is applied to optimize the configuration of atoms by iterative local updates of their positions in order to match the top hat density of states. As observed in Fig. 2.5(a) and (c), good convergence to the target function can be

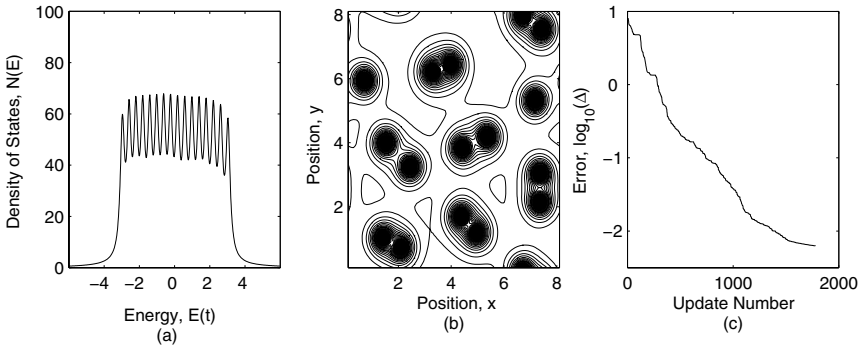

Fig. 2.5. Adaptive quantum design applied to 16 atoms in a two-dimensional box with periodic boundary conditions. The target density of states is a symmetric top hat function with bandwidth $6t$. The atomic configurations are optimized by applying a guided random walk algorithm to the long-range tight-binding model. (a) The best matching solution. (b) Contour plot of the potential of the resulting spatial configuration. (c) Convergence to the target function, $\log_{10}(\Delta)$, with the number of updates.

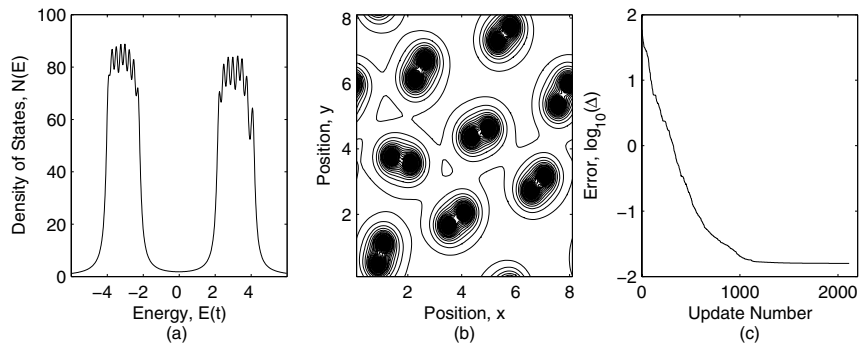

Fig. 2.6. Same as Fig. 2.5, but for a symmetric two-peak target function with peaks of bandwidth $2t$ centered at $-3 \times t$ and $3 \times t$.

achieved for this case after fewer than 2,000 updates. A contour plot of the potential $\sum_{ij} -t_{ij}/|\mathbf{r}_i - \mathbf{r}_j|^{\alpha}$ for the resulting spatial configuration is shown in Fig. 2.5(b). Here, equipotential lines are used to denote the positions and overlapping wave functions of the atomic constituents in the system. For this target function one discovers the formation of dimers with a wide range of inter- and intra-dimer spacings. These self-assembled building blocks have variable directional orientation, and are closely packed.

To explore the generic aspects of this result, let us now turn to the two-peak target function shown in Fig. 2.4(b). This density of states has a gap

in the center of the spectrum, as may be desired for the construction of two-state systems such as q-bits. As shown in Fig. (2.6) (a) and (c), the convergence to this target function is not quite as good as for the top hat function, but saturation occurs already at half the number of iterations compared to the previous example. Interestingly, the optimized spatial configuration that is found in Fig. 2.6(b) also displays a preference for dimer formation.

In the long-range tight-binding system we are studying it is, at least at first sight, not obvious why these target functions prefer dimer building blocks. Let us address this question by examining the individual spectra of the molecular building blocks which were already discussed in Chapter 1 and shown in Fig. 1.3. Each dimer molecule contributes to the density of states a positive and negative pole with energies $E = \pm t_{12}$, where t_{12} is the hopping integral between the two participating atoms. The zero-energy quasiparticle peaks of two isolated atoms are split into bonding and anti-bonding combinations once they form dimer molecules. Therefore, isolated dimers are ideal building blocks for particle–hole symmetric densities of states, such as the top hat and the two-peak target function. The intra-dimer spacing determines the positions of the $E = \pm t_{12}$ poles via Eq. (2.2). With an appropriate distribution of these distances the full target spectrum of the top hat function can be covered. The poles close to the band center ($E = 0$) are provided by the less tightly bound dimers. For the two-peak target function the dimer building blocks required to realize the target spectrum are more tightly bound, and the intra-dimer spacings need to vary less to achieve this target.

The idea that dimers can be used as building blocks for the particle–hole symmetric target functions only applies to isolated dimers, i.e. the dilute limit, or when potential gradients across dimer pairs from the presence of adjacent atoms do not break particle–hole symmetry. The absence of such gradients, even for relatively high atom densities in a long-range interacting system, accounts for the success of dimers in satisfying the target function.

Lower symmetry building blocks, such as trimers and quadrumers, can achieve more complex target functions, in particular those with broken particle–hole symmetry. Due to frustration, trimer molecules intrinsically have asymmetric densities of states with unequal spectral weights on the electron and the hole side of their spectra. Quadrumers have a symmetric spectral response in the absence of longer-range frustrating interactions. As an example of how these building blocks enable more complex target functions, let us consider the asymmetric two-peak density of states given by Eq. (2.5) and shown in Fig. 2.4(c), which has a narrow upper peak and a wider lower peak, separated by a gap. In Fig. 2.7 it is observed that an

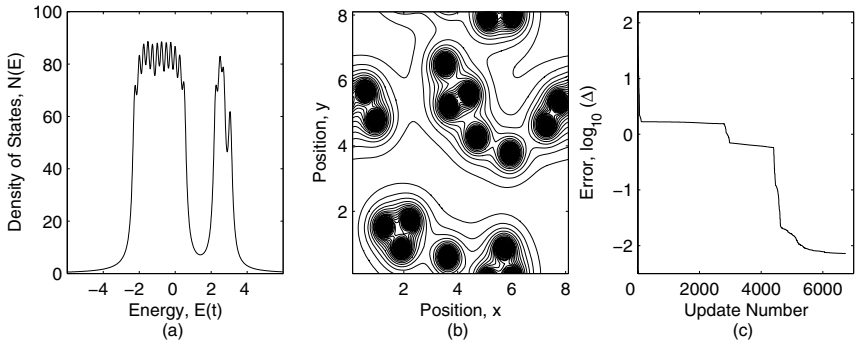

Fig. 2.7. Asymmetric two-peak quasiparticle density of states. The solution contains dimer, trimer, and quadrumer building blocks.

approximate match can be achieved by the adaptive method. As expected, the building blocks for this particle–hole-asymmetric target function are combinations of dimers, trimers, and quadrumers, which partially recombine into larger clusters. Obviously, in order to achieve asymmetric target functions more complex building blocks are required. Especially for systems which contain only a small number of atoms it is important whether the required building blocks are available, and whether there are non-participating unbound atoms that may deteriorate convergence and match to the target function.

Let us finish this section by addressing the convergence properties of the three cases that were discussed. So far, only the simplest guided random walk optimization method was considered in which every "downhill step" is accepted. These "downhill steps" are local updates of individual atomic position that lead to a better match of the spectrum to the target density of states. Especially for the most symmetric target function, this algorithm converges efficiently, with a small remaining error. For the symmetric two-peak function, convergence occurs even faster because the required dimer building blocks are more uniform than for the first case. However, the remaining error is slightly larger, indicating that this may be a metastable solution which could be improved by global updates in which whole subclusters are simultaneously updated. Finally, the convergence plot for the asymmetric two-peak target function (Fig. 2.7(c)) shows several plateaus, indicating metastable configurations that exhibit a high resistance against local updates. For this case, a much larger number of local updates is required to achieve an acceptable match. Hence, more complex target functions clearly call for more sophisticated numerical search tools, including annealing steps, parallelization, and global updating schemes when available.

2.5 Optimal design in discrete configuration space

In the previous section local updates were considered which allow atomic positions to change continuously within a given radius. However, for the case of atoms deposited on a substrate with a given lattice structure, the set of available positions is usually discrete, although it may be very large. This has significant consequences for the adaptive design approach. The search space of solutions is finite in this case, which makes it feasible to study more atoms for similar computational effort relative to the continuous case. At the same time, the discreteness of the lattice can prohibit favorable configurations that are available in the continuous case, thus deteriorating convergence properties of the optimization procedure. In practice, we find that the feasibility of computations for larger numbers of atoms in the discrete case helps to achieve better matches, as long as the lattice spacings remain sufficiently small. In this section, we explore the effects of an underlying grid on the adaptive design procedure. Also, more advanced techniques are implemented for the numerical optimization, including hybrids of the genetic algorithm, simulated annealing, and the guided random walk method.

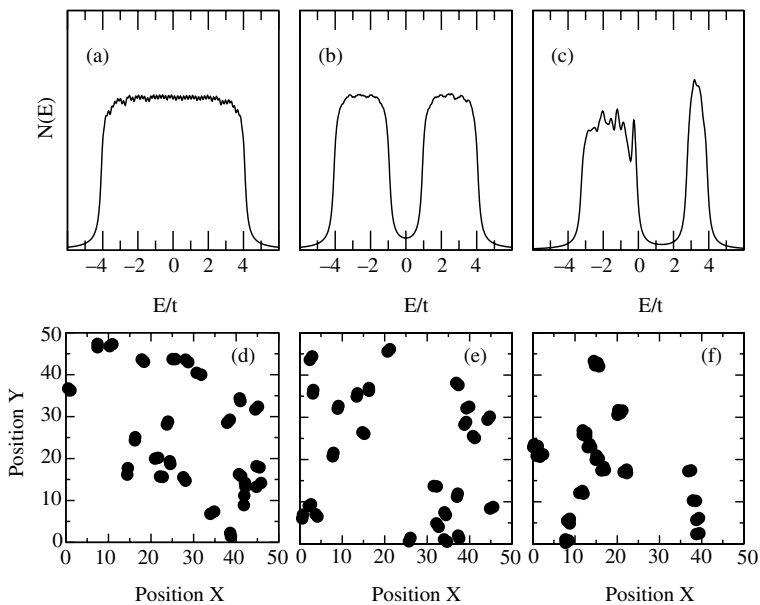

Fig. 2.8. Adaptive design of clusters with 48 atoms on a discrete lattice. The (a) flat, (b) symmetric two-peak, and (c) asymmetric two-peak densities of states are the same as discussed in the previous section. The corresponding atomic configurations are shown in (d), (e), and (f).

In Fig. 2.8, optimal design results are shown for systems with 48 atoms on

a square lattice with spacing 0.01. The length scale is set by the linear size of the two-dimensional box (48×48) to which the particles are confined. These systems are in the dilute limit, and hence the convergence to the target functions, chosen to be the same as in the previous section, is good. Also, because of the larger number of particles, there are fewer finite size effects. For the top hat target function one obtains a small matching error of $\Delta = 0.000643$, for the symmetric two-peak function one finds $\Delta = 0.000605$, and for the asymmetric two-peak function the error is $\Delta = 0.150347$. Thus, analogous to the continuous case, the more symmetric targets are easier to achieve. Again, clustering into dimers is observed for the particle–hole symmetric target functions, whereas trimers are the preferred building blocks for the asymmetric target. Since atomic densities in these examples were chosen to be in the dilute limit, boundary effects and interactions between the building blocks are small. The resulting configurations have the character of liquids, governed by weak interactions between the molecular building blocks, and relatively strong confining forces that lead to the formation of dimers and trimers.

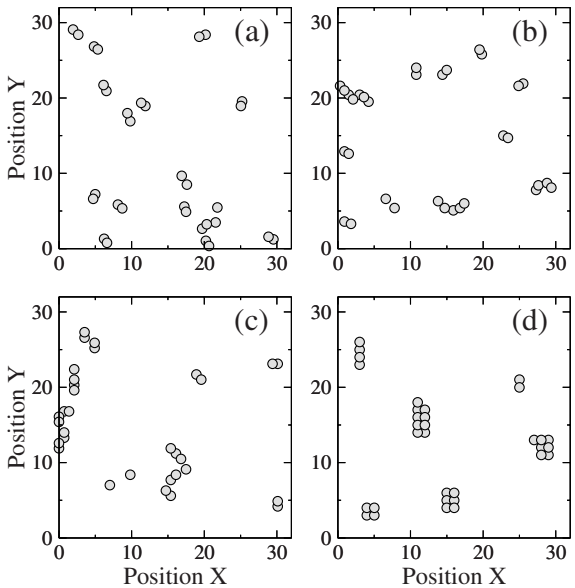

Fig. 2.9. Effect of the coarseness of the underlying lattice on the formation of building blocks. 32 atoms are confined to a 32×32 square box. The target function is the top hat density of states. The grid spacing is varied: (a) 0.01, (b) 0.3, (c) 0.7, and (d) 1.0.

Next, let us explore the dependence of these solutions on the coarseness of the underlying lattice. In Fig. 2.9, the grid spacing is varied over two orders of magnitude from 0.01 up to 1.0. As expected, the convergence to the target top hat function deteriorates dramatically as the substrate is made

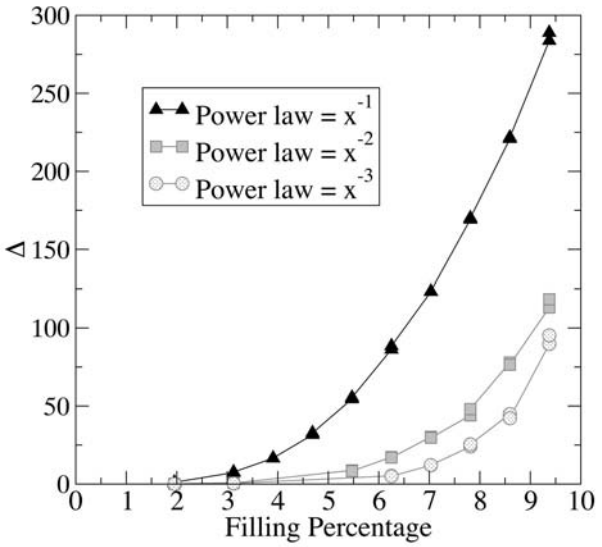

Fig. 2.10. Effect of the density of atoms and the power-law dependence of the tight-binding overlap integral on the convergence of adaptive quantum design.

coarser. For the smallest spacing of 0.01 one converges to a final error of $\Delta = 0.28457$, for a spacing of 0.3 the error is $\Delta = 2.45047$, for a spacing of 0.7 the error becomes $\Delta = 3.81779$, and for the coarsest case with spacing 1.0 the error is $\Delta = 36.3721$, indicating failure of convergence. The corresponding configurations in Fig. 2.9 show a strong dependence of the clustering sizes as the system is trying to cope with less available phase space to match the target function. For the finest grid spacing (Fig. 2.9(a)) one observes almost entirely isolated dimers. As the spacing is increased (Fig. 2.9(b)), a few small strings and groups are formed. At grid spacing 0.7 (Fig. 2.9(c)), the solution is made up mostly of long strings and dimers in close proximity to each other. Ultimately, for the coarsest grid spacing of 1.0 (Fig. 2.9(d)), the final configuration consists of square and rectangular blocks of atoms. This result demonstrates that there is a hierarchy of building blocks. The most suitable building blocks for the given target function are dimers. When these become less available due to lattice constraints, the adaptive method selects higher order solutions, i.e. larger size clusters, in order to cope with the more restricted phase space of possible configurations. Hence, for each level of coarseness, adaptive design discovers solutions that enable – up to a given degree of accuracy – a targeted system response.

By widening the grid spacings and keeping the box size constant, the density of atoms in the system is effectively increased, and simultaneously the available energy levels are spaced further apart. Let us examine these dependencies by varying the power law governing the atomic overlaps (Eq. (2.2)),

and by increasing the number of atoms, i.e. the filling fraction, on a fixed lattice. Results for the achieved convergence are shown in Fig. 2.10. These demonstrate that excellent target matches can be achieved in the dilute limit with filling densities of a few percent. For larger coverages, the numerical search becomes exponentially less effective, indicating increasing frustration effects that need to be addressed by global updating schemes. Higher power laws imply effective shorter-range atomic overlaps, thus rendering the system more dilute. This is reflected in Fig. 2.10, where the departure from the regime of negligible matching errors is pushed toward higher filling percentages as α in Eq. (2.2) is increased.

Some aspects of the long-range tight-binding model on a substrate have already been confirmed experimentally using scanning tunneling microscopy (STM) to precisely position gold atoms on the surface of a nickel-aluminum crystal. In the studies mentioned above, which were performed at the University of California at Irvine by Wilson Ho's group [3, 4], STM measurements show that the splitting in the value of eigenenergies for Au dimers on NiAl depends inversely on Au atom separation corresponding to $\alpha = 1$. Here, the grid spacings are dictated by the 0.29 nm lattice periodicity of available add-atom sites on the NiAl substrate. Surprisingly, the power-law dependence of the effective overlaps t_{ij} between the deposited atoms takes into account interactions with the substrate which are typically difficult to model by first principle computations.

2.6 Optimization and search algorithms

The optimization algorithms used here were based on local updates of atomic positions in order to minimize the error Δ defined in Eq. (2.6). Each atom in the system is visited periodically, and a trial change of its position is attempted. Depending on the response in Δ and on the specific algorithm, this trial step is either accepted or rejected. For the case of continuous configuration space, these local updates are random shifts of positions within a given radius. A hard core constraint is implemented which forbids atoms to be placed on top of each other. For discrete configuration space, a stochastic distribution function is used to decide which sites in the neighborhood of the original position of an atom should be visited in a trial step. While this function is naturally peaked at nearest-neighbor sites, it has to include a finite probability of longer range updates in order to avoid getting stuck in local minima of search space. The particular results discussed in this appendix are obtained for the continuous case.

The Newton–Raphson and Broydn methods are multi-dimensional

generalizations of the one-dimensional secant method [7]. Unfortunately, their global convergence is rather poor for more than 20 variable parameters. Therefore, they are only applicable to the smallest system sizes, and are not useful for the nonlinear multi-parameter searches required for adaptive quantum design.

In the guided random walk or "downhill method", each random step that results in a smaller target error Δ is accepted [8]. Random steps are trial spatial variations about a particle's position. This guided random walk technique is a quickly implemented power horse. However, especially for shallow landscapes of solutions it gets easily stuck in local minima.

Simulated annealing uses an effective temperature representing the likelihood of accepting a step that does not minimize the function. This temperature is lowered slowly – at a rate of 10% – with each iteration. The initial temperature is taken as $T_{\text{init}} = 5t$. This method is better at avoiding local minima than the previous techniques. However, it takes a relatively long time to converge.

The triggered annealing method is a hybrid of the downhill and the simulated annealing method. It implements the downhill method until minimizing steps become hard to find, at which point the simulated annealing method is used to escape from local minima. Parameters are chosen the same as for the simulated annealing method. Triggered annealing tends to converge relatively quickly if there are only a few local minima.

The particle replacement method uses the simple downhill method for guided random updates. In addition, it identifies particles that have not been updated for an extended period, because of being stuck in a local minimum, and assigns them to a new random position within the lattice boundaries. In our implementation, a particle is replaced if there are ten idle iterations without a successful downhill update for that particular particle. Note that this particle replacement update is different from random step updates because it is independent of the previous position of the particle.

In genetic algorithms a population of possible solutions is created. Those that best minimize the function are allowed to take part in creating a new generation of possible solutions [9]. These methods are generally good at avoiding local minima, and are also easily implemented on parallel computers. They typically require more function calls than other search algorithms.

In order to illustrate the efficiency of these various approaches, each method is used to match a flat top hat target function on a one-dimensional lattice with 24 tight-binding atoms and a box size of 96×96. Particles exiting the box on one end enter it from the other side via periodic boundary conditions. The target function (discussed more extensively in the following

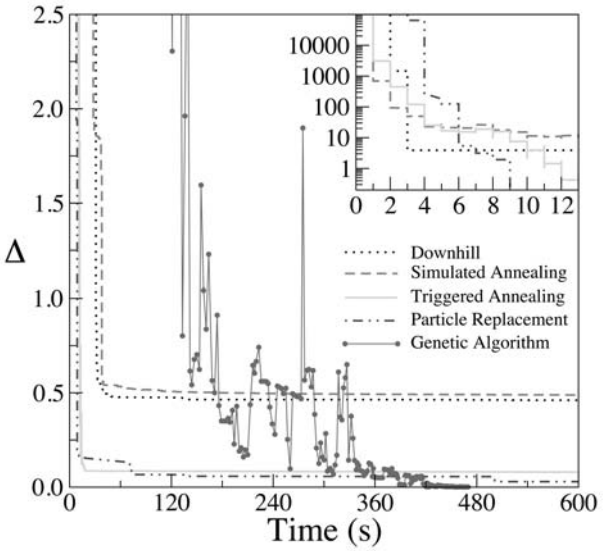

Fig. 2.11. Comparison of the convergence of several optimization algorithms. The error Δ is shown as a function of computer run time. The inset uses a logarithmic scale for Δ.

section) is chosen to be symmetric about $E = 0$, with poles evenly spaced between $E = -4t$ and $4t$. All of these benchmark runs are started with atoms placed randomly along the chain. The computer is a Pentium III 1 GHz with 2 GB pc133 memory, and the additional seven nodes used by the genetic algorithm method are Pentium III 850 MHz processors with 1 GB memory. Each method is run for 600 seconds.

As shown in Fig. 2.11, all methods, with the exception of the genetic algorithm, converge rapidly within the first minute of run time, and show only relatively small corrections afterwards. The two techniques with the fastest convergence (inset of Fig. 2.11) are the downhill and the simulated annealing methods. However, their asymptotic error functions remain relatively large, indicating that they are easily trapped in local minima. In contrast, the triggered annealing and particle replacement methods yield much better matches to the target function, while still converging relatively fast. As shown in the inset, the annealing methods sometimes accept trial steps in the "wrong" direction in order to avoid local minima. Finally, the genetic algorithm takes a relatively long time to converge. However, it yields by far the best match to the target density of states after about 7 minutes run time. In order to ensure best matches to the target, this last method is therefore used whenever the computational effort allows it.

For more information on global optimization methods see Appendix A.

2.7 Summary

In this chapter, it was shown how adaptive quantum design techniques can be applied to tailor the quasiparticle density of states of atomic clusters, modeled by the long-range tight-binding Hamiltonian. Broken symmetry spatial configurations of atoms were optimized to match target spectra. By applying adaptive search algorithms, it was shown that matches to target responses can be achieved by forming hierarchies of molecular building blocks that depend on system constraints. For example, symmetric top hat and two-peak target densities of states can be achieved by forming lattices of weakly interacting dimers. While these are the elementary building blocks for particle–hole symmetric case target functions, more complex molecules, such as trimers and quadrumers, are found to dominate the solutions for asymmetric target functions.

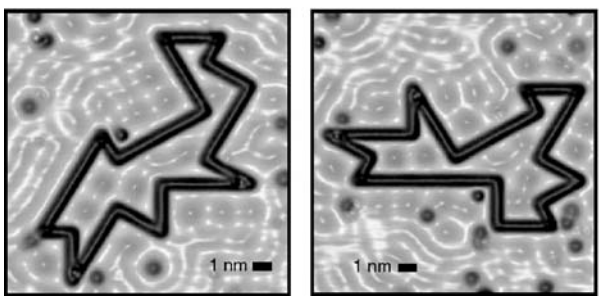

Fig. 2.12. Broken-symmetry isospectral quantum corral structures, used to detect the phases of electronic wave functions. Figure from [10].

From an experimental point of view, it is obvious that broken-symmetry nanostructures can enable quantum responses and functionalities far beyond our imagination. For example, in a series of recent scanning tunneling microscope experiments aperiodic corral structures of carbon monoxide molecules were positioned on a Cu(111) surface [10]. The wave functions of trapped electrons in these structures are reminiscent of standing waves on the surface of a drum (see Fig. 2.12). Making use of the isospectral properties of these corrals, it is possible to extract phase information about the wave functions.

The central task of optimal quantum design is the numerical search for global minima in typically shallow landscapes of configurations with many local minima. Since this procedure typically requires many function calls, an efficient implementation on parallel computers is necessary. In this chapter the complexity of the physical model was minimized in order to limit the computational expense. While the long-range tight-binding model can be

viewed as a semi-realistic testing ground for adaptive quantum design techniques, it is crucial to apply these algorithms to more sophisticated models that include, among other ingredients, orbital directionality, spin degrees of freedom, and electronic correlations. Furthermore, as we will see in later chapters, adaptive design is applicable to related areas in nanotechnology, including in Chapter 3 the design of nano-electronic components [11] and in Chapter 4 nano-photonic components [12–14], and RF systems [15].

2.8 References

1. G.V. Nazin, X.H. Oiu, and W. Ho, *Visualization and spectroscopy of a metal-molecule-metal bridge*, Science **302**, 77–81 (2003).
2. G.V. Nazin, X.H. Oiu, and W. Ho, *Atomic engineering of photon emission with a scanning tunneling microscope*, Physical Review Letters **90**, 216110 1–4 (2003).
3. N. Nilius, T.M. Wallis, M. Persson, and W. Ho, *Distance dependence of the interaction between single atoms: Gold dimers on NiAl(110)*, Physical Review Letters **90**, 196103 1–4 (2003).
4. N. Nilius, T.M. Wallis, and W. Ho, *Development of one-dimensional band structure in artificial gold chains*, Science **297**, 1853–1856 (2002).
5. T.M. Wallis, N. Nilius, and W. Ho, *Electronic density oscillations in gold atomic chains assembled atom by atom*, Physical Review Letters **89**, 236802 1–4 (2002).
6. W.A. Harrison, *Electronic Structure and the Properties of Solids*, W.H. Freeman and Company, San Francisco, California, 1980.
7. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flammery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, United Kingdom, 1986.
8. Y. Chen, R. Yu, W. Li, *et al.*, *Adaptive design of nano-scale dielectric structures for photonics*, Journal of Applied Physics **94**, 6065–6069 (2003).
9. For example, see D. Whitley, *A genetic algorithm tutorial*, Statistics and Computing **4**, 65–85 (1994).
10. C.R. Moon, L.S. Mattos, B.K. Foster, *et al.*, *Quantum phase extraction in isospectral electronic nanostructures*, Science **319**, 782–787 (2008).
11. P. Schmidt, S. Haas, and A.F.J. Levi, *Synthesis of electron transmission in nanoscale semiconductor devices*, Applied Physics Letters **88**, 013502 1–3 (2006).
12. I.L. Gheorma, S. Haas, and A.F.J. Levi, *Aperiodic nanophotonic design*, Journal of Applied Physics **95**, 1420–1426 (2004).
13. L. Sanchis, A. Håkansson, D. López-Zanón, J. Bravo-Abad, and J. Sánchez-Dehesa, *Integrated optical devices design by genetic algorithm*, Applied Physics Letters **84**, 4460–4462 (2004).

14. J. Volk, A. Håkansson, H. T. Miyazaki, *et al.*, *Fully engineered homoepitaxial zinc oxide nanopillar array for near-surface light wave manipulation*, Applied Physics Letters **92**, 183114 1–3 (2008).

15. P. Seliger, M. Mahvash, C. Wang, and A.F.J. Levi, *Optimization of aperiodic dielectric structures*, Journal of Applied Physics **100**, 034310 1–6 (2006).

3 Electron devices and electron transport

K. Magruder, P. Seliger, and A.F.J. Levi

3.1 Introduction

The ability to control the geometry and composition of metal, dielectric and semiconductor materials at the nanoscale has created a significant opportunity for new designs of electronic devices. It is this frontier of electronic device design that we would like to explore in this chapter. An important conclusion we will draw is that technologically significant applications are most likely to be confined to those material systems and fabrication techniques that exhibit the greatest control at the atomic level.

Figure 3.1 categorizes some of the basic geometries commonly considered. Nanodots or quantum dots can be designed to confine the electron wave function in all three dimensions. In this case electron motion on or off the dot is determined either by tunneling, by thermionic emission, or, particularly in the case of a direct band gap semiconductor, by optical transitions. Because of the three-dimensional confinement, it is common to think about quantum dots as *artificial atoms*. However, the electronic properties of quantum dots are very sensitive to their exact dimensions. In the simplest

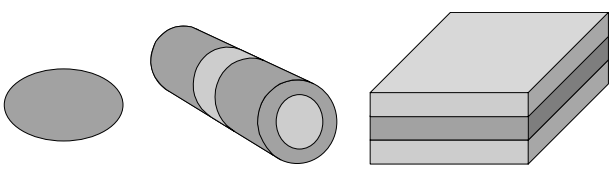

Fig. 3.1. Illustrating the geometry of a semiconductor heterostructure quantum dot, quantum wire, and quantum well. Notice that control of radial and longitudinal material composition allows a quantum dot structure to be embedded in a quantum wire heterostructure.

of models the eigenenergy of a semiconductor conduction band electron in a quantum dot of size L scales nonlinearly as $1/L^2$. Typically, small differences in dot size can give rise to significant variation in, for example, optical absorption. Multiple dots tend to have behavior that reflects the existence of inhomogeneities in size and shape. This lack of control in fabrication makes the analogy to atoms somewhat misleading. It also limits some, but by no means all, potential applications until a time when more precise control over geometry can be achieved.

Nanowires confine electrons in two dimensions. Heterostructures in the nanowire can be used to form quantum dot and shell structures along the length of the wire. While progress in synthesis of semiconductor nanowire shell structures has been rapid in recent years [1–3], charge transport properties, such as electron mobility, remain less than bulk materials. This is most likely due to the presence of impurities and defects incorporated into the structures during growth. Surfaces are particularly prone to defects which, in turn, can dramatically alter electron transport properties. Embedding the semiconductor wire to form a heterostructure shell can mitigate the issue, so long as defects do not exist at the heterointerface and the electron wave function contributing to transport does not penetrate through to the exposed surface of the shell.

In contrast to the challenges facing synthesis of quantum dots and quantum wires, layered semiconductor heterostructures or quantum wells can be designed to confine electrons in one dimension to form a two-dimensional electron gas. Extremely high purity single crystal $Al_\xi Ga_{1-\xi} As/GaAs$

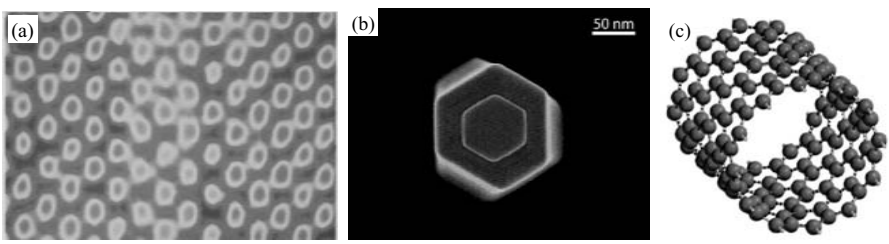

Fig. 3.2. (a) Monolayer control of semiconductor crystal growth illustrated in a transmission electron micrograph (TEM) of an $In_{0.53}Ga_{0.47}As$ quantum well in cross-section that is three monolayers thick and is sandwiched between InP barrier layers. The spots in the image represent tunnels between pairs of atoms. The minimum separation between tunnels in InP is 0.34 nm. (b) High-resolution scanning electron microscope image of the cross-section of a heterostructure nanowire [3] consisting of an InP core and a 1.5 nm thickness InAs tube surrounded by an InP shell. The complete structure is about 125 nm in diameter. Preferential crystal growth on (111) and (110) planes results in a polygonal cross-section. (c) Ball-and-stick model of a single-walled carbon nanotube about 1.2 nm in diameter.

heterostructures have been grown that exhibit remarkably high electron mobility. In what is undoubtedly a triumph of material science, the measured low-temperature electron mobility can be 1,000 times greater than that in the corresponding bulk material [4, 5].

Of the geometries illustrated in Fig. 3.1, the synthesis and fabrication of one-dimensional heterojunction nanowires [1–3], as well as more established techniques such as molecular beam epitaxy (MBE) that allows semiconductor quantum well materials to be controlled with atomic-layer precision in the crystal growth direction [4], have potential to impact future applications.

To illustrate atomic-layer precision in crystal growth, Fig. 3.2(a) shows an $In_{0.53}Ga_{0.47}As$ quantum well in cross-section that is three monolayers thick and is sandwiched between InP barrier-layers. Figure 3.2(b) shows a high-resolution scanning electron microscope image of the cross-section of a heterostructure nanowire [3] about 125 nm in diameter consisting of an InP core and a 1.5 nm thickness InAs tube surrounded by an InP shell.

By way of contrast, Fig. 3.2(c) shows a ball-and-stick model of a single-walled carbon nanotube about 1.2 nm in diameter. Carbon nanotubes were discovered in 1991 by Iijima [6]. They are large macromolecules with the structure of a seamless cylindrical sheet of graphite capped by hemispherical ends. Electrical measurements of individual metallic single-wall carbon nanotubes have revealed remarkable properties. Estimates of low-field room-temperature acoustic-phonon-limited mean-free-path give $l_{\rm ap} \sim 300$ nm and a high-field optic-phonon-limited mean-free-path of $l_{\rm op} \sim 15$ nm [7–9]. These results were obtained by fitting measured current-voltage data from 1.5 nm to 2.5 nm diameter metallic carbon nanotubes as a function of nanotube length and gate voltage to classical Monte Carlo calculations of electron transport. Also of note, extraordinarily high current densities, in some cases estimated to be in excess of 10^9 A cm^{-2}, have been reported [9]. These, and other astonishing claims, have contributed to great excitement surrounding electron transport phenomena in carbon nanotubes. However, some outstanding issues remain to be resolved before the properties of carbon nanotubes can be used in a practical nanoelectronic technology.

Metallic single-wall carbon nanotubes are believed to have two one-dimensional subbands crossing the Fermi energy so that quantum conductance is limited to $2e^2/(\hbar\pi) = 6.45$ kΩ^{-1}. The resulting resistance impacts potential high-frequency performance. For example, assuming that it is possible to design devices with total capacitance of $C = 0.2$ fF [10] then the RC time constant suggests an operating frequency of less than $f = 1/(2\pi RC) = 122$ GHz. This is significantly smaller than the cut-off frequency of conventional CMOS with a minimum feature size of 32 nm.

It is worth mentioning that the *intrinsic* cut-off frequency of a carbon nanotube transistor is believed to be about 50% greater than the equivalent CMOS device [11]. However, long before such a theoretical limit can be approached, there are many practical issues to be resolved. For example, the resistance between a metallic nanotube and bulk metal must be reduced. Typical values for contact resistance are around 40 kΩ which is far greater than the intrinsic nanowire resistance of 6.45 kΩ. Beyond this there are uncertainties relating to manufacturability and reproducibility. So far, carbon nanotubes cannot be fabricated with atomic precision. This means that their exact physical properties are not well controlled. For example, it is not possible to guarantee metallic or semiconducting behavior. Failure to address this as well as other synthesis issues has limited use of carbon nanotubes for device applications.

Rather than dwell on material science challenges facing a specific physical system such as carbon nanotubes, we would like to focus on device design at the nano- and atomic scale in those material growth systems that have successfully shown exquisite control over geometry and material composition. It is this control that creates the opportunity to realize new device designs. We are interested in creating devices with new types of functionality whose operation is determined by new principles. To illustrate the idea that new devices can be created, in the next section we consider the example of a transistor that requires ballistic electron transport to operate.

3.1.1 Example: A transistor that requires ballistic electron transport to operate

The atomic layer-by-layer control of established crystal growth techniques such as MBE can be used to create electronic devices with new principles of operation. This is possible because semiconductor heterostructures with large built-in electric fields can be designed with atomic-layer precision to exploit thermionic emission, electron tunneling, and extreme non-equilibrium electron transport. In fact, nonlinear devices such as transistors can be created that make use of these transport phenomena. The ballistic electron transistor illustrated in Fig. 3.3 is an example of an electronic semiconductor heterostructure device that *requires* ballistic electron transport for its operation [12]. Remarkably, by using an active region that is only 10 nm (33 atomic layers) thick and injecting electrons at high energy, it is possible to create such a transistor that operates satisfactorily at room temperature.

Key features are high-energy electron injection from the emitter into the base region. As illustrated in Fig. 3.3(b), injection energy near 1.3 eV is achieved by using the large conduction band heterojunction offset energy

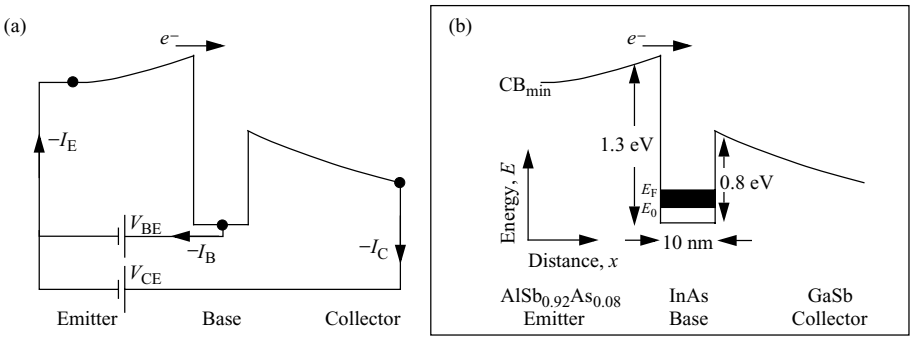

Fig. 3.3. (a) A double heterojunction unipolar transistor. Indicated in the figure are the emitter current I_E, base current I_B, collector current I_C, and voltages V_{BE} and V_{CE} for the transistor biased in the common emitter configuration. The solid dots indicate electrical contacts between the transistor and the leads that connect to the batteries. (b) The conduction band of an $AlSb_{0.92}As_{0.08}$/InAs/GaSb double heterojunction unipolar ballistic electron transistor under bias. The conduction-band minimum CB_{min} is indicated, as are the confinement energy E_0 and the Fermi energy E_F of the occupied two-dimensional electron states in the InAs base. Electrons indicated by e^- are injected from the forward-biased $AlSb_{0.92}As_{0.08}$ emitter into the InAs base region with a large excess kinetic energy [12].

between AlSb and InAs. Because thermal energy at room temperature is near $k_BT = 0.025$ eV, the energy distribution of ballistically injected nonequilibrium injected electrons is well separated from ambient n-type majority carriers in the base. The very thin base and matching of electron velocity across the InAs/GaSb base-collector heterostructure enables more than 90% of the ballistically injected electrons to traverse the base and arrive at the collector terminal of the device. The electron velocity matching condition is illustrated in Fig. 3.4. In the figure electron states used in transmission of an electron of energy E through (100)-oriented semiconductor layers are near the points where the dashed line intersects the solid curves. The group velocity v_{InAs} of ballistic electrons in InAs and v_{GaSb} in GaSb is given by the gradient of the dispersion curve at energy E for InAs and GaSb respectively. Quantum mechanical reflection is determined in part by electron velocity mismatch across the abrupt InAs/GaSb heterointerface. For electron velocities v_{InAs} in InAs and v_{GaSb} in GaSb, quantum mechanical reflection is proportional to $(v_{InAs} - v_{GaSb})^2/(v_{InAs} + v_{GaSb})^2$. There is also a contribution to quantum mechanical reflection from the overlap integral of the cell-periodic part of the electron wave function at the InAs/GaSb heterointerface. In this simplified example, we only consider the envelope wave function and so assume that there is no contribution to quantum mechanical reflection from a mismatch in the character (symmetry) of the electron wave

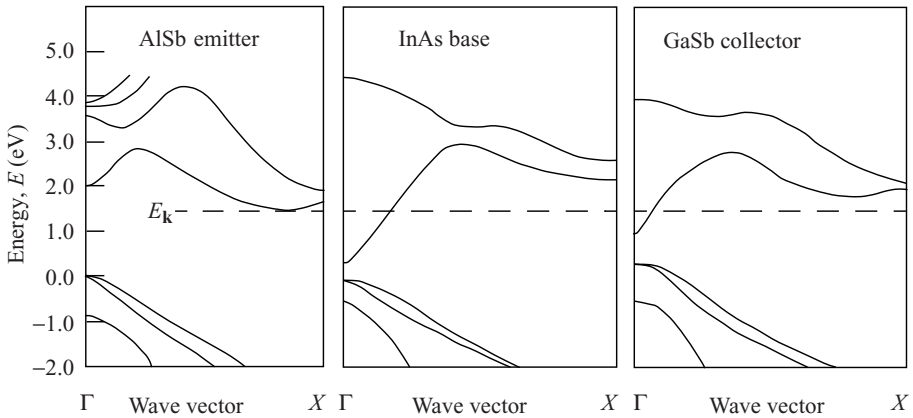

Fig. 3.4. Dispersion curves for electron motion through the three materials forming the AlSb/InAs/GaSb double heterostructure unipolar ballistic electron transistor. The horizontal dashed line indicates the approximate value of energy for an electron moving through the device in the (100) direction. The group velocity of ballistic electrons is given by the gradient of the dispersion curve at energy E. Quantum mechanical reflection of electrons impinging on the InAs/GaSb interface is minimized when the group velocity is matched [12].

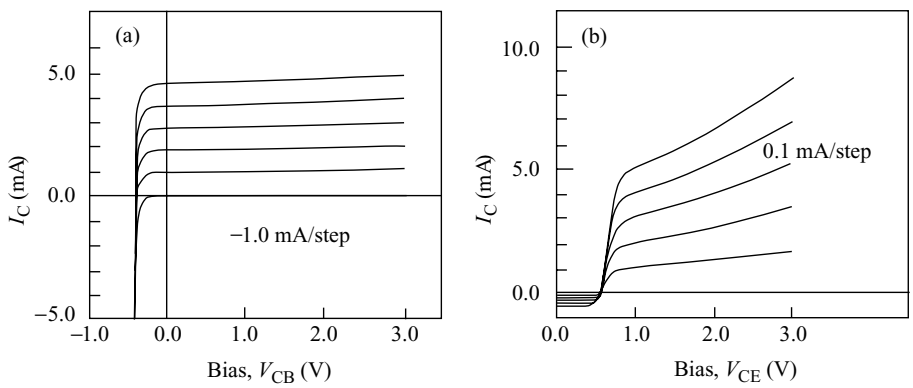

Fig. 3.5. (a) Measured room-temperature ($T = 300$ K) common base current gain characteristics of the device shown schematically in Fig. 3.3(b). Curves were taken in steps beginning with an injected emitter current of zero. The emitter area is 7.8×10^{-5} cm^2, I_C is collector current, and V_{CB} is collector base voltage bias. (b) Measured room-temperature common emitter current gain characteristics of the device in (a). Curves were taken in steps of $I_B = 0.1$ mA, beginning with an injected base current of zero. V_{CE} is collector emitter voltage bias and current gain $\beta \sim 10$ [12].

function across the interface.

The design shown in Fig. 3.3(b) works as a transistor. Figure 3.5(a) shows measured room-temperature ($T = 300$ K) common base current gain

characteristics [12]. Figure 3.5(b) shows measured room-temperature common emitter current gain characteristics of the same device. The measured current gain for this prototype device is $\beta \sim 10$.

The demonstration of a ballistic electron transistor that operates at room temperature is an example of a device that *requires* nanoscale dimensions to operate. The design shown in Fig. 3.3(b) and the experimental results shown in Fig. 3.5 are, however, the result of an ad hoc design process. No systematic design study or search for an optimal design was undertaken and so it is not known if, for example, significantly better designs exist.

There are, of course, a large number of degrees of freedom available that could be used for future designs. It would obviously be helpful to adopt an optimal design strategy to explore the available design space. As a starting point, and very much in the spirit of a prototype problem, in the next section we introduce the physics of electron tunneling and the calculation of tunnel current in a layered semiconductor heterostructure.

3.2 Elastic electron transport and tunnel current

Consider the MBE-grown heterostructure tunnel barrier illustrated in Fig. 3.6(a). The figure shows the local conduction band potential profile, $U(x)$, of an $Al_\xi Ga_{1-\xi}$ heterostructure tunnel barrier in GaAs configured as an $n-i-n$ diode. The average value of $U(x)$ for each atomic monolayer in the (100) crystal growth direction is controlled by the layer's Al concentration, ξ.

In Fig. 3.6(b) an electron of energy E and wave vector $\mathbf{k}$ is shown incident from the left. The undoped $Al_\xi Ga_{1-\xi}As$ barrier is 8 nm thick and has a 0.25 eV conduction band offset energy relative to GaAs corresponding to $\xi = 0.3$. The n-type carrier concentration in the GaAs electrodes is $n = 10^{18}$ cm^{-3}. In the figure a bias voltage of $V_{\text{bias}} = 0.125$ V is applied between the left and right electrodes of the device and a solution to the Poisson equation in the depletion approximation [13] is used to calculate the resulting potential profile.

We would like to calculate the current that flows through the device as a function of V_{bias}. To do this we need to make some approximations that simplify the problem. This will get us some early results and, at the same time, help us make choices on how to improve the model later.

Because we will be considering relatively low-energy electrons near a band minimum, we use an effective electron mass, m. A detailed model might use a more realistic band structure such as the tight-binding model that is popular with some engineers. However, this adds considerable complexity for very little new insight into the problem at hand and so, for now, we choose to

ignore it and use effective electron mass and plane-wave envelope states [14].

If current densities are low, we can assume that self-consistency between solutions of the Schrödinger equation and the Poisson equation gives rise to a small correction and so can be ignored. Inelastic processes are also ignored, further justifying the use of the depletion approximation to calculate the conduction band potential profile. Electrons are assumed independent and noninteracting and so the current may be calculated as a sum over independent channels.

Given the electron wave function, the current due to each of these independent channels can be calculated through the use of the current operator

$$\mathbf{J} = -i\frac{e\hbar}{2m}\left(\Psi^*(x,t)\nabla\Psi(x,t) - \Psi(x,t)\nabla\Psi^*(x,t)\right). \tag{3.1}$$

Since we are only concerned with potentials that vary in the x direction, the Hamiltonian can be separated into components perpendicular and parallel to the interface between the electrode and device region. This yields a separable wave function

$$\Psi(x,y,z,t) = \psi_{\parallel}(y,z)\psi_{\perp}(x)e^{-i\omega t}, \tag{3.2}$$

with energy

$$E = E_{\parallel} + E_{\perp} = \frac{\hbar^2 k_{\parallel}^2}{2m} + \left(\frac{\hbar^2 k_{\perp}(x)^2}{2m} + V(x)\right) = \hbar\omega, \tag{3.3}$$

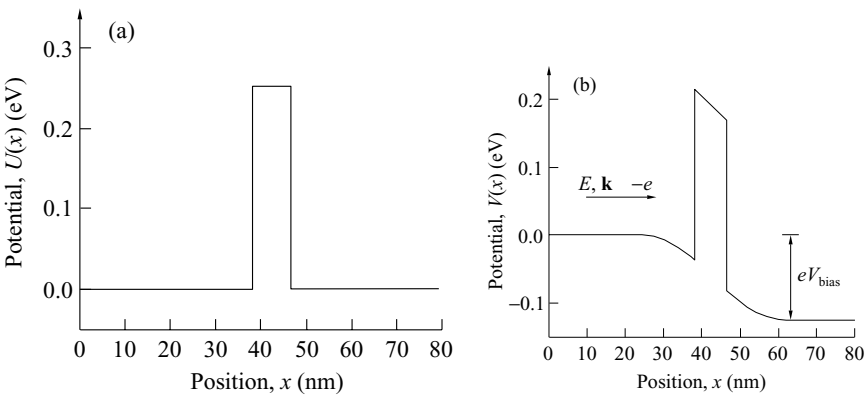

Fig. 3.6. (a) Conduction band potential profile $U(x)$ of an $Al_\xi Ga_{1-\xi}As/GaAs$ heterostructure $n-i-n$ tunnel barrier diode. (b) Conduction band potential profile of the heterostructure tunnel barrier diode in (a) showing an electron of energy E and wave vector $\mathbf{k}$ incident from the left. The undoped $Al_\xi Ga_{1-\xi}As$ barrier is 8 nm thickness and has a 0.25 eV conduction band offset energy relative to GaAs corresponding to $\xi = 0.3$. The n-type carrier concentration in the GaAs is $n = 10^{18}$ cm^{-3} and the diagram shows a bias voltage of $V_{\text{bias}} = 0.125$ V applied between the left and right electrodes.

where

$$k_\perp(x) = \frac{\sqrt{2m(E_\perp - V(x))}}{\hbar}. \tag{3.4}$$

We assume that the current density in the x direction is dependent only on $\mathbf{k}_\perp$ so that Eq. (3.2) through Eq. (3.4) allow us to use the propagation matrix method to calculate the transmission probability [15, 16].

The single electron current density is evaluated at the right-hand side of the system, where it is assumed that no waves are traveling to the left. If the wave function amplitude in this region is c,

$$\psi_\perp(x,t) = ce^{i(k_\perp x - \omega t)}, \tag{3.5}$$

and the current density in the x direction is

$$\mathbf{J} = -i\frac{e\hbar}{2m}\left(\psi_\perp^*(x,t)\nabla\psi_\perp(x,t) - \psi_\perp(x,t)\nabla\psi_\perp^*(x,t)\right) = e\frac{\hbar k_\perp}{m}|c|^2. \tag{3.6}$$

This is the product of the electron charge, the velocity of the electron, and the transmission probability, $|c|^2$.

To extend the result to a three-dimensional current density, we need to multiply by the number of electrons in state $|\mathbf{k}_\perp\rangle$ and then integrate over all $|\mathbf{k}\rangle$ that can contribute to the current. The number of electrons with energy $E_\mathbf{k}$ is given by the product of the Fermi occupation factor

$$f(E_\mathbf{k}) = \frac{1}{1 + e^{((E_\mathbf{k} - \mu)/(k_\mathrm{B}T))}}, \tag{3.7}$$

and the three-dimensional density of states, which in k-space is

$$D_3(\mathbf{k})d^3\mathbf{k} = \frac{d^3\mathbf{k}}{(2\pi)^3}. \tag{3.8}$$

The system is assumed to be in thermal equilibrium and characterized by temperature T and chemical potential μ.

Since the Hamiltonian has been separated into parallel and perpendicular components, the total current density becomes

$$\mathbf{J} = e\int\frac{dk_\perp}{2\pi}\int\frac{d^2\mathbf{k}_\parallel}{(2\pi)^2}\frac{\hbar\mathbf{k}_\perp}{m}|c|^2 f(E_\mathbf{k}). \tag{3.9}$$

Evaluation of the integrals is greatly simplified by converting to energy. However, the vector nature of the current is lost in the process. Therefore we will denote current direction by an arrow superscript. The two-dimensional density of states is a constant in energy, allowing the $k_\parallel$ integral to be evaluated analytically as

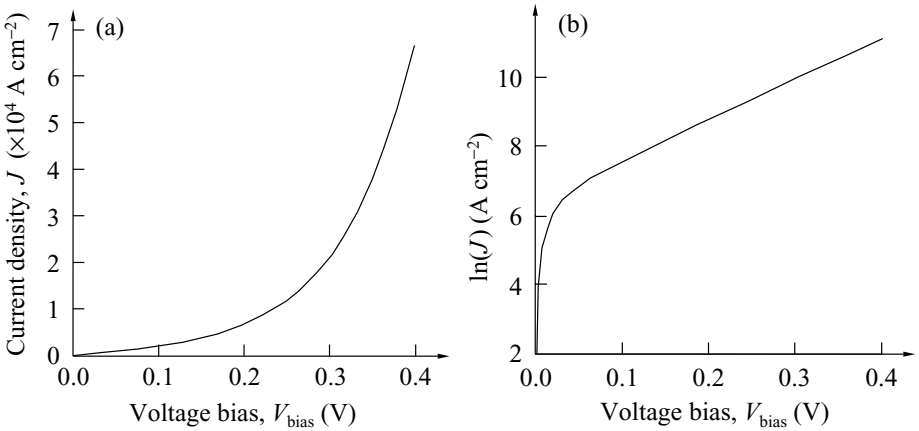

Fig. 3.7. (a) Calculated room-temperature ($T = 300$ K) current–voltage characteristic for the $n-i-n$ tunnel barrier diode illustrated in Fig. 3.6. The 8 nm thickness $Al_\xi Ga_{1-\xi}As$ undoped barrier has a 0.25 eV conduction band offset energy relative to GaAs corresponding to $\xi = 0.3$. The n-type carrier concentration in the GaAs is $n = 10^{18}$ cm^{-3} and a bias voltage of V_{bias} is applied between the left and right electrodes. The effective electron mass is $m = 0.07 \times m_0$. (b) Same as (a) but with current density plotted using a natural logarithm scale.

$$N(k_\perp) = 2\int_0^\infty \frac{2\pi k_\parallel^2 dk_\parallel}{(2\pi)^2} f(E_{\mathbf{k}}) = \frac{m}{\pi\hbar^2}\int_0^\infty \frac{dE_\parallel}{1 + e^{((E_{\mathbf{k}}-\mu)/(k_{\text{B}}T))}}$$
$$= \frac{mk_{\text{B}}T}{\pi\hbar^2}\ln\left(1 + e^{((\mu - E_\perp)/(k_{\text{B}}T))}\right), \tag{3.10}$$

where we have multiplied by 2 to account for electron spin. $N(k_\perp)$ is called the supply function.

Converting the one-dimensional density of states to energy cancels the velocity term in Eq. (3.9) so that the final expression for the current density in one direction is

$$\mathbf{J}^{\rightarrow} = \frac{emk_{\text{B}}T}{2\pi^2\hbar^3}\int_0^\infty |c|^2 \ln\left(1 + e^{((E_\perp - \mu)/(k_{\text{B}}T))}\right) dE_\perp, \tag{3.11}$$

where $E = 0$ eV is chosen as the conduction band minimum.

To drive current through a device a voltage V_{bias} is applied across the terminals, lowering the chemical potential of one of the electrodes by eV_{bias}. Under these circumstances, the supply functions of the electrodes are no longer equal to each other. A greater number of electrons enter the system from the higher potential and the net current that flows is the difference between the left-to-right current and the right-to-left current. Due to time-reversal symmetry for elastic scattering, the transmission probability is the

same in both directions so that the total current density is

$$\mathbf{J} = \frac{emk_{\mathrm{B}}T}{2\pi^2\hbar^3}\int_0^\infty |c|^2 \ln\left(\frac{1+e^{((\mu-E_\perp)/(k_{\mathrm{B}}T))}}{1+e^{((\mu-E_\perp-V_{\mathrm{bias}})/(k_{\mathrm{B}}T))}}\right) dE_\perp. \tag{3.12}$$

The current through the $n-i-n$ tunnel-barrier diode potential of Fig. 3.6 as a function of bias voltage is shown in Fig. 3.7. As expected for a tunnel-barrier diode, current increases exponentially with bias voltage, V_{bias}.

3.3 Local optimal device design using elastic electron transport and tunnel current

In the next few sections we wish to develop methods that enable us to design diodes with desired current–voltage characteristics. Specifically, we are interested in non-exponential behavior. To do this we need to control the design parameters which include the local potential profile, $U(x)$.

3.3.1 Parameterization of the design space

Consider the problem of finding the optimal local potential profile $U(x)$ for an intrinsic region of semiconductor that results in a desired current voltage characteristic in an $n-i-n$ diode. A generic potential profile is illustrated in Fig. 3.8.

The most important design parameters for this optimization problem are the values of the local potential $U(x)$. Each atomic monolayer contains a

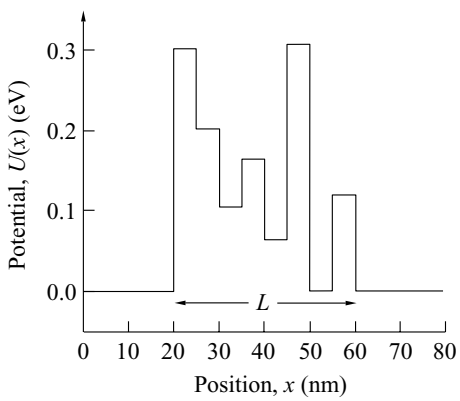

Fig. 3.8. Local conduction band potential profile $U(x)$ in region of thickness L. In this particular case $L = 40$ nm and $U(x)$ consists of 8 layers of equal thickness.

fixed barrier strength, $U(x) = \sum_{l=1}^{N} \chi_l(x) U_l$, where U_l is the barrier potential energy of the lth barrier-layer and $\chi_l(x)$ is the indicator function of the lth barrier-layer. The width of each barrier-layer is an integer multiple of one atomic monolayer. The spacing between adjacent atomic monolayers is material dependent. For (001)-oriented $Al_\xi Ga_{1-\xi}As$ the atomic monolayer thickness is 0.2827 nm. Average local barrier potential energy is determined by the fraction ξ of Al in any particular layer. At room temperature, the value of U for $\xi < 0.42$ is approximately $U(\xi) = 0.8355 \times \xi$ eV. The limitation that $\xi < 0.42$ means that the barrier energy U_l is subject to simple box constraints 0.00 eV $\leq U_l \leq$ 0.35 eV and the dimensions are independent from one another. The number of barrier-layers can be considered an additional aspect of the design problem. Other design parameters include the width L of the region in which $U(x)$ varies and the carrier concentration n in the electrodes.

3.3.2 Mathematical formulation of the design problem

The design problem consists of finding the local potential profile $U(x)$ such that the desired current–voltage characteristic is obtained. The objective current–voltage characteristic is given by J_{obj} and can be any accessible function of voltage bias, V_{bias}. As a specific example, we will consider linear, quadratic, and square root $J_{\text{obj}}(V_{\text{bias}})$. The cost function is formulated in the standard way as a measure of the distance between the desired and the modeled current–voltage behavior

$$J(U) = \sum_{j=1}^{\nu} \left| J_{\text{obj}}(V_{\text{bias}}^j) - J_{\text{sim}}(V_{\text{bias}}^j, U) \right|^2, \tag{3.13}$$

subject to the constraints that

$$J_{\text{sim}}(V_{\text{bias}}^j, U) = \sum_{r=1}^{R} \frac{emk_{\text{B}}T}{2\pi^2\hbar^3} \ln\left(\frac{1 + \exp((\mu - E_r)/k_{\text{B}}T)}{1 + \exp((\mu - E_r - V_{\text{bias}})/k_{\text{B}}T)} \right) T_{\text{sim}}^r \Delta E_r, \tag{3.14}$$

$$T_{\text{sim}}^r(V_{\text{bias}}^j, U) = \left| \frac{1}{A_0^r} \right|^2, \tag{3.15}$$

and

$$\alpha_l^{j,r} = P_l^{j,r}(U_l)\alpha_{l+1}^{j,r}, l = 0, \ldots, NM + 1, \tag{3.16}$$

where P is the propagation matrix (see Chapter 7) for electron wave function of the form $\psi(x) = Ae^{ikx} + Be^{-ikx}$ with terminal condition given by

$$\begin{pmatrix} A^{M,r}_{NM+1} \\ B^{M,r}_{NM+1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{3.17}$$

This corresponds to a wave function with unity amplitude traveling to $+\infty$ and zero amplitude arriving from $+\infty$, i.e., the electron is not reflected after passing through the device. The vector α in Eq. (3.16) is defined by

$$\alpha^{j,r}_l = \begin{pmatrix} A^{M,r}_l \\ B^{M,r}_l \end{pmatrix} = \begin{pmatrix} A^{M,r}_l(V^j_{\text{bias}}, U_l) \\ B^{M,r}_l(V^j_{\text{bias}}, U_l) \end{pmatrix}. \tag{3.18}$$

Even though it may be more intuitive to state the problem as a boundary value problem with a defined incident wave amplitude, the formulation as a terminal value problem significantly speeds up the forward solver. Although the bias voltage and therefore the transmission profile is a continuous function, the objective function is discretized into ν steps. The design parameters are the barrier terms combined into the vector U. The optimal design problem is formulated as a minimization problem

$$\min_U J, \tag{3.19}$$

subject to the constraints Eq. (3.14), Eq. (3.15), Eq. (3.16), and Eq. (3.17). We note that the computation of A^r_0 using Eq. (3.16) is achieved by repeated multiplication so that the computation of A^r_0 consists of forming the propagation matrices $P^{j,r}_l(U_l, V^j_{\text{bias}}), l = 0, \ldots, NM+1$ followed by the matrix multiplication

$$\begin{aligned} \begin{pmatrix} A^{M,r}_l \\ B^{M,r}_l \end{pmatrix} &= \prod_{l=NM+1}^{0} P^{j,r}_l \cdot \begin{pmatrix} A^{M,r}_{NM+1} \\ B^{M,r}_{NM+1} \end{pmatrix} \\ &= P^{j,r}_0 \cdot P^{j,r}_1 \cdot \ldots \cdot P^{j,r}_{NM+1} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \end{aligned} \tag{3.20}$$

where j is the index of the applied voltage bias V^j_{bias} and r is the index for the electron energy E_r. An adjoint based method can be applied to efficiently compute the derivative of the objective function with respect to the design parameters as is shown in the following section.

3.3.3 Derivative of objective function with respect to design parameters

It is possible to compute the derivative of the objective function given in Eq. (3.13) with respect to the design parameters $\frac{\partial J}{\partial U_l}$ without resorting to

computationally costly finite difference approximations. We make use of the adjoint method. The objective function can be reformulated as follows

$$
\begin{aligned}
J(U) &= \sum_{j=1}^{\nu} \left| J_{\mathrm{obj}}(V_{\mathrm{bias}}^j) - J_{\mathrm{sim}}(V_{\mathrm{bias}}^j, U) \right|^2 = \sum_{j=1}^{\nu} \left| J_{\mathrm{obj}}^j - \sum_{r=1}^{R} \left(\mathrm{const}^{j,r} T_{\mathrm{sim}}^r \right) \right|^2 \\
&= \sum_{j=1}^{\nu} \left| J_{\mathrm{obj}}^{j,r} - \sum_{r=1}^{R} \left(\mathrm{const}^{j,r} \left| \frac{1}{A_0^r} \right|^2 \right) \right|^2 \\
&= \sum_{j=1}^{\nu} \left| J_{\mathrm{obj}}^{j} - \sum_{r=1}^{R} \left(\mathrm{const}^{j,r} \frac{1}{(\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r}} \right) \right|^2 ,
\end{aligned} \tag{3.21}
$$

where $(\bar{\alpha}_0^{j,r})^T$ is the complex conjugate transpose of $\alpha_0^{j,r}$ and $Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$.

We compute the derivative

$$
\begin{aligned}
\frac{\partial J}{\partial U_l} &= \sum_{j=1}^{\nu} 2 \left(J_{\mathrm{obj}}^j - \sum_{s=1}^{R} \mathrm{const}^{j,s} \frac{1}{(\bar{\alpha}_0^{j,s})^T Q \alpha_0^{j,s}} \right) \cdots \\
&\quad \sum_{r=1}^{R} \left(\mathrm{const}^{j,r} \frac{2\mathrm{Re}\left[(\bar{\alpha}_0^{j,r})^T Q \partial_{U_l} \alpha_0^{j,r} \right]}{((\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r})^2} \right) \\
&= \sum_{j=1}^{\nu} \mathrm{Re} \left[4 \left(J_{\mathrm{obj}}^j - \sum_{s=1}^{R} \mathrm{const}^{j,s} \frac{1}{(\bar{\alpha}_0^{j,s})^T Q \alpha_0^{j,s}} \right) \cdots \right. \\
&\quad \left. \sum_{r=1}^{R} \left(\mathrm{const}^{j,r} \frac{(\bar{\alpha}_0^{j,r})^T Q}{((\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r})^2} \frac{\partial \alpha_0^{j,r}}{\partial U_l} \right) \right] \\
&= \mathrm{Re} \left[\sum_{j=1}^{\nu} \sum_{i=0}^{NM} \delta_{i,0} 4 \left(J_{\mathrm{obj}}^j - \sum_{s=1}^{R} \mathrm{const}^{j,s} \frac{1}{(\bar{\alpha}_0^{j,s})^T Q \alpha_0^{j,s}} \right) \cdots \right. \\
&\quad \left. \sum_{r=1}^{R} \left(\mathrm{const}^{j,r} \frac{(\bar{\alpha}_0^{j,r})^T Q}{((\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r})^2} \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right) \right] \\
&= \mathrm{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \sum_{i=0}^{NM} \delta_{i,0} 4 \left(J_{\mathrm{obj}}^j - \left(\sum_{s=1}^{R} \mathrm{const}^{j,s} \frac{1}{(\bar{\alpha}_0^{j,s})^T Q \alpha_0^{j,s}} \right) \right) \cdots \right. \\
&\quad \left. \mathrm{const}^{j,r} \frac{(\bar{\alpha}_0^{j,r})^T Q}{((\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r})^2} \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right] ,
\end{aligned} \tag{3.22}
$$

where an artificial summation over i has been introduced and in particular the subscript in the partial derivative has been changed from $\alpha_0^{j,r}$ to $\alpha_i^{j,r}$.

We formulate an initial value problem by

$$\beta_{i+1}^{j,r} = (P_{i-1}^{j,r})^T \beta_i^{j,r} + \delta_{i0} 4 \left(J_{\text{obj}}^j - \sum_{s=1}^{R} \text{const}^{j,s} \frac{1}{(\bar{\alpha}_0^{j,s})^T Q \alpha_0^{j,s}} \right) \dots \\ \text{const}^{j,r} \frac{Q \bar{\alpha}_0^{j,r}}{((\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r})^2}, \tag{3.23}$$

for $i = 0, \dots, NM$ with initial condition $\beta_0^{j,r} = (0\ 0)^T$ and $r = 1, \dots, R$. We recognize the second term on the right-hand side of Eq. (3.23) as a major part of the summand in Eq. (3.22). We continue computing the derivative by substituting Eq. (3.23) in Eq. (3.22) to find

$$\begin{aligned} \frac{\partial J}{\partial U_l} &= \text{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \sum_{i=0}^{NM} \left\{ \delta_{i,0} 4 \left(J_{\text{obj}}^j - \left(\sum_{s=1}^{R} \text{const}^{j,s} \frac{1}{(\bar{\alpha}_0^{j,s})^T Q \alpha_0^{j,s}} \right) \right) \dots \right. \right. \\ &\qquad \left. \left. \text{const}^{j,r} \frac{(\bar{\alpha}_0^{j,r})^T Q}{((\bar{\alpha}_0^{j,r})^T Q \alpha_0^{j,r})^2} \right\} \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right] \\ &= \text{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \sum_{i=0}^{NM} \left\{ \beta_{i+1}^{j,r} - (P_{i-1}^{j,r})^T \beta_i^{j,r} \right\}^T \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right] \\ &= \text{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \left(\sum_{i=0}^{NM} (\beta_{i+1}^{j,r})^T \frac{\partial \alpha_0^{j,r}}{\partial U_l} - \sum_{i=0}^{NM} (\beta_i^{j,r})^T P_{i-1}^{j,r} \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right) \right]. \end{aligned} \tag{3.24}$$

We now take the derivative of Eq. (3.16) to yield

$$\frac{\partial \alpha_i^{j,r}}{\partial U_l} = \frac{\partial P_i^{j,r}}{\partial U_l} \alpha_{i+1}^{j,r} + P_i^{j,r} \frac{\partial \alpha_{i+1}^{j,r}}{\partial U_l}, \tag{3.25}$$

which we in turn substitute into Eq. (3.24)

$$\begin{aligned} \frac{\partial J}{\partial U_l} &= \text{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \left(\sum_{i=0}^{NM} (\beta_{i+1}^{j,r})^T \frac{\partial \alpha_0^{j,r}}{\partial U_l} - \sum_{i=0}^{NM} (\beta_i^{j,r})^T P_{i-1}^{j,r} \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right) \right] \\ &= \text{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \dots \right. \\ &\qquad \left. \left(\sum_{i=0}^{NM} (\beta_{i+1}^{j,r})^T \left(\frac{\partial P_i^{j,r}}{\partial U_l} \alpha_{i+1}^{j,r} + P_i^{j,r} \frac{\partial \alpha_{i+1}^{j,r}}{\partial U_l} \right) - \sum_{i=0}^{NM} (\beta_i^{j,r})^T P_{i-1}^{j,r} \frac{\partial \alpha_i^{j,r}}{\partial U_l} \right) \right] \\ &= \text{Re} \left[\sum_{j=1}^{\nu} \sum_{r=1}^{R} \sum_{i=0}^{NM} (\beta_{i+1}^{j,r})^T \frac{\partial P_i^{j,r}}{\partial U_l} \alpha_{i+1}^{j,r} \right], \end{aligned} \tag{3.26}$$

where the second sum over index i forms a telescoping series with the second term in the first sum over i. We also used the fact that $\frac{\partial \alpha_{NM+1}^{j,r}}{\partial U_l} = 0$ for all j as it is given as the terminal condition in Eq. (3.17).

In summary, the gradient of J can be obtained by solving the original terminal value problem Eq. (3.20) to compute the $\alpha_i^{j,r}$s, the initial value problem Eq. (3.23) to obtain $\beta_i^{j,r}$s, and the partial derivatives of the transmission matrices $\frac{\partial P_i^{j,r}}{\partial U_l}$ for $j = 1, \ldots, \nu$ and each of the barrier-layers $U_l, l = 1, \ldots, N$ as well as $r = 1, \ldots, R$. These derivatives are usually extremely fast to compute because they are simple functions and most of the partial derivatives will be zero. The additional computational cost of obtaining the gradient then consists of solving the initial value problem which takes the same amount of time as solving the terminal value problem from the device simulation. The gradient ∇J is then given by Eq. (3.26) without the rounding errors that would occur in a finite difference approximation of the gradient.

To demonstrate control using the layered semiconductor device illustrated in Fig. 3.8 we first attempt to create a current that varies linearly with the applied V_{bias}. To achieve this, J_{obj} is changed to $J_{\text{obj}}(V_{\text{bias}}) = s_l V_{\text{bias}}$ where the value s_l is, to a certain degree, arbitrary. It is possible and even likely that the achievable device performance varies for different choices of s_l. Unless s_l is fixed by other system constraints the optimization should include the search for an optimal slope parameter s_l, which we consider an *auxiliary* optimization parameter. Incorporating s_l directly into the single objective function search is not trivial because the values of J_{obj} tend to be smaller for smaller s_l. The minimization would inevitably result in the minimum allowable s_l and the corresponding optimal design parameters instead of the slope at which the device performance is optimal. We therefore fix s_l to a value that seems reasonable based on a few simulations. For a fixed auxiliary parameter we proceed with the design optimization by choosing random initial conditions for $U(x)$ followed by a sequential local optimization procedure. If there are multiple active parameter constraints after the optimization we observe that given the chosen auxiliary parameter, s_l is not natural to the device. This suggests that optimal auxiliary parameters yield a design that has the least number of active constraints.

3.3.4 Local optimization

For local optimization and in the absence of any additional knowledge, the initial parameter setting may be randomly chosen from the feasible set. For

the local optimization algorithm we use a variant of the BFGS method which is a quasi-Newton method. The three test problems we consider are a linear, a quadratic, and a square root current current–voltage characteristic.

1. $J(V_{\text{bias}}) = s_l \frac{V_{\text{bias}}}{V_{\text{bias}}^{\max}}$
2. $J(V_{\text{bias}}) = s_q \left(\frac{V_{\text{bias}}}{V_{\text{bias}}^{\max}}\right)^2$
3. $J(V_{\text{bias}}) = s_{\text{sqrt}} \sqrt{\frac{V_{\text{bias}}}{V_{\text{bias}}^{\max}}}$

Note that s_l, s_q, and s_{sqrt} are coefficients of the linear, quadratic, and square-root characteristics, respectively. These coefficients are of engineering interest themselves, as their value influences the accessibility of the objective current–voltage characteristics under given constraints. The constraints for this problem are of the simple box type, where each potential barrier energy can be varied continuously from 0.00 eV to 0.35 eV. For this example we consider 11 potential barriers, each 4 monolayers thick. The initial potential profile $U(x)$ for each local optimization was randomly chosen. The resulting locally optimal potential barrier profiles with corresponding current profile for the linear case are shown in Fig. 3.9.

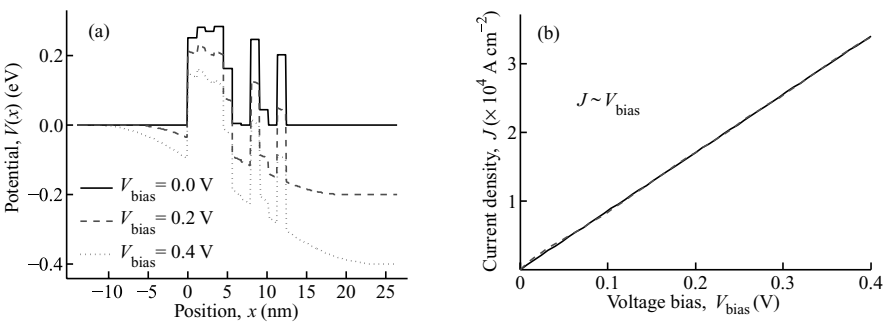

Fig. 3.9. (a) Potential profile with optimized barrier terms for linear current–voltage characteristic at several values of V_{bias}. The optimized potential profile was obtained after a significant number of function and gradient evaluations (>500). (b) The linear current density as a function of V_{bias}. The linear current–voltage characteristic (broken line) does not deviate noticeably from the desired behavior (solid line). In the calculation maximum applied bias is 0.4 V, temperature is $T = 300$ K, carrier density is $n = 10^{18}$ cm^{-3}, effective electron mass is $m^* = 0.07 \times m_0$, relative permittivity is $\epsilon = 13.1$, monolayer thickness is 0.2826×10^{-9} m, each barrier is 4 monolayers thick, there are 11 barriers for a total thickness of 44 monolayers (12.4 nm), minimum barrier energy is 0.0 eV, and maximum possible barrier energy is 0.3 eV.

We see that there are only a few active constraints which could be eliminated for changed s_l. The next design objective included finding a quadratic current–voltage characteristic. The resulting behavior is shown in Fig. 3.10(b) for optimized barrier potentials shown in Fig. 3.10(a). Again the

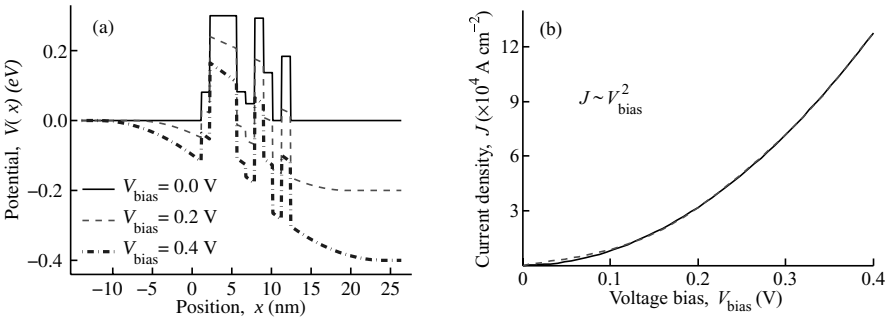

Fig. 3.10. (a) Potential profile with optimized barrier terms for quadratic current–voltage characteristics at several values of V_{bias}. The optimized potential profile was obtained after a significant number of function and gradient evaluations (>500). (b) The quadratic current density as a function of V_{bias}. The quadratic current–voltage characteristic (broken line) does not deviate noticeably from the desired behavior (solid line). The local optimization took about 400 objective function and gradient evaluations. In the calculation maximum applied bias is 0.4 V, temperature is $T = 300$ K, carrier density is $n = 10^{18}$ cm^{-3}, effective electron mass is $m^* = 0.07 \times m_0$, relative permittivity is $\epsilon = 13.1$, monolayer thickness is 0.2826×10^{-9} m, each barrier is 4 monolayers thick, there are 11 barriers for a total thickness of 44 monolayers (12.4 nm), minimum barrier energy is 0.0 eV, and maximum possible barrier energy is 0.3 eV.

local optimization works fairly well in obtaining the desired current–voltage behavior.

A significant increase in difficulty is posed by the square root current–voltage characteristic shown in Fig. 3.11(b). The corresponding optimized barrier potentials are shown in Fig. 3.11(a). The local optimization took 514 function and gradient evaluations until convergence starting at a random parameter setting. We can see that a few constraints are active, i.e. the first two barrier-layers on the left are at their allowed maximum value and three of the barrier-layers in the middle are at their allowed minimum value. Nevertheless the square-root potential is achieved by local optimization only.

Such nonintuitive locally optimal designs are often, at least initially, beyond the engineers' intuition and solely based on optimization by simulation. The locally optimal designs at this point come without a guarantee of global optimality. Hence, there might very well be comparable or even superior designs within the design specifications.

3.3.5 Convergence

We may check the numerical convergence of the forward solver by comparison of the computed current profile for multiple resolutions in space and

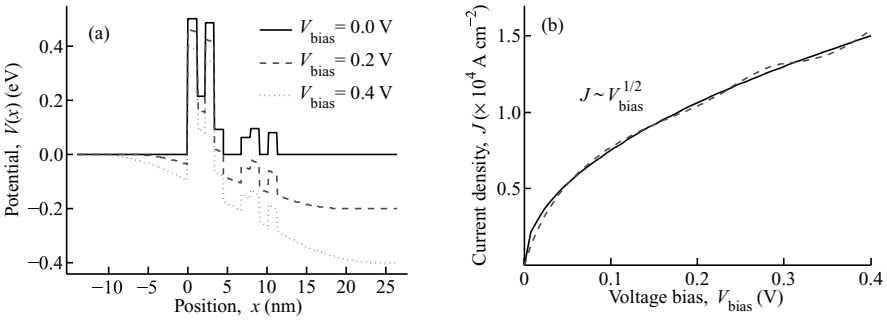

Fig. 3.11. (a) Potential profile with optimized barrier terms for square-root current–voltage characteristics shown at several values of $V_{\rm bias}$. The optimized potential profile was obtained after a significant number of function and gradient evaluations (>500). (b) Current density as a function of $V_{\rm bias}$. The square-root current–voltage characteristic (broken line) deviates noticeably from the desired behavior (solid line). The square-root behavior is only approximately achieved in comparison to the linear or squared behavior. In the calculation maximum applied bias is 0.4 V, temperature is $T = 300$ K, carrier density is $n = 10^{18}$ cm^{-3}, effective electron mass is $m^* = 0.07 \times m_0$, relative permittivity is $\epsilon = 13.1$, monolayer thickness is 0.2826×10^{-9} m, each barrier is 4 monolayers thick, there are 11 barriers for a total thickness of 44 monolayers (12.4 nm), minimum barrier energy is 0.0 eV, and maximum possible barrier energy is 0.5 eV.

energy. Figure 3.12 plots the relative error between the simulated quadratic current profile computed for the highest resolution and the simulated current at lower resolutions. The highest spatial resolution used was $n_x = 32$ grid points per monolayer (0.2826 nm) and the highest energy resolution was $n_{\Delta E} = 32$ grid points per energy interval $\Delta E \approx k_{\rm B}T/4$. The legend in the plot lists the increasing number of grid points in the two dimensions from $n_x = n_{\Delta E} = 1$ to $n_x = n_{\Delta E} = 16$. The plot shows the relative error between the current profile computed at the indicated resolution and the maximum resolution $\left(C(n_x = n_{\Delta E} = 32) - C(n_x, n_{\Delta E})\right)/C(n_x = n_{\Delta E} = 32)$. With $n_x = n_{\Delta E} = 4$ the error is essentially negligible on a linear scale and the convergence of the numerical forward solver is apparent.

3.3.6 Natural objective functions and efficient parallel search

Alternatives to the least squares performance measure can be considered. For the linear current–voltage characteristic a measure of linearity can be used. In the electron-tunneling problem we may be more interested in current for $V_{\rm bias}$ near 0 V rather than the maximum $V_{\rm bias}^{\rm max}$. In this case simple bias voltage dependent weights can be added to the cost function.

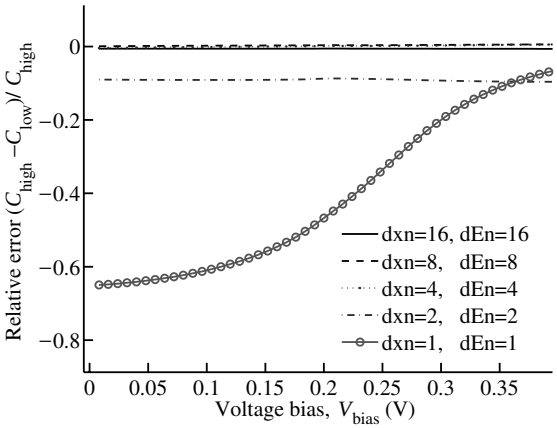

Fig. 3.12. Convergence of the simulated current profile with increasing number of grid points in both the spatial dimension (dxn) as well as number of grid points in energy dimension (dEn). The relative error between 16 grid points per spatial monolayer (dxn = 16) and 16 grid points per energy unit (dEn = 16) compared to the highest resolution computation (dxn = 32, dEn = 32) is not noticeable on a linear scale.

We may be interested in a linear current–voltage characteristic without fixing auxiliary parameters such as the slope and other coefficients beforehand. This would reduce the amount of engineering intuition in the design process and allow a computer to take on more of the design process. A computer could automatically determine the particular slope of the linear current–voltage characteristic that most naturally fits the given device. Finding such a *natural* auxiliary parameter is also necessary to determine the capabilities of a given design. It is easy to evaluate the objective function for many different auxiliary parameters once the forward problem has been solved. Therefore the parallel search does not add significantly to the computational effort. For a linear current–voltage characteristic the objective function can be evaluated for multiple slopes simultaneously without sacrificing additional compute time. To implement a parallel search and find a natural auxiliary parameter value several cost functions with different auxiliary parameter values need to be compared. We have found that the least number of active constraints is a characteristic of a natural design. An active constraint represents a maximum or minimum possible parameter value and indicates that device performance could be improved by relieving this active limitation. If there are no active constraints no such simple improvement can be made, suggesting a natural state or natural design for the given auxiliary parameters.

3.4 Inelastic electron transport

Beyond elastic scattering of electrons from a static potential profile, we are interested in inelastic electron transport in which electrons can gain or lose energy by collisions with other electrons or lattice vibrations. With this in mind, we now turn our attention to the calculation of current in the presence of inelastic scattering. Consider the inelastic system shown in Fig. 3.13, where the potential barrier is an inelastic "black box". We will only consider one-dimensional inelastic systems, so that an electron in initial state $|k\rangle$ with energy E_k enters from the left and inelastically scatters into a final state $|k'\rangle$ with energy $E'_{k'}$. When the inelastic collision occurs, the electron may either emit a phonon or absorb a phonon of energy $\hbar\omega_0$ and momentum $\hbar q$, provided the system is not at absolute zero. If temperature $T = 0$ K then there will not be any real phonons present for the electron to absorb, and phonons may only be emitted.

To conserve energy and momentum, the electron's final energy is $E' = E - n\hbar\omega_0$, where n is the number of phonons that have been excited, and its final momentum is $\hbar k' = \hbar k - \sum_j \hbar q_j$, where the sum is over all phonons the electron has interacted with. We choose the conventions for n and q such that positive n denotes a net phonon emission and $-\hbar q_j$ is the momentum that the jth phonon carries away.

To determine the current through this system we must first decide how we will approach the inelastic scattering problem. If the inelastic scattering is coherent, we could assume that the electrons exist in independent plane wave states and that these states can be solved using an appropriate Schrödinger

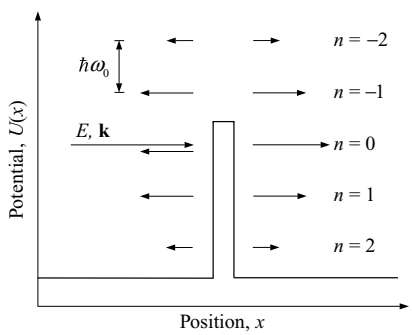

Fig. 3.13. Reflection and transmission due to an inelastic system. An electron initially in state $|k\rangle$ with energy E inelastically scatters into final state $|k'\rangle$ with energy $E' = E - n\hbar\omega_0$. Positive n denotes net phonon emission. If $T = 0$ K, the electron may only lose energy.

equation. Once we have the wave functions, the transmission probability for each channel can be calculated and used to find the current.

In this picture we have assumed that the electron states at all times and positions have a definite phase with respect to the incident electron. In other words, it is assumed that there are no scattering mechanisms unaccounted for that could decohere the electrons and inhibit our ability to predict the behavior of the system. This is the coherent scattering regime which will be explored in Section 3.4.2.

Conversely, we could consider a situation in which there is so much scattering that the electron's motion through the potential is not dependent upon its phase. This is called the incoherent, or sequential, tunneling regime, in which the electrons are treated as localized particles. We explore this regime in the next section.

3.4.1 Incoherent transport and rate equations

Incoherent sequential resonant tunneling and the application of classical rate equations to the transport properties of weakly coupled quantum systems were considered by Luryi in 1985 [17] and have since been studied by many [18–20]. A typical system, shown schematically in Fig. 3.14(a), contains a quantum dot (QD) or a single electronic level of a molecule weakly coupled to two electrodes. A ball-and-stick model of copper phthalocyanine (CuPc, chemical formula $C_{32}H_{16}CuN_8$), a molecule that has been used in experiments [21], is shown in the inset. One can envisage a metal electrode, for example a scanning tunneling microscope (STM) tip, placed in proximity to a molecule that is on the surface of a metal. The molecule is held against this surface by van der Waals forces, so that when a potential is applied across the tip and metal, current flows through the molecule. When set up in such a fashion, the molecule acts as a QD.

Electron tunneling rates Γ_L and Γ_R describe the process in which electrons are able to hop on or off the QD from the left and right electrodes, respectively. It is assumed that the electron tunneling to the QD through the vacuum occurs at a much slower rate than any other rate in the system, making the QD weakly coupled and justifying the scattering rate approach. This also justifies two important assumptions. First, the electrodes remain in thermal equilibrium at all times, allowing the use of Fermi distribution functions. Second, the electron's lifetime on the QD is long enough that it is able to excite the number of phonons required to reach the electronic level of the QD. The long average time between tunneling events allows the QD to dissipate these phonons before the next electron arrives.

Electron transport across the QD is limited by both the coulomb blockade

effect associated with the electronic level and a low tunneling rate. This allows one to think of the electrons with energy $E_0 \pm n\hbar\omega_0$ in the electrodes as particles with a high scattering rate in the electrodes waiting for their turn to hop on to the QD and the battery as a voltage source. With increasing applied voltage bias, additional phonon assisted pathways begin to open by which electrons can flow as current. Figure 3.14(b) shows two such pathways. An electron with energy E_0 may hop directly on to the electronic level of the QD while an electron with energy $E_0 + \hbar\omega_0$ may hop on to the QD and subsequently emit a phonon in order to reach the same electronic level. Due to the low rate at which electrons are able to tunnel on to the QD, it is assumed that the opening of additional channels does not change the tunneling rates of already open channels. This lack of feedback between inelastic channels means that the total tunneling rate on to the QD is the sum of the tunneling rates of all open channels, and we expect that each time a new inelastic channel opens the current will increase in a step-like fashion.

The transport equations can be simplified provided that the QD vibrations are decoupled from the surrounding environment. This assumption will not alter the qualitative nature of the calculation, as the primary effects of the environmental coupling are usually small changes to the current amplitude and phonon frequencies. Additionally it does not prevent the phonons from dissipating between scattering events. The total current through the weakly

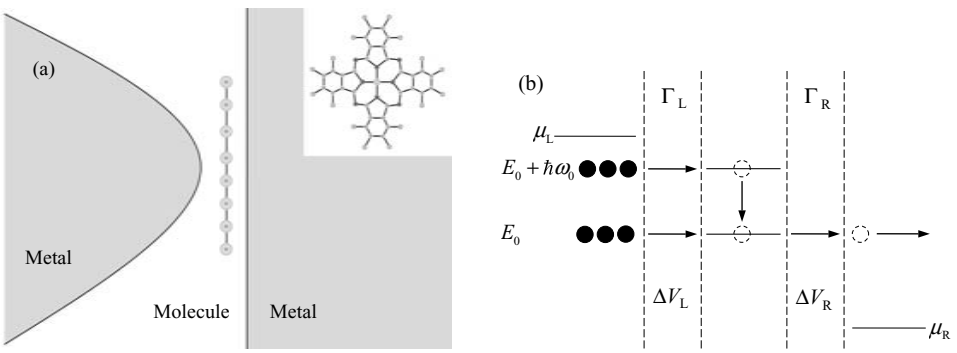

Fig. 3.14. (a) A quantum dot (QD) device consisting of a molecule sandwiched between a metal tip electrode and a planar metal electrode. Shown in the inset is a ball-and-stick model of a CuPc ($C_{32}H_{16}CuN_8$) molecule that has been used in single molecule experiments [21]. (b) Energy diagram of device shown in (a). The electron may either hop directly on to the molecular electronic state with eigenenergy E_0 or may first hop on to the dot with energy $E_0 + \hbar\omega_0$ and then emit a phonon to reach the electronic level. Coupling rates Γ_α and chemical potentials μ_α are shown where α is L for the left electrode and R for the right electrode. There is no change in potential across the QD so that $eV_{\text{bias}} = \Delta V_L + \Delta V_R$.

coupled system is given by

$$I = \frac{2e\Gamma_{\mathrm{L}}\Gamma_{\mathrm{R}}\widetilde{n}_{\mathrm{R}}\widetilde{n}_{\mathrm{L}}\left(\exp\left(\frac{E_0-\mu_{\mathrm{L}}}{k_{\mathrm{B}}T}\right)-\exp\left(\frac{E_0-\mu_{\mathrm{R}}}{k_{\mathrm{B}}T}\right)\right)}{\Gamma_{\mathrm{L}}\widetilde{n}_{\mathrm{L}}\left(2+\exp\left(\frac{E_0-\mu_{\mathrm{L}}}{k_{\mathrm{B}}T}\right)\right)+\Gamma_{\mathrm{R}}\widetilde{n}_{\mathrm{R}}\left(2+\exp\left(\frac{E_0-\mu_{\mathrm{R}}}{k_{\mathrm{B}}T}\right)\right)}, \tag{3.27}$$

where the 2 in the numerator accounts for spin and

$$\widetilde{n}_{\mathrm{L,R}} = \sum_{n=-\infty}^{\infty} P_n(g) f(E_0 + n\hbar\omega_0 - \mu_{\mathrm{L,R}}), \tag{3.28}$$

is a weighted sum of Fermi distribution functions, $f(E)$, that when multiplied by the tunneling rates $\Gamma_{\mathrm{L,R}}$ gives the tunneling rate due to all open inelastic channels. The function

$$P_n(g) = e^{nb-g\coth(b)} I_n\left(\frac{g}{\sinh(b)}\right), \tag{3.29}$$

is a Franck–Condon factor where $b = \frac{\hbar\omega_0}{2k_{\mathrm{B}}T}$, I_n is the modified Bessel function of the first kind, and g is the coupling strength. This function is related to the overlap between initial and final wave functions and gives the probability that an electron is able to excite n phonons and scatter into the QD electronic state.

The step-like behavior predicted for the current can be clearly seen in

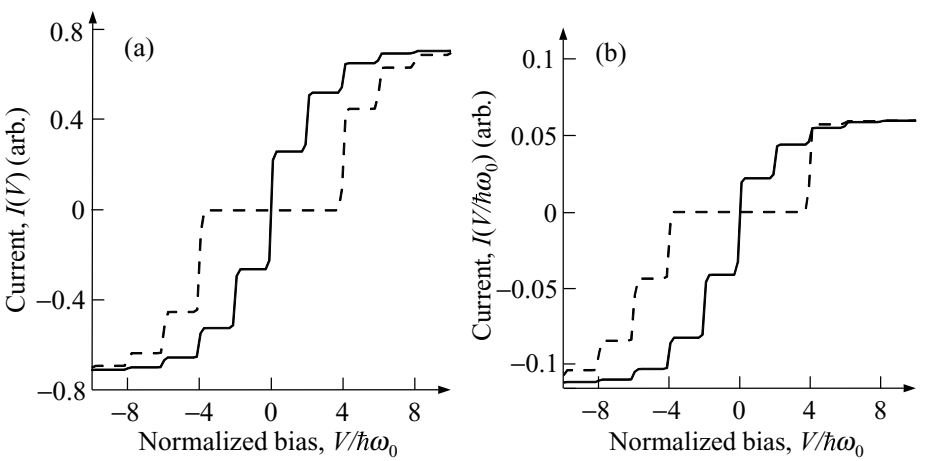

Fig. 3.15. (a) Current through a quantum dot system calculated using the rate equation method with $E_0 = 0$ (solid), $E_0 = 2 \times \hbar\omega_0$ (dashed), and $\Gamma_{\mathrm{L}} = \Gamma_{\mathrm{R}}$. Increasing the QD electronic level prevents current from flowing until the chemical potential is greater than E_0. (b) Current through same system as (a) but with asymmetric scattering rates $\Gamma_{\mathrm{L}} = 0.05 \times \Gamma_{\mathrm{R}}$. The total current is attenuated by the slower scattering rate and is no longer symmetric about $V_{\mathrm{bias}} = 0$. The remaining model parameters are $g = 1$ and $k_{\mathrm{B}}T = 0.02 \times \hbar\omega_0$.

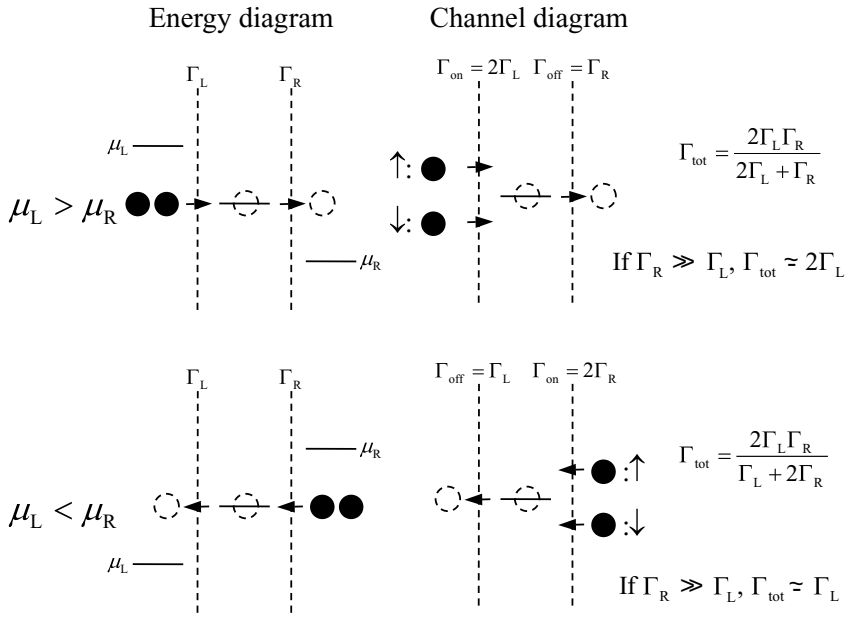

Fig. 3.16. Simplified picture of transport through a QD. The electrode with the higher chemical potential supplies electrons to the QD via two channels due to electron spin. Only one electron may occupy the QD at a time, so that there is only one channel available to leave the QD to the electrode with the lower chemical potential. More current will flow when the slower tunneling rate has more channels available, leading to current asymmetry with respect to bias polarity.

Fig. 3.15. When the electronic energy level of the QD is increased above zero, current may not flow until the chemical potential of one of the leads surpasses E_0. Once it does, electrons from filled states are able to begin hopping on to the QD and current flows. By the time the chemical potential passes E_0, several inelastic channels may already be able to contribute to current, resulting in fewer steps when $E_0 \neq 0$.

In Fig. 3.15(b) the effects of asymmetry in the scattering rates can be seen. In this case $\Gamma_L = 0.05 \times \Gamma_R$, causing the magnitude of the current to be attenuated and dependent upon the sign of the bias. To see why the current is asymmetric we can use the simplified model of the QD system that is shown in Fig. 3.16. We ignore for the moment the inelastic scattering, temperature and other effects, and consider only electrons and rates.

We assume that the flow of electrons is restricted to one direction which is determined by the polarity of the bias and that only an elastic channel is available for current flow. The two parallel paths through which an electron can hop on to the QD are for the two electron spin states. Each of these

paths has the same tunneling rate Γ_{in}, so that the total tunneling rate on to the quantum dot is $\Gamma_{on} = 2\Gamma_{in}$. Since there is only one path out of the QD, the total tunneling rate off is $\Gamma_{off} = \Gamma_{out}$. The tunneling rates Γ_{in} and Γ_{out} are set equal to either Γ_L or Γ_R, depending on the polarity of the bias.

The total time required for an electron to transit the QD system and contribute to current is the sum of the time spent waiting to tunnel on to the QD, τ_{on}, and the time spent waiting to tunnel off of the QD, τ_{off}. Expressing $\tau_{tot} = \tau_{on} + \tau_{off}$ in terms of tunneling rates,

$$\Gamma_{tot} = \frac{1}{\frac{1}{\Gamma_{on}} + \frac{1}{\Gamma_{off}}} = \frac{2\Gamma_{in}\Gamma_{out}}{2\Gamma_{in} + \Gamma_{out}}. \tag{3.30}$$

This equation is remarkably similar to (3.27).

As an example of how asymmetric tunneling rates can affect the current we consider the case where $\Gamma_R >> \Gamma_L$, similar to what is shown in Fig. 3.15. If the potential forces $\mu_L > \mu_R$, $\Gamma_{on} = 2\Gamma_L$ and $\Gamma_{off} = \Gamma_R$. This gives $\Gamma_{tot} \approx 2\Gamma_L$, as shown in the top of Fig. 3.16. Conversely, when $\mu_R > \mu_L$, $\Gamma_{on} = 2\Gamma_R$, $\Gamma_{off} = \Gamma_L$, and $\Gamma_{tot} \approx \Gamma_L$. This case is shown in the bottom half of Fig. 3.16. Since the current through the QD is limited by the slower of the two rates, more current flows when the slower rate has access to multiple channels.

This description, where a quantum system is reduced to a set of scattering rates, has been developed to model the step changes in current that are seen in single molecule experiments [22]. By adding varying levels of sophistication, simulations have been able to reproduce the qualitative, and to some extent quantitative, behavior of the current–voltage characteristic [21]. This model however *requires* that *thermal equilibrium* be reached in the weakly coupled resonant state.

As an example of a practical device that requires *non-equilibrium* conditions, consider the tunnel emitter ballistic electron transistor shown in Fig. 3.17(a) [23]. There is almost no collector current for emitter electron injection energies less than the conduction band heterojunction potential barrier energy between the InAs base and GaSb collector. For base-emitter voltage bias $V_{BE} \geq 0.8$ V, ballistically injected electrons that traverse the base are collected and flow as collector current, I_C. Emitter current I_E that does not contribute to collector current appears as base current. At a base-emitter voltage bias $V_{BE} \sim 1.7$ V almost 90% of injected electrons are collected. The energy of the ballistically injected electrons is $E \sim eV_{BE}$. The measured ratio $\alpha = I_C/I_E$ as a function of V_{BE} is shown in Fig. 3.17(b). As may be seen, α varies with electron injection energy due to the electron velocity mismatch at the base-collector heterointerface.

This device cannot be modeled by using the rate equations of Eq. (3.27)

through Eq. (3.29) due to the non-equilibrium carrier distribution in the base. If this distribution were to reach equilibrium, the only current flowing into the collector would be the reverse bias current between the base and collector and the device would cease to function as a transistor. We can, however, still use rates to model the performance of this device. By estimating the rates with which electrons are able to hop into and out of the base region, a Monte Carlo calculation can be performed to simulate the non-equilibrium electron transport that determines behavior of this device.

If, instead of a single QD electronic level, electrons can tunnel incoherently from a continuum of occupied left-hand metal electrode states into unoccupied right-hand electrode states via emission of molecular vibrations or tunnel barrier phonons then the number of available states determines the inelastic contribution to total current. In this case there is a step change in conductance as electron energy increases above the threshold to excite the vibrational mode. The derivative of conductance, d^2I/dV_{bias}^2, as a function of applied voltage bias, V_{bias}, gives a measure of the vibrational spectrum of molecules [24, 25] and tunnel barrier phonons [26, 27].

Before moving on to the coherent regime, it seems reasonable to ask why we can treat this system in such a classical way. For a QD experiment utilizing gold electrodes, the mobility-derived mean free path for an electron in the metal is $l_{\text{k}} \sim 40$ nm. The de Broglie wavelength at the Fermi energy $\lambda_{\text{F}} = 0.5$ nm corresponding to a Fermi wave vector $k_{\text{F}} = 1.2 \times 10^{10}$ m^{-1}. Clearly we can treat these electrons as quantum mechanical waves inside the conductor bulk because they satisfy $k_{\text{F}} l_{\text{k}} \gg 1$. However, between the

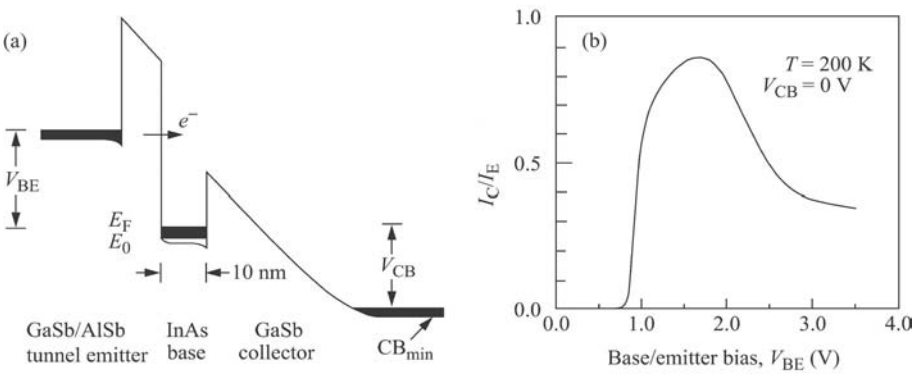

Fig. 3.17. (a) Energy diagram of a tunnel emitter hot electron transistor. Electrons must tunnel through the barrier between emitter and base before attempting to transit the base. (b) Measured $\alpha = I_{\text{C}}/I_{\text{E}}$ through device shown in (a) as a function of base-emitter voltage, V_{BE}. Current through the device is controlled by the impedance mismatch and conduction band offset between the collector and base [23].

metal conductor surface and the QD is a large potential barrier that restricts the flow of the electrons. Since the scattering rate within the electrode is assumed to be much larger than the tunneling rate, the conductor electrons may be viewed as repeatedly attempting to tunnel through the vacuum. A majority of the time the electrons are scattered back into the bulk electrode. Those few electrons that are transmitted rapidly decohere and contribute to an incoherent tunnel current.

3.4.2 Coherent inelastic electron transport

Suppose we are able to completely describe the motion of an electron through a quantum system in a coherent manner. This would allow a wave-like description of the electron in which it is able to interfere with itself and interact with its surroundings quantum mechanically. Preserving the wave nature when considering inelastic scattering leads to drastically different behavior than the sequential tunneling regime described in Section 3.4.1. In the following sections we consider the transport properties of an inelastic system in terms of coherent plane wave electron states.

The time-independent Hamiltonian [28, 29] we consider for an electron in one-dimension is

$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \hbar\omega_0 \hat{b}^\dagger \hat{b} + V(x) + \delta(x)\left(W_0 + W_1\left(\hat{b}^\dagger + \hat{b}\right)\right), \tag{3.31}$$

where $\delta(x)$ couples the localized phonon and the incident electron at $x = 0$, $\hbar\omega_0$ is the phonon energy, and $\hat{b}^\dagger$ and $\hat{b}$ are the phonon creation and annihilation operators, respectively. In this expression, W_0 is the static magnitude of the delta and W_1 is the electron–phonon coupling constant, both having units of energy times length. Note that $V(x)$ is the surrounding static potential the electron experiences.

We consider the case when the potential $V(x)$ is a step at $x = 0$, so that the potential is V_j for negative x and V_{j+1} for positive x, as shown in Fig. 3.18. The system is at temperature $T = 0$ K. This prevents the electron from absorbing a phonon and having energy greater than its injection energy. The electron can, however, excite a phonon at the delta barrier and lose energy. The fact that the electron can exit the system with different energy levels along with the requirement for unitarity creates very rich and intriguing transmission spectra as a function of system parameters.

To solve the Schrödinger equation we expand the wave function in terms of the oscillator basis

$$\langle x|\Psi\rangle = \sum_{n=0}^{n=\infty} \psi_n(x)|n\rangle, \tag{3.32}$$

with the number of excited phonons given by n. We assume that the electrons are made up of plane wave states so that the wave functions in regions j and $j+1$ are

$$\psi_n^j(x) = a_n e^{ik_n^j x} + b_n e^{-ik_n^j x}, \tag{3.33}$$
$$\psi_n^{j+1}(x) = c_n e^{ik_n^{j+1} x} + d_n e^{-ik_n^{j+1} x}, \tag{3.34}$$

where

$$k_n^j = \frac{\sqrt{2m\left(E - n\hbar\omega_0 - V_j\right)}}{\hbar}. \tag{3.35}$$

Integrating the Schrödinger equation about $x = 0$ and ensuring continuity of the wave function, the solutions for a_n and b_n are

$$a_n = i\frac{mW_1}{\hbar^2 k_n^j}\sqrt{n}(c_{n-1} + d_{n-1}) + \left(\frac{1}{2}\left(1 + \frac{k_n^{j+1}}{k_n^j}\right) + i\frac{mW_0}{\hbar^2 k_n^j}\right) c_n$$
$$+ \left(\frac{1}{2}\left(1 - \frac{k_n^{j+1}}{k_n^j}\right) + i\frac{mW_0}{\hbar^2 k_n^j}\right) d_n + i\frac{mW_1}{\hbar^2 k_n^j}\sqrt{n+1}W_1(c_{n+1} + d_{n+1}), \tag{3.36}$$

and

$$b_n = -i\frac{mW_1}{\hbar^2 k_n^j}\sqrt{n}(c_{n-1} + d_{n-1}) + \left(\frac{1}{2}\left(1 - \frac{k_n^{j+1}}{k_n^j}\right) - i\frac{mW_0}{\hbar^2 k_n^j}\right) c_n$$
$$+ \left(\frac{1}{2}\left(1 + \frac{k_n^{j+1}}{k_n^j}\right) - i\frac{mW_0}{\hbar^2 k_n^j}\right) d_n - i\frac{mW_1}{\hbar^2 k_n^j}\sqrt{n+1}W_1(c_{n+1} + d_{n+1}). \tag{3.37}$$

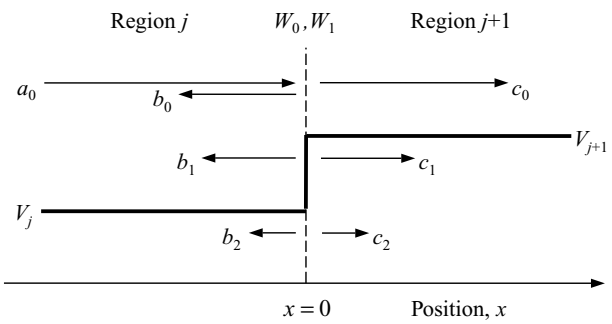

Fig. 3.18. Electron wave of amplitude a_0 incident on static potential step (thick solid line) in the presence of an Einstein phonon located at $x = 0$ (dashed line). The subscript indicates how many phonons have been excited by the electron. W_0 is the static delta-potential amplitude, W_1 is the dynamic delta-potential amplitude. The system is at temperature $T = 0$ K, so that there are no excited phonons present before the electron enters the system.

Note that W_1 couples the inelastic channels and each channel is directly dependent on those channels that have one more or one less phonon. Thus, this system can only excite one phonon of energy $\hbar\omega_0$ at a time, forbidding the excitation of phonons with energy that is a multiple of $\hbar\omega_0$.

The expressions for a_n and b_n can be combined into a matrix equation of the form

$$\begin{bmatrix} a_0 \\ b_0 \\ a_1 \\ b_1 \\ \vdots \\ a_N \\ b_N \end{bmatrix} = \mathbf{P}_{\text{inelastic}} \begin{bmatrix} c_0 \\ d_0 \\ c_1 \\ d_1 \\ \vdots \\ c_N \\ d_N, \end{bmatrix}, \tag{3.38}$$

which can be solved using the propagation matrix method. The boundary conditions for the system are $d_n = 0$, since we assume that no waves enter from the right, and $a_n = \delta_{0,n}$ since the temperature is $T = 0$ K. Once the c_n are obtained, the total transmission for the system may be calculated by

$$T(E) = \sum_{n=0}^{(E-eV_{\text{right}})/(\hbar\omega_0)} \frac{k_n^{\text{right}}}{k_0^{\text{left}}} |c_n|^2, \tag{3.39}$$

where the left and right indicate that the variable should be evaluated at the left or right edge of the system. The term above the summation ensures that only those modes that are able to propagate out of the right-hand side of the system can contribute to the total transmission, and the k normalization guarantees unitarity.

3.4.3 Coherent current continuity

To ensure current is conserved, we must determine the current density at the left and right sides of the system and verify that they are equal. To calculate the current density we use Eq. (3.1)

$$\mathbf{J} = -i\frac{e\hbar}{2m} \left(\Psi^*(x,t)\nabla\Psi(x,t) - \Psi(x,t)\nabla\Psi^*(x,t)\right). \tag{3.40}$$

This time, however, the total wave function is more complicated since we must, at least in principle, include an infinite number of inelastic channels.

A generic inelastic wave function is

$$\Psi(x,t) = \sum_{n=0}^{N} \left(a_n e^{i(k_n x - \omega_n t)} + b_n e^{-i(k_n x + \omega_n t)} \right) \tag{3.41}$$

$$+ \sum_{n=N+1}^{\infty} \left(a_n e^{-\kappa_n x} e^{-i\omega_n t} + b_n e^{\kappa_n x} e^{-i\omega_n t} \right), \tag{3.42}$$

where N is the number of phonons excited by the lowest energy propagating state, $k_n = i\kappa_n$ is given by Eq. (3.35), and

$$\omega_n = \frac{E - n\hbar\omega_0}{\hbar}. \tag{3.43}$$

Although the wave numbers and coefficients are dependent on x through $V(x)$, we will restrict ourselves to evaluating the current density far enough away from the device region that the potential is constant.

Substituting Eq. (3.42) into Eq. (3.40) will yield many cross terms with a time dependence of $e^{i(\omega_n - \omega_m)t}$. These terms will average to zero over time, and therefore do not contribute to the total current density. The remaining terms are

$$\mathbf{J} = \frac{e\hbar}{m} \sum_{n=0}^{N} k_n \left(|a_n|^2 - |b_n|^2 \right). \tag{3.44}$$

Since all of the cross-channel terms average out to zero the inelastic channels are completely independent of each other. The only location where they are not independent is at the inelastic delta barrier.

In addition to the current due to propagating states, Eq. (3.44) could have another term containing the current due to decaying states. However, we have assumed that the current is measured far away from the device region containing the localized phonon and the potential is constant. Thus, these terms will never contribute to the current and are omitted.

The current densities at the left- and right-hand sides of the system are

$$\mathbf{J}_{\text{left}} = e\frac{\hbar k_0^{\text{left}}}{m} - \frac{e\hbar}{m} \sum_{n=0}^{N} k_n^{\text{left}} |b_n|^2, \tag{3.45}$$

and

$$\mathbf{J}_{\text{right}} = \frac{e\hbar}{m} \sum_{n=0}^{N} k_n^{\text{right}} |c_n|^2, \tag{3.46}$$

where we have used the boundary conditions $a_n = \delta_{0,n}$ and $d_n = 0$. Setting these two equations equal to each other gives the unitarity condition

$$1 = \sum_{n=0}^{N} \left(\frac{k_n^{\text{left}}}{k_0^{\text{left}}} |b_n|^2 + \frac{k_n^{\text{right}}}{k_0^{\text{left}}} |c_n|^2 \right). \tag{3.47}$$

One may now see why the inelastic transmission expression contains the k normalization term and is unitary. With this in hand we now look at a few examples of inelastic systems.

3.4.4 Examples of calculating coherent inelastic electron transmission

The addition of electron–phonon interactions creates a fascinating array of phenomena that can provide insight into inelastic transport. As a first example, we look at a system composed of a dynamic delta barrier with a constant background potential. We choose $\hbar\omega_0 = 0.5$ eV, $W_0 = 1$ eV nm, and $W_1 = 0.6$ eV nm. Figure 3.19 shows the transmission of this system compared to the elastic case along with the transmission due to each inelastic channel.

In Fig. 3.19(a) the dramatic effect that inelastic processes have on the transmission spectrum is evident. The cusps are a classic sign of a unitary system [28, 30], and Fig. 3.19(b) clearly shows that the cusps are due to the opening of a new inelastic channel.

As a new channel opens, unitarity requires that the amplitudes of the higher energy states be reduced in order to feed into the new state. At

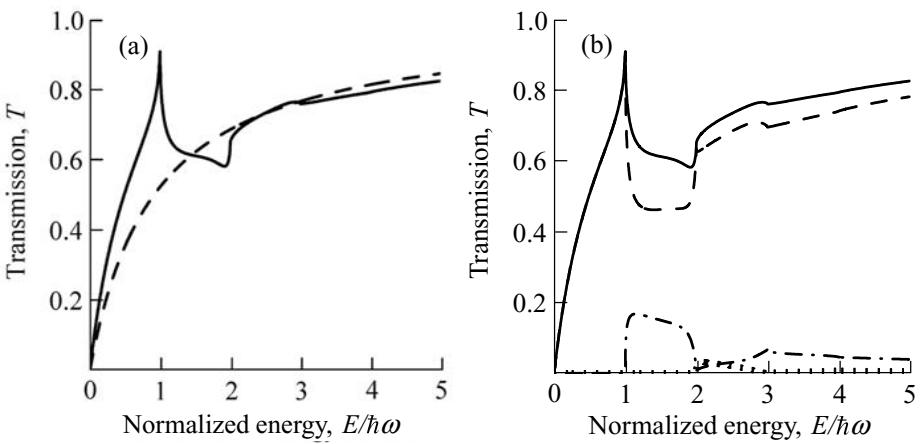

Fig. 3.19. (a) Calculated total coherent electron transmission in the presence of inelastic phonon scattering (solid) and total elastic transmission in the absence of any phonon scattering (dashed). (b) Transmission of total (solid), zero phonon channel (dashed), one phonon channel (dash-dotted), and two phonon channel (dotted). The background potential is constant, and the model parameters are $m = 0.07 \times m_0$, $\hbar\omega_0 = 0.5$ eV, $W_0 = 1$ eV nm, $W_1 = 0.6$ eV nm, and $T = 0$ K.

$E/(\hbar\omega_0) = 1$ in the figure, we can see that the $n = 1$ channel is beginning to propagate. The amplitude of the high-velocity elastic channel decreases as the amplitude of the slower velocity inelastic channel increases, and the result is a lower total transmission. This creates the cusp-like appearance.

Near $E/(\hbar\omega_0) = 2$ there is an increase in total transmission as the $n = 2$ channel opens up. This reduces the amplitude of the $n = 1$ channel, which in turn increases the amplitude of the elastic channel, increasing the overall transmission and creating an "S" shape in the total transmission. This effect was described by Breit [30], whereby the opening of a new inelastic channel alternates between either an "S" shape or a cusp.

To explain all of the features of Fig. 3.19, consider the oscillatory nature of the localized phonon in addition to the unitarity of the system. The oscillating phonon creates a variation in the amplitude of the delta barrier, which in turn perturbs electron transmission. Since tunneling exponentially increases with decreasing potential barrier energy, the electron feels a delta barrier that is effectively lower than the true amplitude of the barrier. The result is the enhanced transmission seen in Fig. 3.19 that occurs before the excitation of the first real phonon.

As a second example consider Fig. 3.20. The phonon energy is $\hbar\omega_0 = 36$ meV, corresponding to the LO phonon in GaAs. The static component of the delta barrier is zero in this case, so that elastic transmission is unity for all $E > 0$. The dynamic component is $W_1 = 0.6$ eV nm. With this particular set of parameters, the first inelastic threshold displays an "S" behavior. This causes the inelastic transmission to continue increasing past the first threshold until it reaches the second, where a cusp occurs.

Next, consider the wave functions of the inelastic channels. This time a potential barrier with an applied bias is considered. We take the same barrier used in Fig. 3.6(a) but use a doping concentration of 10^{17} cm^{-3} so that the depletion region is more pronounced. Using the same delta parameters as Fig. 3.19, a phonon energy of $\hbar\omega_0 = 36$ meV, and an injection energy of $E = 25$ meV, we get the probability densities shown in Fig. 3.21.

Figure 3.21(b) shows detail of the potential of Fig. 3.21(a) along with the normalized electron probability densities up to the fourth inelastic channel. The delta potential is located on the left edge of the tunnel barrier and we can see that inelastic scattering causes charge density to accumulate in the depletion region. Below the potential, the decaying states are localized around the delta and are able to tunnel out of the system due to the applied bias.

One might now ask why one may treat the electron as a delocalized wave and why these coherent inelastic effects do not have a strong effect on current

in most semiconductor devices. To answer the first question one could again turn to the electron wavelength and mean free path. The wavelength of an electron in the conduction band of GaAs with carrier concentration 10^{18} cm^{-3} at the Fermi level $E_F = 52$ meV is $\lambda_F = 20$ nm and the Fermi wave vector is $k_F = 3 \times 10^8$ m^{-1}. The room temperature mobility-derived mean-free-path is $l_k \sim 40$ nm and because $k_F l_k \gg 1$ electrons have wave character in the electrodes. Since the electron wavelength is on the order of the physical dimensions of the quantum device, the electron can be treated quantum mechanically provided the scattering rate is low enough to retain coherence.

Inelastic scattering in semiconductors has been measured in resonant tunnel diodes [31]. This effect shows up as an additional current peak that is shifted in energy by $\hbar\omega_0$. This is possible because electrons moving via a resonant state experience an enhanced probability of inelastic scattering compared to electron motion via a non-resonant state.

The coherent inelastic transmission spectra we have described cause current flowing through the system to *decrease* when real phonons are excited. This behavior is caused by coherent interference between elastic and inelastic scattering.

However, experiments involving inelastic scattering have shown that the excitation of phonons causes the transmission to *increase*. This behavior is seen in both the single molecule experiments [21] and in inelastic electron

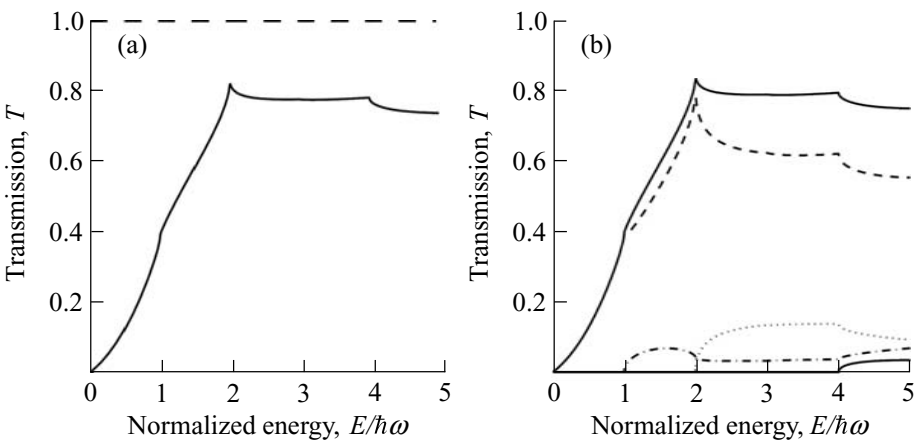

Fig. 3.20. (a) Calculated inelastic (solid) and elastic (dashed) coherent transmission through a dynamic delta barrier. Since there is no static delta, elastic transmission in the absence of phonons is unity. (b) Transmission of total (solid), zero phonon channel (dashed), one phonon channel (dash-dotted) and two phonon channel (dotted), and four phonon channel (lower solid). The three phonon channel cannot be seen on this scale. The background potential is constant, and the model parameters are $m = 0.07 \times m_0$, $\hbar\omega_0 = 36$ meV, $W_0 = 0$ eV nm, $W_1 = 0.6$ eV nm, and $T = 0$ K.

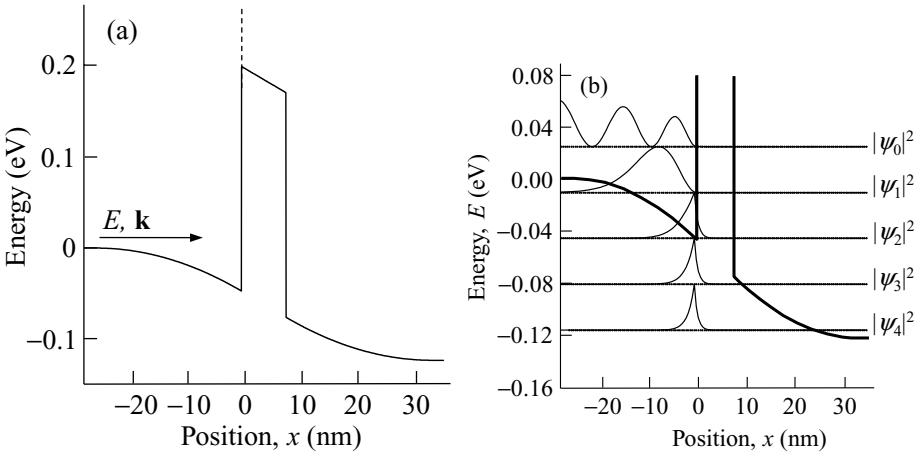

Fig. 3.21. (a) 8 nm thick tunnel barrier of energy 0.25 eV with an Einstein phonon located at $x = 0$ (dotted line) and a 0.125 V bias applied. The electron carrier concentration is 10^{17} cm^{-3} on either side of the undoped barrier. The potential profile is calculated by solving the Poisson equation. (b) Potential (thick solid line) from (a) with normalized elastic and inelastic modulus of wave function squared. The horizontal lines indicate the energy of each channel. The electron effective mass is $m = 0.07 \times m_0$, injected electron energy is $E = 25$ meV, and temperature $T = 0$ K. The incident electron interacts with 36 meV energy Einstein phonons with delta-potential parameters of $W_0 = 1$ eV nm and $W_1 = 0.6$ eV nm at position $x = 0$.

tunneling spectroscopy (IETS) experiments [24–27]. In the experiments, any number of electrons can be supplied by the source so that when an inelastic channel opens more electrons are able to flow through the system. Calculation of current flow in this incoherent semi-classical system typically involves the use of perturbative methods.

A theory capable of properly describing the transition from the coherent quantum transport regime to the incoherent semi-classical behavior has yet to be developed. In any such theory, the role of electron scattering in the electrodes will be critical in distinguishing between the two behaviors.

3.5 Summary

Progress in material science has enabled fabrication of electronic devices with nanoscale and sometimes atomically precise dimensions. Ad hoc designs have exploited these new degrees of freedom to create devices, such as the ballistic electron transistor, that operate using physical processes that *require* nanoscale active regions. An efficient optimal device design strategy in combination with development of realistic physical models holds the

promise of creating a vast number of device configurations with customized functionality. These new configurations are, at least initially, likely to be nonintuitive in nature and so not explored by a conventional ad hoc design methodology. Expanded opportunities in device design exist if better, physically realistic, models of electron transport in nanoscale structures can be developed. Of particular interest is the ability to correctly describe the transition from coherent inelastic electron transport to the incoherent semiclassical regime.

3.6 References

1. A.J. Mieszawska, R. Jalilian, G.U. Sumanasekera, and F.P. Zamborini, *The synthesis and fabrication of one-dimensional nanoscale heterojunctions*, Small **3**, 722–756 (2007).
2. X. Jiang, Q. Xiong, S. Nam, *et al.*, *InAs/InP radial nanowire heterostructures as high electron mobility devices*, Nano Letters **7**, 3214–3218 (2007).
3. P. Mohan, J. Motohisa, and T. Fukui, *Fabrication of InP/InAs/InP core-multishell heterostructure nanowires by selective area metalorganic vapor phase epitaxy*, Applied Physics Letters **88**, 133105 1–3 (2006).
4. L. Pfeiffer, K.W. West, H.L. Stormer, and K.W. Baldwin, *Electron mobilities exceeding* 10^7 $cm^2/V\ s$ *in modulation-doped GaAs*, Applied Physics Letters **55**, 1888–1890 (1989).
5. E.H. Hwang and S. Das Sarma, *Limit to two-dimensional mobility in modulation-doped GaAs quantum structures: How to acheive a mobility of 100 million*, Physical Review B **77**, 235437 1–6 (2008).
6. S. Iijima, *Helical microtubules of graphitic carbon*, Nature **354**, 56–58 (1991).
7. A. Javey, J. Guo, M. Paulsson, *et al.*, *High-field quasiballistic transport in short carbon nanotubes*, Physical Review Letters **92**, 106804 1–4 (2004).
8. J-Y. Park, S. Rosenblatt, Y. Yaish, *et al.*, *Electron–phonon scattering in metallic single-walled carbon nanotubes*, Nano Letters **4**, 517–520 (2004).
9. Z. Yao, C.L. Kane, and C. Dekker, *High-field electrical transport in single-wall carbon nanotubes*, Physical Review Letters **84**, 2941–2944 (2000).
10. S. Ilani, L.A.K. Donev, M. Kindermann, and P.L. McEuen, *Measurement of the quantum capacitance of interacting electrons in carbon nanotubes*, Nature Physics **2**, 687–691 (2006).
11. J. Guo, A. Javey, G. Bosan, and M. Lundstrom, *Assessment of high-frequency performance potential of carbon nanotube transistors*, IEEE Transactions on Nanotechnology **4**, 715–721 (2005).
12. A.F.J. Levi and T.H. Chiu, *Room-temperature operation of hot-electron transistors*, Applied Physics Letters **51**, 984–986 (1987).

13. For example, S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, pp. 333–338, Prentice Hall, Englewood Cliffs, New Jersey, 1989.

14. For example, G. Bastard, *Superlattice band structure in the envelope-function approximation*, Physical Review B **24**, 5693–5697 (1981).

15. E.O. Kane, Basic concepts of tunneling, in *Tunneling Phenomena in Solids*, ed. E. Burstein and S. Lundqvist, pp. 1–11, Plenum Press, New York, New York, 1969.

16. A.F.J. Levi, *Applied Quantum Mechanics*, pp. 171-217, Cambridge University Press, Cambridge, United Kingdom, 2006.

17. S. Luryi, *Frequency limit of double-barrier resonant-tunneling oscillator*, Applied Physics Letters **47**, 490–492 (1985).

18. D.H. Davies, S. Hershfield, P. Hyldgaard, and J.W. Wilkins, *Current and rate equation for resonant tunneling*, Physical Review B **47**, 4603–4618 (1993).

19. S. Braig and K. Flensberg, *Vibrational sidebands and dissipative tunneling in molecular transistors*, Physical Review B **68**, 205324 1–10 (2003).

20. A. Mitra, I. Aleiner, and A.J. Millis, *Phonon effects in molecular transistors: Quantal and classical treatment*, Physical Review B **69**, 245302 1–21 (2004).

21. See, for example, S.W. Wu, G.V. Nazin, X. Chen, X.H. Qiu, and W.Ho, *Control of relative tunneling rates in single molecule bipolar electron transport*, Physical Review Letters **93**, 236802 1–4 (2004).

22. See, for example, H. Park, J. Park, A.K.L. Lim, *et al.*, *Nanomechanical oscillations in a single-C_{60} transistor*, Nature **407**, 57–60 (2000).

23. T.H. Chiu and A.F.J. Levi, *Electron transport in an AlSb/InAs/GaSb tunnel emitter hot-electron transistor*, Applied Physics Letters **55**, 1891–1893 (1989).

24. R.C. Jaklevic and J. Lambe, *Molecular vibration spectra by electron tunneling*, Physical Review Letters **17**, 1139–1140 (1966).

25. P.K. Hansma, *Tunneling Spectroscopy*, Plenum Press, New York, New York, 1982.

26. A.F.J. Levi, W.A. Phillips, and C.J. Adkins, *Phonon structure of amorphous SiOx by inelastic tunnelling spectroscopy*, Solid State Communications **45**, 43–45 (1983).

27. M.C. Payne, A.F.J. Levi, W.A. Phillips, J.C. Inkson, and C.J. Adkins, *Phonon structure of amorphous germanium by inelastic electron tunnelling spectroscopy*, Journal of Physics C **17**, 1643–1653 (1984).

28. B.Y. Gelfand, S. Schmitt-Rink, and A.F.J. Levi, *Tunneling in the presence of phonons: A solvable model*, Physical Review Letters **62**, 1683–1686 (1989).

29. T. Brandes and J. Robinson, *Transmission through a quantum dynamical delta barrier*, Physica Status Solidi B **234**, 378–384 (2002).

30. G. Breit, *Energy dependence of reactions at thresholds*, Physical Review **107**, 1612–1615 (1957).

31. V.J. Goldman, D.C. Tsui, J.E. Cunningham, *Evidence for LO-phonon-emission-assisted tunneling in double-barrier heterostructures*, Physical Review B **36**, 7635–7637 (1987).

4 Aperiodic dielectric design

Philip Seliger

4.1 Introduction

The spatial arrangement of nanoscale dielectric scattering centers embedded in an otherwise uniform medium can strongly influence propagation of an incident electromagnetic (EM) wave. Exploiting this fact, one may iteratively solve a parameter optimization problem [1] to find a spatial arrangement of identical, non-overlapping scattering cylinders so that the scattered EM-wave closely matches a desired target-response.* Of course, the efficiency of adaptive algorithms used to find solutions may become an increasingly critical issue as the number of design parameters such as scattering centers increases. However, even for modest numbers of scattering cylinders this method holds the promise of creating nano-photonic device designs that outperform conventional approaches based on spatially periodic photonic crystal (PC) structures.

The configurations considered here consist of either lossless dielectric rods in air or circular holes in a dielectric similar to the majority of quasi two-dimensional (2D) PCs reported in the literature by, for example, [2, 3]. To confirm the validity of the 2D simulations they are compared to full three-dimensional (3D) simulations as well as measurements.

In Section 4.2 the forward problem and a semi-analytic method to compute the electromagnetic field distribution is introduced. The semi-analytic Fourier-Bessel series solutions and a guided random walk routine can be used for electromagnetic device design and the optimization algorithm is described in Section 4.3. The results of these design problems are presented in Section 4.4. The inefficiencies of the optimization routine in conjunction

* The problem considered here differs somewhat from conventional inverse problems which usually involve calculations on experimentally measured data.

with the particular implementation of the forward solver soon became apparent. The presented types of aperiodic design demand highly efficient forward solvers as well as optimization routines. A second, finite-difference (FD) forward solver modeling a specific experimental setup is introduced in Section 4.6. The advantage of the FD-solver is two-fold. Firstly, the time it takes to compute one EM-field distribution is nearly independent of the number of scattering centers. Secondly, a local optimization method for finding locally optimal design parameters was partially built into the forward solver using the adjoint method, which is described in Section 4.8. Verification of the second forward solver and optimal aperiodic designs are presented in Section 4.9. The performance of other aperiodic structures is also given at the end of Section 4.9.

4.2 Calculation of the scattered field

The optimization process is an iterative procedure, in which the scattered field from a trial configuration of cylindrical rods or circular holes is compared with that of the target or objective function. Computation of the scattered field is commonly referred to as the forward problem. A well-known method to compute 2D EM-field distributions uses Bessel function Fourier series and is introduced first in this section. With the capability to simulate the behavior of a device one may continue to define a measure of device performance for a desired electromagnetic response. Next, an optimization scheme that can be used for the device design is introduced.

The electromagnetic field solver is based on the analytic solution of the Helmholtz equation by separation of variables in polar coordinates. A typical problem is a set of N long, parallel, lossless circular dielectric rods distributed in a uniform medium and illuminated by an electromagnetic wave perpendicular to the axis of the cylinders. The natural geometry of the system and efficiency considerations lead one to use a 2D electromagnetic field solver. When analyzing scattering from one cylinder, the solution, expressed as a Fourier–Bessel series, is found by imposing the continuity of the electric and magnetic field components at the rod surface. However, when studying scattering from two or more cylinders, multiple scattering results in an additional linear system that has to be solved in order to find the Fourier–Bessel coefficients [4]. For a given number of Bessel functions included in the truncated Fourier series, this linear system has a reduced form which is conveniently described using the scattering matrix method [5–9]. The input wave can, in principle, be of arbitrary shape so long as it is expressed as a Fourier–Bessel series. In our simulations the input beam is a gaussian

and both TE and TM polarizations were considered. The calculated Fourier–Bessel coefficients for the gaussian beam conform [10, 11]. Additional details on the electromagnetic solver may be found in the following section.

4.2.1 Fourier–Bessel based electromagnetic solver

For brevity, only TM polarized electromagnetic waves are presented. For the TE case the equations are similar with H replacing E. The total field is written as the sum of the incident field E_{inc} and the field scattered from the N cylinders E_{sc}.

$$E = E_{\text{inc}} + \sum_{i=1}^{N} E_{\text{sc}}^{i}. \tag{4.1}$$

The actual incident field on a cylinder labeled with index j is

$$E_{\text{inc}}^{j} = E_{\text{inc}} + \sum_{i \neq j}^{N} E_{\text{sc}}^{i}, \tag{4.2}$$

[4, 9]. The Helmholtz equation must be solved for the total field, $\nabla^2 E + k^2 E = 0$, where $k = k_0$ in the region outside the cylinders and $k = k_1$ inside the cylinders. A method used to solve this equation is separation of variables in polar coordinates. All field quantities should be expressed in the form of Fourier–Bessel series with the coefficients α_m and β_m determined from the boundary conditions. Hence, $E = \sum_{m=-\infty}^{\infty} \alpha_m Z_m(k\rho) e^{im\theta} + \sum_{m=-\infty}^{\infty} \beta_m \widetilde{Z}_m(k\rho) e^{im\theta}$, where Z_m and $\widetilde{Z}_m$ are two conjugate cylindrical functions. These functions are either the first-order Bessel functions J_m and Y_m or the second-order Bessel functions (Hankel functions) $H_m^{(1)}$ and $H_m^{(2)}$. The valid pair of functions depends on the boundary conditions. Outside the cylinders the asymptotic behavior determines which functions are used. Since only the $H_m^{(2)}$ functions behave as an outward-propagating, cylindrical wave, the field of the scattered wave E_{sc} has to be written using only Hankel functions of the second kind, $H_m^{(2)}$ of the $\{H_m^{(1)}, H_m^{(2)}\}$ pair. Hence the scattered field is $E_{\text{sc}} = \sum_{m=-\infty}^{\infty} b_m H_m^{(2)}(k_0\rho) e^{im\theta}$. Inside the cylinders the $\{J_m, Y_m\}$ pair of functions is chosen because both Hankel functions are singular at the origin and are therefore unphysical. The singularities stem from the Y_m part of the Hankel function ($H_m^{(1)} = J_m + iY_m$, $H_m^{(2)} = J_m - iY_m$) and thus only the J_m functions are kept for the Fourier series inside a cylinder. In this case the total internal field is $E_{\text{tot}}^{\text{int}} = \sum_{m=-\infty}^{\infty} a_m J_m(k_1\rho) e^{im\theta}$.

After expressing the incident field as a Fourier–Bessel series in polar coordinates, the electric and magnetic field continuity conditions must be satisfied on the boundary of each cylinder. These boundary conditions yield

a system of equations for the unknown Fourier–Bessel coefficients of the scattered field outside the cylinders and total field inside the cylinders. The linear system of equations can be simplified by using the relationship between the Fourier–Bessel coefficients of a field incident on a cylinder and the coefficients for the scattered and internal fields [5–8].

4.3 Optimization

As a measure of the quality of a given design a cost function also referred to as an error is defined. The cost function measures the difference between the desired Poynting vector and simulated Poynting vector. The cost function is defined by

$$D^* = \sum_{i=0}^{N_p} \left| \frac{S_n(\alpha_i)}{N_s} - \frac{T(\alpha_i)}{N_t} \right|^\gamma , \tag{4.3}$$

and discussed in detail in Section 4.3.1. The real part of the simulated Poynting vector $S_n(\alpha)$ normal to a given observation line at angle α is compared to the desired value $T(\alpha)$. Note that N_s and N_t are appropriate normalization values. A simple yet effective optimization method is the guided random walk. The positions of individual cylinders are randomly changed by a small amount, the scattered fields in the modified configuration are calculated. If the resulting simulated Poynting vectors are closer to the target values the difference Eq. (4.3) is decreased and the new configuration is accepted, otherwise the modified configuration is rejected. Next, another random change is attempted unless the optimization is terminated. The implementation of the adaptive algorithm also includes various types of collective motion such as moving more than one cylinder per iteration, and moving or rotating all the cylinders.

It is observed that for starting configurations in which cylinders are randomly positioned the convergence rate towards the target is approximately the same. On the other hand, if one starts with a quasi-optimal configuration the number of iterations needed can be small.

In general the cost function Eq. (4.3) may be computed for any desired function of the electromagnetic field distribution that can be expressed as an objective or target function T. A typical target function may involve redirecting and reshaping the input beam. However, more complex systems involving mode converters could also be considered.

The adaptive algorithm described above always results in an electromagnetic field distribution that approximates the target function. The closeness

of an achieved design to the desired field distribution depends on the number of scatterers N as well as their refractive index. The reason for this may be found by considering an analogy with the multipole expansion of fields where using higher order multipole moments assures a better approximation at the expense of increased computational effort.

4.3.1 Cost function

Optimization is based on the minimization of the functional Eq. (4.3). This is defined as the residual error between the calculated angular distribution of the normal component of the Poynting vector $\mathbf{S}$ and a distribution expressed as a target function T. The cost or error function is computed starting from the difference in intensity between the target and the result. The cost function is calculated along an observation line (often a circle around the group of cylinders). This line is divided into small portions and the normal component of the $\mathbf{S}$ vector is calculated in the center of each segment.

Consider the target function $T(\alpha)$ to be the angular distribution of intensity exiting the circular observation region and $S_{\mathrm{n}}(\alpha)$ the simulated normal component of the real part of the Poynting vector $S_{\mathrm{n}}(\alpha) = \mathbf{S}(\alpha) \cdot \mathbf{n}(\alpha)$, where $\mathbf{n}$ is the unit vector normal to the observation surface. In this space of functions defined on $\alpha \in [0°, 360°]$ and having real values a difference D between the modeled $S_{\mathrm{n}}(\alpha)$ and the desired target $T(\alpha)$ is defined by

$$D = \frac{1}{2\pi} \int_0^{2\pi} |S_{\mathrm{n}}(\alpha) - T(\alpha)|^{\gamma} \, d\alpha. \tag{4.4}$$

To properly evaluate the difference between the target and simulation the functions T and S must be similarly normalized.

In general, the exponent γ in Eq. (4.4) can take any value. Choosing $\gamma = 1$ assures that each improvement is considered with the same weight. When choosing $\gamma > 1$, improvements made in regions where the target and the results are very different influence the integral D more than a few smaller improvements in other regions. This means that $\gamma = 1$ tends to ensure a uniform convergence while $\gamma > 1$ favors reduction of major differences between target and result. The greater the numerical value of γ, the more important this effect becomes, while for $0 < \gamma < 1$ the effect is reversed. And finally, a negative exponent γ tends to push the solution further away from the target, in a manner which depends on the numerical value of γ. The standard choice is $\gamma = 2$ yielding the D to equal the L_2-norm. When $\gamma = 1$ Eq. (4.4) is equal to the L_1-norm. This compromises the differentiability of the cost functional but can improve the convergence of the optimization algorithm in particular near a locally optimal design. If $\gamma(\alpha)$ is angle dependent the

cost function can be used to weight the error control differently with α. If the derivative is used in the optimization scheme, $\gamma > 1$ is recommended to maintain differentiability.

One can use different exponents for different stages of the iterative process. For example $\gamma = 1$ could be used at the beginning of the convergence procedure to avoid local minima. Later, the value of γ could be increased to accelerate convergence towards a minimum. Furthermore, if one decides that this minimum is not sufficiently close to the target function, application of a negative exponent would repel the iterations from this local minimum to some intermediate point where $\gamma > 1$ iterations can be restarted in the search for a better local minimum.

To compute $S_n(\alpha)$ the real part of the Poynting vector is used. For most scattering directions, $S_n(\alpha)$ is computed using the total field. The exception is the region where the input beam enters the scattering region. In the first example (the top hat target function) only the scattered field is used to compute $S_n(\alpha)$. In the second example (the $\cos^2$ target function) the difference between the Poynting vector of the incident field and the Poynting vector of the total field is used. The latter is the general expression for the backscattered power, which is valid even when evanescent fields or lossy media are present. When the medium is not lossy and the observation contour is far enough from the evanescent near-field the two approaches give identical results.

The differences D^* are computed numerically by dividing the observation line into small and equal subintervals and the integral is replaced by a sum over these portions as shown in Eq. (4.3). Different normalization methods can be used such as normalization to the maximum value (which is nonlinear), normalization to the sum of all values, or normalization to the sum of the squares. Since in this case the functions involve intensity, it is preferred to normalize the sum, *i.e.* normalize the total power.

It is worth mentioning that when optimizing for a modal shape distribution both amplitude and phase can be controlled in a similar fashion. A better choice for cost function would be the overlap integral between the actual field distribution and the target modal field distribution. In this case the fitness function would be maximized.

4.4 Results

To illustrate the design approach, consider an input beam of gaussian profile scattered by an angle of 45°. The scattering angle is defined with respect to the original direction of propagation of the wave, so that backscatter

corresponds to a scattering angle of 180°. The target functions that are considered are a top hat distribution of the optical intensity with respect to the scattering angle and a cosine squared ($\cos^2$) distribution of the intensity. We note that a modal field distribution target function requires amplitude as well as phase to be specified.

As an initial demonstration the top hat intensity function is chosen as a target because it is difficult to achieve (even approximately) in conventional optical systems and serves as a good test for the optimal design approach. Applications of such an intensity distribution include guaranteeing the uniform illumination of the active area of a photodetector. On the other hand, the $\cos^2$ target intensity distribution approximates the transverse spatial mode intensity typically found in a waveguide and so could be considered a step in the design of a waveguide coupler.

4.4.1 Top hat objective function

Optimization of a top hat intensity distribution target function is performed starting from a configuration of $N = 56$ dielectric rods (represented in Fig. 4.1(a) by the small circles), each having refractive index $n_r = 1.5$, and diameter $d = 0.4$ μm. The medium surrounding the rods is air and the structure is illuminated by a TM polarized (electric field along the z direction) gaussian beam of width $2\sigma = 4$ μm, wavelength $\lambda = 1$ μm, and propagating along the positive x direction (from left to right in Fig. 4.1(a)). The initial configuration of the rods and intensity distribution are illustrated in Fig. 4.1(a), where the arrows represent (in arbitrary units) the real part of the Poynting vectors. The target function window (represented in Fig. 4.1(a) by a missing arc in the 7 μm radius observation circle) extends from 30° to 60°. Figure 4.1(b) shows normalized intensity (real part of the normal component of total field Poynting vectors directed outwards) as a function of angle on a radius of 7 μm from the center of the symmetric array. Clearly, for the initial configuration, the overlap with the top hat target function (broken line) is poor.

Figure 4.2(a) shows the spatial distribution of the $N = 56$ rods with the real part of Poynting vectors after 9,700 iterations of the adaptive search algorithm. Figure 4.2(b) shows the corresponding angular distribution of the intensity. In Fig. 4.2(c) the distribution of the electric field (relative magnitude, with 1 corresponding to the maximum magnitude in the incident gaussian beam) is displayed and Fig. 4.2(d) shows the relative error versus the number of iterations (the errors are normalized with respect to the initial value). Note that for this system with $2 \times N$ positional degrees of freedom, the error is not saturated even after 9,700 iterations (Fig. 4.2(d)) and the

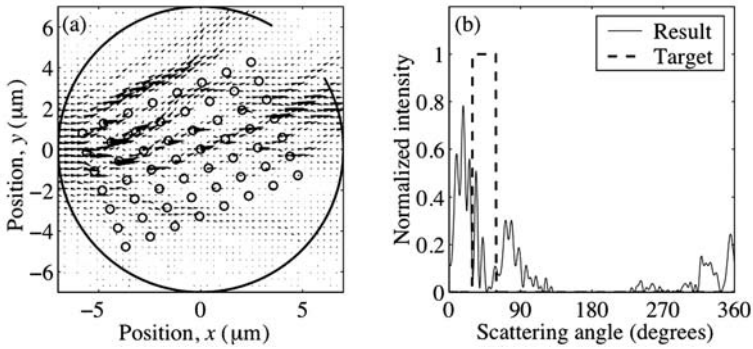

Fig. 4.1. Starting configuration for the top hat target function example. 56 dielectric rods with diameter $d = 0.4$ μm, index $n_r = 1.5$, in air. The incident beam, which propagates along the x axis (left to right), is a TM polarized gaussian beam with beamwidth $2\sigma = 4$ μm, and wavelength $\lambda = 1$ μm. (a) Positions of the cylinders and distribution of the Poynting vector field. The observation circle has a radius of 7 μm and it is represented without the target window between 30° and 60°. (b) Computed angular intensity distribution (continuous line) at radius 7 μm compared with the target distribution (dashed line).

structure can be further optimized by performing additional iterations. The error is computed using the method described in Section 4.3.1 with exponent $\gamma = 2$ on the observation circle of radius 7 μm.

For comparison, Fig. 4.3 shows the results of using a PC with the same number of dielectric rods, $N = 56$. Clearly, the number of rods is not sufficient to redirect the entire beam in the 45° (±15°) direction. Much of the scattered field falls outside the target area, and the error with respect to the target function is unacceptably large. The spatial symmetry of the PC excludes the realization of something similar to the top hat target function. Only by breaking this symmetry may one come close to the desired device response. In general, broken symmetry enables and increases functionality.

4.4.2 Cosine squared objective function

For a $\cos^2$ objective or target function a different setup is used. In this case $N = 26$ cylinders with lower refractive index (SiO_2, $n_r = 1.45$) are embedded in a higher refractive index material (Si, $n_r = 3.5$). The 3D equivalent of the modeled situation is a Si slab with cylindrical perforations embedded in SiO_2. The incident wave is a TE polarized (magnetic field along the z direction) gaussian beam of width $2\sigma = 1.5$ μm and wavelength $\lambda = 1.5$ μm. In a

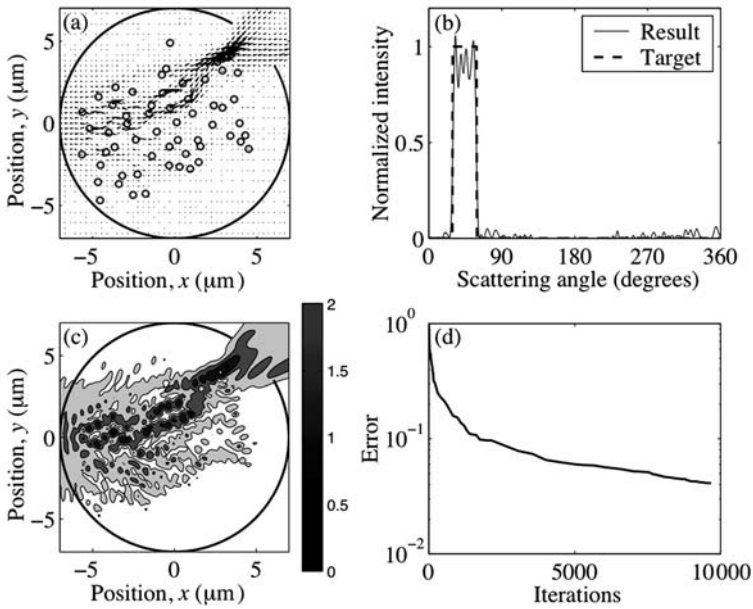

Fig. 4.2. Optimized configuration for the top hat function example. 56 dielectric rods with diameter, $d = 0.4$ μm, index $n_r = 1.5$, in air. TM polarization gaussian beam with incident beamwidth $2\sigma = 4$ μm, and wavelength $\lambda = 1$ μm. (a) Positions of the cylinders and distribution of the Poynting vector field after 9,700 iterations. The observation circle has a radius of 7 μm and it is represented without the target window between 30° and 60°. (b) Computed angular intensity distribution after 9,700 iterations (continuous line) at radius 7 μm compared with the target distribution (dashed line). (c) Contour plot of the electric field magnitude in relative units after 9,700 iterations. (d) Evolution of the error.

similar manner to the previous example, the initial configuration and intensity distribution are illustrated in Fig. 4.4. Figure 4.5 shows the optimized configuration, Poynting vectors, relative error versus number of iterations, and relative magnitude of the magnetic field (with 1 corresponding to the maximum magnitude in the original gaussian beam). This time the diameter d can vary in addition to the positions of the $N = 26$ cylinders. This design therefore has $3N$ degrees of freedom. The values of the diameters are constrained to the range $0.2 \leq d \leq 0.5$ μm. The error function D is computed using the metric from Section 4.3.1 with exponent $\gamma = 1$ on a 6 μm radius circle. Notice that due to the small number of cylinders the actual number of degrees of freedom is smaller (78) relative to the previous calculations with a top hat target function (112) and the optimization saturates after a relatively small number of iterations.

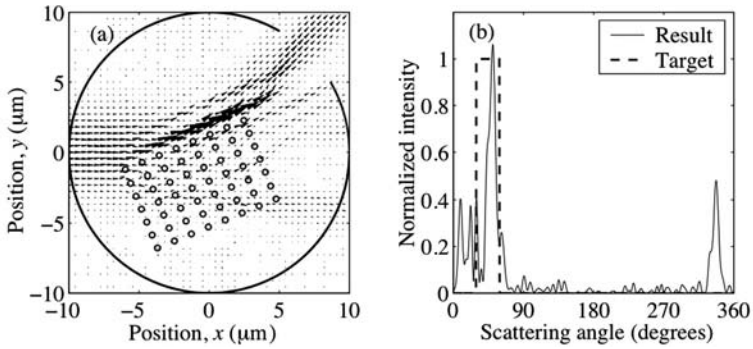

Fig. 4.3. Comparison with a periodic structure (PC) with 56 dielectric rods ($n_r = 1.5$) in air. The lattice constant and the angle are chosen so that for the incident wavelength $\lambda = 1\,\mu m$ the Bragg diffraction condition is satisfied for 45°. (a) Positions of the cylinders and distribution of the Poynting vector field. The incident beam is propagating along the x axis (left to right). The observation circle has a radius of 7 μm and it is represented without the target window between 30° and 60°. (b) Computed angular intensity distribution (continuous line) at radius 7 μm compared with the target distribution (dashed line).

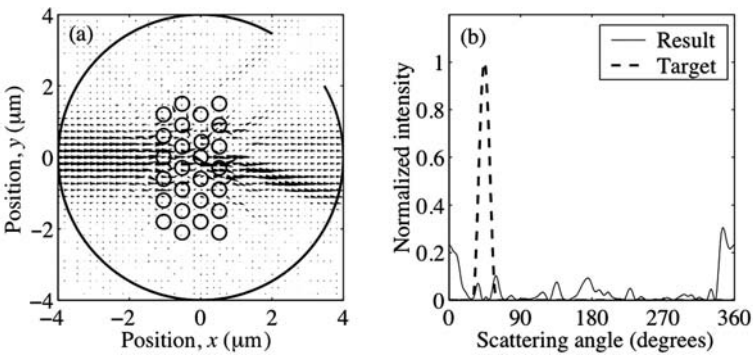

Fig. 4.4. Initial configuration for the cosine squared target function example. Si ($n_r = 3.5$) with 26 cylindrical holes filled with SiO_2, TM gaussian beam incidence, beamwidth $2\sigma = 1.5$ μm, and wavelength $\lambda = 1.5$ μm. (a) Positions of the cylinders and distribution of the Poynting vector field. The incident beam is propagating along the x axis (left to right). The observation circle has a radius of 6 μm and it is represented without the target window between 30° and 60°. (b) Computed angular intensity distribution (continuous line) at radius 6 μm compared with the target distribution (dashed line).

4.4.3 Computing resources

Optimization work on 1D problems [12] requires relatively insignificant computational resources. Optimization algorithms using 2D electromagnetic

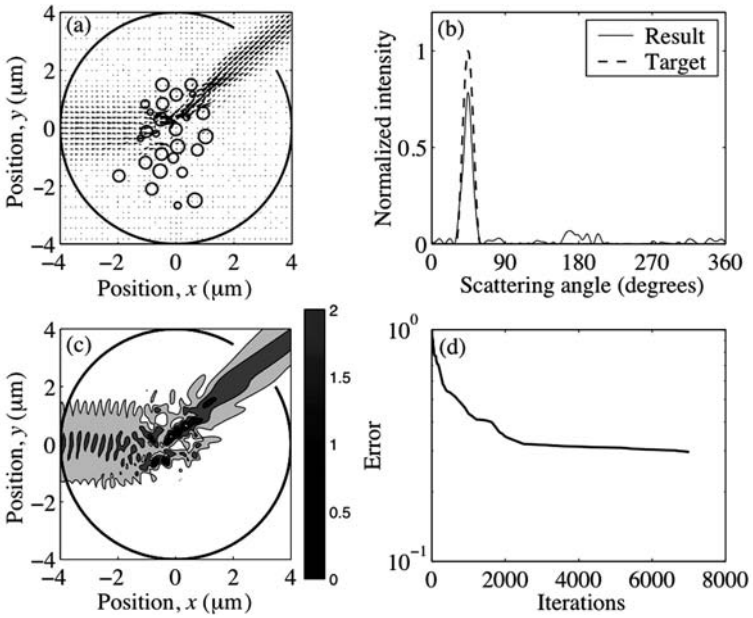

Fig. 4.5. Optimized configuration for the cosine squared target function example. Si ($n_r = 3.5$) having 26 cylindrical holes filled with SiO_2, TM gaussian beam incidence, beamwidth $2\sigma = 1.5$ μm, and wavelength $\lambda = 1.5$ μm. (a) Positions of the cylinders and distribution of the Poynting vector field. The incident beam is propagating along the x axis (left to right) after 7,000 iterations. The observation circle has a radius of 6 μm and it is represented without the target window between 30° and 60°. (b) Computed angular intensity distribution after 7,000 iterations (continuous line) at radius 6 μm compared with the target distribution (dashed line). (c) Contour plot of the magnetic field magnitude in relative units after 7,000 iterations. (d) Evolution of the error as a function of iteration number.

solvers are computationally more intensive than those for 1D structures. Realistic simulations in 3D are even more compute intensive and often require parallel computing. Each of the simulations (top hat and $\cos^2$ target) discussed here took 4 days to complete (9,700 iterations in the first case and 7,600 in the second) using a Pentium IV processor with a 3 GHz clock frequency, 533 MHz memory bus, and 1 GB RDRAM.

The compute time for the forward problem solver is dominated by the solution of a linear system with a full matrix of complex numbers and thus is strongly dependent on the number of cylinders N in the design and the number of Bessel function terms N_b retained in the Bessel–Fourier series. The size of the system is $N_m = N(2N_b + 1)$ and the solver routine (iteratively refined LU decomposition [13]) time is proportional to the cube of the matrix size ($O(N_m^3)$). When using an incident plane wave or a cylindrical wave the number of Bessel functions needed is very small, however, an appropriately

accurate approximation of the gaussian beam shape requires a large number of Bessel function terms in the series expression causing a corresponding increase in compute time.

4.4.4 Comparison with 3D Simulation

A comparison with 3D electromagnetic simulations serves to confirm the accuracy of solutions obtained with the 2D simulator. Also, because 3D simulations are much more time consuming than 2D computations, one might adopt the 2D optimization result as a starting point for a limited number of 3D iterations which are then used to refine the optimal design.

Comparisons were made between our 2D electromagnetic simulations and 3D Finite Integration Technique (FIT) simulations obtained using a commercially available package, CST Microwave Studio [14].

In a realistic structure one might anticipate the infinitely long cylindrical hole structure used in the $\cos^2$ target simulation to be replaced with SiO_2 filled holes in a Si slab itself embedded in SiO_2. The input beam might be launched into this slab from a ridge waveguide. In our simulations, the wavelength of the light is $\lambda = 1.5$ μm and the polarization is TE. The Si slab is 0.6 μm thick and the effective index of the fundamental mode of this slab waveguide is the same as the index of the material surrounding the cylinders in the 2D simulation. The diameter of the holes is $d = 0.4$ μm and the mode size of the ridge waveguide (Fig. 4.6) is approximately 1.5 μm and so the same as the width of the gaussian beam used for the 2D simulations. The differences between the two simulations (Fig. 4.6(b) for the 3D and Fig. 4.6(d) for the 2D simulation results) are primarily due to the nonuniformity of the field and structure in the z direction but also because of the difference between the gaussian beam and the ridge mode and the fundamentally 3D discontinuity between the ridge waveguide and the slab. These differences translate into a 30% decrease in peak intensity and 25% decrease in the total power directed in the desired direction for the 3D simulation compared to the 2D computation (Fig. 4.6(c)).

4.4.5 Sensitivity analysis

The influence of small changes in the wavelength of the incident beam on error is analyzed. Using the optimized configuration from the $\cos^2$ target example (Fig. 4.5(a)) the effect of changing the frequency of the input light is simulated. The relative deviation from the minimum error function value is plotted in Fig. 4.7 versus the electromagnetic wave frequency shift. Frequency variations of $\Delta f = 200$ GHz only change the error by 0.3% of its

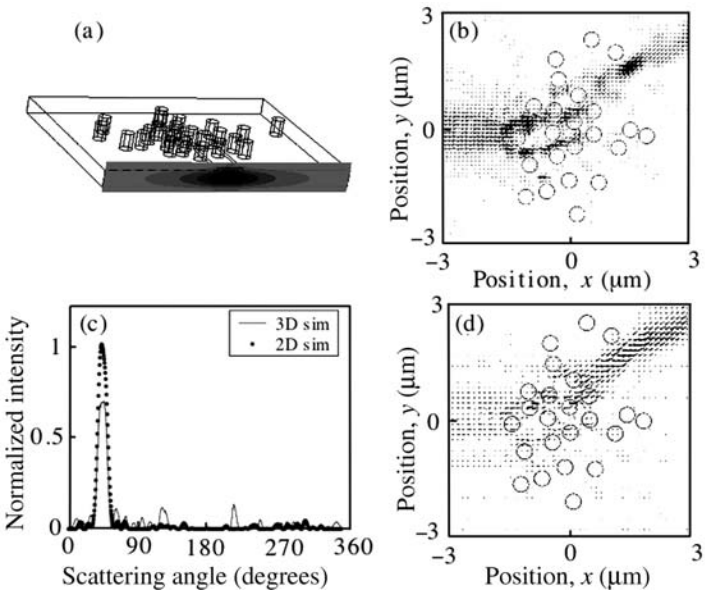

Fig. 4.6. Comparison between 3D and 2D simulations. (a) The simulated 3D structure and the modal field of the input ridge waveguide. The thickness of the Si slab in the 3D simulation is 0.6 μm and the horizontal mode size of the ridge waveguide is approximately $w = 1.5$ μm; the diameter of the holes is $d = 0.4$ μm. The incident light is TE polarized and has wavelength $\lambda = 1.5$ μm. (b) 3D Simulation results showing the Poynting vector field in the middle, horizontal plane of the Si slab. (c) Comparison between the angular intensity distribution for the 3D and 2D simulations. (d) 2D simulation results of the Poynting vector field.

minimum value. Thus an optical beam centered at wavelength $\lambda = 1.5$ μm modulated at very high speed will behave essentially as the simulated monochromatic wave at $\lambda = 1.5$ μm (200 THz). Even a 1 THz deviation in frequency changes the error by only 6–7%.

Another important issue is the sensitivity of the aperiodic nano-photonic design to slight misplacements and variations introduced by the fabrication process. As an initial study the sensitivity of the design to changes in the position of the cylinders is explored. Other parameters that are influenced by fabrication processes such as the diameter and detailed shape of the cylinders, index of refraction, direction of the input ridge waveguide with respect to the cylinders were not analyzed.

To estimate the sensitivity to position, for the same $\cos^2$ target function example, the following method is used. First, one by one the cylinders are displaced by a small fixed distance Δ in the positive and negative x and y directions and the change in error is evaluated for each cylinder (the maximum error created out of the four displacements). The displacement values are

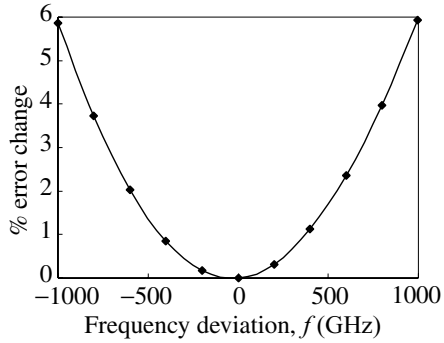

Fig. 4.7. Sensitivity analysis with respect to the frequency of the incident light for the optimized structure from Fig. 4.5. The percentage change in error function is shown when the frequency of the light is modified from the original value $f_0 = 200$ THz .

small compared to the initial diameters of the cylinders ($d = 400$ nm). Five values for the displacement Δ are considered: 10, 20, 30, 40, and 50 nm. Next the ten displaced cylinders with the greatest influence on the error function are selected. As expected, the most influential cylinders are located in regions associated with high field intensity.

The selected cylinders are all individually moved randomly by the same step size in the positive and negative x and y directions (each movement has the same 1/4 probability). The movement of each cylinder is independent of the movements of the other selected cylinders. This way a number of perturbed configurations can be generated. The error function was evaluated for ten of these modified configurations and Fig. 4.8 illustrates the dependence on the size of the displacement.

This very simplified method for estimating the sensitivity with respect to position is chosen because a more complete approach would involve independent displacements of each cylinder in random directions and with variable distances and hence give rise to significant computational effort. The plot in Fig. 4.8 suggests that 10 nm precision in fabrication may be needed to ensure less than 10% decrease in performance for devices operating at wavelength $\lambda = 1.5$ μm.

The designs that have been presented so far utilize a common scattering element. The result is an impressive variety of possible configurations and capabilities. Meanwhile, progress has been made in manufacturing these devices. Volk *et al.* [15] present the implementation of an aperiodic structure at the nm scale using lithographic fabrication methods. The aperiodic design uses dielectric, cylindrically shaped nano-pillars that operate as a lens. The lens is used to focus polarized laser light. Figure 4.9 shows scanning electron

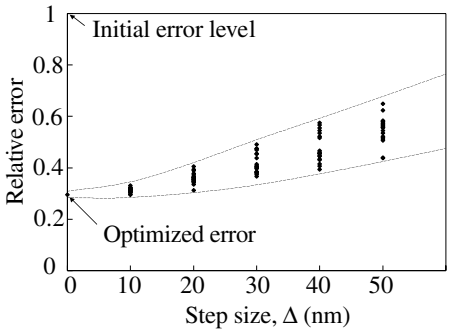

Fig. 4.8. Sensitivity analysis of the positioning of the scattering cylinders for the optimized structure from Fig. 4.5. The ten most sensitive cylinders are randomly moved along the x and y axis by a step of various lengths, ± 10, ± 20, ± 30, ± 40 and ± 50 nm shown on the horizontal axis. A sample of ten error functions for different randomly perturbed cylinder configurations are plotted for each step size.

microscope images of the nano-photonic device.

The necessity of improving the materials as well as the manufacturing techniques on the nm scale is apparent and is discussed in [15]. Nevertheless, the algorithms for optimizing such structures deserve equal attention.

In Sections 4.5 through 4.8 the focus is on the choice of forward solver and algorithm that is used to find feasible aperiodic designs. Experiments testing the optimized aperiodic designs are described in Section 4.9. Because the Helmholtz equation is linear the devices were built on a macroscopic scale and the measurements were performed using electromagnetic millimeter waves. In this manner the cost and complications of lithography are avoided.

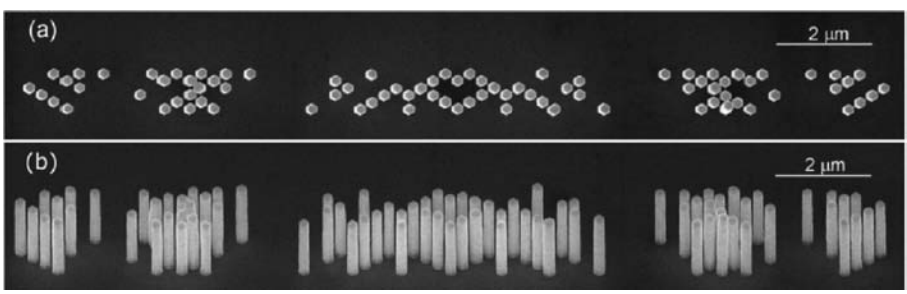

Fig. 4.9. (a) Top and perspective view (b) scanning electron microscopic (SEM) images taken on the homoepitaxial CnO nano-pillar array. The highly uniform 2.1 μm long prism shaped vertical nanocrystals are collectively aligned according to the crystal orientation of the wurtzite-type ZnO substrate [15]. The existence of these planar manufacturing techniques highlights the importance of efficiently producing feasible designs that can be implemented and tested.

4.5 Efficient local optimization using the adjoint method

The study of photonic crystals (PCs) [16] is inspired, in part, by a desire to seek new compact designs for optical and RF components. Much work has focused on two-dimensional (2D) periodic dielectric structures due to availability of planar fabrication techniques. However, there are a number of fundamental issues that appear to be impediments to adoption of PCs as a technology. These include the fact that the inherent spatial periodicity of the PC structure results in limited functionality. Often one must break spatial symmetry to obtain a useful device response. For example, waveguides are typically created by introducing a line defect and filters might make use of one or more point defects. Hence, one may make the observation that usually a desired functionality requires breaking the underlying spatial symmetry of the periodic dielectric structure. Even in situations where one wishes to access properties intrinsic to periodic dielectrics such as nonlinear dispersion, coupling electromagnetic radiation from free-space and finite-size effects presents significant challenges [17].

One approach that attempts to circumvent such difficulties is application of optimization techniques to PCs [18–21]. On the one hand, such numerical studies are usually limited to a finite number of identical dielectric scatterers whose broken symmetry spatial distribution is restricted to periodic PC lattice positions. On the other hand, they benefit from the fact that a less biased search of solution space can result in nonintuitive optimized designs. Our initial approach [22] has been to retain identical dielectric scatterers but to remove all bias to periodic PC inspired designs. In this way adaptive algorithms can seek optimal solutions in a much larger space of aperiodic dielectric structures and hence, at least in principle, access a larger range of functionalities.

To handle the large number of possible aperiodic configurations, efficient global optimization is a necessity. After demonstrating the design by simulation, laboratory measurements using mm-wave electromagnetic (EM) radiation at frequency $f_0 = 37.5$ GHz corresponding to free-space wavelength $\lambda_0 = 8$ mm were used to demonstrate an aperiodic test design. To efficiently simulate aperiodic designs an alternative implementation of the forward solver using finite difference equations is introduced. The FD method is sufficiently accurate for the purpose of the test design. In the case of the distributed parameter system that is introduced in Section 4.6, the time it takes to complete the forward solver once does not depend on the number of design parameters nor the particular configuration. This must be viewed

in contrast to the Fourier–Bessel series expansions discussed in Section 4.2.1 where the computation increased dramatically with each additional scattering cylinder. Naturally, finite difference solvers have their limitations, such as a limited size of problems that a model can handle accurately and a lack of analytical results.

4.6 Finite difference frequency domain electromagnetic solver

Note that the objective or target response is in general a function of the EM field. For the following variation of the prototype problem the objective is the relative EM power distribution along a measurement curve. Even though the objective response along the measurement curve is specified in a limited region of the device, the forward problem is solved over the entire modeling domain when using the finite difference solver. Efficiency considerations led to the implementation of a 2D finite difference EM field solver.

As mentioned previously, electromagnetic wave scattering from nonmagnetic, lossless, dielectric is determined by the Helmholtz equation. Because the Helmholtz equation is linear the device behavior scales with frequency and hence may be applied to the design of both RF and nanophotonic devices. For reasons of cost and ease of implementation the experiments were performed at mm-wave frequencies using a configuration approximating a 2D geometry. Figure 4.10(a) shows the basic experimental arrangement in which a $f_0 = 37.5$ GHz RF signal is introduced into a waveguide whose 7 mm $\times$ 3.5 mm aperture is attached to a metal horn. The EM power distribution is detected using a probe that can move to angle θ on a circular path with radius $r_s = 60$ mm. This defines the measurement curve, s. To maintain the 2D nature of the EM experiment, the total structure is sandwiched between two metal plates separated by 3.5 mm $< \lambda_0/2$. The prototype problem introduced in Section 4.4 is modified slightly to model the experimental setup. The modified objective response seeks to scatter incident EM radiation into a top hat function with the peak on the measurement curve occurring over the angular range from 30° to 60° with respect to the incident direction of propagation.

The forward problem simulates the propagation of the EM wave over the domain illustrated in Fig. 4.10(b). This EM scattering domain contains the dielectric scatterers whose spatial arrangement is optimized. Since stationary solutions at a single frequency are sought, Maxwell's equations in the frequency domain are considered and therefore the time variable can be eliminated. When only dielectric material is present, and no induced or other currents flow, Maxwell's equations simplify further. The magnetic field can

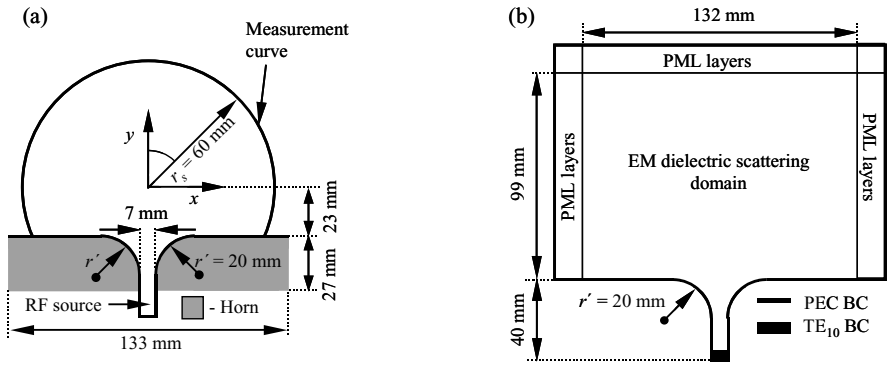

Fig. 4.10. (a) Top view of experimental layout showing physical dimensions. EM power from an RF source of frequency $f_0 = 37.5$ GHz is fed via a waveguide and horn. The measurement curve is indicated. In the experiments, EM power is measured as a function of angle θ on this curve. (b) Domain decomposition of the finite difference (FD) EM simulation. PML layers of finite thickness are truncated with a PEC boundary condition. The waveguide introduces the EM beam as a fixed TE10 mode EM field on the indicated boundary.

be eliminated to yield

$$\nabla \times \left(\mu_{\mathrm{r}}{}^{-1} \nabla \times \mathbf{E}\right) - \omega^2 \epsilon_0 \mu_0 \epsilon_{\mathrm{r}} \mathbf{E} = -\mathbf{i}\omega\mu_{\mathbf{0}}\mathbf{J}, \tag{4.5}$$

where E is the electric field, ϵ_0 is the permittivity of free-space, ϵ_{r} is the relative permittivity, μ_0 is the permeability of free-space, and μ_{r} is the relative permeability. One boundary condition (BC) is that inside metal $\mathbf{E} = \mathbf{0}$, so metal is treated as a perfect electric conductor (PEC). For waves leaving the scattering domain that are not bounded by metal it is required that $\mathbf{E} \to \mathbf{0}$ with distance to the wave source. The decay of E must conform with the radiation condition [23]. The source of the EM radiation is modeled as the excited TE10 mode of the waveguide. This is modeled using a Dirichlet BC in the finite difference implementation (TE10 BC in Fig. 4.10(b)). In the described scenario only the z-component of electric field propagates in the slab waveguide and Eq. (4.5) reduces to the scalar partial differential equation

$$\left(\frac{\partial}{\partial x}\left(\frac{1}{\mu_{\mathrm{r},y}}\frac{\partial}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{1}{\mu_{\mathrm{r},x}}\frac{\partial}{\partial y}\right) + \omega^2 \epsilon_0 \mu_0 \epsilon_{\mathrm{r}}(x,y)\right) E_z = i\omega\mu_0 J_z = 0, \tag{4.6}$$

where $J_z = 0$ because zero currents are assumed, induced or otherwise.

The partial differential equation (PDE) Eq. (4.6) may be used to describe scattering of an EM wave by dielectric cylinders that are represented by a relative permittivity distribution $\epsilon_{\mathrm{r}}(x, y)$. Consider a 2D configuration of Teflon scattering cylinders, each with a diameter 3.175 ± 0.025 mm. Teflon

is modeled in Eq. (4.6) as a lossless dielectric with a real valued relative permittivity $\epsilon_r = 2.05$ and relative permeability $\mu_r = 1$. In this situation Eq. (4.6) is identical to the scalar Helmholtz equation over the EM dielectric scattering domain indicated in Fig. 4.10(b).

The EM waves can propagate into open space beyond the EM dielectric scattering domain. Since the total energy is finite, the magnitude of EM waves must tend to zero as the propagation distance tends to infinity. Simulation of the open space by a bounded domain is achieved using perfectly matched layers (PML) consisting of an artificial material with varying complex permeability and permittivity in Eq. (4.6) [24, 25]. The PMLs are designed to simulate a perfectly absorbing medium.

Numerically, the scattering problem is solved by approximating the PDE in Eq. (4.6) with the finite difference equation

$$\begin{aligned}\frac{1}{\Delta x}\left(\frac{1}{\mu^y_{i+\frac{1}{2},j}}\frac{E_{i+1,j}-E_{i,j}}{\Delta x}-\frac{1}{\mu^y_{i-\frac{1}{2},j}}\frac{E_{i,j}-E_{i-1,j}}{\Delta x}\right)\cdots\\ +\frac{1}{\Delta y}\left(\frac{1}{\mu^x_{i,j+\frac{1}{2}}}\frac{E_{i,j+1}-E_{i,j}}{\Delta y}-\frac{1}{\mu^x_{i,j-\frac{1}{2}}}\frac{E_{i,j}-E_{i,j-1}}{\Delta y}\right)\cdots\\ +\omega^2\epsilon_{i,j}E_{i,j}=0,\end{aligned} \tag{4.7}$$

for the electric field $E_{i,j}$ at grid point (x_j, y_i). Note that $\epsilon_{i,j}$ is the approximated value of the permittivity, $\mu_{i\pm1/2,j\pm1/2}$ is the permeability on a staggered grid and (x_j, y_i) are evenly spaced square-grid points with step size $\Delta x = \Delta y \leq \lambda_0/20$. Equation (4.7) is satisfied at all interior points of the scattering domain including the PML layers. Dirichlet boundary conditions are satisfied at boundary points as shown in Fig. 4.10(b).

When values are ordered into a column vector, Eq. (4.7) can be written in matrix notation as:

$$\mathbf{L}\cdot E = b_{\mathrm{BC}}, \tag{4.8}$$

where $\mathbf{L}$ is the complex-valued sparse FD matrix including the formulation of the PML absorbing boundary, $E_{i,j}$ are the unknown complex electric field values at the FD grid points (x_i, y_j), and b_{BC} contains the boundary conditions of the FD equation.

The power of EM waves at any point on the measurement curve can be derived from the solution of Eq. (4.8). More precisely, the power of EM waves is first computed at all grid points. The power along the measurement curve is then obtained by interpolation. For a point (x, y) on the measurement curve defined by angle θ, the nearest four grid points are denoted by $(x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2)$ and the modeled EM power is approximated

by

$$v(\theta) = v(x, y) = \Sigma_{i=1,j=1}^{2} \delta_{i,j} v_{i,j} / \Sigma_{k=1,l=1}^{2} \delta_{k,l}, \tag{4.9}$$

where $v_{i,j} = E^*(x_i, y_j)E(x_i, y_j)/2$. The gaussian weights at position (x, y) are given by

$$\delta_{i,j} = \exp\left[-\left(\frac{(x - x_i)^2}{\Delta x}\right) - \left(\frac{(y - y_i)^2}{\Delta y}\right)\right], \tag{4.10}$$

where $i = 1, 2, j = 1, 2$, and $\Delta x, \Delta y$ is the size of the FD grid in the x and y direction, respectively. Choosing the commonly used gaussian weights is a matter of convenience and is not based on electromagnetic theory or analysis. Were it a primary source of inaccuracy in the model it would require modification. A sparse projection row vector $W(\theta)$ with only four non-zero entries that are associated with neighboring grid points yields the power projected on to the target function at scattering angle θ

$$v(\theta) = W(\theta) \cdot \mathrm{diag}(E^*) \cdot E/2. \tag{4.11}$$

Comparison of calculated EM power profile along the measurement curve with experimental results serves to verify the proper operation and test the accuracy of the FD method.

4.7 Cost functional

As a measure of the fitness of a given configuration with respect to the design objectives a scalar cost functional J is defined, similar to the error function stated in Section 4.3.1. The cost functional is

$$J(p) = \sum_{i=1}^{M} \Delta\theta \; w^2(\theta_i) |v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)|^{\gamma(\theta_i)}, \tag{4.12}$$

subject to the constraints that $v_{\mathrm{mod}}(\theta_i)$ represents the voltage that satisfies the discretized Maxwell's equations given by Eq. (4.8) and Eq. (4.11). Again, the cost functional measures how well the modeled EM power profile v_{mod} along the measurement curve matches the desired profile v_{obj}. In the design problem v stands for relative power but it can represent other features of the EM-field, such as the Poynting vector or complex electric field if control of the phase is desired. In our prototype problem the cost function is discretized into M points along the measurement curve.

Minimizing the cost functional over possible parameter settings p yields an optimized design. In particular, p contains the coordinates of the scattering cylinder centers and explicitly determines the FD matrix $\mathbf{L}(p)$ in Eq. (4.8).

Other design considerations such as robustness against small variations in parameters could also be translated into additional terms in J at increased computational cost. The linear weight $w(\theta_i)$ and the exponent $\gamma(\theta_i)$ allow flexibility in placing emphasis on design aspects of the power profile along the measurement curve. For example, focusing 95% of the power on to the measurement curve might demand compromises in the shape of the focused power profile. The modeled power distribution might not form a perfect top hat. In this case one might increase the linear weight w and possibly lower the value of the exponent γ for angles within the top hat peak to improve the solution. Further weights for normalization can be introduced as desired. Note that J is only indirectly a function of the design parameters p which explicitly appear in the FD matrix $\mathbf{L}(p)$. The EM power $v(\theta_i)$ is related to p in a highly nonlinear fashion through the solution of the constraint given by Eq. (4.8) and Eq. (4.11). Finding the gradient of the cost functional with respect to design parameters can be performed at minimal computational cost as is described in Section 4.8.

4.8 Gradient-based optimization using the adjoint method

Efficient local optimization techniques require evaluation of the gradient of the cost functional with respect to the design parameters. Since the relative permittivity coefficient has finite jump discontinuities where the refractive index changes from air to Teflon, the FD matrix $\mathbf{L}(p)$ is not a differentiable function of the cylinder positions, and therefore the power profile is not differentiable with respect to the cylinder positions. To remedy this non-differentiability in our implementation, permittivity at a grid point is taken to be the average value of $\epsilon_{\mathrm{r}}(x, y)$ over a circle of radius $R \approx \Delta x$ centered at the grid point. This effectively smoothes out the physical jump discontinuities in $\epsilon_{\mathrm{r}}(x, y)$. The smoothing in the finite difference method allows analytic computation of the gradient using the adjoint method. As R tends to zero, the model approaches the physical discontinuity in permittivity. The gradient of the cost functional $\nabla_p J = \left(\frac{\partial J}{\partial p_1}, \frac{\partial J}{\partial p_2}, ..., \frac{\partial J}{\partial p_{2N}}\right) \in \mathrm{R}^{2N}$ is given by

$$\begin{aligned}\partial_{p_i} J_0 = &- \sum_{i=1}^{M} \Delta\theta m^2(\theta_i)\gamma(\theta_i)\mathrm{sgn}\left[v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)\right]\ldots \\ &\left|v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)\right|^{\gamma(\theta_i)-1} \frac{dv_{\mathrm{mod}}}{dp_i} \\ = &-\mathrm{Re}\left[\sum_{i=1}^{M} \Delta\theta m^2(\theta_i)\gamma(\theta_i)\mathrm{sgn}\left[v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)\right]\ldots\right.\end{aligned}$$

$$
\begin{aligned}
&\left|v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)\right|^{\gamma(\theta_i)-1} W(\theta_i) \cdot \mathrm{diag}(E^*) \cdot \partial_{p_i} E\Big] \\
&= -\mathrm{Re}\left[\left(\mathbf{L}^T \cdot h\right)^T \cdot \partial_{p_i} E\right] = -\mathrm{Re}\left[h^T \cdot \mathbf{L} \cdot \partial_{p_i} E\right] \\
&= \mathrm{Re}\left[h^T \cdot (\partial_{p_i} \mathbf{L}) \cdot E\right],
\end{aligned}
$$

$$
\frac{\partial J_0}{\partial p_i} = \mathrm{Re}\left[h^T \cdot (\partial_{p_i} \mathbf{L}(p)) \cdot E\right], \tag{4.13}
$$

which can be evaluated using h. To solve for h the adjoint equation must be solved, which is given by

$$
\begin{aligned}
\mathbf{L}^T(p) \cdot h = \mathrm{diag}(E^*) \cdot \textstyle\sum_{i=1}^{M} m^2(\theta_i)\gamma(\theta_i)\mathrm{sgn}\left[v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)\right] \ldots \\
\left|v_{\mathrm{obj}}(\theta_i) - v_{\mathrm{mod}}(\theta_i)\right|^{\gamma(\theta_i)-1} \Delta\theta \; W^T(\theta_i).
\end{aligned} \tag{4.14}
$$

In Eq. (4.14), superscript T indicates the transpose and superscript $*$ indicates the complex conjugate. The cost of computing the matrix derivatives Eq. (4.13) grows linearly in the number of design parameters but evaluation of these derivatives adds little to the overall computational effort because the matrix derivatives are sparse, simple, and explicit functions of p. Most of the time is spent on solving the linear system Eq. (4.14) at the same computational effort as solving the forward simulation a single time. This must be seen in contrast to approximating the gradient with finite differences using $N + 1$ forward solves.

For local optimization a modified gradient method is implemented. Cylinder overlap is not allowed which does represent an important constraint during the iterative optimization procedure. In our implementation, colliding cylinders are moved apart a distance $\lambda_0/8$ prior to continued optimization.

4.9 Results and comparison with experiment

It is important to verify numerical simulations by performing experiments. The detected power profile along the measurement curve is compared to the calculated power profile. As an initial test, a 5×5 finite-sized periodic array of cylindrical dielectric scatterers with lattice constant equal to the free-space wavelength λ_0 was studied. This is described in Section 4.9.1. Section 4.9.2 describes results of an optimized aperiodic structure.

4.9.1 Finite-sized periodic structure

Figure 4.11(a) is a photograph of 25 Teflon cylinders arranged in a 5×5 finite-sized periodic array and attached to a metal slab that forms the lower half of a waveguide with upper metal plate removed. The Teflon cylinders have a measured diameter of 3.175 ± 0.025 mm and $\epsilon_{r_z} = 2.05$. Experiments to measure EM power are performed with the upper half of the metal waveguide attached. EM power reaching the measurement curve is detected using a small dipole antenna feeding a narrow-band amplifier.

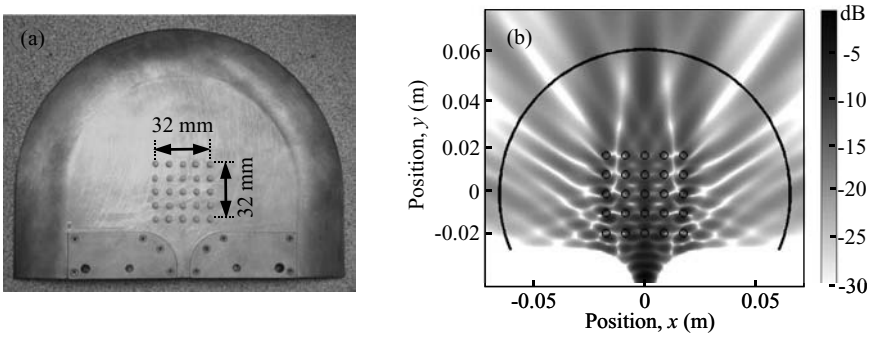

Fig. 4.11. (a) Top view of 25 Teflon cylinders arranged in a 5×5 finite-sized periodic array and attached to a metal slab that forms the lower half of a waveguide. The upper metal plate of the EM waveguide is removed. (b) Calculated relative EM power at frequency $f_0 = 37.5$ GHz in the dielectric scattering domain. The grey scale is in dB. EM radiation emerging from the metal horn is incident on a 5×5 finite-sized lattice of dielectric cylinders. The dielectric cylinders are 3.175 ± 0.025 mm in diameter and are placed symmetrically about the origin with lattice spacing (lattice constant) $a = \lambda_0$. The relative permittivity of the Teflon cylinders is $\epsilon_{r_z} = 2.05$. The solid line is the measurement curve.

Figure 4.11(b) shows the power distribution at frequency $f_0 = 37.5$ GHz calculated using the method described in Section 4.6. In Fig. 4.11(b), the grey scale indicates relative EM power measured in units of dB. The calculations show a diffraction pattern that is symmetric. This is to be expected for the periodic array. In addition, there is interference between EM waves emanating from the dielectric array and subsequently reflected from the metal horn. Figure 4.12(a) shows a comparison between calculated and detected EM power as a function of angle θ on the measurement curve, s. The relative power scale is linear. As may be seen, agreement between calculated and measured data is good with all the main features appearing in both data. Figure 4.12(b) is the same data as in (a) but with relative EM power plotted on a logarithmic scale. Here, the excellent agreement for the three main peaks and the –30 dB minima at $\pm 40°$ are apparent. A slight asymmetry in the measured data also exists whose origin is likely due to dielectric

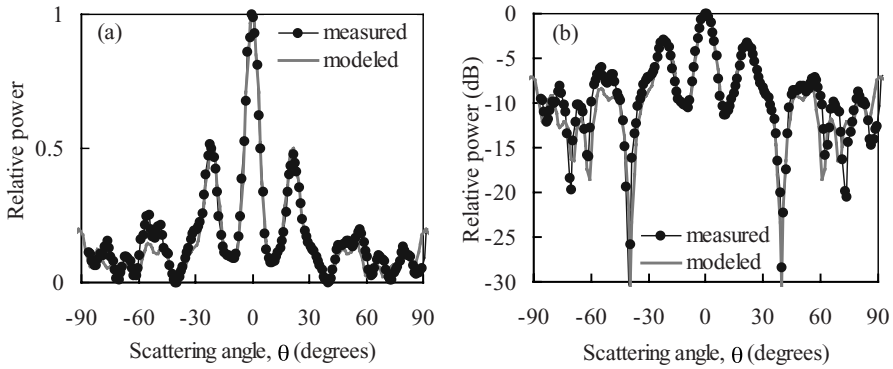

Fig. 4.12. (a) Calculated and detected relative EM power as a function of angle on the measurement curve. (b) Same as (a) but relative power is displayed on a logarithmic scale (dB).

cylinder placement errors that are measured to have a standard deviation less than 0.1 mm. The disagreement between modeled and measured power increases with increasing positive and negative scattering angle.

The overall agreement between calculated and measured results provides the confidence to consider the optimization of aperiodic dielectric structures.

4.9.2 Aperiodic dielectric structure for a top hat objective function

An objective response is sought in which the maximum amount of incident EM radiation is scattered into a top hat function whose peak occurs in the angular range 30° to 60° and is defined along a measurement curve. The measurement curve is shown in Fig. 4.10(a). Symmetrically placed structures, such as the 5×5 finite-sized periodic array are unable to provide the desired functionality. However, one anticipates that breaking the symmetry of the spatial arrangements of dielectric scatterers will provide a better solution.

Rather than attempt to adapt ad hoc PC-inspired designs, a randomized gradient descent algorithm is used to find the optimal spatial configuration of 50 identical, dielectric Teflon cylinders. The result is a nonintuitive aperiodic distribution.

Figure 4.13(a) is a photograph of the experimental arrangement. As may be seen, the positions of the Teflon cylinders do not have any obvious spatial symmetry. Figure 4.13(b) shows the calculated relative EM power at frequency $f_0 = 37.5$ GHz. The grey scale is in dB and the solid line represents the measurement curve. EM radiation from the metal horn is incident on the 50 dielectric cylinders in the dielectric scattering domain. Figure 4.13(c)

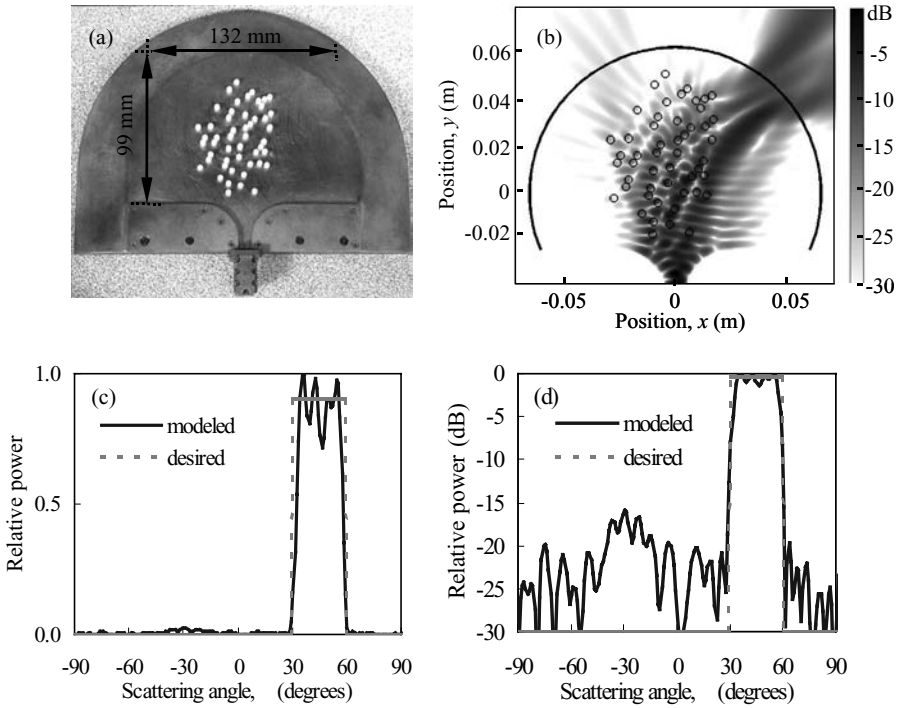

Fig. 4.13. (a) Top view of aperiodic dielectric array attached to lower metal plate that forms the lower half of an EM waveguide. The upper metal plate of the EM waveguide is removed. (b) Calculated relative EM power at frequency $f_0 = 37.5$ GHz. The grey scale is a logarithmic scale (dB). EM radiation from the metal horn is incident on 50 dielectric cylinders in the scattering domain. The cylinders have a diameter of 3.175 ± 0.025 mm and a relative permittivity $\epsilon_r = 2.05$. The cylinder positions are optimized to focus EM power on the measurement curve under a top hat function which peaks between angles 30° and 60°. The measurement curve traced by the power probe is shown as a solid line. (c) The desired and modeled power profile along the measurement curve. 95% of the calculated EM power reaching the measurement curve is focused under the top hat peak. The ripples in the top hat's power as a function of θ are 1.45 dB peak-to-peak. (d) Same as (c) but power is displayed on a logarithmic scale (dB).

shows the desired (broken line) and modeled (solid line) power profile along the measurement curve. Note that 95.0% of the calculated EM power reaching the measurement curve is focused under the top hat peak. The ripples in the top hat's power as a function of θ are 1.45 dB peak-to-peak. Figure 4.13(d) is the same as (c) but power is displayed on a logarithmic scale (dB).

The implementation of the aperiodic scattering structure differs slightly from the design. Precise placement of the Teflon cylinders is more difficult than measuring their positions. The sufficient accuracy of the forward solver can nevertheless be observed. Figure 4.14(a) shows calculated and detected

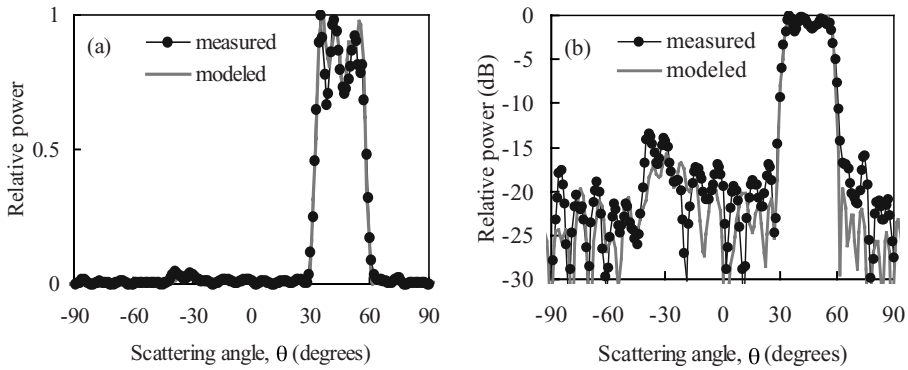

Fig. 4.14. (a) Calculated and measured EM power profile along measurement curve. The relative power scale is linear. (b) Same as (a) but relative power on a logarithmic scale (dB). Ripples at the top hat peak are approximately 1.77 dB peak-to-peak. The portion of the measured power focused between scattering angles 30° and 60° is 92.4%.

relative EM power profile along the measurement curve. The relative power scale is linear. As with the finite-sized periodic array, agreement between experiment and calculation is good. Figure 4.14(b) is the same as (a) but the relative power scale is logarithmic and measured in dB. Ripples at the top hat peak are approximately 1.77 dB peak-to-peak. This is only 0.32 dB greater than the calculated value. The portion of the measured power focused between scattering angles in the range is 92.4%. This is slightly less than the calculated value of 95.0%. The measured relative EM power reflected back into the waveguide is neglected by the forward solver but S_{11} was measured to be –20 dB.

4.9.3 Sensitivity analysis

Practical device design must be robust against variations in spatial configuration introduced during manufacture. A degradation of performance from design to implementation can be observed between Fig. 4.13 and Fig. 4.14. The robustness of an optimal design is closely associated with performance degradation as a function of perturbation of the design parameters. However, a simple Euclidean norm is often not an appropriate measure of the size of a perturbation. Performance degradation can greatly vary for perturbations with the same Euclidean distance to the optimal design. When the cost functional is twice differentiable, the direction of maximal sensitivity is characterized by the eigenvector associated with the largest eigenvalue of the Hessian matrix. However, for our prototype problem, evaluation of the Hessian matrix can be computationally challenging, in particular when constraints inhibit the free perturbation of the cylinder positions.

Instead of evaluating the local Hessian, the robustness of the 5×5 finite-sized periodic array and the optimized aperiodic structure with respect to a uniform random perturbation are compared. The random perturbations of maximum size L are centered around each scattering site. As a measure of the robustness of the power profile we evaluated the standard deviation $\sigma(J)$ of a sample of cost function values. The cost function values were evaluated at randomly perturbed parameter settings. The parameter settings p^{pert} are generated by changing each cylinder location of the locally optimal design p^0 by a length L multiplied by a uniformly distributed random number. The results of numerical simulations shown in Fig. 4.15 indicate that the aperiodic structure is more sensitive than the 5×5 finite-sized periodic array. With increasing L, the standard deviation of J for the aperiodic lattice increases faster than for the 5×5 finite-sized periodic array. The standard deviation initially increases as $\sigma(J) = 0.0524\, L/\lambda_0$ for the 5×5 finite-sized periodic array while for the aperiodic structure with 50 cylinders $\sigma(J) = 0.1369\, L/\lambda_0$. The greater sensitivity of the aperiodic design suggests competition between device performance and robustness. Hard to achieve functionality such as the top hat objective function described in Section 4.9.2 results in a higher sensitivity to small perturbations than the 5×5 finite-sized periodic array. In addition, the aperiodic design contains twice the number of scattering cylinders as the 5×5 PC array.

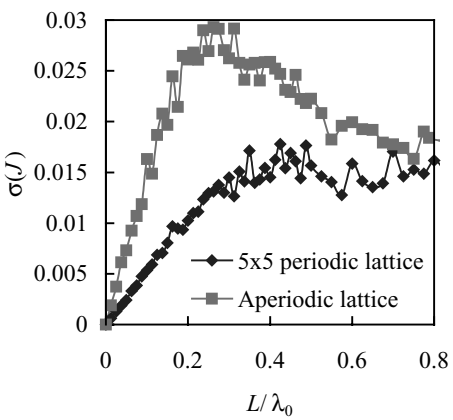

Fig. 4.15. Sample deviation $\sigma(J)$ as a function of L. The change in power profile was measured using the cost functional for perturbations of the locally optimal design as shown in Fig. 4.14(a). Each cylinder coordinate was perturbed randomly on $[-L, L]$ following a uniform distribution. Each point represents the standard deviation of a sample of 200 randomly perturbed parameter settings. The 5×5 finite-sized periodic array increases with L until it reaches its limiting value for randomly distributed scattering cylinders, while the aperiodic lattice peaks before settling down to its limiting value.

Figure 4.15 shows saturation of $\sigma(J)$ for large L. One may investigate this asymptotic behavior by considering the statistical properties of the power profile for perturbations of size $L_{\max}$ of completely random, non-overlapping, configurations of cylinders in the scattering domain. The limiting value for $\sigma(J)$ is 0.017, similar to the values for $L/\lambda_0 = 0.8$ shown in Fig. 4.15. The inherent robustness of aperiodic designs versus PC crystals deserves further investigation.

4.9.4 Further discussion

Using a computationally efficient gradient-based optimizer and finite difference forward solver the spatial arrangement of identical, parallel, dielectric cylinders may be configured to closely match a desired EM power response. There is good agreement between calculations and experiments for an objective function in which EM radiation propagating in a slab waveguide is scattered into a top hat function that peaks in the angular range 30° to 60° on a measurement curve. The spatial arrangement of scattering cylinders is aperiodic and nonintuitive. The methodology can extend the functionality of RF and nano-photonic devices beyond that of PC inspired designs.

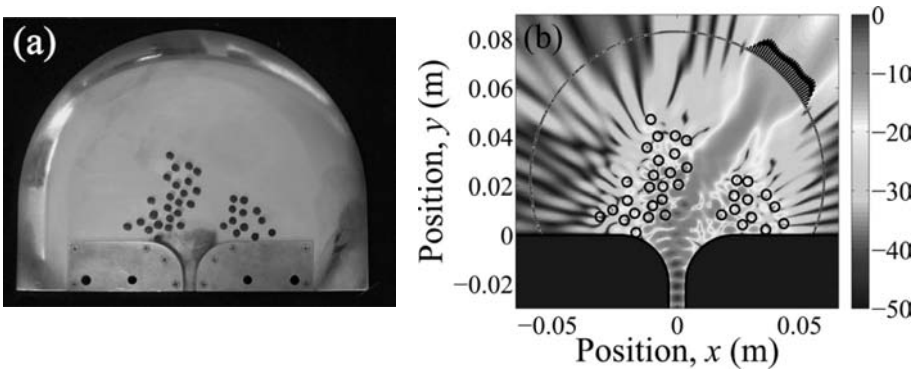

Fig. 4.16. (a) Alternative design to bend an incident EM-beam uniformly into 30° to 60° using dielectric cylinders with higher refractive index. The cylinders have a radius $r = 2$ mm and are made from Ultem2300 and have a relative permittivity $\epsilon_r = 3.365$ at $\lambda_0 = 8$ mm. (b) Simulated power distribution for the compact alternative design shown in (a). The simulated power distribution shows that the aperiodic array channels and confines the EM beam. Many degrees of freedom allow even more compact designs.

The space of aperiodic designs is vastly larger than that of photonic crystals, which is a strict subset. Searching for optimal designs in such high-dimensional, nonlinear, and non-convex space is extremely time consuming and remains an active research area. For example the prototype problem of bending and shaping an incident beam yields multiple designs displaying

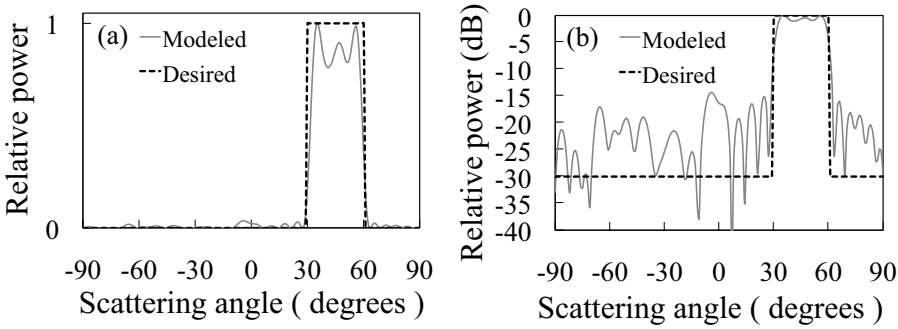

Fig. 4.17. (a) The simulated performance of the alternative design shown in Fig. 4.16. (b) Simulated power profile as in (a) on a logarithmic scale (dB). The design's performance using 36 Ultem cylinders is nearly as good as using 50 Teflon cylinders.

nearly equal performance. Using only 36 cylinders made of the dielectric material Ultem2300 yields an entirely different design with almost equivalent performance. The Ultem cylinders have a diameter of 4 mm and a slightly higher relative permittivity $\epsilon_r = 3.365$ than Teflon. One alternative design is shown in Fig. 4.16, illustrating a different mode of operation. Instead of guiding the EM beam the cylinders are arranged to form a sort of aperiodic mirror that blocks the light from propagating through the structure. Figure 4.17 shows the resulting simulated power profile along the measurement curve. The design successfully focuses in excess of 90% of RF power into the desired angular window but fails to suppress the leakage to below –30 dB outside the angular window. No doubt this design can be improved upon by adding more cylinders to prevent power leakage. It is important to notice the variety of design options at the engineer's disposal. Varying cylinder positions, cylinder radii, refractive index of all or individual cylinders allow vast numbers of design possibilities from fairly basic building blocks.

The seemingly simple objective of rotating and shaping the power of an incident EM wave is itself not easily achieved using photonic crystals. A possible solution using periodic structures is the PC-waveguide shown in Fig. 4.18. Using the higher refractive index of Ultem there exists a PC structure that has a band gap at the operating frequency. The solution for using PCs is simply to form a PC-waveguide from the horn to the measurement curve. This design may perform well with respect to the required amount of relative power focused into the desired angular window. Much more functionality cannot be expected for designs that are constrained to periodic structures. The power profile for the PC-waveguide is shown in Fig. 4.19. The majority of the power is delivered to the desired area, but there is literally no control over the shape of the power profile along the measurement curve.

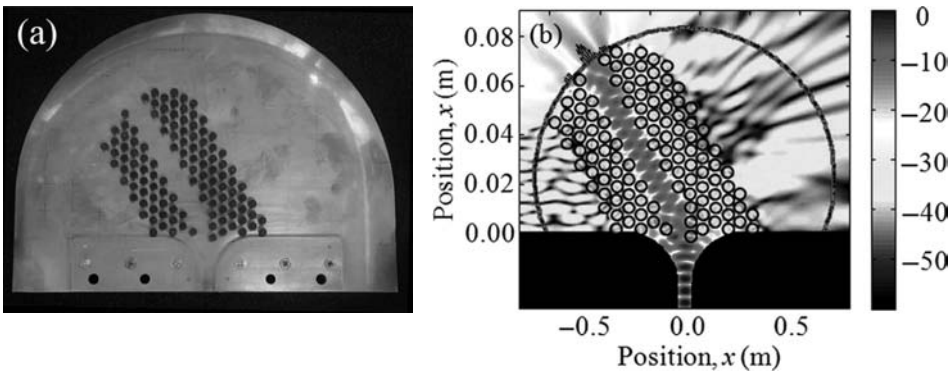

Fig. 4.18. (a) Periodic photonic crystal design to bend light at a −45° angle. (b) Simulation of the EM power distribution over the modeling domain for the PC waveguide shown in (a). Using Ultem cylinders, the finite-sized PC contains the EM wave reasonably inside the waveguide.

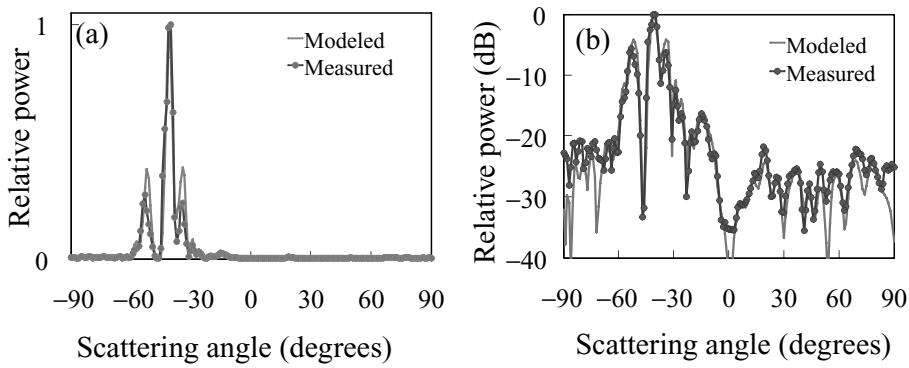

Fig. 4.19. (a) Power profile of finite-sized PC waveguide with aperiodic truncation displayed in Fig. 4.18. The design guides the incident EM wave and confines most of the power reaching the measurement curve at a −45° angle. (b) Power profile shown in (a) on a logarithmic scale (dB).

PC designs exist in one, two, and three dimensions with varying functionality. One advantage of PCs is that under the assumption of the infinite extent of the PC structures analytical results are available. Inaccuracies and errors are introduced from the fact that realistic designs should be as compact as possible. Truncating PCs gracefully to finite size is a proposed application of aperiodic device design. An aperiodic truncation of the PC-waveguide is shown in Fig. 4.20.

The performance of a finite-size PC with aperiodic aperture improves dramatically. Besides channeling a high percentage of the power into the

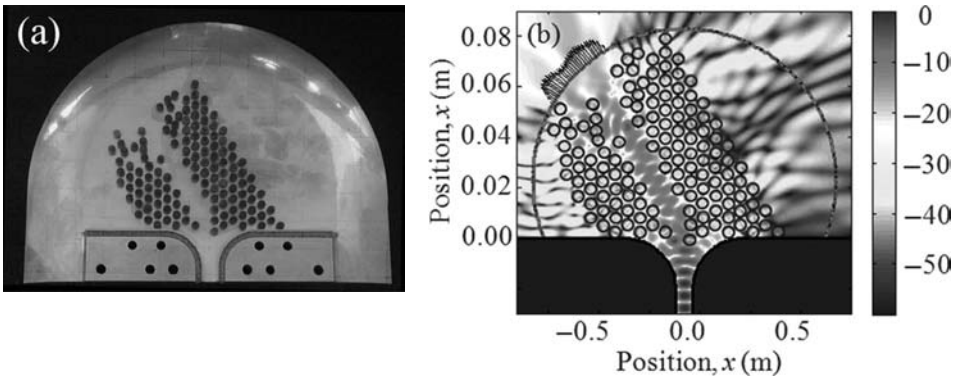

Fig. 4.20. (a) Design of a finite-sized, PC-waveguide with aperiodic aperture made with Ultem cylinders. (b) Simulation of the power distribution for the design shown in (a). The PC-waveguide confines the light and an aperiodic truncation shapes the EM beam as it exits the waveguide. Aperiodic designs solve some problems that are associated with finite-sized photonic crystals.

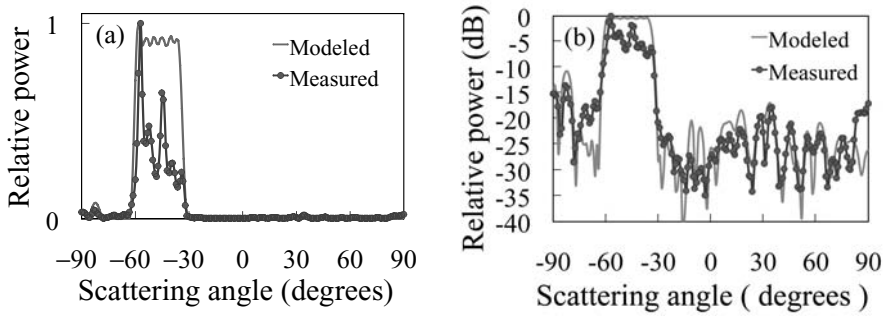

Fig. 4.21. (a) Power profile of the PC waveguide with aperiodic aperture shown in Fig. 4.20. In addition to guiding most of the EM power the aperiodic aperture shapes the EM beam as it exits the waveguide. The difference between simulation and measurement most likely stems from placement errors of the Ultem cylinders. (b) Same power profile as in (a) on a logarithmic scale (dB).

angular range 30° to 60° (Fig. 4.21), the power is now much more uniformly distributed over the desired region. In addition, the horn-PC-waveguide transition may be improved using aperiodic design to decrease the impedance mismatch and therefore input reflection.

Other optical components could include frequency demultiplexers. To decrease the size of the search space, the scattering cylinders are constrained to lattice sites of a PC crystal pattern. In this manner a binary optimization algorithm searches the design space by either occupying a lattice site or not. The thus achieved frequency demultiplexer is shown in Fig. 4.22.

The next step in this design process would be to investigate the

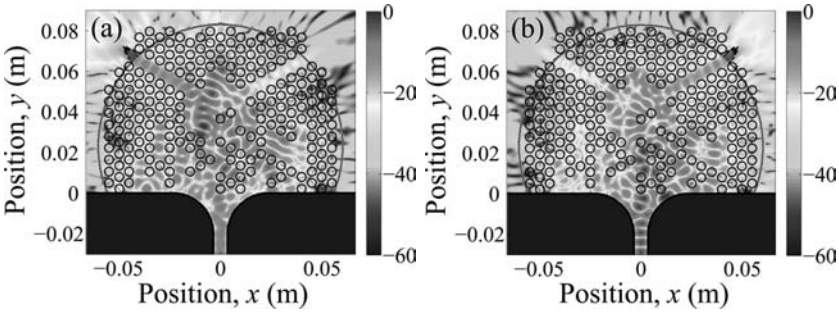

Fig. 4.22. (a) Aperiodic frequency splitter with cylinders occupying lattice sites. At 37.8 GHz most of the energy is channeled into the left waveguide. The power is scaled to the maximum power and the shading is on a logarithmic scale (dB). (b) At 38.3 GHz the majority of the EM power is channeled into the right waveguide.

performance improvement that can be achieved by allowing the cylinders to move to off-lattice sites.

4.9.5 Comparison with photonic crystal inspired devices

In recent years, spatially periodic dielectric structures have been studied and applied to both optics and microwaves [26–28]. It has been shown that introduction of point and line defects in PCs can be used to filter, demultiplex, and guide electromagnetic waves [3, 2, 29, 30]. However, there are numerous unresolved design issues with PC-inspired devices that limit prospects of adoption as practical components. For example, when coupling between standard fibers or waveguides and PC waveguides, the back reflection is either unacceptably large [31] or requires use of relatively large coupling regions [32, 33]. As another example, it is well known that the necessarily finite size of the PC can have a dramatic and detrimental impact on device performance [34]. Solutions to these and similar problems are stymied by the limited number of degrees of freedom inherent to PC design. Rather than struggle for solutions within the constraints of spatial symmetry imposed by PC structures, our approach is based on breaking the underlying spatial symmetries and thereby exposing larger numbers of degrees of freedom with which to design and optimize nano-photonic RF devices.

Our experience so far indicates that optimization of such systems is best achieved using numerical adaptive design techniques. This is because solutions, such as that illustrated in Fig. 4.22, have such a high degree of broken symmetry it is unlikely analytic methods or conventional intuition could usefully be applied.

4.9.6 Summary

In this chapter it has been shown that aperiodic nano-photonic RF dielectric structures designed using adaptive algorithms can be tailored to closely match desired electromagnetic transmission and scattering properties. It is the broken symmetry of the structure that allows more degrees of freedom and the possibility of better optimization compared to symmetric crystal structures.

In general the frequency response and spatial configuration of this system can have very complicated forms. It is the large number of degrees of freedom that allow one to tailor the response to the desired objective or target. The configuration space is even more complex if arbitrary shapes and materials with electromagnetic loss or gain are considered.

4.10 References

1. Rangarajan K. Sundaram, *A First Course in Optimization Theory*, Cambridge University Press, Cambridge, United Kingdom, 1996.
2. J. Smajic, C. Hafner, and D. Erni, *Design and optimization of an achromatic photonic crystal bend*, Optics Express **11**, 1378–1384 (2003).
3. S. Fan, S.G. Johnson, J.D. Joannopoulos, C. Manolatou, and H.A. Haus, *Waveguide branches in photonic crystals*, Journal of the Optical Society of America B **18**, 162–165 (2001).
4. D. Felbacq, G. Tayeb, and D. Maystre, *Scattering by a random set of parallel cylinders*, Journal of the Optical Society of America A **11**, 2526–2538 (1994).
5. G. Guida, D. Maystre, G. Tayeb, and P. Vincent, *Mean-field theory of two-dimensional metallic photonic crystals*, Journal of the Optical Society of America B **15**, 2308–2315 (1998).
6. G. Tayeb and D. Maystre, *Rigorous theoretical study of finite-size two-dimensional photonic crystals doped by microcavities*, Journal of the Optical Society of America A **14**, 3323–3332 (1997).
7. B. Gralak, S. Enoch and G. Tayeb, *Anomalous refractive properties of photonic crystals*, Journal of the Optical Society of America A **17**, 1012–1020 (2000).
8. J. Yonekura, M. Ikeda, and T. Baba, *Analysis of finite 2-D photonic crystals of columns and lightwave devices using the scattering matrix method*, IEEE Journal of Lightwave Technology **17**, 1500 (1999).
9. B.C. Gupta and Z. Ye, *Disorder effects on the imaging of a negative refractive lens made by arrays of dielectric cylinders*, Journal of Applied Physics **94**, 2173–2176 (2003).
10. S. Kozaki, *Scattering of a Gaussian beam by a homogeneous dielectric cylinder*, Journal

of Applied Physics **53**, 7195–7200 (1982).

11. Z. Wu and L. Guo, *Electromagnetic scattering from a multilayered cylinder arbitrarily located in a Gaussian beam, a new recursive algorithm*, Progress in Electromagnetics Research **18**, 317–333 (1998).

12. Y. Chen, Y. Rong, W. Li, *et al.*, *Adaptive design of nano-scale dielectric structures for photonics*, Journal of Applied Physics **94**, 6065–6068 (2003).

13. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes*, Cambridge University Press, Cambridge, United Kingdom, 2002.

14. http://www.cst.de/

15. J. Volk, A. Håkansson, H.T. Miyazaki, *et al.*, *Fully engineered homoepitaxial zinc oxide nanopillar array for near-surface light wave manipulation*, Applied Physics Letters **92**, 183114–183116 (2008).

16. J.D. Joannopoulos, R.D. Mead, and J.N. Winn, *Photonic Crystals*, Princeton University Press, Princeton, New Jersey, 1995.

17. For example, S.J. McNab, N. Moll, and Y.A. Vlasov, *Ultra-low loss photonic integrated circuit with membrane-type photonic crystal waveguides*, Optics Express **11**, 2927–2939 (2003). Y. Du and A.F.J. Levi, *Accessing transmission-mode dispersion in superprisms*, Solid State Electronics **47**, 1369–1377 (2003).

18. J.M. Geremia, J. Williams, and H. Mabuchi, *Inverse-problem approach to designing photonic crystals for cavity QED experiments*, Physical Review E **66**, 066606 (2002).

19. L. Sanchis, A. Håkansson, D. López-Zanón, J. Bravo-Abad, and J. Sánchez-Dehesa, *Integrated optical devices design by genetic algorithm*, Applied Physics Letters **84**, 4460–4462 (2004).

20. A. Håkansson, J. Sánchez-Dehesa, and L. Sanchis, *Inverse design of photonic crystal devices*, IEEE Journal of Selected Areas in Communications **23**, 1365–1371 (2005).

21. Y. Jiao, S. Fan, and D.A.B. Miller, *Demonstration of systematic photonic crystal device design and optimization by low-rank adjustments: an extremely compact mode separator*, Optics Letters **30**, 141–143 (2005).

22. I.L. Gheorma, S. Haas, and A.F.J. Levi, *Aperiodic nanophotonic design*, Journal of Applied Physics **95**, 1420–1426 (2004).

23. P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, United Kingdom, 2003.

24. D.M. Kingsland, J. Gong, J.L. Volakis, and J.F. Lee, *Performance of an anisotropic artificial absorber for truncating finite-element meshes*, IEEE Transactions on Antennas and Propagation **44**, 975–982 (1996).

25. J.-P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, Journal of Computational Physics **114**, 185–200 (1994).

26. J.D. Joannopoulos, R.D. Meade, and J.N. Winn, *Photonic Crystals*, Princeton University Press, Princeton, NJ, 1995.

27. C.A. Kyriazidou, H.F. Contopanagos, and N.G. Alexopoulos, *Monolithic waveguide*

filters using printed photonic-bandgap materials, IEEE Transactions on Microwave Theory Techniques **49**, 297–307 (2001).

28. F. Gadot, A. Ammouche, A. de Lustrac, *et al.*, *Photonic band gap materials for devices in the microwave domain*, IEEE Transactions on Magnetics **34**, 3028–3031 (1998).

29. M. Koshiba, *Wavelength division multiplexing and demultiplexing with photonic crystal waveguide couplers*, IEEE Journal of Lightwave Technology **19**, 1970 (2001).

30. S. Fan, H. Haus, P. Villeneuve, and J. Joannopoulos, *Channel drop filters in photonic crystals*, Optics Express **3**, 4–11 (1998).

31. T.D. Happ, M. Kamp, and A. Forchel, *Photonic crystal tapers for ultracompact mode conversion*, Optics Letters **26**, 1102–1104 (2001).

32. A. Mekis and J.D. Joannopoulos, *Tapered couplers for efficient interfacing between dielectric and photonic crystal waveguides*, IEEE Journal of Lightwave Technology **19**, 861 (2001).

33. Y. Xu, R.K. Lee, and A. Yariv, *Adiabatic coupling between conventional dielectric waveguides and waveguides with discrete translational symmetry*, Optics Letters **25**, 755–757 (2000).

34. Y.H. Ye, D.Y. Jeong, T.S. Mayer, and Q.M. Zhang, *Finite-size effect on highly dispersive photonic-crystal optical components*, Applied Physics Letters **82**, 2380–2382 (2003).

5 Design at the classical–quantum boundary

Rodrigo Muniz and Stephan Haas

5.1 Introduction

In this chapter we explore systems whose description lies at the boundary between classical and quantum theory. There are of course many ways to approach this problem. Here, we choose to study the interaction of classical light with small metal particles of arbitrary shape. Specifically, we consider a physical model that is capable of observing the transition from bulk material properties to nanoscale structures, for which quantum effects dominate. We then explore the landscape of possible physical responses of such systems, using optimal design techniques to train our intuition.

The prevalent classical model describing the interaction of visible and infrared electromagnetic radiation with nanoscale metallic clusters is based on Mie theory [1]. This local continuum field model which uses empirical values of a bulk material's linear optical response has been used to describe plasmon resonances in nanoparticles [2–4]. However, such a semi-empirical continuum description necessarily breaks down beyond a certain level of coarseness introduced by atomic length scales. Thus, it cannot be used to describe the interface between quantum and classical macroscopic regimes. Moreover, extensions of Mie theory to inhomogeneous cluster shapes are commonly restricted to low-order harmonic expansions (e.g. elliptical distortions) and so do not exhaust the full realm of possible geometric configurations. In addition, near-field applications, such as surface enhanced Raman scattering [5], are most naturally described using a real-space theory that includes the *non-local* electronic response of inhomogeneous structures, again beyond the scope of Mie theory.

In the following section we describe a microscopic approach that demonstrates the breakdown of this concept at atomic scales, whereas for large cluster sizes the classical predictions for the plasmon resonances are reproduced.

To illustrate this approach, in the subsequent section we apply it to a simple system of two atoms with variable separation, which allows a fully analytical treatment. In the next section, we then examine plasmonic resonances in clusters with a small number of atoms. Then we show how for larger clusters this approach is useful in describing the coexistence of excitations of quantum and classical character in inhomogeneous nanoscale structures. Finally, we explore nonintuitive aspects of optimal design and conclude by discussing possible future research directions.

5.2 Non-local linear response theory

To capture the main single-particle and collective aspects of light–matter interaction in inhomogeneous nanoscale systems we adopt the linear response approximation [6], valid for sufficiently low intensities of electromagnetic radiation.

The starting point of this approach is the determination of the eigenenergies E_i and wave functions $\Psi_i(\mathbf{r})$ of the Hamiltonian for the electrons in the nanostructure. These can be obtained from several models of varying complexity, ranging from effective mass and tight-binding Hamiltonians to density functional theory and exact diagonalization or other sophisticated numerical solutions of Hubbard-type models. In the following discussion, we limit ourselves to the effective mass and the tight-binding models, noting that response optimization will require multiple evaluations of E_i and $\Psi_i(\mathbf{r})$. While desirable, the additional complexity of dealing with more sophisticated models would limit us to very small clusters and would also make the optimization iterations numerically very expensive.

In the tight-binding model, one considers electrons hopping between atomic sites with a matrix element $t_{i,j}$, which generally decreases rapidly in magnitude with the inter-atomic separation. The wave functions $\Psi_i(\mathbf{r})$ are obtained as a linear combination of orbitals

$$\Psi_i(\mathbf{r}) = \sum_{i,j} \alpha_{ij} \varphi(\mathbf{r} - \mathbf{R_j}), \tag{5.1}$$

where $\varphi(\mathbf{r} - \mathbf{R_j})$ is the wave function of an orbital around an atom localized at position $\mathbf{R_j}$ and α_{ij} are the coefficients of the eigenvector (with energy E_i) of the Hamiltonian, which has the matrix elements

$$\langle \varphi(\mathbf{r} - \mathbf{R_i}) | H | \varphi(\mathbf{r} - \mathbf{R_j}) \rangle = \begin{cases} \mu_i & i = j \\ -t_{i,j} & i \neq j. \end{cases} \tag{5.2}$$

The resulting tight-binding Hamiltonian,

$$H = -\sum_{i,j}(t_{i,j}c_i^\dagger c_j + h.c.) + \sum_i \mu_i c_i^\dagger c_i, \tag{5.3}$$

contains an on-site potential V_i which can be varied to mimic the effect of different atoms in the system. The Hamiltonian matrix can be numerically diagonalized using the Householder method to first obtain a tridiagonal matrix and then a QL algorithm for the final eigenvectors and eigenvalues [7]. Although the tight-binding approach is rather crude, it is attractive because of the relative computational ease with which its wave functions and energies can be computed.

The same holds for the effective mass model, in which one starts from the Schrödinger equation for noninteracting electrons with mass m_e and charge e moving in potential $V(\mathbf{r})$, given by

$$H\Psi_i(\mathbf{r}) = \left(-\frac{\hbar^2}{2m_e}\nabla^2 + V(\mathbf{r})\right)\Psi_i(\mathbf{r}) = E_i\Psi_i(\mathbf{r}). \tag{5.4}$$

The wave functions obtained from the Hamiltonian are inserted into the Poisson equation which in turn determines the local potential due to the spatial distribution of the positive background charges. Using the jellium approximation, the resulting potential is implicitly given by

$$\nabla^2 V(\mathbf{r}) = 4e\pi\rho(\mathbf{r}), \tag{5.5}$$

where the density of the positive background charge $\rho(\mathbf{r})$ satisfies the condition of neutrality inside the nanostructure so that

$$\int \rho(\mathbf{r})d\mathbf{r} = N_{\mathrm{el}}, \tag{5.6}$$

where N_{el} is the number of electrons.

Once the electronic energy levels and wave functions have been obtained, it is possible to calculate the dielectric susceptibility $\chi(\mathbf{r},\mathbf{r}',\omega)$ using a real-space formulation of the random phase approximation [8, 9]

$$\chi(\mathbf{r},\mathbf{r}',\omega) = \sum_{i,j}\frac{f(E_i)-f(E_j)}{E_i - E_j - \omega - i\gamma}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}')\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}). \tag{5.7}$$

The induced charge density distribution function is then obtained by

$$\rho_{\mathrm{ind}}(\mathbf{r},\omega) = \int \chi(\mathbf{r},\mathbf{r}',\omega)(\phi_{\mathrm{ind}}(\mathbf{r}',\omega) + \phi_{\mathrm{ext}}(\mathbf{r}',\omega))d\mathbf{r}', \tag{5.8}$$

where the induced potential is given by

$$\phi_{\mathrm{ind}}(\mathbf{r},\omega) = \int \frac{\rho_{\mathrm{ind}}(\mathbf{r}',\omega)}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'. \tag{5.9}$$

One can avoid the large memory requirement to store $\chi(\mathbf{r},\mathbf{r}',\omega)$ by calculating the induced charge density distribution iteratively via

$$\rho_{\mathrm{ind}}(\mathbf{r},\omega) = \sum_{i,j} \frac{f(E_i) - f(E_j)}{E_i - E_j - \omega - i\gamma} \psi_i^*(\mathbf{r})\psi_j(\mathbf{r}) \times \int \psi_i(\mathbf{r}'), \phi_{\mathrm{tot}}(\mathbf{r}',\omega)\psi_j^*(\mathbf{r}')d\mathbf{r}' \tag{5.10}$$

with $\phi_{\mathrm{tot}}(\mathbf{r}',\omega) = \phi_{\mathrm{ind}}(\mathbf{r}',\omega) + \phi_{\mathrm{ext}}(\mathbf{r}',\omega)$. The integrals can be evaluated numerically. Equations (5.9) and (5.10) are solved self-consistently by iterating $\phi_{\mathrm{ind}}(\mathbf{r},\omega)$ and $\rho_{\mathrm{ind}}(\mathbf{r},\omega)$. For sufficiently small clusters, this procedure typically converges in 3 to 8 steps when starting with $\phi_{\mathrm{ind}}(\mathbf{r},\omega) = 0$, depending on the proximity to a resonance and on the value of the damping constant γ. A much better performance can be achieved when the initial $\phi_{\mathrm{ind}}(\mathbf{r},\omega)$ is taken as the solution of a previously solved nearby frequency. Upon its convergence, the frequency and spatial dependence of the induced electric field and the induced energy are obtained using

$$\mathbf{E}_{\mathrm{ind}}(\mathbf{r},\omega) = -\nabla\phi_{\mathrm{ind}}(\mathbf{r},\omega), \tag{5.11}$$

and

$$W_{\mathrm{ind}}(\omega) = \frac{1}{2}\int |\mathbf{E}_{\mathrm{ind}}(\mathbf{r},\omega)|^2 d\mathbf{r}. \tag{5.12}$$

The observed resonances in the induced energy and charge density distribution at certain driving frequencies of the applied electric field correspond to collective modes of the cluster.

5.3 Dielectric response of a diatomic molecule

Let us now explicitly see how this approach works for a two-atom system, in which case the integrals can be performed analytically. For a diatomic system the tight-binding Hamiltonian is given by

$$H = \begin{pmatrix} \mu & -t \\ -t & \mu \end{pmatrix}. \tag{5.13}$$

Diagonalizing this matrix yields the following eigenvalues and corresponding eigenstates:

$$E_0 = \mu - t, \quad |\psi_0\rangle = \frac{|\varphi_{\mathrm{a}}\rangle + |\varphi_{\mathrm{b}}\rangle}{\sqrt{2}}, \tag{5.14}$$

$$E_1 = \mu + t, \quad |\psi_1\rangle = \frac{|\varphi_{\mathrm{a}}\rangle - |\varphi_{\mathrm{b}}\rangle}{\sqrt{2}}. \tag{5.15}$$

Here $|\varphi_{\rm a}\rangle$ and $|\varphi_{\rm b}\rangle$ are taken to be $1s$-orbitals wave functions, centered at the atoms a and b respectively,

$$\varphi_{\rm a}(\mathbf{r}) = \frac{1}{\sqrt{\pi a_{\rm B}^3}} e^{-|\mathbf{r}-\mathbf{R}_a|/a_{\rm B}}, \tag{5.16}$$

$$\varphi_{\rm b}(\mathbf{r}) = \frac{1}{\sqrt{\pi a_{\rm B}^3}} e^{-|\mathbf{r}-\mathbf{R}_b|/a_{\rm B}}, \tag{5.17}$$

where $a_{\rm B}$ is an effective Bohr radius. Within the random phase approximation [9], the susceptibility of the system only has two terms contributing to the sum over the states,

$$\chi(\mathbf{r},\mathbf{r}',\omega) = \left(\frac{f(E_1)-f(E_0)}{E_1-E_0-\omega-i\gamma} + \frac{f(E_0)-f(E_1)}{E_0-E_1-\omega-i\gamma}\right) \times \psi_0^*(\mathbf{r})\psi_0(\mathbf{r}')\psi_1^*(\mathbf{r}')\psi_1(\mathbf{r}). \tag{5.18}$$

Therefore the induced charge is given by

$$\rho_{\rm ind}(\mathbf{r},\omega) = \psi_0^*(\mathbf{r})\psi_1(\mathbf{r})\frac{2(E_1-E_0)(f(E_1)-f(E_0))}{(E_1-E_0)^2-(\omega+i\gamma)^2} \times \int d\mathbf{r}'\ \psi_0(\mathbf{r}')\phi_{\rm tot}(\mathbf{r}',\omega)\psi_1^*(\mathbf{r}'). \tag{5.19}$$

Let us now define

$$\alpha \equiv \frac{2\Delta E \Delta f_E}{(\Delta E)^2-\omega^2+\gamma^2-i2\omega\gamma}\int d\mathbf{r}'\ \psi_0(\mathbf{r}')\phi_{\rm tot}(\mathbf{r}',\omega)\psi_1^*(\mathbf{r}'), \tag{5.20}$$

where $\Delta E = E_1 - E_0$ and $\Delta f_E = f(E_1) - f(E_0)$. Using the fact that for s-orbitals

$$\psi_0(\mathbf{r})\psi_1(\mathbf{r}) = \frac{(\varphi_{\rm a}(\mathbf{r})+\varphi_{\rm b}(\mathbf{r}))}{\sqrt{2}}\frac{(\varphi_{\rm a}(\mathbf{r})-\varphi_{\rm b}(\mathbf{r}))}{\sqrt{2}} = \frac{\varphi_{\rm a}^2(\mathbf{r})-\varphi_{\rm b}^2(\mathbf{r})}{2}, \tag{5.21}$$

we can simply write

$$\rho_{\rm ind}(\mathbf{r},\omega) = \frac{\alpha}{2}(\varphi_{\rm a}^2(\mathbf{r})-\varphi_{\rm b}^2(\mathbf{r})), \tag{5.22}$$

which in turn allows the calculation of $\phi_{\rm ind}(\mathbf{r},\omega)$ via

$$\phi_{\rm ind}(\mathbf{r},\omega) = \int d\mathbf{r}'\ \frac{\rho_{ind}(\mathbf{r}',\omega)}{|\mathbf{r}-\mathbf{r}'|} = \frac{\alpha}{2}\int d\mathbf{r}'\ \frac{\varphi_{\rm a}^2(\mathbf{r}')-\varphi_{\rm b}^2(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{5.23}$$

Once this potential is calculated we are able to compute

$$\langle\psi_0|\phi_{\rm tot}|\psi_1\rangle = \int d\mathbf{r}'\ \psi_0(\mathbf{r}')\phi_{tot}(\mathbf{r}',\omega)\psi_1^*(\mathbf{r}'), \tag{5.24}$$

and plug it back into Eq. (5.20) to obtain a linear equation for α.

Using elliptical coordinates with $z_a - z_b = R$, this integral can be performed analytically, yielding

$$\langle\psi_0|\phi_{\text{tot}}|\psi_1\rangle = -\frac{\mathcal{E}R}{2} + \frac{5\alpha}{16a_{\text{B}}} - \frac{\alpha}{2R} + \alpha\left(\frac{13}{16a_{\text{B}}} + \frac{3R}{8a_{\text{B}}^2} + \frac{1}{2R} + \frac{R^2}{12a_{\text{B}}^3}\right)e^{-2R/a_{\text{B}}}, \tag{5.25}$$

where it is assumed that the external electric field of magnitude $\mathcal{E}$ is applied along the z direction, i.e. along the line connecting the two atoms. Plugging this result back into Eq. (5.20), one finds

$$\left(\frac{\alpha}{\mathcal{E}R}a\right)^{-1} = \frac{5}{8a_{\text{B}}} - \frac{1}{R} + \left(\frac{13}{8a_{\text{B}}} + \frac{3R}{4a_{\text{B}}^2} + \frac{1}{R} + \frac{R^2}{6a_{\text{B}}^3}\right)e^{-2R/a_{\text{B}}} - \frac{\Delta E^2 - \omega^2 + \gamma^2 - i2\omega\gamma}{\Delta E \Delta f_E}. \tag{5.26}$$

This enables us to determine a closed analytical form for the physical observables, i.e. we can express the induced charge and induced potential in terms of $r_a = |\mathbf{r} - \mathbf{R}_a|$ and $r_a = |\mathbf{b} - \mathbf{R}_b|$. Namely,

$$\rho_{\text{ind}}(\mathbf{r},\omega) = \frac{\alpha}{2\pi a_{\text{B}}^3}\left(e^{-\frac{2r_a}{a_{\text{B}}}} - e^{-\frac{2r_b}{a_{\text{B}}}}\right), \tag{5.27}$$

$$\phi_{\text{ind}}(\mathbf{r},\omega) = \frac{\alpha}{2}\left[\frac{1}{r_a} - \left(\frac{1}{a_{\text{B}}} + \frac{1}{r_a}\right)e^{-\frac{2r_a}{a_{\text{B}}}} - \frac{1}{r_b} + \left(\frac{1}{a_{\text{B}}} + \frac{1}{r_b}\right)e^{-2r_b/a_{\text{B}}}\right], \tag{5.28}$$

where the ω dependency is contained in α. The corresponding total induced energy is given by

$$W_{\text{ind}}(\omega) = \alpha^2\left[\frac{5}{16a_{\text{B}}} - \frac{1}{2R} + \left(\frac{13}{16a_{\text{B}}} + \frac{3R}{8a_{\text{B}}^2} + \frac{1}{2R} + \frac{R^2}{12a_{\text{B}}^3}\right)e^{-2R/a_{\text{B}}}\right]. \tag{5.29}$$

In Fig. 5.1, the calculated induced charge density and magnitude of the induced electric field are plotted as a function of frequency of the external electric field and the separation between the two atoms. This simple system only has a dipole resonance at the frequency which separates the bonding and anti-bonding energy eigenstates. As the distance between the atoms is increased, the tight-binding matrix element is expected to fall off rapidly. Here, we parameterize this effect with a generic [10] power-law, $t \propto R^{-3}$. Consequently the resonance frequency drops off with decreasing inter-atomic separation. Furthermore, the magnitude of the induced electric field increases with inter-atomic distance, and hence with the cross-section of the diatomic "antenna" structure.

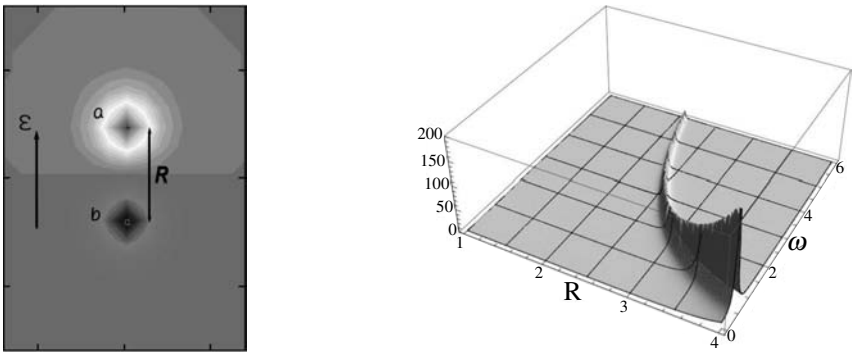

Fig. 5.1. Dielectric response of a diatomic molecule. Left: induced charge density with the external electric field in the vertical direction, and $R = 4a_{\mathrm{B}}$. Right: total induced energy as a function of ω and R, with the tight-binding hopping parameter changing as $t \propto R^{-3}$.

5.4 Dielectric response of small clusters

Next, we explore the collective electromagnetic response in atomic clusters of various sizes and geometries. Our aim is to understand their dielectric response based on the fully quantum-mechanical description given above, which captures accurately their relevant collective modes. The electronic energy levels and wave functions, calculated within the tight-binding model, are used to determine the non-local dielectric response function. It is expected that the system shape, the electron filling and the driving frequency of the external electric field strongly control the resonance properties of the collective excitations in the frequency and spatial domains.

Let us first focus on the dielectric response function in linear chains of atoms, with the intent to identify the basic features of their collective excitations [11]. The frequency dependence of the induced energy in such systems, exposed to a driving electric field along the chain direction, is shown in Fig. 5.2(a). It exhibits a series of resonances, which increase in number for chains with increasing length. As observed in the spatial charge density distribution, e.g. shown for the 5-atom chain in Fig. 5.2(b), the lowest peak corresponds to a dipole resonance. When increasing the system size, the dipole peak moves to lower frequencies, which is the expected finite-size scaling behavior. The resonances at higher frequency correspond to higher harmonic charge density distributions. For example, in Fig. 5.2(c), we show the charge density distribution corresponding to the highest frequency resonance of the 6-atom chain. In contrast to the dipole resonance, these modes

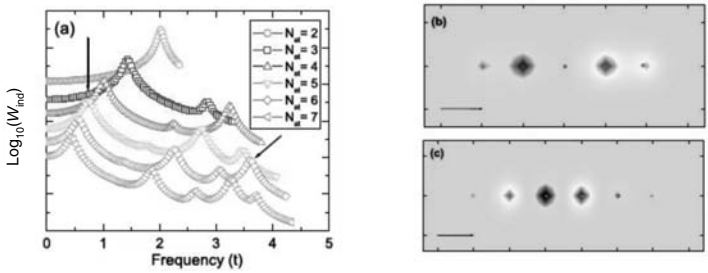

Fig. 5.2. Longitudinal modes in atomic chains. (a) Decimal logarithm of the total induced energy (artificially offset) as a function of the frequency of an external electric field which is applied along the direction of the chain. The resonance peaks correspond to different modes. (b) Induced charge density distribution for the lowest energy mode at $\hbar\omega = 0.73{\times}t$ in the 5-atom chain. (c) Induced charge density distribution for the highest-energy mode at $\hbar\omega = 3.56{\times}t$ in the 6-atom chain.

show a rapidly oscillating charge density distribution, and thus have the potential to provide spatial localization of collective excitations in more sophisticated structures. While an extension to much larger chains is numerically prohibitive within the current method, the finite-size scaling of the observed dielectric response of these clusters is consistent with the 1D bulk expectation of a dominant low-energy plasmon mode, coexisting with a particle-hole continuum of much weaker spectral intensity.

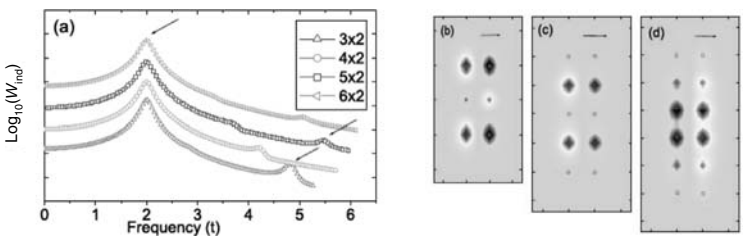

Fig. 5.3. Transverse modes in coupled chain structures. (a) Logarithm of the total induced energy (artificially offset) as a function of the frequency of an external electric field applied transversely to the chain. The low-energy mode is central, the analog of a bulk plasmon, and the high-energy mode is located at the surface, the analog of a surface plasmon. (b) Induced charge density distribution for the mode at $\hbar\omega = 4.91 \times t$ in the 3-atom double chain. (c) Induced charge density distribution for $\hbar\omega = 5.41 \times t$ in the 5-atom double chain. (d) Induced charge density distribution for $\hbar\omega = 1.93 \times t$ in the 6-atom double chain.

In order to study the transverse collective modes we apply an external electric field perpendicular to ladder structures made of coupled linear chains

of atoms.* Figure 5.3(a) shows that for every chain size there are two resonance peaks for the total induced energy, the higher energy is an end mode, as shown in Figs. 5.3(b) and (c) for the 3- and 5-atom double chains respectively, whereas the lower energy peak corresponds to a central mode, as displayed in Fig. 5.3(d) for the 6-atom double chain. It is also confirmed that as the length of the chain is increased, the central mode gets stronger relative to the end mode, which is the expected behavior for bulk versus surface excitations. These results are in agreement with the findings of [12, 13].

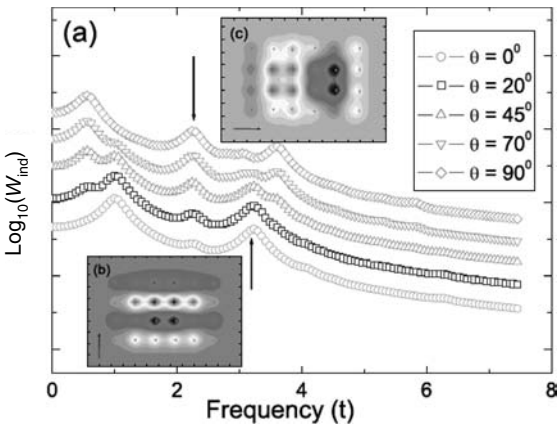

Fig. 5.4. Dependence on the direction of the applied electric field. (a) Logarithm of the total induced energy (artificially offset) as a function of frequency of external electric fields applied to a 4×6-rectangle at different incident angles. $\theta = 0^{\circ}$ when the field is parallel to the 4-atoms edge, and $\theta = 90^{\circ}$ when it is parallel to the 6-atoms edge. (b) Induced charge distribution for $\theta = 0^{\circ}$ and $\hbar\omega = 3.15 \times t$. (c) Induced charge density distribution for $\theta = 90^{\circ}$ and $\hbar\omega = 2.15 \times t$.

Let us next examine what happens when the direction of the external electric field is varied. Figure 5.4 shows the dielectric response of a 4×6-atom rectangular structure for different electric field incidence angles. When the field is parallel to one of the edges ($\theta = 0^{\circ}$ or $\theta = 90^{\circ}$), the response is essentially that of a single chain with the same length, shown in Fig. 5.2(a). Also, the induced spatial charge density modulation is analogous to those of the correspondent linear chain. This may be seen in Fig. 5.4(b) and (c). At intermediate angles the response is a superposition of the two above cases, changing gradually from one extreme to the other as the angle is

* Within this approach, at least two coupled chains are necessary to visualize charge redistributions along the transverse direction, since charge fluctuations within the orbitals are not accounted for.

changed. Notice for instance that as the angle increases, the peak at the same frequency of the 4-atoms dipole resonance diminishes, while simultaneously another resonance is formed at the frequency of the dipole mode of a 6-atoms chain when the angle is tuned from $\theta = 0^{\circ}$ to $\theta = 90^{\circ}$. For $\theta = 0^{\circ}$ there is only the peak at the frequency of the 4-atom chain dipole resonance, whereas for $\theta = 90^{\circ}$ only the dipole peak corresponding to the 6-atoms dipole frequency is present. The superposition of the response from each direction is a consequence of the linear response approximation employed, since the response is a linear combination of those obtained from each direction component of the external field.

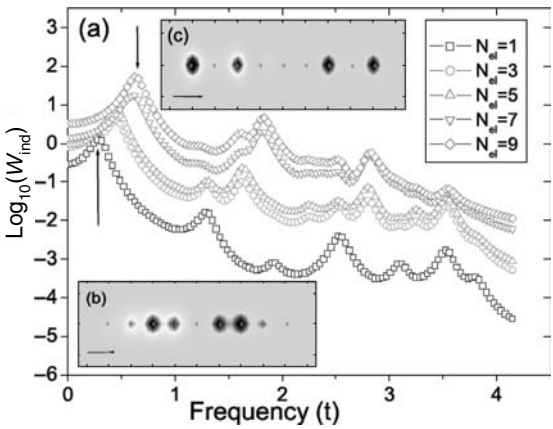

Fig. 5.5. Variation of the number of electrons. (a) Logarithm of the total induced energy as a function of the external electric field frequency. The number of electrons N_{el} in a 9-atom chain is varied. (b) Induced charge density distribution for $N_{\mathrm{el}} = 1$ at $\hbar\omega = 0.36 \times t$. (c) Induced charge density distribution for $N_{\mathrm{el}} = 9$ at $\hbar\omega = 0.53 \times t$.

Next, let us analyze the dependence of the resonance modes on the number of electrons in the cluster. Figure 5.5(a) shows significant changes in the response of a 9-atom chain with the external field applied along its direction. In particular, it is observed that the response is stronger when there are more electrons in the sample, a quite obvious fact since there are more particles contributing to the collective response. Moreover, the resonance frequencies of lower modes increase with the number of electrons, which can be understood as a consequence of the one-dimensional tight-binding density of states being smallest at the center of the band. Hence the energy levels around the Fermi energy are more sparse in the finite system, and therefore the excitations require larger frequencies at half-filling. The same does not hold for higher frequency modes since these correspond to transitions between the lowest and highest levels for any number of electrons in the

sample. Therefore these modes have the same frequency, independent of the electronic filling. Higher filling also allows the induced charge density to concentrate closer to the boundaries of the structure, as a comparison between Fig. 5.5(b) and (c) demonstrates. Figure 5.5(b) shows that a 9-atom chain with $N_{\mathrm{el}} = 1$ electron has its induced charge density localized around the center of the chain. In contrast, Fig. 5.5(c) displays the induced charge density localized at the boundaries of the same structure with $N_{\mathrm{el}} = 9$. This concentration closer to the surface happens because higher energy states have a stronger charge density modulation than the lower energy ones. Therefore the induced charge density is more localized for higher fillings, because at low fillings the excitations responsible for the induced charge density are between the more homogeneous lower energy levels. This can be interpreted as a finite-size rendition of the fact that by increasing the electronic filling one obtains the classical response with all the induced charge density on the surface of the object.

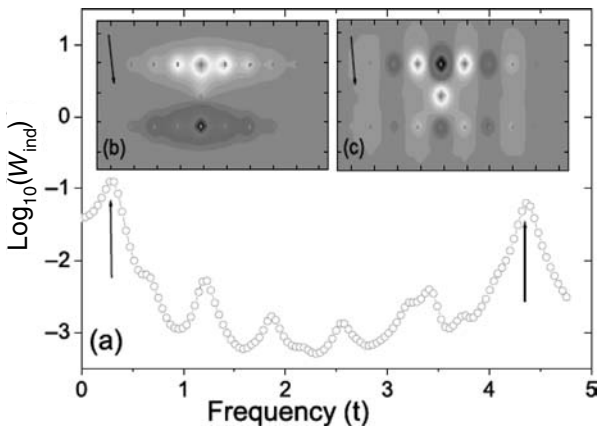

Fig. 5.6. Connection between two chains, $N_{\mathrm{el}} = 1$. (a) Logarithm of the total induced energy as a function of the frequency of an external electric field applied to two 8-atom chains with an extra atom connecting them at the center. (b) Induced charge density distribution for $\hbar\omega = 2.28{\times}t$. (c) Induced charge density distribution for $\hbar\omega = 4.36{\times}t$.

Access to high-energy states is very important for achieving spatial localization of the induced charge density, as the next example shows. In order to find a structure with spatially localized plasmons we consider two parallel 8-atoms chains connected to each other by an extra atom at the center. When an external electric field is applied transversely to the chains, the electrons

are stimulated to hop between them, but this is only realizable through the connection, therefore the plasmonic excitation is sharply localized around it. Figure 5.6(a) shows the response of two 8-atoms chains, Fig. 5.6(b) and (c) show respectively the induced charge density for the dipole and the highest modes. It is seen that the induced charge density of the lowest frequency mode is spread along the chains, whereas the high-frequency plasmon is more localized, since it corresponds to excitations to the highest energy state that has a large charge modulation, as pointed out before.

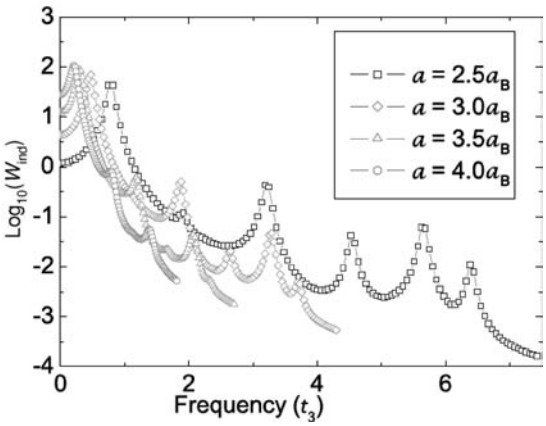

Fig. 5.7. Variation of the distance between neighbor atoms a. Logarithm of the total induced energy as a function of the frequency of an external electric field applied to 7-atom chains with different spacing between atoms a in units of the Bohr radius a_{B}. The frequency unit is t_3, the tight-binding hopping parameter for $a = 3a_{\mathrm{B}}$.

Let us finally analyze the dependence of the various dielectric response modes on the inter-atomic distance. The dipole moment of the chain is proportional to its length, and consequently also proportional to the distance between atoms. Hence one would naively expect that the strength of the dielectric response is strictly proportional to the atomic spacing. However, higher frequency modes require that the electrons are able to hop quickly along the chain in order to produce the fast charge oscillations of the mode. Hence the oscillator strength of the high-frequency modes is suppressed for systems where electrons cannot move fast enough. In the tight-binding model, the hopping rate stems from the overlap of the atomic orbitals on different sites, which decreases with increasing spacing between atoms [10]. Therefore the high-frequency modes are suppressed for chains with large inter-atom spacing, because the hopping is so weak that it overcomes the

gain coming from a larger dipole moment. On the other hand the oscillator strength of the slow modes increases for larger spacings, because they do not require fast motion along the chain. In this case, the contribution from a larger dipole moment dominates over the suppression due to the smaller hopping rates. This fact is demonstrated in Fig. 5.7, where the responses of 7-atom chains with different atomic spacings are shown. The tight-binding hopping parameter t changes with the atomic spacing a, and here we considered a generic [10] power-law dependence $t \sim a^{-3}$. The figure clearly shows that the strength of the slowest mode is reduced for shorter spacings. The opposite is true for the two fastest modes, while the intermediate modes have nearly no change.

5.5 Dielectric response of a metallic rod

Turning now to metallic structures, described within the effective mass approach, we illustrate the application of inhomogeneous linear response to an infinitely long elliptic rod illuminated by an external field. Assume that the rod is aligned in the z direction and the electric field polarization is along the $(1, 1, 0)$ direction. For the wave functions we assume periodic boundary conditions along the z direction. To quantify the response of arbitrary geometries to the applied field, we calculate the energy of the induced field, defined by

$$W_{\mathrm{ind}}(\omega) = (1/2) \int |\mathbf{E}_{\mathrm{ind}}(\mathbf{r}, \omega)|^2 \, d\mathbf{r}. \tag{5.30}$$

In Fig. 5.8 the logarithm of the energy of the induced field ($\log_{10}(W_{\mathrm{ind}})$) is displayed as a function of the applied external field photon energy $\hbar\omega$ and the characteristic system size R for the aspect ratio (a) $a : b = 1 : 1.3$ and (b) $a : b = 1 : 2$. For sufficiently large rod sizes, one clearly observes two plasmon resonances, ω_+ and ω_-, consistent with earlier predictions based on Mie theory [4]. Our method confirms that these resonances occur at $\omega_+ = \omega_{\mathrm{p}}(b/(a+b))^{1/2}$ and $\omega_- = \omega_{\mathrm{p}}(a/(a+b))^{1/2}$, where ω_{p} is the bulk plasmon frequency $\omega_{\mathrm{p}} = \sqrt{4\pi e^2 \rho / m_{\mathrm{e}}}$. Because the scale in Fig. 5.8 is a logarithmic scale, it is clear that the spectral intensity for large rod sizes is orders of magnitude greater compared to smaller rod sizes. Most importantly, it is evident from this figure that the classical picture of two well-defined resonances breaks down below a characteristic system size. For sufficiently small rod sizes, the two macroscopic resonances split into multi-level molecular excitations, with the overall spectral weight shifting towards lower energies. For the chosen parameters this transition occurs at $R_{\mathrm{c}} \approx 6.5\ L$. Below R_{c},

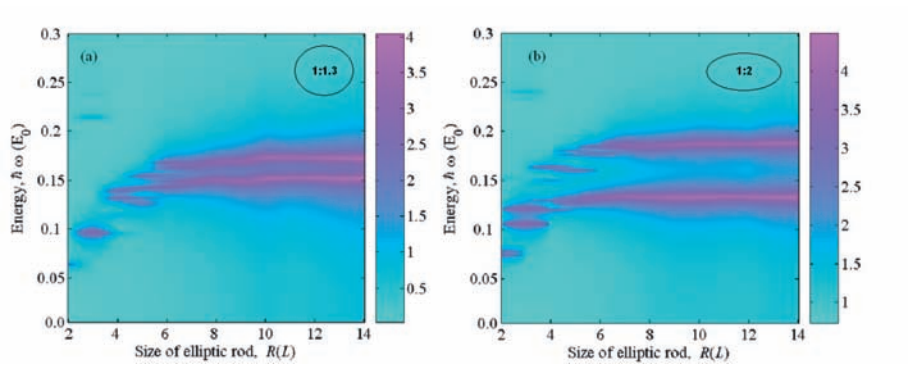

Fig. 5.8. (a) Logarithm of the energy of the induced electric field ($\log_{10}(W_{\rm ind})$) (see text) in an elliptic metallic rod due to applied external field $\mathbf{E}_{\rm ext}$ along the $(1,1,0)$ direction as a function of face surface size R and photon energy $\hbar\omega$ of the external field in units of E_0. Aspect ratio (semi-minor to semi-major axis) $a:b=1:1.3$. Temperature $T=0$ K, carrier density $\rho=(1/480)\ L^{-3}$, and $\gamma=10^{-3}\ E_0$. (b) Same as (a) but with $a:b=1:2$. Ref. [17].

the spectral intensity of the energy levels is reduced because fewer electrons participate in the individual resonances.

The physics determining the value of $R_{\rm c}$ may be illustrated by considering an infinite cylindrical rod of radius R and electron density ρ. In the classical regime, the observed collective oscillations are only weakly damped, indicating that they are well separated from the quasi-continuum of single-particle excitations. This leads to the condition $\omega_{\rm p}R/v_{\rm F}\gg 1$, stating that the plasmon phase velocity is greater than the Fermi velocity $v_{\rm F}$ of the electrons [14]. For $r<R$ the electrons are trapped in a harmonic potential due to the uniform positive background [15] and the characteristic collective frequency is $\omega_{\rm p}=\sqrt{4\pi e^2\rho/(2m_{\rm e})}$. Estimating the Fermi velocity using a bulk value $v_{\rm F}=(3\pi^2\rho)^{1/3}\hbar/m_{\rm e}$ one obtains $R\gg R_{\rm c}=\frac{\pi^{1/6}3^{1/3}}{\sqrt{2}}\frac{\hbar}{em_{\rm e}^{1/2}\rho^{1/6}}\approx 3.4\ L$ for an electron density $\rho=(1/480)\ L^{-3}$, in reasonable, but only approximate, agreement with the calculated threshold $R_{\rm c}\approx 6.5\ L$ shown in Fig. 5.8.

For system sizes below $R_{\rm c}$ the dominant excitations are observed to shift towards lower energies as the positive background potential becomes increasingly anharmonic. The harmonicity criterion for excitations can be expressed as $\sigma_0/R\le 1$, where $\sigma_0=(\frac{\hbar}{m_{\rm e}\omega_{\rm p}})^{1/2}$ is the classical turning point for an electron in the ground state of the harmonic potential. This is equivalent to the condition $R\ge R_{\rm c}=(\frac{\hbar^2}{2m_e\rho e^2})^{1/4}$, yielding $R\ge 3\ L$.

The nature of the excitations changes as the characteristic size R crosses the quantum threshold $R_{\rm c}$. In Fig. 5.9, we show logarithm of the calculated energy of the induced field ($\log_{10}(W_{\rm ind})$) as a function of external field frequency. As shown in Fig. 5.9(a), for relatively large rod sizes, the two distinct

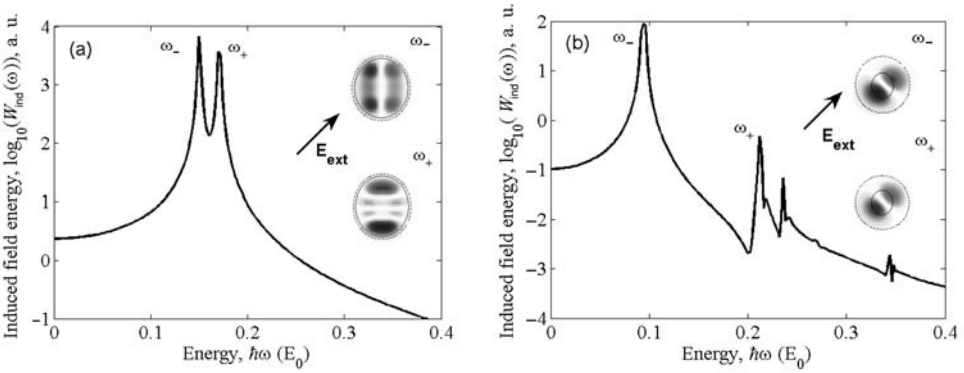

Fig. 5.9. (a) Logarithm of the energy of the induced electric field ($\log_{10}(W_{\rm ind})$) in an elliptic rod of size $R = 11\ L$ with aspect ratio $a : b = 1 : 1.3$ as a function of photon's energy $\hbar\omega$ of the external field in units of E_0. Carrier density $\rho = (1/480)\ L^{-3}$ and $\gamma = 10^{-3}\ E_0$. The direction of the external field is indicated by the arrow. Inset: induced charge density at the resonant frequencies ω_+ and ω_-. The boundary of the rod is shown using the solid line, and the dotted line shows the set of classical turning points, corresponding to the positive background potential. (b) Same as (a), but for size $R = 3\ L$. Ref. [17].

plasmon resonances labeled ω_- and ω_+ correspond to two orthogonal bipolar charge distributions, as indicated in the inset. The spatial orientations of these induced resonances are aligned with the semi-major and semi-minor axes, and do not depend on the direction of the incident field. In contrast, for system sizes less than $R_{\rm c}$ (Fig. 5.9(b)), the excitation spectrum consists of several lower-intensity modes, dominated by a low-frequency resonance. However, such spatial characteristics of these modes are different from the classical limit, i.e. they are *aligned* with the incident field, indicating that Mie theory breaks down in the quantum regime.

5.6 Response of inhomogeneous structures

An advantage of the self-consistent non-local response theory described here is that it extends naturally to inhomogeneous structures with nanoscale and atomically sharp edges and corners. This is of great practical interest to nano-photonic applications since such features are commonly associated with substantial local field enhancements [5]. However, the intuition driving

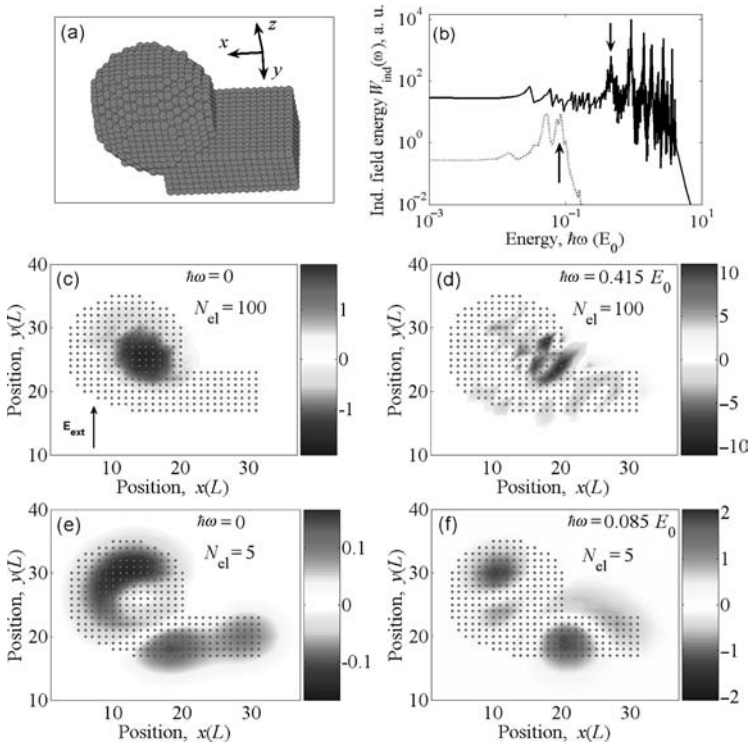

Fig. 5.10. (a) Three-dimensional conjoined sphere and rectangular bar. (b) Logarithm of the induced field energy in the structure for $N_{\rm el} = 100$ (solid line) and $N_{\rm el} = 5$ (dotted line) as a function of the external field frequency. The material volume is $V_{\rm mat} = 5516\ L^3$. The arrows in the figure show the frequencies chosen for results of calculations shown in (d) and (f). Induced charge density is calculated for a nanoscale asymmetric structure for the indicated charge densities and frequencies of the external field. Temperature $T = 0$ K and $\gamma = 10^{-3}\ E_0$. (c) $N_{\rm el} = 100$, $\hbar\omega = 0$, the direction of the external field is indicated by the arrow. (d) $N_{\rm el} = 100$, $\hbar\omega = 0.415\ E_0$. (e) $N_{\rm el} = 5$, $\hbar\omega = 0$. (f) $N_{\rm el} = 5$, $\hbar\omega = 0.085\ E_0$. Ref. [17].

such expectations is usually derived from classical continuum field theory which may not be applicable in this regime. When investigating structures at the microscopic level one needs to account quantitatively for changes in the response spectrum. One would also like to describe the breakdown of the continuum picture and quantum-mechanical discretization effects which may ultimately lead to new and interesting functionalities that are accessible by nanoscale engineering.

To explore this we consider a representative asymmetric nanostructured system of material volume $V_{\text{mat}} = 5516\ L^3$ consisting of a conjoined sphere and rectangular bar (Fig. 5.10(a)) whose response strongly depends on the carrier concentration (Fig. 5.10(b)). Plots of induced charge distributions in response to an external electric field are shown for different carrier concentrations and frequencies of the incident radiation.

For $N_{\text{el}} = 100$ in the low-frequency limit (Fig. 5.10(c)), this leads to a dipole-like anisotropic response for which the induced field is not collinear to the external field $\mathbf{E}_{\text{ext}}$ (shown by the arrow). Note, the induced charge is localized within the boundaries of the nanostructure, and follows the nanostructure's physical bounds given by the positive charge distribution, in agreement with expectations from classical field theory. At this relatively high carrier concentration and at a high external field frequency of $\omega = 0.415\ E_0/\hbar$, one observes a complex local enhancement of the induced charge density (Fig. 5.10(d)). For the low carrier concentration ($N_{\text{el}} = 5$) shown in Fig. 5.10(e) and (f) the induced charge density arises from excitation of low-energy eigenstates and the overall response is weaker. Note, the induced charge is not localized within the boundaries of the nanostructure. Varying the carrier concentration hence acts as a "switch" that can activate different resonant regions in a broken-symmetry atomic-scale structure.

At higher frequencies, the system response can be even more complex. For example, in the case of low carrier concentrations and resonance frequency $\omega = 0.085\ E_0/\hbar$, the induced dipole field is rotated with respect to the static limit (Fig. 5.10(f)). Hence, by tuning the frequency of the external field, discrete quantum states within anisotropic nanostructures can be either accessed or avoided. This control exposes many quantum functionalities that are beyond conventional Mie theory.

As an example, let us consider the induced electric field at the interface between a sharp tip and a flat surface illuminated by plane-wave radiation incident in the z direction (Fig. 5.11(a)). The total material volume of the system is $V_{\text{mat}} = 33404\ L^3$.

In the static limit, a moderate enhancement of field intensity is found in close proximity to the tip. Results of calculating logarithm of the induced field intensity ($\log_{10}(|\mathbf{E}_{\text{ind}}(\mathbf{r})|^2)$) are shown in Fig. 5.11(c). While most of the induced charge density accumulates at the surface, there is considerable penetration into the bulk. In contrast, when the same structure is illuminated at resonance, the induced plasmonic response can be much more intense than the low-frequency limit. Here, the induced field intensity increases by two orders of magnitude, while the spatial dispersion of the "hot spot", as measured by full width half maximum, remains approximately

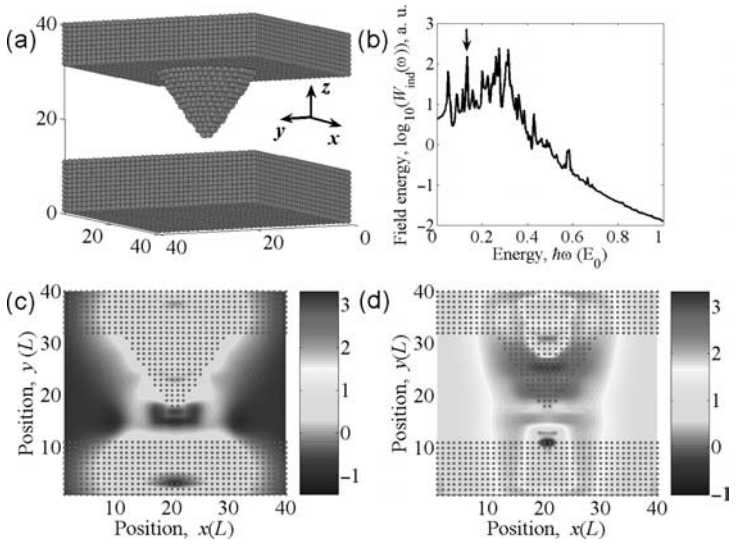

Fig. 5.11. (a) Three-dimensional image of a structure consisting of a sharp tip and two parallel flat plates of material volume $V_{\rm mat} = 33404\ L^3$ containing $N_{\rm el} = 207$ electrons. The incident electric field is along the $(0, 0, 1)$ direction, temperature $T = 0$ K, and $\gamma = 10^{-3}\ E_0$. (b) Logarithm of the induced electric field energy in the structure as a function of the field frequency, the arrow shows the frequency chosen for results of calculations shown in (d). (c) Logarithm of the intensity of the induced electric field ($\log_{10}(|\mathbf{E}_{\rm ind}(\mathbf{r})|^2)$) in static limit, $\hbar\omega = 0$. (d) At resonance, $\hbar\omega = 0.135\ E_0$. Ref. [17].

the same (see Fig. 5.11(d)). This demonstrates that custom-designed atomic-scale tip structures may be used to control the near field, but fully quantum-mechanical modeling is needed to quantitatively account for local enhancements, as well as screening and focusing effects.

In summary, a non-local response theory describing the interaction of electromagnetic radiation with inhomogeneous nanoscale structures has been developed and implemented. This model may be used to describe metals and semiconductors in the metallic regime. For simple geometries, semi-empirical continuum-field Mie theory is found to break down beyond a critical coarseness. At this level of coarse graining, the response is no longer exclusively determined by particle shape. In more complex inhomogeneous geometries excitations of quantum and classical character can coexist. For objects with nanoscale sharp features, a real-space response theory uncovers new functionalities, such as local resonances that are activated by tuning the carrier concentration or the frequency of the incident field.

5.7 Optimization

Next we describe studies to explore optimal design of nanoscale metallic structures to control electromagnetic field intensity on subwavelength scales. We are motivated by the nonintuitive nature of quantum response and the potential for applications such as surface enhanced Raman scattering [5].

5.7.1 Static response

To gain better understanding and intuition into the dielectric response of nanoscale metallic clusters in the quantum limit, let us first consider systems consisting of two identical nanospheres with a total number of electrons $N_{\rm el}$ and separated by an adjustable distance D (top view shown in Fig. 5.12). The nanospheres are placed in a static electric field, and the z direction of the external field is aligned along the line connecting the sphere centers. Our goal is to maximize the intensity of the induced electric field $W_{\rm ind} = \int_{V_0} |\mathbf{E}(\mathbf{r})|^2 d\mathbf{r}$ in an objective volume of radius $R_{\rm V} = L$, centered between the two clusters, by varying the cluster separation D. In the regime of large electron densities, classical theory predicts that $W_{\rm ind}$ diverges as the spheres approach each other, i.e. $W_{\rm ind} \to \infty$ as $D \to 0$, and hence the spherical clusters would need to be as close as possible to each other to maximize the induced field in the target area. As seen in Fig. 5.12, this is no longer true for small carrier concentrations (here $N_{\rm el} = 20$), in which case quantum fluctuations strongly influence the electromagnetic response. For sufficiently small separations (Fig. 5.12(a) and (b)) the entire system responds as a single dipole (with small corrections at the interface between the two clusters). The charge density distribution depicted in Fig. 5.12(a) shows charge polarization (lower: positive, upper: negative) along the applied field, whereas the corresponding induced electric field in Fig. 5.12(b) remains relatively homogeneous throughout the entire system. Remarkably, as shown in Fig. 5.12(c) and (d), there are *finite optimum separations* between the spheres which maximize the induced field at the center between them. As will become apparent, the physical reason for this is that quantum wave functions constrain accessibility to geometric features. For the parameters chosen in this example, $D_{\rm opt} \approx 4L$ occurs at the threshold separation distance beyond which the two clusters cease to act as a single dipole. It is evident from Fig. 5.12(d) that at this resonance condition the overall induced field intensity is highly inhomogeneous and peaks at two orders of magnitude larger compared to off-resonance conditions. Moreover, there can be further such resonances, e.g. at $D_{\rm opt} \approx 7L$ for the present parameters, which maximize the induced

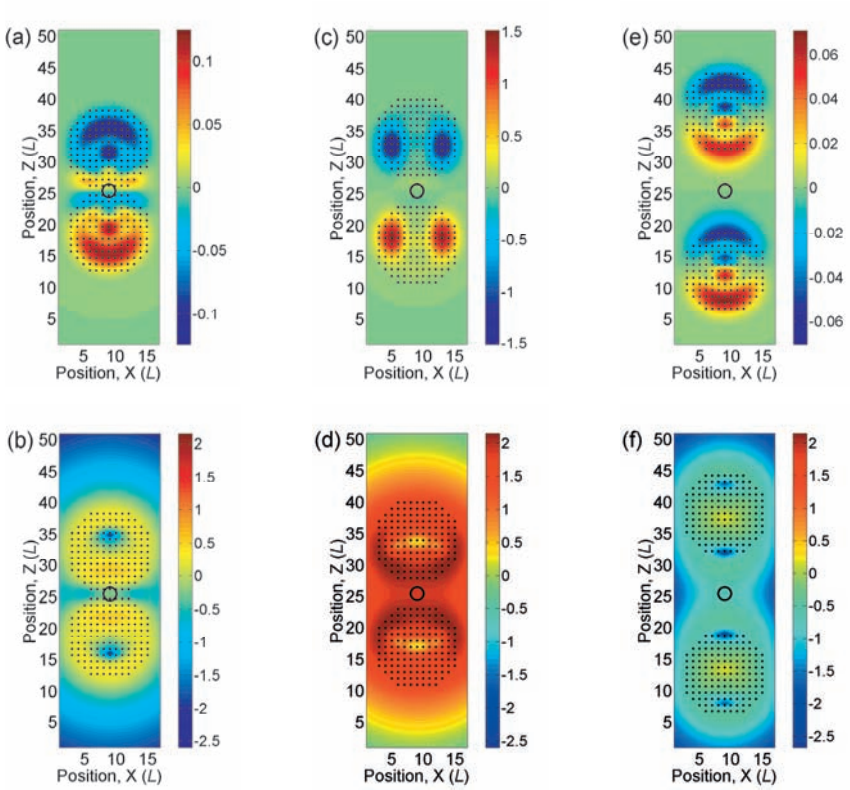

Fig. 5.12. The direction of the static external field $\mathbf{E}_{\rm ext}(\omega = 0)$ points along the z axis. Induced charge density (upper row) and corresponding induced electric field $\log_{10}(|\mathbf{E}_{\rm ind}(x, 0, z)|^2)$ (lower row) in systems of two conducting spherical clusters with radii $R = 7L$ (shown by black dots), where L is the lattice spacing. The plots show top views of the three-dimensional systems. The relative distance D between the two closest points of the spheres is varied from left to right: (a),(b): $D = 0$; (c),(d): $D = 4L$, (e),(f): $D = 13L$. $N_{\rm el} = 20$ for the system and damping constant $\gamma = 2 \times 10^{-3}\ E_0$. The objective volume over which the field intensity is to be maximized is indicated as a circle of radius L at the system center. Ref. [18].

field intensity in between the nanospheres. As observed in Fig. 5.12(e) and (f), for larger distances D, one can ultimately treat the spheres as independent dipoles, for which the induced field energy $W_{\rm ind}$ scales as D^{λ}, with $\lambda = -8$, which can be verified numerically.

5.7.2 Dynamic response and screening

Although the characteristic Thomas–Fermi screening length is known to increase with decreasing carrier concentration, this quantity only describes the screening of slowly varying potentials. The lower the carrier concentration,

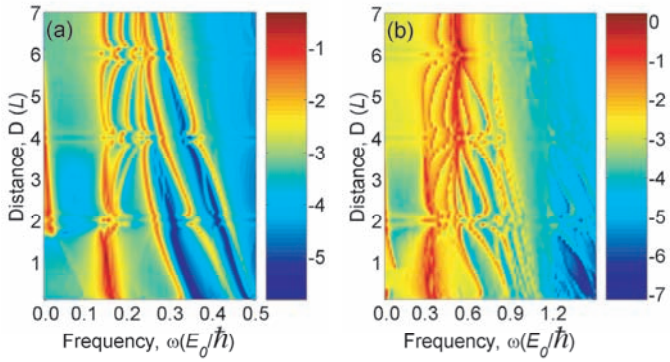

Fig. 5.13. Landscapes of the induced electric field energy $\log_{10}(W_{\rm ind})$ in systems of two conducting nanospheres with radii $R = 7L$ as a function of the frequency of the external field ω and the cluster separation distance D, measured between the two closest points of the nanospheres. The direction of the external field points along the line connecting the sphere centers, Damping constant $\gamma = 2 \times 10^{-3} E_0$. (a) $N_{\rm el} = 20$, (b) $N_{\rm el} = 100$. Ref. [18].

the worse the system screens rapidly varying potentials. For relatively small electron densities the screening length becomes comparable with the distance between the spheres D. Thus, a strong sensitivity of the response of the system to the carrier concentration suggests strong effects of carrier screening. This effect will be most pronounced in the region between the spheres, where the potential undergoes significant changes.

To illustrate how the carrier concentration in the nanospheres dramatically changes the dynamic dielectric response of the system, we show in Fig. 5.13 plots of $W_{\rm ind}$ as a function of the frequency of the external field and relative distance D between the spheres. In Fig. 5.13(a) we consider the case of the same low carrier concentration as in Fig. 5.12. There are $N_{\rm el} = 20$ electrons in the system, with a characteristic Fermi wavelength $\lambda_{\rm F} \approx 3L$, which is the same order of magnitude as the radii of the spherical clusters. The various observed resonances correspond to excitations of different geometric modes available in the discrete spectrum of the system (dipole–dipole, quadruple–quadruple, etc). For the parameters chosen in Fig. 5.13(a), the dominant geometric resonances occur at frequencies less than $\hbar\omega = 0.2E_0$. Interestingly, there is no zero-frequency peak at $D \approx 0$, which is in stark contrast to the case of denser fillings that correspond to the classical limit, e.g. as shown in Fig. 5.13(b). At low fillings, the delocalized charge density response results in less efficient screening in the region between the clusters, and hence the system of two clusters responds as a whole. This significantly reduces the magnitude of the induced charge densities near the

closest surfaces of the spheres and limits the maximum possible value of W_{ind}. Moreover, quantum discreteness of the energy levels results in a non-monotonic dependence of W_{ind} on D, which in turn leads to the *non-zero optimal distance* D_{opt}, discussed above.

Note that for the parameter set chosen here, at a finite frequency $\omega = 0.1\, E_0/\hbar$ the optimal distance is near $D = 0$, i.e. similar to the static response of the classical system. In Fig. 5.13(b) we consider the same system parameters but at a higher carrier concentration, i.e. $N_{\mathrm{el}} = 100$ electrons. The corresponding characteristic Fermi wavelength is $\lambda_{\mathrm{F}} \approx 1L$, which renders the dielectric response of the system to be much closer to the classical limit. In this case many more geometric resonances are observed compared to the low-filling regime (Fig. 5.13(a)). Also, in contrast to the quantum limit these resonances depend more strongly on changes in the relative distance D and converge into a single dominant peak at $\hbar\omega \approx 0.72E_0$ for $D \geq 7L$. Also note the large maximum of W_{ind} at $D \approx 0$ and $\omega \approx 0$, as is expected for the static limit in the classical regime.

5.7.3 Optimal static response

The previous example illustrates that there can be significant differences between the dielectric responses in the classical and quantum regimes. Let us now explore how the quantum functionality of such structures can, at least in principle, be used for the design of nanoscale devices. In the following, we pursue an optimal design problem of a toy system with multiple adjustable parameters, using a numerical global optimization technique based on the genetic algorithm [16]. Specifically, we wish to optimize a system containing five point-like charges with $q = +4e$ each. In order to reduce the complexity of the problem the charges are placed on a line along the z axis, and we optimize the z coordinates of the placed charges. This reduces the optimization problem to five parameters. A static external electric field is applied along the z axis. In order to discretize the numerical search space, the positions of the charges are restricted to be on a lattice with lattice constant L. We search for optimal configurations of the charges that maximize the induced field intensity in a target region V_0, located at the center of the system at $z = L_{\mathrm{tot}}/2$. Here, L_{tot} is the total length of the optimization box along the z direction. The total number of electrons in the system $N_{\mathrm{el}} = 20$ is chosen to insure the system's response to be in the quantum regime.

In Fig. 5.14(a) the intensity of the induced field is shown for the case of all the point charges placed at the center of the target area, as would be suggested by classical intuition. While the induced field is indeed largest at the system center, the overall intensity is relatively small compared to

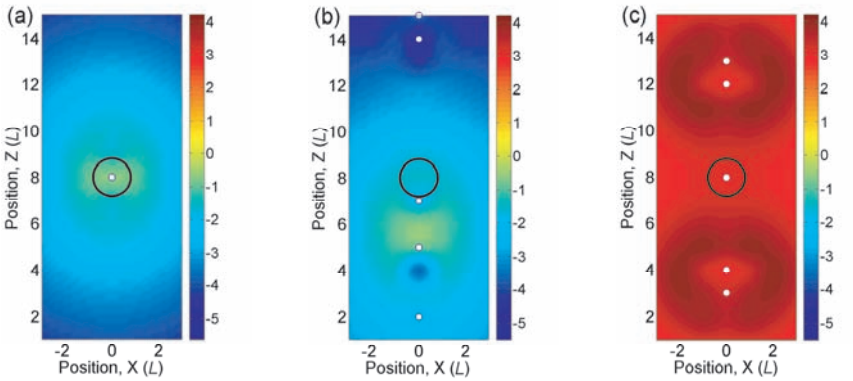

Fig. 5.14. Induced field intensity $\log_{10}(|\mathbf{E}_{\mathrm{ind}}(x,0,z)|^2)$ for different configurations of scatterers: five point-like charges with $q = +4e$ (marked by white dots), each is placed along a line on the z axis, parallel to the external field. $N_{\mathrm{el}} = 20$ and $\gamma = 2 \times 10^{-3} E_0$. (a) All charges are placed at the center, (b) A random configuration, (c) Optimized configuration with objective to yield maximum intensity of induced field. Ref. [18].

optimized configurations. For purposes of comparison, in Fig. 5.14(b) we also show the intensity of the induced field for a random configuration of point charges. In contrast to Fig. 5.14(a,b), Fig. 5.14(c) displays the induced field for a numerically *optimized configuration* of point charges. In agreement with the examples in Fig. 5.12 and 5.13(a), the optimal distance between the placed charges is finite. The optimization algorithm finds a compromise between the distance to the system center and the inter-particle distances of the point charges that maximizes the induced charge density. This is achieved via maximization of the induced charge localization in the quantum system, leading to the most efficient screening near the target volume. Thus, we find that using a genetic algorithm it is possible to create highly efficient optimized structures with broken spatial symmetries which function as a subwavelength lens for electromagnetic radiation.

It is also worth mentioning that the optimal configuration in Fig. 5.14(c) has an inversion symmetry about the system center which arises naturally during the optimization procedure. Comparing the optimized result with the classical configuration (Fig. 5.14(a)) and the random configuration (Fig. 5.14(b)), it is evident that the optimal configuration leads to a field intensity in the focal region that is four orders of magnitude larger. We have performed further optimizations for the positions of two to eight point-like charges with the same objective functionality. Significant even-odd effects are observed in the optimal arrangements. For even numbers of charges, none of the charges should be placed in the system center, whereas for odd

numbers of charges the optimal configuration consists of two symmetrically arranged equal groups of charges, and one charge is placed in the center of the target area.

5.7.4 Optimal dynamic response

We now consider the case of time-varying fields, with the goal to design a "frequency splitter" in the subwavelength limit. In this example we again allow the point-like charges to be placed along the z axis, and search for optimal spatial configurations of the charges which maximize the induced field intensity in a target volume centered at $z = 2L_{\mathrm{tot}}/3$ for a field frequency ω_1, and in a second target volume centered at $z = L_{\mathrm{tot}}/3$ for a second field frequency ω_2. Numerical optimization was performed for ten moving positive background charges with $q = +2e$ each and $N_{\mathrm{el}} = 20$ electrons in the system. In Fig. 5.15, we show the induced field intensity for optimized configurations of charges with two different frequencies (a) $\hbar\omega_1 = 0.09E_0$ and (b) $\hbar\omega_2 = 0.130E_0$. The selectivity of this device can be quantified by the induced field energies W_1 and W_2 in the target volumes 1 and 2 correspondingly, and their ratio at two distinct frequencies ω_1 and ω_2. We find that for the optimized configuration the ratio $W_1/W_2 \approx 0.09$ at ω_1, and $W_2/W_1 \approx 0.07$ at ω_2.

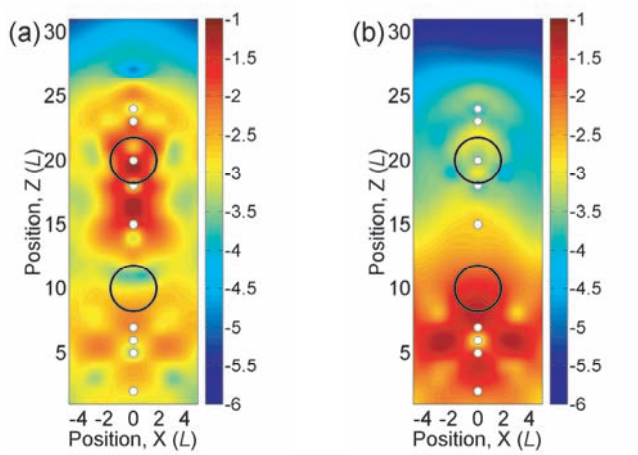

Fig. 5.15. "Frequency splitter": induced field intensities $\log_{10}(|\mathbf{E}_{\mathrm{ind}}(x,0,z)|^2)$ for optimized configurations with $N_{\mathrm{el}} = 20$ electrons and 10 moving scattering centers (marked by white dots), $\gamma = 2 \times 10^{-3} E_0$. The external field points along the z direction. (a) $\hbar\omega_1 = 0.09E_0$, (b) $\hbar\omega_2 = 0.130E_0$. Ref. [18].

5.8 Summary and outlook

The dielectric properties of nanoclusters in their quantum regime offer a richness of functionalities which is absent in the classical case. These include a highly non-trivial screening response and dependence on the frequency of the driving field. Intuition based on classical field theory, e.g. divergence of induced field in the static limit as the distance between the clusters decreases, breaks down, and can thus not be relied on for the design of new atomic-scaled devices. In particular, one cannot expect to induce localized charge density distributions with a characteristic length scale much smaller than the typical Fermi wavelength of the system and collective excitation can be dramatically modified. Moreover, in the quantum regime the delocalized induced charge densities can provide *increased robustness* of the optimized quantity, and thus decrease the complexity of optimal design. Quantum mechanical effects are also found to set boundaries on the maximum values of target quantities, i.e. induced field intensity in the system. Using genetic search algorithms, we have seen that optimal design can lead to field intensities orders of magnitude larger than "simple" guesses based on intuition derived from classical theory.

5.9 References

1. G. Mie, *Beiträge zur optik trüber medien, speziell kolloidaler metallösungen*, Annalen der Physik **330**, 377–445 (1908).

2. D.M. Wood and N.W. Ashcroft, *Quantum size effects in the optical properties of small metallic particles*, Physical Review B **25**, 6255–6274 (1982); M.J. Rice, W.R. Schneider, and S. Strässler, *Electronic polarizabilities of very small metallic particles and thin films*, Physical Review B **8**, 474–482 (1973).

3. Q.P. Li and S. DasSarma, *Elementary excitation spectrum of one-dimensional electron systems in confined semiconductor structures: Zero magnetic field*, Physical Review B **43**, 11768–11786 (1991).

4. D.R. Fredkin and I.D. Mayergoyz, *Resonant behavior of dielectric objects (electrostatic resonances)*, Physical Review Letters **91**, 253902 1–4 (2003).

5. S. Nie and S.R. Emory, *Probing single molecules and single nanoparticles by surface-enhanced Raman scattering*, Science **275**, 1102–1106 (1997).

6. S. Pokrant and K.B. Whaley, *Tight-binding studies of surface effects on electronic structure of CdSe nanocrystals: the role of organic ligands, surface reconstruction, and inorganic capping shells*, European Physical Journal D **6**, 255–267 (1998); P. Chen and K.B. Whaley, *Magneto-optical response of CdSe nanostructures*, Physical Review B

70, 045311 1-12 (2004); J. Schrier and K.B. Whaley, *Atomistic theory of coherent spin transfer between molecularly bridged quanatum dots*, Physical Review B **72**, 085320 1–8 (2005); S. Lee, P. von Allmen, F. Oyafuso, G. Klimeck, and K.B. Whaley, *Effect of electron-nuclear spin interaction on electron-spin qubits localized in self-assembled quantum dots*, Journal of Applied Physics **97**, 043706 1–8 (2005).

7. W.H. Press, B.P. Flanney, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge, United Kingdom, 1988.
8. J. Lindhard, *On the properties of a gas of charged particles*, Kongelige Danske Videnskab, Selskab, Matemat-Fysis Meddel **28** n. 8 (1954).
9. H. Ehrenreich and M.H. Cohen, *Self-consistent field approach to the many-electron problem*, Physical Review **115**, 786–790 (1959).
10. W.A. Harrison, *Electronic Structure and the Properties of Solids*, W. H. Freeman, San Francisco, California, 1980.
11. R.A. Muniz, S. Haas, A.F.J. Levi, and I. Grigorenko, *Plasmonic excitations in tight binding nanostructures*, Physics Review B **80**, 045413 (1–6) (2009).
12. J. Yan, Z. Yuan, and S. Gao, *End and central plasmon resonances in linear atomic chains*, Physical Review Letters **98**, 216602 1–4 (2007).
13. Ke Zhao, M. Claudia Troparevsky, Di Xiao, Adolfo G. Eguiluz, and Zhenyu Zhang, *Electronic coupling and optimal gap size between two metal nanoparticles*, Physical Review Letters **102**, 186804 1–4 (2009).
14. D. Pines, P. Nozieres, *The Theory of Quantum Liquids*, **1**, 55, W.A. Benjamin, Inc., New York, New York, 1966.
15. V.V. Kresin, *Collective resonances and response properties of electrons in metal cluster*, Physics Reports **220**, 1–52 (1992).
16. J. Thalken, W. Li, S. Haas, and A.F.J. Levi, *Adaptive design of excitonic absorption in broken-symmetry quantum wells*, Applied Physics Letters **85**, 121–123 (2004).
17. I. Grigorenko, S. Haas, and A.F.J. Levi, *Electromagnetic response of broken-symmetry nano-scale clusters*, Physical Review Letters **97**, 036806 1–4 (2006).
18. I. Grigorenko, S. Haas, A.V. Balatsky, and A.F.J. Levi, *Optimal control of electromagnetic field using metallic nanoclusters*, New Journal of Physics **10**, 043017 1–4 (2008).

6 Robust optimization in high dimensions

Omid Nohadani and Dimitris Bertsimas

6.1 Introduction

Optimization has a distinguished history in engineering and industrial design. Most approaches, however, assume that the input parameters are precisely known and that the implementation does not suffer any errors. Information used to model a problem is often noisy, incomplete or even erroneous. In science and engineering, measurement errors are inevitable. In business applications, the cost and selling price as well as the demand for a product are, at best, expert opinions. Moreover, even if uncertainties in the model data can be ignored, solutions cannot be implemented to infinite precision, as assumed in continuous optimization. Therefore, an "optimal" solution can easily be sub-optimal or, even worse, infeasible.

There has been evidence illustrating that if errors (in implementation or estimation of parameters) are not taken into account during the design process, the actual phenomenon can completely disappear. A prime example is optimizing the truss design for suspension bridges. The Tacoma Narrows bridge was the first of its kind that was optimized to divert the wind above and below the roadbed [1]. Only a few months after its opening in 1940, it collapsed due to moderate winds which caused twisting vibrational modes. In another example, Ben-Tal and Nemirovski demonstrated that only 5% errors can entirely destroy the radiation characteristics of an otherwise optimized phased locked and impedance matched array of antennas [2]. Therefore, taking errors into account during the optimization process is a first-order effect.

Traditionally, sensitivity analysis was performed to study the impact of perturbations on specific designs and to find solutions that are least sensitive among a larger set of optima. While these approaches can be used to compare different designs, they do not intrinsically find one that is less sensitive, that

is, they do not improve the robustness directly.

Stochastic optimization is the traditional approach to address optimization under uncertainty [3, 4]. The algorithm takes a probabilistic approach. The probability distribution of the uncertainties is estimated and incorporated into the model using:

1. chance constraints (i.e. a constraint which is violated less than $p\%$ of the time) [5],
2. risk measures (e.g. standard deviations, value-at-risk and conditional value-at-risk) [6–10], or
3. a large number of scenarios emulating the distribution [11, 12].

However, the actual distribution of the uncertainties is seldom available. An illustrative example is the demand for a product over the coming week. Any specified probability distribution is, at best, an expert's opinion. Furthermore, even if the distribution is known, solving the resulting problem remains a challenge [13]. For example, a chance constraint is usually "computationally intractable" [14].

Alternatively, in structural optimization, a robust design is achieved through a multi-criteria optimization problem where a minimization of both the expected value and the standard deviation of the objective function is sought using a gradient-based method [15]. Other approaches incorporate uncertainties and perturbations through tolerance bands and margins in the respective multi-objective function while taking constraints into account by adding a penalty term to the original constraints [16].

Robust optimization is another approach towards optimization under uncertainty. Adopting a min-max approach, a robust optimal design is one with the best worst-case performance. Despite significant developments in the theory of robust optimization, particularly over the past decade, a gap remains between the robust techniques developed to date, and problems in the real-world. Most current robust methods are restricted to convex problems such as linear, convex quadratic, conic-quadratic, linear discrete problems [2, 17–19] and convex constrained continuous minimax problems [20]. More recently, a linearization of the uncertainty set allows one to reduce the dependence of the constraints on the uncertain parameters and can provide robust solutions to nonlinear problems [21]. Furthermore, Zhang successfully formulated a general robust optimization approach for nonlinear problems with parameter uncertainties involving both equality and inequality constraints [22]. This approach provides first-order robustness at the nominal value.

However, an increasing number of engineering design problems, besides being non-convex, involve the use of computer-based simulations. In

simulation-based applications, the relationship between the design and the outcome is not defined as functions used in mathematical programming models. Instead, that relationship is embedded within complex numerical models such as partial differential equation (PDE) solvers [23, 24], response surface, radial basis functions [25], and kriging metamodels [26]. Consequently, robust techniques found in the literature cannot be applied to these important sets of practical problems.

In this chapter, we discuss an approach to robust optimization that is applicable to problems whose objective functions are non-convex and given by a numerical simulation driven model, thus directly relevant to engineering design. This technique requires only a subroutine which provides the value of the objective function. Because of its generality, this method is applicable to a wide range of practical problems. Obviously, the method becomes more efficient if gradient information is given as well. We showcase its practicability on different design examples, each of which creates its own perspective on robust optimization techniques.

We emphasize that the robust local search we will discuss in the following improves the robustness directly by reducing the costs of possible worst-case scenarios that may occur when designs are implemented with errors. Moreover, it is analogous to local search techniques, such as gradient descent, which entails finding descent directions and iteratively taking steps along these directions to optimize the nominal cost. This robust local search iteratively takes appropriate steps along descent directions for the robust problem, in order to find robust designs. This analogy continues to hold through the iterations; the robust local search is designed to terminate at a robust local minimum, a point where no improving direction exists. We introduce descent directions and the local minimum of the robust problem; the analogies of these concepts in the optimization theory are important, well studied, and form the building blocks of powerful optimization techniques, such as steepest descent and subgradient techniques. Our proposed framework has the same potential, but for the richer robust problem.

In general, there are two common forms of perturbation: (i) *implementation errors*, which are caused in an imperfect realization of the desired decision variables, and (ii) *parameter uncertainties*, which are due to modeling errors during the problem definition, such as noise. Note that our discussion on parameter errors also extends to other sources of errors, such as deviations between a computer simulation and the underlying model (e.g. numerical noise), stochastic errors, or the difference between the computer model and the meta-model, as discussed by [27]. Even though perturbations (i) and (ii) have been addressed as sources of uncertainty, the case where

both are simultaneously present has not received appropriate attention. For the ease of exposition, in Section 6.2 we first introduce a robust optimization method for generic unconstrained and non-convex problems, in order to minimize the worst-case cost under implementation errors. The case of a restricted search space is illustrated in Section 6.2.3 using an example of robust optimization in electromagnetic scattering problems. If additional information on the restriction can be exploited, the search efficiency can be increased significantly. This is discussed in Section 6.2.4 based on a recent application in ultrafast optics. In Section 6.3, we discuss the presence of constraints, which can be non-convex or convex. The example of a constrained polynomial optimization problem is discussed in Section 6.3.3. We further generalize the method to the case where both implementation errors and parameter uncertainties are present.

6.2 Unconstrained robust optimization

In this section, we illustrate the robust non-convex optimization for problems with implementation errors, as was introduced in Ref. [28, 29]. We discuss the notion of the descent direction for the robust problem, which is a vector that points away from all the worst implementation errors. Consequently, a robust local minimum is a solution at which no such direction can be found.

6.2.1 Problem definition

The nominal cost function, possibly non-convex, is denoted by $f(\mathbf{x})$, where $\mathbf{x} \in \mathrm{R}^n$ is the design vector. The *nominal optimization problem* is

$$\min_{\mathbf{x}} f(\mathbf{x}). \tag{6.1}$$

For the ease of exposition, we first discuss the robust optimization method for implementation errors only, as they may occur during the fabrication process. When implementing $\mathbf{x}$, additive implementation errors $\Delta\mathbf{x} \in \mathrm{R}^n$ may be introduced due to an imperfect realization process, resulting in a design $\mathbf{x} + \Delta\mathbf{x}$. Here, $\Delta\mathbf{x}$ is assumed to reside within an uncertainty set

$$\mathcal{U} := \{\Delta\mathbf{x} \in \mathrm{R}^n \mid \|\Delta\mathbf{x}\|_2 \leq \Gamma\}. \tag{6.2}$$

Note that $\Gamma > 0$ is a scalar describing the size of perturbation against which the design needs to be protected. While our approach applies to other norms $\|\Delta\mathbf{x}\|_p \leq \Gamma$ in Eq. (6.2) (p being a positive integer, including $p = \infty$), we present the case of $p = 2$. We seek a robust design $\mathbf{x}$ by minimizing the

worst-case cost

$$g(\mathbf{x}) := \max_{\Delta\mathbf{x}\in\mathcal{U}} \; f(\mathbf{x}+\Delta\mathbf{x}). \tag{6.3}$$

The worst-case cost $g(\mathbf{x})$ is the maximum possible cost of implementing $\mathbf{x}$ due to an error $\Delta\mathbf{x}\in\mathcal{U}$. Thus, the *robust optimization problem* is given through

$$\min_{\mathbf{x}} g(\mathbf{x}) \equiv \min_{\mathbf{x}} \max_{\Delta\mathbf{x}\in\mathcal{U}} f(\mathbf{x}+\Delta\mathbf{x}). \tag{6.4}$$

In other words, the robust optimization method seeks to minimize the worst-case cost. When implementing a certain design $\mathbf{x}=\hat{\mathbf{x}}$, the possible realization due to implementation errors $\Delta\mathbf{x}\in\mathcal{U}$ lies in the set

$$\mathcal{N} := \{\mathbf{x} \mid \|\mathbf{x}-\hat{\mathbf{x}}\|_2 \le \Gamma\}. \tag{6.5}$$

We call $\mathcal{N}$ the *neighborhood* of $\hat{\mathbf{x}}$; such a neighborhood is illustrated in Fig. 6.1(a). A design $\mathbf{x}$ is a *neighbor* of $\hat{\mathbf{x}}$ if it is in $\mathcal{N}$. Therefore, $g(\hat{\mathbf{x}})$ is the maximum cost attained within $\mathcal{N}$. Let $\Delta\mathbf{x}^*$ be one of the worst implementation errors at $\hat{\mathbf{x}}$, $\Delta\mathbf{x}^* = \arg\max_{\Delta\mathbf{x}\in\mathcal{U}} f(\hat{\mathbf{x}}+\Delta\mathbf{x})$. Then, $g(\hat{\mathbf{x}})$ is given by $f(\hat{\mathbf{x}}+\Delta\mathbf{x}^*)$. Since we seek to navigate away from all the worst implementation errors, we define the *set of worst implementation errors* at $\hat{\mathbf{x}}$

$$\mathcal{U}^*(\hat{\mathbf{x}}) := \left\{ \Delta\mathbf{x}^* \mid \Delta\mathbf{x}^* = \arg\max_{\Delta\mathbf{x}\in\mathcal{U}} \; f(\hat{\mathbf{x}}+\Delta\mathbf{x}) \right\}. \tag{6.6}$$

6.2.2 Robust local search algorithm

Given the set of worst implementation errors, $\mathcal{U}^*(\hat{\mathbf{x}})$, a descent direction can be found efficiently by solving the following second-order cone program (SOCP):

$$\begin{aligned} &\min_{\mathbf{d},\beta} \beta \\ &s.t.\ \|\mathbf{d}\|_2 \le 1, \\ &\quad \mathbf{d}'\Delta\mathbf{x}^* \le \beta \quad \forall \Delta\mathbf{x}^* \in \mathcal{U}^*(\hat{\mathbf{x}}), \\ &\quad \beta \le -\epsilon, \end{aligned} \tag{6.7}$$

where ϵ is a small positive scalar. A feasible solution to Problem (6.7), $\mathbf{d}^*$, forms the maximum possible angle $\theta_{\max}$ with all $\Delta\mathbf{x}^*$. An example is illustrated in Fig. 6.1(b). This angle is always greater than 90° due to the constraint $\beta \le -\epsilon < 0$. When ϵ is sufficiently small, and Problem (6.7) is infeasible, then $\hat{\mathbf{x}}$ is a good estimate of a robust local minimum. Note that the constraint $\|\mathbf{d}^*\|_2 = 1$ is automatically satisfied if the problem is feasible.

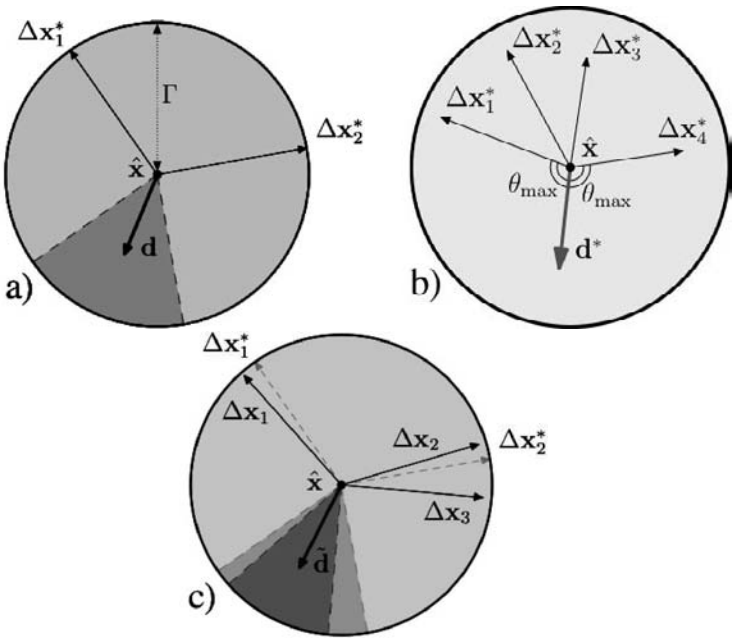

Fig. 6.1. (a) A two-dimensional illustration of the neighborhood. For a design $\hat{\mathbf{x}}$, all possible implementation errors $\Delta\mathbf{x} \in \mathcal{U}$ are contained in the shaded circle. The bold arrow $\mathbf{d}$ shows a possible descent direction and thin arrows $\Delta\mathbf{x}_i^*$ represent worst errors. (b) The solid arrow indicates the optimal direction $\mathbf{d}^*$ which makes the largest possible angle $\theta_{\max} = \cos^{-1}\beta^* \geq 90^\circ$ with all $\Delta\mathbf{x}^*$. (c) Without knowing all $\Delta\mathbf{x}^*$, the direction $\tilde{\mathbf{d}}$ points away from all $\Delta\mathbf{x}_j \in \mathcal{M} = \{\Delta\mathbf{x}_1, \Delta\mathbf{x}_2, \Delta\mathbf{x}_3\}$, when all $\mathbf{x}_i^*$ lie within the cone spanned by $\Delta\mathbf{x}_j$.

Such an SOCP can be solved efficiently using both commercial and non-commercial solvers.

Consequently, if we have an oracle returning $\mathcal{U}^*(\mathbf{x})$, we can iteratively find descent directions and use them to update the current iterates. In most engineering design problems, however, we cannot expect to find $\Delta\mathbf{x}^*$. Therefore, an alternative approach is required. In Ref. [28], it was argued that descent directions can be found without knowing the worst implementation errors $\Delta\mathbf{x}^*$ exactly. As illustrated in Fig. 6.1(c), finding a set $\mathcal{M}$, such that all the worst errors $\Delta\mathbf{x}^*$ are confined to the sector demarcated by $\Delta\mathbf{x}_i \in \mathcal{M}$, would suffice. The set $\mathcal{M}$ does not have to be unique. If this set satisfies condition:

$$\Delta\mathbf{x}^* = \sum_{i|\Delta\mathbf{x}_i \in \mathcal{M}} \alpha_i \Delta\mathbf{x}_i, \tag{6.8}$$

the cone of descent directions pointing away from $\Delta\mathbf{x}_i \in \mathcal{M}$ is a subset of the cone of directions pointing away from $\Delta\mathbf{x}^*$. Because $\Delta\mathbf{x}^*$ usually reside

among designs with nominal costs higher than the rest of the neighborhood, the following algorithm summarizes a strategy for the robust local search:

Algorithm 1

Step 0. *Initialization: Let* $\mathbf{x}^1$ *be an arbitrarily chosen initial decision vector. Set* $k := 1$.

Step 1. *Neighborhood exploration* :

Find $\mathcal{M}^k$*, a set containing implementation errors* $\Delta\mathbf{x}_i$ *indicating where the highest cost is likely to occur within the neighborhood of* $\mathbf{x}^k$*. For this, we conduct multiple gradient ascent sequences. The results of all function evaluations* $(\mathbf{x}, f(\mathbf{x}))$ *are recorded in a history set* $\mathcal{H}^k$*, combined with all past histories. The set* $\mathcal{M}^k$ *includes elements of* $\mathcal{H}^k$ *which are within the neighborhood and have highest costs.*

Step 2. *Robust local move* :

(i) Solve an SOCP (similar to problem (6.7), but with the set $\mathcal{U}^*(\mathbf{x}^k)$ *replaced by set* $\mathcal{M}^k$*); terminate if the problem is infeasible.*

(ii) Set $\mathbf{x}^{k+1} := \mathbf{x}^k + t^k\mathbf{d}^*$*, where* $\mathbf{d}^*$ *is the optimal solution to the SOCP.*

(iii) Set $k := k + 1$*. Go to step 1.*

Note that while we employ a geometric approach to motivate for the descent direction $\mathbf{d}$, it is actually determined based on a directional derivative framework, as was discussed in detail in Ref. [28] along with a thorough mathematical analysis, proofs of convergence, and an extension to problems with parameter uncertainties. With this new framework, the method above can serve as a building block for other gradient-based optimization routines, allowing a larger range of applicability.

In the following, we demonstrate two example applications of the robust local search algorithm for non-convex search spaces. In the first case of an electromagnetic multi-scattering problem, the search space in the uncertainty set $\mathcal{U}$ in Eq. (6.2) is restricted. In the second application in designing chirped mirrors for ultrafast optics, the search space is initially restricted. However, using additional features of the problem, we show how to transform into an unrestricted space, where the optimization can be conducted significantly more efficiently leading not only to a more robust solution, but also to a higher manufacturing yield.

6.2.3 Example in electromagnetic scattering

The search for attractive and novel materials in controlling and manipulating electromagnetic field propagation has identified a plethora of unique characteristics in photonic crystals (PCs). Their novel functionalities are based on diffraction phenomena, which require periodic structures. Upon breaking the spatial symmetry, new degrees of freedom are revealed which allow for additional functionality and, possibly, for higher levels of control. More recently, unbiased optimization schemes were performed on the spatial distribution (aperiodic) of a large number of identical dielectric cylinders [30, 31]. While these works demonstrate the advantage of optimization, the robustness of the solutions still remains an open issue. In this section, we apply the robust optimization method to electromagnetic scattering problems with large degrees of freedom, and report on novel results when this technique is applied to optimization of aperiodic dielectric structures.

6.2.3.1 Problem description

The incoming electromagnetic field couples in its lowest mode to the perfectly conducting metallic waveguide. Figure 6.2(a) sketches the horizontal set-up. In the vertical direction, the domain is bounded by two perfectly conducting plates, which are separated by less than 1/2 the wavelength, in order to warrant a two-dimensional wave propagation. Identical dielectric cylinders are placed in the domain between the plates. The sides of the domain are open in the forward direction. In order to account for a finite total energy and to warrant a realistic decay of the field at infinity, the open sides are modeled by perfectly matching layers [32, 33]. The objective of the optimization is to determine the position of the cylinders such that the forward electromagnetic power matches the shape of a desired power distribution, as shown in Fig. 6.2(b).

As in the experimental measurements, the frequency is fixed to $f = 37.5$ GHz [31]. Furthermore, the dielectric scatterers are nonmagnetic and lossless. Therefore, stationary solutions of the Maxwell equations are given through the two-dimensional Helmholtz equations, taking the boundary conditions into account. This means that only the z-component of the electric field E_z can propagate in the domain. The magnitude of E_z in the domain is given through the partial differential equation (PDE)

$$(\partial_x(\mu_{r_y}^{-1}\partial_x) + \partial_y(\mu_{r_x}^{-1}\partial_y))E_z - \omega_0^2\mu_0\epsilon_0\epsilon_{r_z}E_z = 0, \tag{6.9}$$

with μ_r the relative and μ_0 the vacuum permeability. Note that ϵ_r denotes the relative and ϵ_0 the vacuum permittivity. Equation (6.9) is numerically

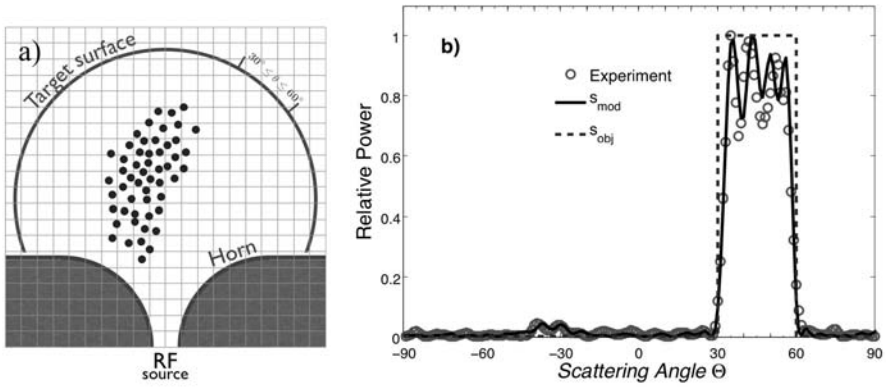

Fig. 6.2. (a) Schematic setup: the RF-source couples to the waveguide. Circles denote the positions of scattering cylinders for a desired top hat power profile. The unshaded grid depicts the domain. (b) Comparison between experimental data (circles) [31] and modeled predictions.

determined using an evenly meshed square-grid (x_i, y_i). The resulting finite-difference PDE approximates the field $E_{z,i,j}$ everywhere inside the domain including the dielectric scatterers. The imposed boundary conditions (Dirichlet conditions for the metallic horn and perfectly matching layers) are satisfied. This linear equation system is solved by ordering the values of $E_{z,i,j}$ of the PDE into a column vector. Hence, the finite-difference PDE can be rewritten as

$$\mathbf{L} \cdot \mathbf{E}_z = \mathbf{b}, \tag{6.10}$$

where $\mathbf{L}$ denotes the finite-difference matrix, which is complex-valued and sparse, $\mathbf{E}_z$ describes the complex-valued electric field that is to be computed, and $\mathbf{b}$ contains the boundary conditions. With this, the magnitude of the field at any point of the domain can be determined by solving the linear system of Eq. (6.10).

The power at any point on the target surface $(x(\theta), y(\theta))$ for an incident angle θ is computed through interpolation using the nearest four mesh points and their standard gaussian weights $\mathbf{W}(\theta)$ with respect to $(x(\theta), y(\theta))$ as

$$s_{\mathrm{mod}}(\theta) = \frac{\mathbf{W}(\theta)}{2} \cdot \mathrm{diag}(\mathbf{E_z}) \cdot \mathbf{E_z}. \tag{6.11}$$

In the numerical implementation, we exploited the sparsity of $\mathbf{L}$, which improved the efficiency of the algorithm significantly. In fact, the solution of a realistic forward problem ($\sim 70,000 \times 70,000$ matrix), including 50 dielectric scatterers, requires about 0.7 second on a commercially available

Intel Xeon 3.4 GHz. Since the size of $\mathbf{L}$ determines the size of the problem, the computational efficiency of our implementation is independent of the number of scattering cylinders.

To verify this finite-difference technique for the power along the target surface (radius = 60 mm from the domain center), we compared our simulations with experimental measurements from Ref. [31] for the same optimal arrangement of 50 dielectric scatterers ($\epsilon_r = 2.05$ and 3.175 ± 0.025 diameter). Figure 6.2(b) illustrates the good agreement between experimental and model data on a linear scale for an objective top hat function.

In the optimization problem, the design vector $\mathbf{x} \in \mathrm{R}^{100}$ describes the positions of the 50 cylinders. For a given $\mathbf{x}$ in the domain, the power profile s_{mod} over discretized angles on the target surface, θ_k, is computed. We can thus evaluate the objective function

$$f_{\mathrm{EM}}(\mathbf{x}) = \sum_{k=1}^{m} |s_{\mathrm{mod}}(\theta_k) - s_{\mathrm{obj}}(\theta_k)|^2. \tag{6.12}$$

Note that $f(\mathbf{x})$ is not a direct function of $\mathbf{x}$ and not convex in $\mathbf{x}$. Furthermore, using the adjoint technique, our implementation provides the cost function gradient $\nabla_{\mathbf{x}} f_{\mathrm{EM}}(\mathbf{x})$ at no additional computational expense. We refer interested readers to Ref. [31] for a more thorough discussion of the physical problem.

Because of the underlying Helmholtz equation, the model scales with frequency and can be extended to nano-photonic designs. While degradation due to implementation errors is already significant in laboratory experiments today, it will be amplified under nanoscale implementations. Therefore, there is a need to find designs that are robust against implementation errors. Thus, analogous to the general robust problem in Eq. (6.4), the specific robust optimization problem is defined as

$$\min_{\mathbf{x}\in\mathcal{X}} g_{\mathrm{EM}}(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\Delta\mathbf{x}\in\mathcal{U}} f_{\mathrm{EM}}(\mathbf{x} + \Delta\mathbf{x}). \tag{6.13}$$

In this setting, $\Delta\mathbf{x}$ represents displacement errors of the scattering cylinders.

6.2.3.2 Computation results

We first construct the uncertainty set $\mathcal{U}$ to include most of the errors expected, in analogy to Eq. (6.2). In laboratory experiments, the implementation errors $\Delta\mathbf{x}$ are observed to have a standard deviation of 40 μm. Therefore, to define an uncertainty set incorporating 99% of the perturbations, i.e. $P(\Delta\mathbf{x} \in \mathcal{U} = 99\%)$, we define

$$\mathcal{U} = \{\Delta\mathbf{x} \mid \|\Delta\mathbf{x}\|_2 \leq \Gamma = 550\ \mu\mathrm{m}\}, \tag{6.14}$$

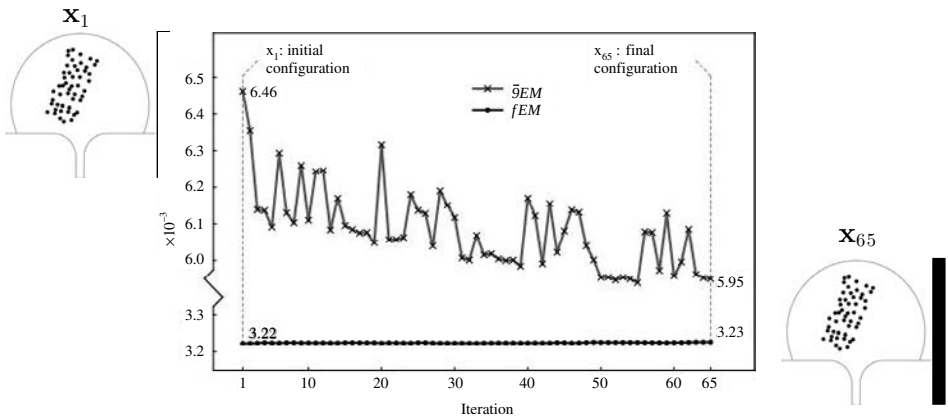

Fig. 6.3. Performance of the robust local search algorithm for the electromagnetic scattering problem. The initial cylinder configuration, $\mathbf{x}_1$, and final configuration, $\mathbf{x}_{65}$, are shown outside the line plot. While the differences between the two configurations seem negligible, the worst-case cost of $\mathbf{x}_{65}$ is 8% lower than that of $\mathbf{x}_1$. The nominal costs of the two configurations are practically the same.

where Δx_i is assumed to be independently and normally distributed with mean 0 and standard deviation 40 μm.

Note that $\mathcal{U}$ defines a restricted search space, both with respect to the condition $\|\Delta\mathbf{x}\|_2 \leq \Gamma$, as well as with respect to the non-overlapping condition. The latter is imposed to maintain the dimensionality of the problem by requiring the distance between two cylinders to be at least $\lambda/2$. Therefore, the inner maximization problem to compute $g_{\mathrm{EM}}(\mathbf{x})$ becomes a constraint maximization problem that is computationally more intensive than conventional unconstraint optimization problems.

The standard procedure used to address Problem (6.13) is to find an optimal design minimizing Eq. (6.12). Subsequently, the sensitivity of the optimal design to implementation errors will be assessed through random simulations. However, because the problem is highly non-convex and of high dimension, there is, to the best of our knowledge, no approach to find a design that is less sensitive to implementation errors without using safety margins. Nevertheless, a design that minimizes $f_{\mathrm{EM}}(\mathbf{x})$ locally is readily found by applying a gradient-related descent algorithm [29, 31]. We applied the discussed robust local search algorithm using such a local minimum as the starting point.

Figure 6.3 shows that the robust local algorithm finds a final design $\mathbf{x}^{65}$ with a worst-case cost that is 8% lower than that of $\mathbf{x}^1$. Throughout the iterations, the nominal cost remains practically the same. The worst-case cost of $\mathbf{x}^{65}$ was estimated with 110,000 neighbors in its neighborhood.

Note, however, that these 110,000 neighbors were not evaluated in a single neighborhood exploration. Instead, more than 95% of them were performed in the iterations leading up to the iteration 65. The estimated worst-case cost at each iteration also shows a downward trend: as the iteration count increases, the knowledge about the problem grows and more robust designs are discovered.

Since we can only estimate the worst-case cost, there is always a chance for late discoveries of worst implementation errors. Therefore, the decrease of the estimated worst-case cost may not be monotonic. Finally, note that the initial design $\mathbf{x}^1$ may already have inherent robustness to implementation errors because it is a local minimum, i.e. with all factors being equal, a design with a low nominal cost will have a low worst-case cost.

We also observed that the neighborhood exploration strategy is more efficient in assessing the worst-case cost when compared to random sampling as is the standard today in perturbation analysis. For example, when estimating $\tilde{g}_{\mathrm{EM}}(\mathbf{x}^1)$, the best estimate attained by random sampling after 30,000 function evaluations using the numerical solver is 96% of the estimate obtained by our algorithm using only 3,000 function evaluations. This is not surprising since finding the optimal solution to the inner optimization problem by random searches is usually inferior to applying gradient-related algorithms using multiple starting points.

In the following, we illustrate an example in ultrafast optics that deals with optimizing double chirped mirrors with a large number of coating layers. In this application of the robust optimization, the search space is initially restricted. However, using additional features of the problem, we demonstrate how to transform the search space into an unrestricted space, where the optimization can be conducted significantly more efficiently leading not only to a more robust solution, but also to a higher manufacturing yield.

6.2.4 Example in chirped mirrors

The dispersion compensating mirror, first proposed in 1994 [34], has become an enabling technology for modern ultrafast lasers. Solid-state mode-locked lasers can only operate at or below few-cycle pulse widths when the total cavity dispersion is reduced to nearly zero, with only a small amount (on the order of a few fs^2) of residual second-order dispersion. While prisms can be used to compensate for second- and third-order cavity dispersion, their relatively high loss and inability to compensate for arbitrary dispersion limits their use; pulse durations below ten femtoseconds were not possible directly from oscillators until the development of high performance double-chirped mirror pairs [35–37].

As bandwidths increase, so does the number of layers required to produce a mirror with the high reflectivity needed for an intra-cavity mirror. For bandwidths exceeding an octave, mirror pairs with over 200 total layers are generally required. The sensitivity of a dielectric stack to manufacturing errors increases with the number of layers, and dispersion compensating mirrors push the limits of manufacturing tolerances, requiring layer precisions on the order of a nanometer. Currently, this challenges even the best manufacturers.

While the nominal optimization of layer thickness has led to successful design of dispersion-compensating dielectric mirrors allowing dispersion and reflectivity control over nearly an octave bandwidth, in practice the performance for such complicated mirrors is limited by the manufacturing tolerances of the mirrors. Small perturbations in layer thickness not only result in sub-optimal designs but, due to the nonlinear nature of mode-locking, such perturbation may completely destroy the phenomenon.

Despite the fact that manufacturing errors often limit the performance of thin-film devices [38], there has been little work on optimizing thin-film designs to mitigate the effects of errors. Some previous work in designing fault-tolerant mirrors has focused on optimizing first-order tolerances, a method readily available in commercial thin-film design codes [39].

Here, we tailor the discussed approach to robust optimization such that it probes the exact merit function in a bounded space of potential thickness errors. While this results in a much more computationally involved optimization, the result is arguably more robust to significant perturbation as the full structure of the merit function is considered in a neighborhood around a nominal solution. Furthermore, the robustness is guaranteed to be equal to or better than that obtained with nominal optimization and, in the case where it is equal, no sacrifice in nominal optimality will be made.

Other prior work was done by Yakovlev and Tempea [40], who employed stochastic global optimization to achieve robustness of the final solution by virtue of the fact that they optimized a Monte Carlo computed integral over a neighborhood around a nominal design. Their method does not suffer the limitations of first-order tolerances, and was able to produce mirrors with significant improvement over nominally optimized designs, demonstrating conclusively that robustness can be greatly improved at the design level by proper optimization.

As in the previous example in Section 6.2.3, the robust design of double chirped mirrors involves objectives and constraints that are not explicitly given and highly non-convex. Thus, no internal structure can be exploited. Here, we illustrate a tailored version of the deterministic robust optimization method, as introduced in Section 6.2. This method provides for designs

which are intrinsically protected against potentially significant layer thickness perturbations occurring during manufacture. We cast the algorithm specifically for double chirped mirrors and tailor the parameters to this particular problem. First, we discuss the optical properties of these mirrors and define a cost function based on reflectivity and group delay. We continue with the introduction of the concept of the uncertainty set as well as a novel method to identify worst-case designs within this set. Once these configurations are found, we show how an update direction can be found which eliminates these worst cases using the robust local search, as introduced in Section 6.2. Furthermore, we demonstrate the performance of the nominal and robust solutions for a large range of perturbation and propose a technique to increase the manufacturing yield.

6.2.4.1 Computation of cost function

The merit function for a chirped mirror is typically composed of two terms, one representing the performance of the reflectivity (which is ideally one) and another which quantifies the deviation of the dispersion from ideal. We employ a cost or merit function that is given as

$$f(\mathbf{x}) = \sum_k w_r(\lambda_k)\left[R(\lambda_k;\mathbf{x}) - 1\right]^4 + \sum_k w_d(\lambda_k)\left[\tau_g(\lambda_k;\mathbf{x}) - \hat{\tau}_g(\lambda_k) + \tau_0(\mathbf{x})\right]^2, \tag{6.15}$$

where $R(\lambda;\mathbf{x})$ is the wavelength domain reflectivity of the total mirror pair described by layer thicknesses $\mathbf{x}$, $\tau_g(\lambda;\mathbf{x})$ is the group delay (GD) of the pair, $\hat{\tau}_g(\lambda)$ is the ideal GD, and the $w_{r,d}(\lambda)$ are weighting functions. To account for an irrelevant offset between the computed and ideal group delay curves, we include a constant offset, $\tau_0(\mathbf{x})$, that minimizes the error. For the reflectivity errors, we use the fourth power of the error to approximate a Chebychev norm, though a standard squared error can also be used.

The computation of reflectivity from a thin-film stack is done using transfer matrix methods [41]. In a standard nominal optimization, the merit function and its gradient must be evaluated thousands of times over hundreds of wavelengths. In a robust optimization, the computational burden is even greater, with the merit function typically computed on the order of a million times. Any discrepancy in the gradient will hinder the convergence rate. Thus, it is imperative that the merit function be computed efficiently and accurately. We employ the methods described in [42, 43], where the group delay is computed in an approximate analytic form that allows for a significant reduction in computational complexity. The approximation simply neglects

the local change in wavelength of the Fresnel reflections between each layer. For chirped mirrors, the approximation error is negligible, as demonstrated and explained in [42]. The gradient of the group delay is computed analytically in a self-consistent manner with the approximation, resulting in an optimization that converges quickly, in terms of both iterations and total processing time.

6.2.4.2 Problem statement

This design problem consists of a double chirped mirror pair with 208 layers for use in a few-cycle Titanium:sapphire mode-locked laser [44]. The initial design was computed using the analytic method of [37]. The materials used were SiO_2 and TaO_5, with the dispersion of each modeled using Sellmeier coefficients obtained from fits to manufacturers' index data. The total reflection dispersion of the pair is specified to compensate for 2.2 mm of Titanium:sapphire, 2 meters of air, and 8 mm of barium fluoride in a cavity containing six mirrors. The group delay and reflectivity are optimized over 156 wavelengths, uniformly spaced from 650 to 1200 nanometers. This discretization was empirically found to be sufficient to avoid narrow resonances "leaking" through the grid. The angle of incidence is taken to be six degrees, and the polarization is assumed to be transverse in the magnetic field (TM). The reflectivity and group delay are optimized as in Eq. (6.15), with constant weightings $w_r = 1$ and $w_d = 10^{-8}\,\mathrm{fs}^{-2}$. Figure 6.4 illustrates the experimental arrangement of the Titanium:sapphire mode-locked laser using nominally optimized double chirped mirrors.

6.2.4.3 Implementation errors

For this application, we model manufacturing errors as independent random sources of additive noise, since any known systematic errors, such as miscalibration, can be best addressed in the actual production. As empirically supported, the layer-thickness errors can be regarded as not correlated. Therefore, we assume that when manufacturing a mirror with layer thicknesses given by $\mathbf{x}$, statistically independent additive implementation errors $\Delta\mathbf{x} \in \mathrm{R}^n$ may be introduced due to variation in the coating process, resulting in actual thicknesses $\mathbf{x} + \Delta\mathbf{x}$. We assume a mean of zero and a variance on each layer that is motivated by actual manufacturing errors. Here, $\Delta\mathbf{x}$ resides within an uncertainty set

$$\mathcal{U} := \{\Delta\mathbf{x} \in \mathrm{R}^n \mid \|\Delta\mathbf{x}\|_2 \leq \Gamma\}. \tag{6.16}$$

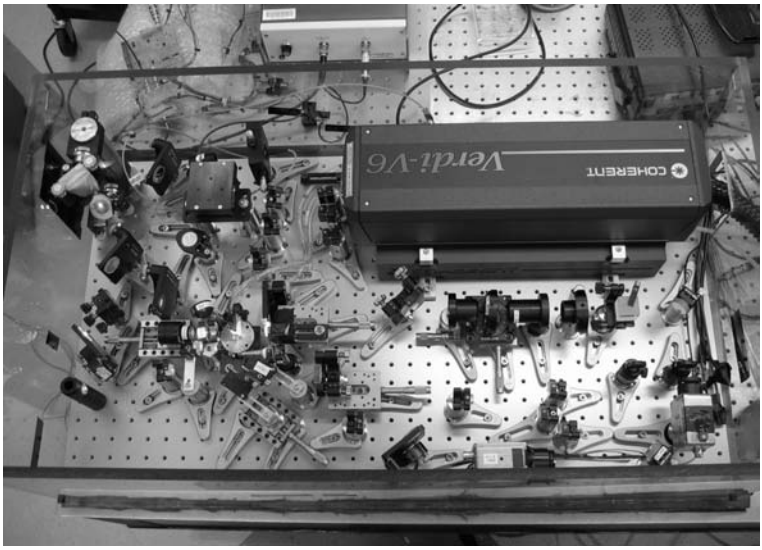

Fig. 6.4. The experimental arrangement. Double chirped mirrors are utilized to compensate the dispersion over an entire octave. (With permission of The Research Laboratory for Electronics at MIT.)

Note that $\Gamma > 0$ is a scalar describing the size of perturbation against which the design needs to be protected. For this paper, we took the manufacturing uncertainty to be normally distributed with a standard deviation of $\sigma = 0.5$ nm. To maintain 95% cumulative confidence to capture all errors within $\mathcal{U}$ for this 208-dimensional problem, we chose $\Gamma = 0.0075$ μm.

6.2.4.4 Restricted search space

To protect a design against errors, it is helpful to utilize available understanding of possible errors. For example, if there are worst-case scenarios in the respective neighborhood that are very rare according to our assumed layer perturbation distribution, there is no need for them to be considered during the inner maximization problem Eq. (6.4). By excluding these rare events from $\mathcal{U}$, we are able to protect the design against realistic and statistically relevant errors only, without needlessly sacrificing nominal performance to guard against rare errors. Moreover, this approach leads to a reduction of the size of the respective search space and, thus, to an increase of the computational efficiency.

It is well known that the reflection coefficients of thin-film stacks are closely related to the Fourier transform of the layer thicknesses [45]. Thus, one promising class of rare perturbations to eliminate from consideration is those which have strong correlations between the layers. These errors involve,

for example, shifting of all the thicknesses in one direction, which results in a spectral shift regardless of the design. Even though such errors may occur in actual manufacturing due to systematic issues, there is little or nothing that can be done to deal with them by design optimization, and to attempt to do so will only result in a highly compromised design. We thus restrict ourselves to considering only statistically independent random perturbations to the layers. In this context, the probability of errors occurring with high correlation between the layers is negligible, and thus we should not concern ourselves with protecting against them. We therefore seek a class of errors which restrict the allowable correlation between layers, i.e. we restrict the maximum variation in the amplitude of the Fourier components of the error vector. A straightforward way to do this is to restrict the search to the class of error vectors with minimum coherence, requiring all Fourier components to have a uniform amplitude.

In addition to the above, this choice of subset is justified empirically. Monte Carlo simulations reveal that the set of perturbations with uniform amplitude in the Fourier domain with uniformly distributed phases has virtually identical statistics to the general uncertainty set $\mathcal{U}$ defined in Eq. (6.16). The cumulative probability distribution of the reduced set never exceeds the full set by more than 4%. This confirms that our worst-case search over the reduced subset will not miss anything statistically relevant in the full set, and thus robustness is not compromised by using this set.

In the restricted space, the components of $\Delta\mathbf{x}$ can be written as

$$\Delta x_j = \frac{\Gamma}{\lfloor N/2 \rfloor} \sum_{k=1}^{\lfloor N/2 \rfloor} \cos\left(\frac{2\pi k j}{N} + \phi_k\right), \tag{6.17}$$

where ϕ_k is the phase of the kth Fourier component of $\Delta\mathbf{x}$ and N is the number of layers. We furthermore assume the constant (zero frequency) component is zero, which corresponds to the aforementioned pathological case of all layers shifting a similar amount. Using *Parseval's theorem*, i.e. the sum of the square of a function is equal to the sum of the square of its transform, we can verify that the magnitude of the errors remains on the shell of the original uncertainty set $\mathcal{U}$,

$$\|\Delta\mathbf{x}\|_2^2 = \sum_{k=1}^{N} |\Delta x_k|^2 = \Gamma^2. \tag{6.18}$$

Using this transformation, we search over the phases ϕ_k for worst-case neighbors. Therefore, the search space dimensionality is reduced to $\lfloor N/2 \rfloor$, hence the efficiency of this algorithm increases by N^2. Most importantly, since the maximization problem is over the free phase-space on the shell

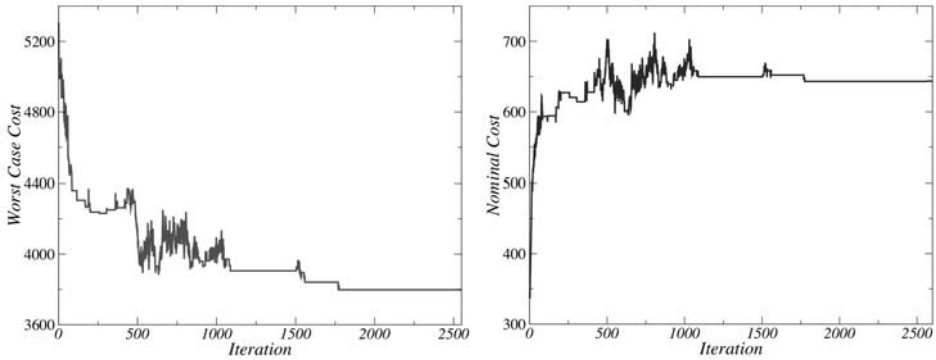

Fig. 6.5. The robust optimization algorithm improves (left) the worst-case cost in the neighborhood of the current design. Discoveries of new bad neighbors cause the small peaks. The price of robustness is an increase in the nominal cost (right).

and the magnitudes of these vectors are constant, the advantages of an unconstrained search can be exploited. Consequently, we obtain the set of local maxima in the phase space using standard gradient-based optimization. Furthermore, we obtain a set of true bad neighbors, which is significantly smaller in size ($\ll 500$) than had we left the search space more general. Since this size determines the number of constraints in the problem, we experience a significant speed up in this part of the algorithm as well.

6.2.4.5 Results

Starting from a nominally optimized solution, we apply the robust optimization algorithm to successively decrease the worst-case cost, as in Eq. (6.4). The performance is shown in Fig. 6.5. The significant improvement of robustness comes at the price of a small increase in the nominal cost. The algorithm converges to the robust local minimum, at which point no descent direction can be found.

The reflectivity and group delay of the robust and nominal optimum are shown in Fig. 6.6. While both solutions satisfy the design objectives, the robust design is significantly more protected against possible errors. The unavoidable "Price of Robustness" through a decrease in the nominal performance of the robust solution is apparent, with increased ripple in the group delay and reflectivity. This price is especially apparent in the bottom plot of Fig. 6.6, which compares the total group delay error for the robust and nominally optimized mirror pairs. However, as will be shown in Fig. 6.8 and 6.9, the robust solution performs better when the layer perturbations are taken into account. Even though the nominally optimized design is able

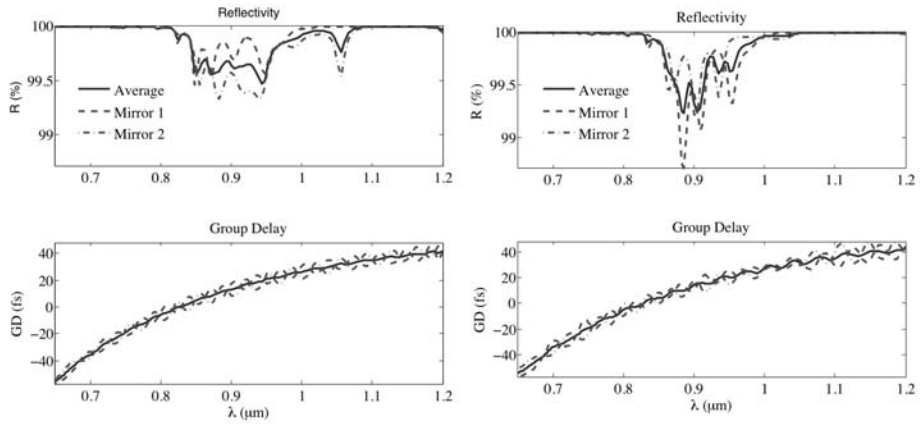

Fig. 6.6. Reflectivity and group delay for each chirped mirror in the pair: (Left) nominally optimal design; (Right) robustly optimal configuration.

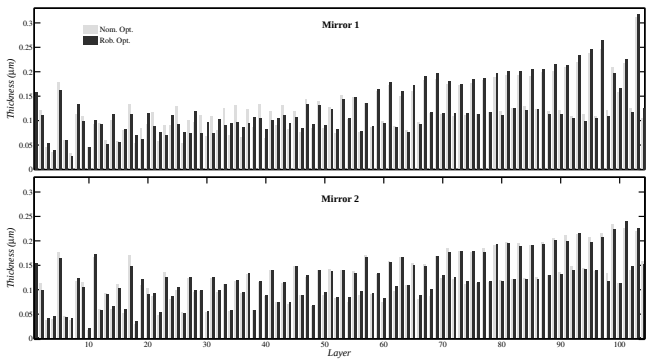

Fig. 6.7. Layer thicknesses of nominal optimum and robust optimum of the mirror pair.

to achieve GD errors of less than one femtosecond, it turns out that the half nanometer layer perturbations we took as our assumed manufacturing tolerances result in GD errors on the order of plus or minus five femtoseconds.

In Fig. 6.7, we show the layer thicknesses for the mirror pair after both nominal optimization and robust optimization. The general structure of the mirror is preserved in the robust optimum solution, in keeping with the observation that its nominal performance is not degraded significantly. The larger variations are found in the first several layers, which perform impedance matching into the chirped stack, suggesting that they are the most sensitive to perturbation. This is consistent with the fact that any spurious reflections off the front surface of the mirror will significantly degrade the GD performance.

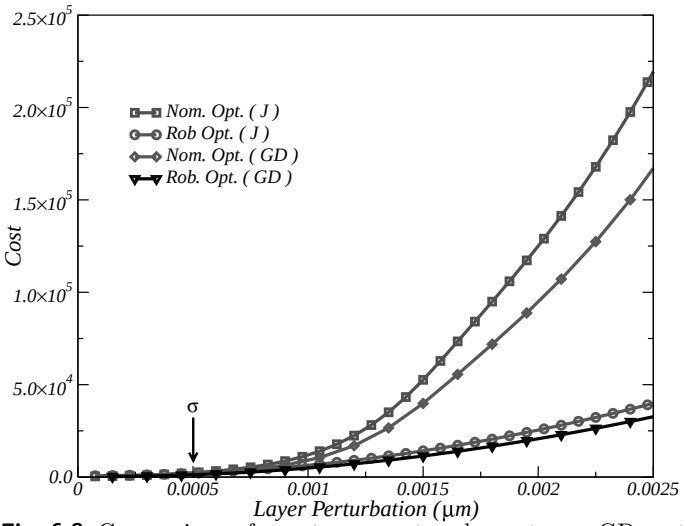

Fig. 6.8. Comparison of worst-case cost and worst-case GD cost of two designs, the nominal and robust optimum, for increasing size of possible perturbations or errors.

While we intended to match the size of the uncertainty to the reported manufacturing and measurement errors, the value of Γ might not fully reflect the actual errors. Therefore, our algorithm seeks to find robust solutions with stable performance even beyond predicted errors. To illustrate these effects, we varied the size of the uncertainty set and evaluated the worst possible neighbor within this neighborhood. The worst-case scenarios of the nominal optimum and robust optimum, in both cost as well as the optical properties, are compared for increasing neighborhood size in Fig. 6.8.

The worst-case performance of both the nominal and robust designs behaves fairly similarly within a small range of perturbations, which is in fact comparable to Γ. However, once the size of possible errors increases, the worst-case cost of the nominal design drastically rises, showing that this design would lose its phenomena completely.

Since any manufacturing process is to some extent statistical, it is essential for a design to give a high manufacturing yield. Our robust optimization method not only minimizes the worst-case performance, but also addresses these statistical effects. This is demonstrated in Fig. 6.9. A series of Monte Carlo simulations, each with 10^6 randomly sampled designs with normally distributed layer perturbations, was performed with varying standard deviations at the robust and the nominal optimum. The mean μ and the 95th percentile P_{95} of the distribution for each perturbation size are plotted to illustrate the center and the actual width of this statistical process. While both designs are similarly distributed within the expected errors σ, they deviate significantly beyond this mark. In fact, the mean and more importantly

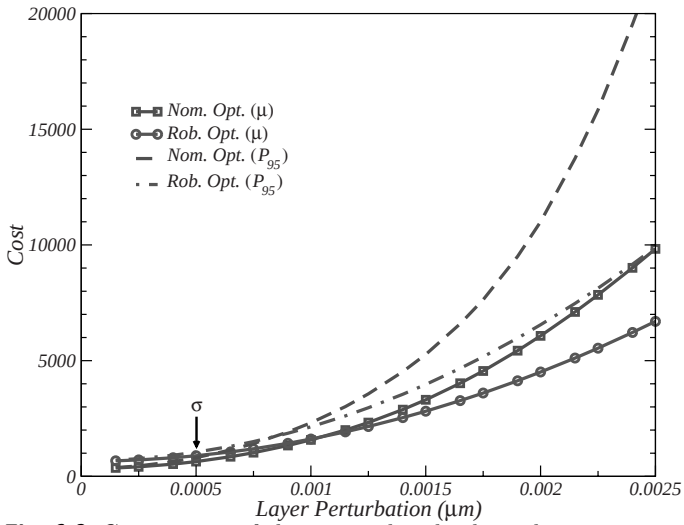

Fig. 6.9. Comparison of the nominal and robust design: mean and 95th percentile of the cost distribution of 10^6 randomly sampled designs for varying perturbation sizes.

the spread of the distribution for the nominal optimum design increases rapidly beyond σ, while the robust optimum is more moderate. Moreover, the mean of the nominal optimum at all perturbation sizes is within the distribution (P_{95}) of the robust optimum, demonstrating that the manufacturing yield of the robust solution remains high and provides performances comparable to the nominal design, even beyond the assumed errors. Since the notion of the actual manufacturing errors is often somewhat uncertain, our method can provide a robust solution despite these uncertainties.

We have applied the method to a demanding optimization of a 208 layer chirped mirror pair with nearly an octave of bandwidth. To avoid taking into account extremely rare potential errors, we perform this search over the space of all errors on the shell of our neighborhood whose components are minimally coherent. This avoids taking into account rare but highly significant errors, such as those associated with certain types of systematic manufacturing errors, that would otherwise dominate the optimization. This modification allows an unconstrained inner maximization over a reduced search space, thus improving the efficiency.

Furthermore, the robust solution compared with that obtained using standard optimization techniques is found to achieve improved statistical performance for layer errors of half a nanometer. Moreover, the fault tolerance of the robust solution increases significantly relative to the nominally optimized mirror as the error variance increases, demonstrating that the robust solution is not tied to the particular manufacturing error variance assumed during optimization (see Fig. 6.9). Therefore, our robust design provides

for a high manufacturing yield even when errors occur that are larger than originally assumed.

This current arrangement of layers, as shown in Fig. 6.7, is currently being manufactured by a third party. Therefore, experimental verification of these results is not possible yet. However, the numerical solver as well as the Monte Carlo error simulations we employed have been verified previously [43]. Hence, we are confident that the presented results on robustness will be verified as well, once the manufactured mirrors are delivered.

Note that in this initial demonstration, we performed the optimization on a fixed number of layers. However, the robust optimization problem can be viewed as providing a new cost function which takes into account robustness and, thus, can be used within other refinement algorithms, such as needle optimization [46], that allow for changing layer counts.

6.3 Constrained robust optimization

In this section, we generalize the unconstrained robust optimization that we introduced in Section 6.2 to accommodate constraints. We will discuss cases where the explicit knowledge of the constraints allows us to significantly improve the optimization by exploiting the structure of the constraints imposed. Our overall goal is to maintain the generality of the method in order to be applicable to a large variety of engineering design problems. For this, we shall continue regarding the problem at hand given to us via a numerical simulation that returns the function value and its gradient for a given design variable [47]. First, we only consider implementation errors as the source of uncertainty. Once this framework is set, we provide an extension to cases where both implementation errors and parameter uncertainties are present.

To illustrate the performance of the constrained robust optimization method, we will provide an example application in constrained polynomial optimization. Furthermore, we demonstrate that when the search space is restricted, one can exploit the structure by using a more efficient optimization.

6.3.1 Constrained problem under implementation errors

To incorporate constraints to the discussion in Section 6.2, we start off analogously to the nominal problem as stated in Eq. (6.1).

6.3.1.1 Problem definition

Consider the nominal optimization problem

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}) \\ &s.t.\ h_j(\mathbf{x}) \leq 0,\ \forall j, \end{aligned} \tag{6.19}$$

where the objective function and the constraints may be non-convex. To find a design which is robust against implementation errors $\Delta\mathbf{x}$, we formulate the robust problem

$$\begin{aligned} &\min_{\mathbf{x}} \max_{\Delta\mathbf{x}\in\mathcal{U}} f(\mathbf{x}+\Delta\mathbf{x}) \\ &s.t.\ \max_{\Delta\mathbf{x}\in\mathcal{U}} h_j(\mathbf{x}+\Delta\mathbf{x}) \leq 0,\ \forall j, \end{aligned} \tag{6.20}$$

where the uncertainty set $\mathcal{U}$ is given by

$$\mathcal{U} := \{\Delta\mathbf{x} \in \mathrm{R}^n \mid \|\Delta\mathbf{x}\|_2 \leq \Gamma\}. \tag{6.21}$$

A design is robust if, and only if, no constraints are violated for any errors in $\mathcal{U}$. Of all the robust designs, we seek one with the lowest worst-case cost $g(\mathbf{x})$. When a design $\hat{\mathbf{x}}$ is implemented with errors in $\mathcal{U}$, the realized design falls within the neighborhood

$$\mathcal{N} := \{\mathbf{x} \mid \|\mathbf{x}-\hat{\mathbf{x}}\|_2 \leq \Gamma\}. \tag{6.22}$$

Figure 6.10 illustrates the neighborhood $\mathcal{N}$ of a design $\hat{\mathbf{x}}$ along with the constraints. Note that $\hat{\mathbf{x}}$ is robust if, and only if, none of its neighbors violates any constraints. Equivalently, there is no overlap between the neighborhood of $\hat{\mathbf{x}}$ and the shaded regions $h_j(\mathbf{x}) > 0$ in Fig. 6.10.

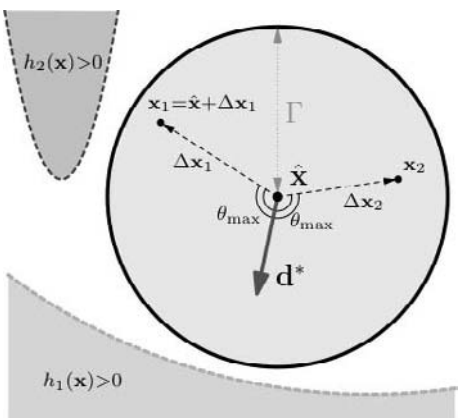

Fig. 6.10. A two-dimensional illustration of the neighborhood $\mathcal{N}$ in the design space $\mathbf{x}$. The shaded regions $h_j(\mathbf{x}) > 0$ contain designs violating the constraints j. Note that h_1 is a convex constraint but not h_2.

6.3.1.2 Robust local search for problems with constraints

When constraints do not come into play in the vicinity of the neighborhood of $\hat{\mathbf{x}}$, the worst-case cost can be reduced iteratively, using the robust local search algorithm for the unconstrained problem, as discussed in Section 6.2. The additional procedures for the robust local search algorithm that are required when constraints are present, are:

(i) Neighborhood search: To determine if there are neighbors violating constraint h_j, the constraint maximization problem,

$$\max_{\Delta\mathbf{x}\in\mathcal{U}} h_j(\hat{\mathbf{x}}+\Delta\mathbf{x}), \tag{6.23}$$

is solved using multiple gradient ascents from different starting designs. Gradient ascents are used because Problem (6.23) is not a convex optimization problem, in general. We shall consider in Section 6.3.1.3 the case where h_j is an explicitly given convex function and, consequently, Problem (6.23) can be solved using more efficient techniques. If a neighbor has a constraint value exceeding zero, for any constraint, it is recorded in a history set $\mathcal{Y}$.

(ii) Check feasibility under perturbations: If $\hat{\mathbf{x}}$ has neighbors in the history set $\mathcal{Y}$, then it is not feasible under perturbations. Otherwise, the algorithm treats $\hat{\mathbf{x}}$ to be feasible under perturbations.

(iii)a. Robust local move if $\hat{\mathbf{x}}$ is not feasible under perturbations: Because constraint violations are more important than cost considerations, and because we want the algorithm to operate within the feasible region of a robust problem, nominal cost is ignored when neighbors violating constraints are encountered. To ensure that the new neighborhood does not contain neighbors in $\mathcal{Y}$, an update step along a direction $\mathbf{d}^*_{feas}$ is taken. This is illustrated in Fig. 6.11a. Here, $\mathbf{d}^*_{feas}$ makes the largest possible angle with all the vectors $\mathbf{y}_i-\hat{\mathbf{x}}$. Such a $\mathbf{d}^*_{feas}$ can be found by solving the SOCP

$$\begin{aligned}&\min_{\mathbf{d},\beta}\ \beta\\ &s.t.\ \|\mathbf{d}\|_2\le 1,\\ &\quad \mathbf{d}'\left(\tfrac{\mathbf{y}_i-\hat{\mathbf{x}}}{\|\mathbf{y}_i-\hat{\mathbf{x}}\|_2}\right)\le\beta,\quad \forall\mathbf{y}_i\in\mathcal{Y},\\ &\quad \beta\le-\epsilon.\end{aligned} \tag{6.24}$$

As shown in Fig. 6.11a, a sufficiently large step along $\mathbf{d}^*_{feas}$ yields a robust design.

(iii)b. Robust local move if $\hat{\mathbf{x}}$ is feasible under perturbations: When $\hat{\mathbf{x}}$ is feasible under perturbations, the update step is similar to that for an unconstrained problem, as in Section 6.2. However, ignoring designs that violate constraints and lie just beyond the neighborhood might lead to a non-robust design. This issue is taken into account when determining an

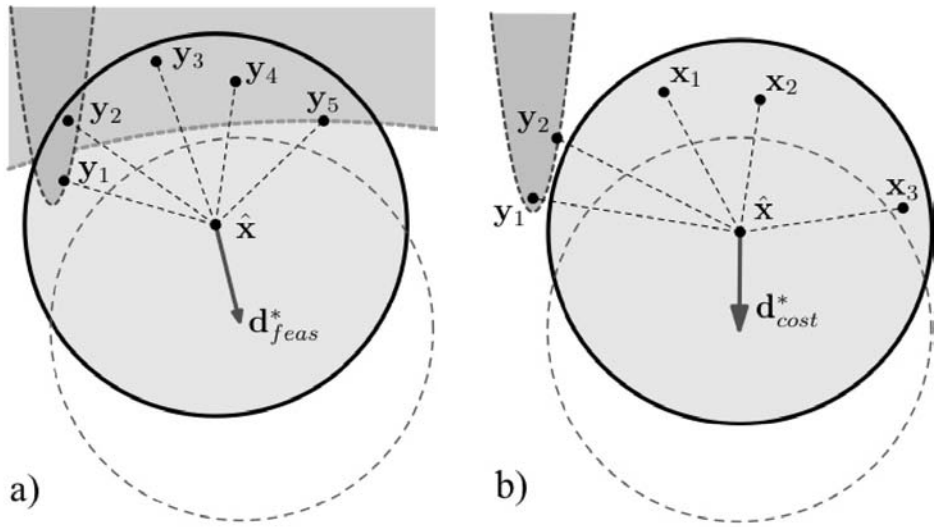

Fig. 6.11. A two-dimensional illustration of the robust local move: (a) when $\hat{\mathbf{x}}$ is non-robust; the upper shaded regions contain constraint-violating designs, including infeasible neighbors $\mathbf{y}_i$. Vector $\mathbf{d}^*_{feas}$ points away from all $\mathbf{y}_i$. (b) When $\hat{\mathbf{x}}$ is robust; $\mathbf{x}_i$ denotes a bad neighbor with high nominal cost, while $\mathbf{y}_i$ denotes an infeasible neighbor lying just outside the neighborhood. The circle with the broken circumference denotes the updated neighborhood.

update direction $\mathbf{d}^*_{cost}$, as illustrated in Fig. 6.11b. This update direction $\mathbf{d}^*_{cost}$ can be found by solving the SOCP

$$\begin{aligned} &\min_{\mathbf{d},\beta} \beta \\ &s.t.\ \|\mathbf{d}\|_2 \leq 1, \\ &\quad \mathbf{d}'\left(\tfrac{\mathbf{x}_i - \hat{\mathbf{x}}}{\|\mathbf{x}_i - \hat{\mathbf{x}}\|_2}\right) \leq \beta, \quad \forall \mathbf{x}_i \in \mathcal{M}, \\ &\quad \mathbf{d}'\left(\tfrac{\mathbf{y}_i - \hat{\mathbf{x}}}{\|\mathbf{y}_i - \hat{\mathbf{x}}\|_2}\right) \leq \beta, \quad \forall \mathbf{y}_i \in \mathcal{Y}_+, \\ &\quad \beta \leq -\epsilon, \end{aligned} \tag{6.25}$$

where $\mathcal{M}$ contains neighbors with highest cost within the neighborhood, and $\mathcal{Y}_+$ is the set of known infeasible designs lying in the slightly enlarged neighborhood $\mathcal{N}_+$,

$$\mathcal{N}_+ := \{\mathbf{x} \mid \|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq (1+\delta)\Gamma\}, \tag{6.26}$$

δ being a small positive scalar for designs that lie just beyond the neighborhood, as illustrated in Fig. 6.11b. Since $\hat{\mathbf{x}}$ is robust, there are no infeasible designs in the neighborhood $\mathcal{N}$. Therefore, all infeasible designs in $\mathcal{Y}_+$ lie at a distance between Γ and $(1+\delta)\Gamma$.

Termination criteria:

We shall first define the robust local minimum for a problem with constraints:

Definition 1

$\mathbf{x}^*$ *is a robust local minimum for the problem with constraints if*

(i) *Feasible under perturbations*
$\mathbf{x}^*$ *remains feasible under perturbations,*

$$h_j(\mathbf{x}^* + \Delta\mathbf{x}) \leq 0, \quad \forall j, \forall \Delta\mathbf{x} \in \mathcal{U}, \text{ and} \tag{6.27}$$

(ii) *No descent direction*
there are no improving directions $\mathbf{d}^*_{cost}$ *at* $\mathbf{x}^*$.

Given the above definition, we can only terminate at step (iii)b, where $\mathbf{x}^*$ is feasible under perturbations. Furthermore, for there to be no direction $\mathbf{d}^*_{cost}$ at $\mathbf{x}^*$, it must be surrounded by neighbors with high cost and infeasible designs in $\mathcal{N}_+$.

6.3.1.3 Enhancements when constraints are convex

In this section, we review the case when $\mathbf{h}_i$ is explicitly given as a convex function. If Problem (6.23) is convex, it can be solved with techniques that are more efficient than multiple gradient ascents. Table 6.1 summarizes the required procedures for solving Problem (6.23). For symmetric

Table 6.1. *Algorithms to solve Problem (6.23). Note that* $\mathbf{Q}$ *is symmetric. LP abbreviates a linear program and SDP a semi-definite program.*

$\mathbf{h}_i(\mathbf{x})$	Problem (6.23)	Required computation
$\mathbf{a}'\mathbf{x} + b$	$\mathbf{a}'\hat{\mathbf{x}} + \Gamma\|\mathbf{a}\|_2 + b \leq 0$	solve LP
$\mathbf{x}'\mathbf{Q}\mathbf{x} + 2\mathbf{b}'\mathbf{x} + c$	single trust region problem	1 SDP in the worst case
$-\mathbf{h}_i$ is convex	convex problem	1 gradient ascent

constraints, the resulting single trust region problem can be expressed as $\max_{\Delta\mathbf{x}\in\mathcal{U}} \Delta\mathbf{x}'\mathbf{Q}\Delta\mathbf{x} + 2(\mathbf{Q}\hat{\mathbf{x}} + \mathbf{b})'\Delta\mathbf{x} + \hat{\mathbf{x}}\mathbf{Q}'\hat{\mathbf{x}} + 2\mathbf{b}'\hat{\mathbf{x}} + c$.

The possible improvements to the robust local search are:

(i) Neighborhood search: Solve Problem (6.23) with the corresponding method of Table 6.1 instead of multiple gradient ascents in order to improve the computational efficiency.

(ii) Check feasibility under perturbations: If $h_j^{rob}(\hat{\mathbf{x}}) \equiv \max_{\Delta\mathbf{x}\in\mathcal{U}} h_j(\hat{\mathbf{x}} + \Delta\mathbf{x}) > 0$, $\hat{\mathbf{x}}$ is not feasible under perturbations.

(iii) Robust local move: To warrant that all designs in the new neighborhood are feasible, the direction should be chosen such that it points away from the infeasible regions. The corresponding vectors describing the closest points in $h_j^{rob}(\hat{\mathbf{x}})$ are given by $\nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}})$ as illustrated in Fig. 6.12. Therefore, $\mathbf{d}$ has to satisfy

$$\mathbf{d}'_{feas} \nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}}) < \beta \|\nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}})\|_2,$$

and

$$\mathbf{d}'_{cost} \nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}}), < \beta \|\nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}})\|_2,$$

in SOCP (6.24) and SOCP (6.25), respectively. Note that $\nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}}) = \nabla_{\mathbf{x}} h(\hat{\mathbf{x}} + \Delta\mathbf{x}_j^*)$, which can be evaluated easily.

In particular, if h_j is a linear constraint, then $h_j^{rob}(\mathbf{x}) = \mathbf{a}'\mathbf{x} + \Gamma\|\mathbf{a}\|_2 + b \leq 0$ is the same for all $\mathbf{x}$. Consequently, we can replace the constraint $\max_{\Delta\mathbf{x}\in\mathcal{U}} h_j(\mathbf{x} + \Delta\mathbf{x}) = \max_{\Delta\mathbf{x}\in\mathcal{U}} \mathbf{a}'(\mathbf{x} + \Delta\mathbf{x}) \leq 0$ with its robust counterpart $h_j^{rob}(\mathbf{x})$. Here, $h_j^{rob}(\mathbf{x})$ is a constraint on $\mathbf{x}$ without any uncertainties, as illustrated in Fig. 6.12.

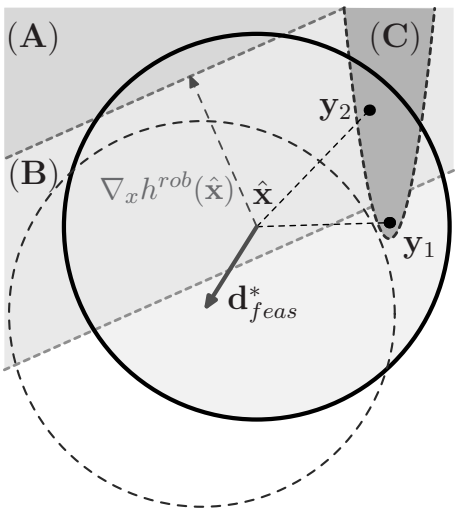

Fig. 6.12. A two-dimensional illustration of the neighborhood when one of the violated constraints is a linear function. (A) denotes the infeasible region. Because $\hat{\mathbf{x}}$ has neighbors in region (A), $\hat{\mathbf{x}}$ lies in the infeasible region of its robust counterpart (B). $\mathbf{y}_i$ denotes neighbors which violate a non-convex constraint, shown in region (C). $\mathbf{d}^*_{feas}$ denotes a direction which would reduce the infeasible region within the neighborhood and points away from the gradient of the robust counterpart and all bad neighbors $\mathbf{y}_i$. The dashed circle represents the updated neighborhood.

6.3.1.4 Constrained robust local search algorithm

In this section, we utilize the methods outlined in Sections 6.3.1.2 and 6.3.1.3 to formalize the overall algorithm:

Algorithm 2 *[Constrained robust local search]*
Step 0. *Initialization: Set* $k := 1$. *Let* $\mathbf{x}^1$ *be an arbitrary decision vector.*
Step 1. *Neighborhood search:*

i. Find neighbors with high cost through $n + 1$ *gradient ascents sequences, where* n *is the dimension of* $\mathbf{x}$. *Record all evaluated neighbors and their costs in a history set* $\mathcal{H}^k$, *together with* $\mathcal{H}^{k-1}$.
ii. Let $\mathcal{J}$ *be the set of constraints to the convex constraint maximization Problem (6.23) that are convex. Find optimizer* $\Delta\mathbf{x}_j^*$ *and highest constraint value* $h_j^{rob}(\mathbf{x}^k)$, *for all* $j \in \mathcal{J}$, *according to the methods listed in Table 6.1. Let* $\bar{\mathcal{J}} \subseteq \mathcal{J}$ *be the set of constraints which are violated under perturbations.*
iii. For every constraint $j \notin \mathcal{J}$, *find infeasible neighbors by applying* $n + 1$ *gradient ascents sequences on Problem (6.23), with* $\hat{\mathbf{x}} = \mathbf{x}^k$. *Record all infeasible neighbors in a history set* $\mathcal{Y}^k$, *together with set* $\mathcal{Y}^{k-1}$.

Step 2. *Check feasibility under perturbations:* $\mathbf{x}^k$ *is not feasible under perturbations if either* $\mathcal{Y}^k$ *or* $\bar{\mathcal{J}}$ *is not empty.*
Step 3. *Robust local move:*

i. If $\mathbf{x}^k$ *is not feasible under perturbations, solve SOCP (6.24) with additional constraints* $\mathbf{d}'_{feas}\nabla_{\mathbf{x}}h_j^{rob}(\mathbf{x}^k) < \beta\|\nabla_{\mathbf{x}}h_j^{rob}(\mathbf{x}^k)\|_2$, *for all* $j \in \bar{\mathcal{J}}$. *Find direction* d_{feas}^* *and set* $\mathbf{x}^{k+1} := \mathbf{x}^{k+1} + t^k d_{feas}^*$.
ii. If $\mathbf{x}^k$ *is feasible under perturbations, solve SOCP (6.25) to find a direction* d_{cost}^*. *Set* $\mathbf{x}^{k+1} := \mathbf{x}^{k+1} + t^k d_{feas}^*$. *If no direction* d_{cost}^* *exists, reduce the size of* $\mathcal{M}$; *if the size is below a threshold, terminate.*

In Steps 3(i) and 3(ii), t^k is the minimum distance chosen such that the undesirable designs are excluded from the neighborhood of the new iterate $\mathbf{x}^{k+1}$. Finding t^k requires solving a simple geometric problem. For more details, refer to [28].

6.3.2 Generalization to include parameter uncertainties

6.3.2.1 Problem definition

Consider the nominal problem:

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}, \bar{\mathbf{p}}) \\ &s.t.\ h_j(\mathbf{x}, \bar{\mathbf{p}}) \leq 0,\ \ \forall j, \end{aligned} \tag{6.28}$$

where $\bar{\mathbf{p}} \in \mathrm{R}^m$ is a coefficient vector of the problem parameters. For our purpose, we can restrict $\bar{\mathbf{p}}$ to parameters with perturbations only. For example, if Problem (6.28) is given by

$$\begin{aligned} &\min_{\mathbf{x}} 4x_1^3 + x_2^2 + 2x_1^2 x_2 \\ &s.t.\ 3x_1^2 + 5x_2^2 \leq 20, \end{aligned} \tag{6.29}$$

then we can extract $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\bar{\mathbf{p}} = \begin{pmatrix} 4 \\ 1 \\ 2 \\ 3 \\ 5 \\ 20 \end{pmatrix}$. Note that uncertainties can even be present in the exponent, e.g. 3 in the monomial $4x_1^3$.

In addition to implementation errors, there can be perturbations $\Delta\mathbf{p}$ in parameters $\bar{\mathbf{p}}$ as well. The true, but unknown, parameter, $\mathbf{p}$, can then be expressed as $\bar{\mathbf{p}} + \Delta\mathbf{p}$. To protect the design against both types of perturbation, we formulate the robust problem

$$\begin{aligned} &\min_{\mathbf{x}} \max_{\Delta\mathbf{z} \in \mathcal{U}} f(\mathbf{x} + \Delta\mathbf{x}, \bar{\mathbf{p}} + \Delta\mathbf{p}) \\ &s.t. \max_{\Delta\mathbf{z} \in \mathcal{U}} h_j(\mathbf{x} + \Delta\mathbf{x}, \bar{\mathbf{p}} + \Delta\mathbf{p}) \leq 0,\ \ \forall j, \end{aligned} \tag{6.30}$$

where $\Delta\mathbf{z} = \begin{pmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{p} \end{pmatrix}$. Here, $\Delta\mathbf{z}$ lies within the uncertainty set

$$\mathcal{U} = \left\{\ \Delta\mathbf{z} \in \mathrm{R}^{n+m}\ \ |\ \ \|\Delta\mathbf{z}\|_2\ \leq\ \Gamma\ \right\}, \tag{6.31}$$

where $\Gamma > 0$ is a scalar describing the size of perturbations we want to protect the design against. Similar to Problem (6.20), a design is robust only if no constraints are violated under the perturbations. Among these robust designs, we seek to minimize the worst-case cost

$$g(\mathbf{x}) := \max_{\Delta\mathbf{z} \in \mathcal{U}}\ f\left(\mathbf{x} + \Delta\mathbf{x}, \bar{\mathbf{p}} + \Delta\mathbf{p}\right). \tag{6.32}$$

6.3.2.2 Generalized constrained robust local search algorithm

Problem (6.30) is equivalent to the following problem with implementation errors only,

$$\begin{aligned} &\min_{\mathbf{z}} \max_{\Delta\mathbf{z}\in\mathcal{U}} f(\mathbf{z}+\Delta\mathbf{z}) \\ &s.t. \max_{\Delta\mathbf{z}\in\mathcal{U}} h_j(\mathbf{z}+\Delta\mathbf{z}) \le 0, \ \forall j, \\ &\qquad \mathbf{p} = \bar{\mathbf{p}}, \end{aligned} \tag{6.33}$$

where $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{p} \end{pmatrix}$. The idea behind generalizing the constrained robust local search algorithm is analogous to the approach described in Section 6.2 for the unconstrained problem, discussed in [28]. Consequently, the necessary modifications to Algorithm 2 are:

(i) Neighborhood search : Given $\hat{\mathbf{x}}$, $\hat{\mathbf{z}} = \begin{pmatrix} \hat{\mathbf{x}} \\ \bar{\mathbf{p}} \end{pmatrix}$ is the decision vector. Therefore, the neighborhood can be described as

$$\mathcal{N} := \{\mathbf{z} \mid \|\mathbf{z}-\hat{\mathbf{z}}\|_2 \le \Gamma\} = \left\{ \begin{pmatrix} \mathbf{x} \\ \mathbf{p} \end{pmatrix} \mid \left\| \begin{pmatrix} \mathbf{x}-\hat{\mathbf{x}} \\ \mathbf{p}-\bar{\mathbf{p}} \end{pmatrix} \right\|_2 \le \Gamma \right\}. \tag{6.34}$$

(ii) Robust local move : Let $\mathbf{d}^* = \begin{pmatrix} \mathbf{d}_x^* \\ \mathbf{d}_p^* \end{pmatrix}$ be an update direction in the $\mathbf{z}$ space. Because $\mathbf{p}$ is not a decision vector but a given system parameter, the algorithm has to ensure that $\mathbf{p} = \bar{\mathbf{p}}$ is satisfied at every iterate. Thus, $\mathbf{d}_p^* = \mathbf{0}$.

When finding the update direction, the condition $\mathbf{d}_p = \mathbf{0}$ must be included in either of SOCP (6.24) and (6.25) along with the feasibility constraints $\mathbf{d}'\nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}}) < \beta\|\nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{x}})\|_2$. As discussed earlier, we seek a direction $\mathbf{d}$ that points away from the worst-case and infeasible neighbors. We achieve this objective by maximizing the angle between $\mathbf{d}$ and all worst-case neighbors as well as the angle between $\mathbf{d}$ and the gradient of all constraints. For example, if a design $\mathbf{z}$ is not feasible under perturbations, the SOCP is given by

$$\begin{aligned} \min_{\mathbf{d}=(\mathbf{d}_x,\mathbf{d}_p),\beta} \ & \beta \\ s.t. \ & \|\mathbf{d}\|_2 \le 1, \\ & \mathbf{d}'(\mathbf{z}_i - \hat{\mathbf{z}}) \le \beta\|\mathbf{z}_i - \hat{\mathbf{z}}\|_2, && \forall \mathbf{y}_i \in \mathcal{Y}, \\ & \mathbf{d}'\nabla_{\mathbf{z}} h_j^{rob}(\hat{\mathbf{z}}) < \beta\|\nabla_{\mathbf{z}} h_j^{rob}(\hat{\mathbf{z}})\|_2, && \forall j \in \bar{\mathcal{J}}, \\ & \mathbf{d}_p = \mathbf{0}, \\ & \beta \le -\epsilon. \end{aligned} \tag{6.35}$$

Here, $\mathcal{Y}^k$ is the set of infeasible designs in the neighborhood. Since the p-component of $\mathbf{d}$ is zero, this problem reduces to the following:

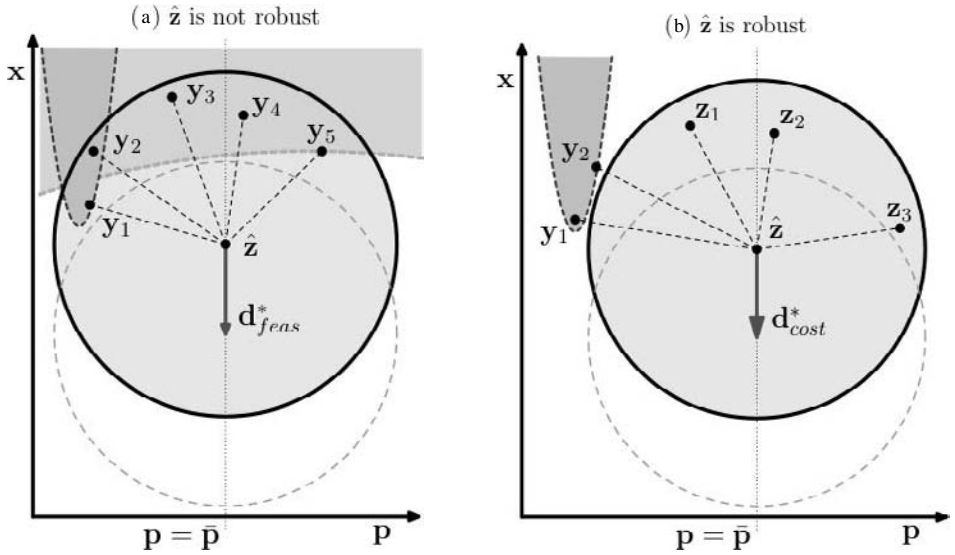

Fig. 6.13. A two-dimensional illustration of the robust local move for problems with both implementation errors and parameter uncertainties. The neighborhood spans the $\mathbf{z} = (\mathbf{x}, \mathbf{p})$ space: (a) the constrained counterpart of Fig. 6.11(a) and (b) the constrained counterpart of Fig. 6.11(b). Note that the direction found must lie within the hyper-planes $\mathbf{p} = \bar{\mathbf{p}}$.

$$
\begin{aligned}
&\min_{\mathbf{d}_x, \beta} \ \beta \\
&s.t. \ \|\mathbf{d}_x\|_2 \leq 1, \\
&\quad \mathbf{d}_x' \left(\mathbf{x}_i - \hat{\mathbf{x}}\right) \leq \beta \|\mathbf{z}_i - \hat{\mathbf{z}}\|_2, \qquad \forall \mathbf{y}_i \in \mathcal{Y}, \\
&\quad \mathbf{d}_x' \nabla_{\mathbf{x}} h_j^{rob}(\hat{\mathbf{z}}) < \beta \|\nabla_{\mathbf{z}} h_j^{rob}(\hat{\mathbf{z}})\|_2, \quad \forall j \in \bar{\mathcal{J}}, \\
&\quad \beta \leq -\epsilon.
\end{aligned}
\tag{6.36}
$$

A similar approach is carried out for the case of $\mathbf{z}$ being robust. Consequently, both $\mathbf{d}^*_{feas}$ and $\mathbf{d}^*_{cost}$ satisfy $\mathbf{p} = \bar{\mathbf{p}}$ at every iteration. This is illustrated in Fig. 6.13.

Now, we have arrived at the constrained robust local search algorithm for Problem (6.30) with both implementation errors and parameter uncertainties:

Algorithm 3 *[Generalized constrained robust local search]*

Step 0. *Initialization*: *Set* $k := 1$. *Let* $\mathbf{x}^1$ *be an arbitrary initial decision vector.*

Step 1. *Neighborhood search*: *Same as Step 1 in Algorithm 2, but over the neighborhood (6.34).*

Step 2. *Check feasibility under perturbations*: $\mathbf{z}^k$, *and equivalently* $\mathbf{x}^k$, *is*

feasible under perturbations, if $\mathcal{Y}^k$ *and* $\bar{\mathcal{J}}^k$ *are empty.*

Step 3. Robust local move*:*

i. If $\mathbf{z}^k$ *is not feasible under perturbations, find a direction* d^*_{feas} *by solving SOCP (6.36) with* $\hat{\mathbf{z}} = \mathbf{z}^k$*. Set* $\mathbf{z}^{k+1} := \mathbf{z}^{k+1} + t^k d^*_{feas}$*.*

ii. If $\mathbf{x}$ *is feasible under perturbations, solve the SOCP*

$$
\begin{aligned}
&\min_{\mathbf{d}_x,\beta} \beta \\
&s.t. \|\mathbf{d}_x\|_2 \leq 1, \\
&\quad \mathbf{d}_x'\left(\mathbf{x}_i - \mathbf{x}^k\right) \leq \beta \left\| \begin{pmatrix} \mathbf{x}_i - \mathbf{x}^k \\ \mathbf{p}_i - \bar{\mathbf{p}} \end{pmatrix} \right\|_2, \quad \forall \mathbf{z}_i \in \mathcal{M}^k, \mathbf{z}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{p}_i \end{pmatrix}, \\
&\quad \mathbf{d}_x'\left(\mathbf{x}_i - \mathbf{x}^k\right) \leq \beta \left\| \begin{pmatrix} \mathbf{x}_i - \mathbf{x}^k \\ \mathbf{p}_i - \bar{\mathbf{p}} \end{pmatrix} \right\|_2, \quad \forall \mathbf{y}_i \in \mathcal{Y}^k_+, \mathbf{y}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{p}_i \end{pmatrix}, \\
&\quad \mathbf{d}_x' \nabla_{\mathbf{x}} h^{rob}_j(\mathbf{z}^k) < \beta \|\nabla_{\mathbf{z}} h^{rob}_j(\mathbf{z}^k)\|_2, \quad \forall j \in \bar{\mathcal{J}}_+, \\
&\quad \beta \leq -\epsilon,
\end{aligned}
\tag{6.37}
$$

to find a direction d^*_{cost}*. Note that* $\mathcal{Y}^k_+$ *is the set of infeasible designs in the enlarged neighborhood* $\mathcal{N}^k_+$ *as in Eq. (6.26).* $\bar{\mathcal{J}}_+$ *is the set of constraints which are not violated in the neighborhood of* $\hat{\mathbf{x}}$*, but are violated in the slightly enlarged neighborhood* $\mathcal{N}_+$*. Set* $\mathbf{z}^{k+1} := \mathbf{z}^{k+1} + t^k d^*_{feas}$*. If no direction* d^*_{cost} *exists, reduce the size of* $\mathcal{M}$*; if the size is below a threshold, terminate.*

We have finished introducing the robust local search method with constraints. In the following sections, we will present an application to showcase the performance of this method.

6.3.3 Example in polynomial optimization

Here, we provide an example application in constrained polynomial optimization. Our primary goal is to develop intuition and illustrate how the constrained robust local search algorithm from Section 6.3 performs. Moreover, we provide a modified version of the application for which the constraints are convex. In this case, we can use an alternative method by replacing the constraints by their robust counterparts, allowing for a more efficient optimization.

6.3.3.1 Problem description

The first problem is sufficiently simple, so as to develop intuition into the algorithm. Consider the nominal problem

$$\begin{aligned} &\min_{x,y} f_{\text{poly}}(x,y), \\ &s.t.\ h_1(x,y) \le 0, \\ &\quad\ h_2(x,y) \le 0, \end{aligned} \tag{6.38}$$

where

$$\begin{aligned} f_{\text{poly}}(x,y) &= 2x^6 - 12.2x^5 + 21.2x^4 + 6.2x - 6.4x^3 - 4.7x^2 + y^6 - 11y^5 + 43.3y^4 \\ &\quad -10y - 74.8y^3 + 56.9y^2 - 4.1xy - 0.1y^2x^2 + 0.4y^2x + 0.4x^2y, \\ h_1(x,y) &= (x-1.5)^4 + (y-1.5)^4 - 10.125, \\ h_2(x,y) &= -(2.5-x)^3 - (y+1.5)^3 + 15.75. \end{aligned}$$

Given implementation errors $\mathbf{\Delta} = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$ such that $\|\mathbf{\Delta}\|_2 \le 0.5$, the robust problem is

$$\begin{aligned} &\min_{x,y} \max_{\|\mathbf{\Delta}\|_2 \le 0.5} f_{\text{poly}}(x+\Delta x, y+\Delta y), \\ &s.t. \max_{\|\mathbf{\Delta}\|_2 \le 0.5} h_1(x+\Delta x, y+\Delta y) \le 0, \\ &\quad\ \max_{\|\mathbf{\Delta}\|_2 \le 0.5} h_2(x+\Delta x, y+\Delta y) \le 0. \end{aligned} \tag{6.39}$$

To the best of our knowledge, there are no practical ways to solve such a robust problem, given today's technology [48]. If the relaxation method for polynomial optimization problems is used, as in Ref. [49], Problem (6.39) leads to a large polynomial SDP problem, which cannot be solved in practice today [50, 48]. In Fig. 6.14, a contour plot of the nominal and the estimated worst-cost surface along with their local and global extrema are shown to generate intuition for the performance of the robust optimization method. The computation takes less than 10 minutes to terminate on an Intel Xeon processor with 3.4 GHz clock. This is fast enough for a prototype-problem. Three different initial designs with their respective neighborhoods are sketched as well.

6.3.3.2 Computation results

For the constrained Problem (6.38), the non-convex cost surface and the feasible region are shown in Fig. 6.14(a). Note that the feasible region is not convex, because h_2 is not a convex constraint. Let $g_{\text{poly}}(x,y)$ be the worst-case cost function given as

$$g_{\text{poly}}(x,y) := \max_{\|\mathbf{\Delta}\|_2 \le 0.5} f_{poly}(x+\Delta x, y+\Delta y).$$

Figure 6.14(b) shows the worst-case cost estimated by using sampling on the cost surface f_{poly}. In the robust Problem (6.39), we seek to find a

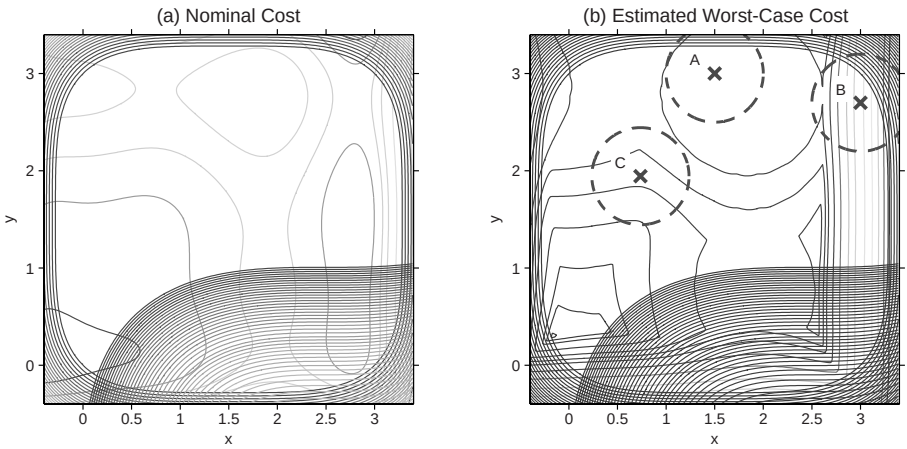

Fig. 6.14. Contour plot of (a) the nominal cost function and (b) the estimated worst-case cost function. The shaded regions denote designs which violate at least one of the two constraints, h_1 and h_2. While point A and point B are feasible, they are not feasible under perturbations due to their infeasible neighbors. Point C, on the other hand, remains feasible under perturbations.

design which minimizes $g_{\text{poly}}(x, y)$, such that its neighborhood lies within the unshaded region. An example of such a design is the point C in Fig. 6.14(b).

Two separate robust local searches were carried out from initial designs A and B. The initial design A exemplifies initial configurations whose neighborhood contains infeasible designs and is close to a local minimum. The design B represents only configurations whose neighborhood contains infeasible designs. Figure 6.15 shows that, in both instances, the algorithm terminated at designs that are feasible under perturbations and have significantly lower worst-case costs. However, it converged to different robust local minima in the two instances, as shown in Fig. 6.15(c). The presence of multiple robust local minima is not surprising because $g_{\text{poly}}(x, y)$ is non-convex. Figure 6.15(c) also shows that both robust local minima I and II satisfy the terminating conditions as stated in Section 6.3.1.2:

(i) Feasible under perturbations: Both their neighborhoods do not overlap with the shaded regions.

(ii) No direction $\mathbf{d}^*_{cost}$ found: Both designs are surrounded by bad neighbors and infeasible designs lying just outside their respective neighborhoods. Note that for robust local minimum II, the bad neighbors lie on the same contour line, even though they are apart, indicating that any further improvement is restricted by the infeasible neighboring designs.

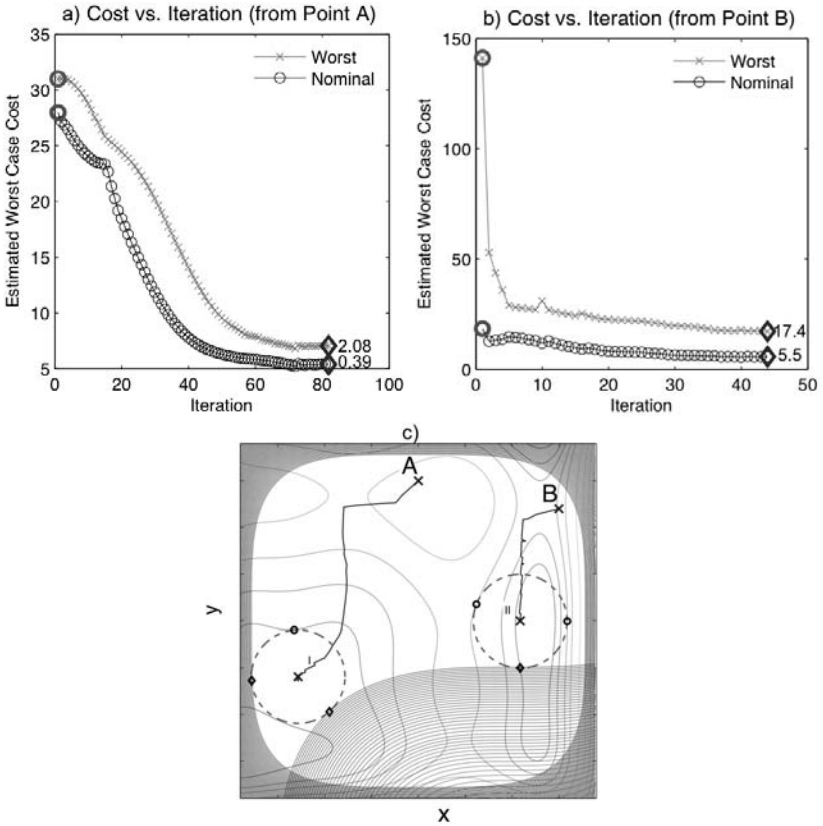

Fig. 6.15. Performance of the robust local search algorithm from two different starting points A and B. The circle marker indicates the starting and the diamond marker the final design. (a) Starting from point A, the algorithm reduces the worst-case cost and the nominal cost. (b) Starting from point B, the algorithm converges to a different robust solution, which has a significantly larger worst-case cost and nominal cost. (c) The broken circles sketch the neighborhood of minima. For each minimum, (i) there is no overlap between its neighborhood and the shaded infeasible regions, and (ii) there is no improving direction because it is surrounded by neighbors of high cost (bold circle) and infeasible designs (bold diamond) residing just beyond the neighborhood. Two bad neighbors of minimum II (started from B) share the same cost, since they lie on the same contour line.

6.3.3.3 When constraints are linear

In Section 6.3.1.3, we argued that the robust local search can be more efficient if the constraints are explicitly given as convex functions. To illustrate this, suppose that the constraints in Problem (6.38) are linear and given by

$$h_1(x, y) = 0.6x - y + 0.17,$$
$$h_2(x, y) = -16x - y - 3.15. \tag{6.40}$$

As shown in the first row of Table 6.1 for linear constraints, the robust counterparts of the constraints in Eq. (6.40) using $a_1 = (0.6, -1)'$, $a_2 = (-16, -1)'$, and $\gamma = 0.5$ are

$$\begin{aligned} h_1^{rob}(x, y) &= 0.6x - y + 0.17 + 0.5831 \ \leq \ 0, \\ h_2^{rob}(x, y) &= -16x - y - 3.15 + 8.0156 \ \leq \ 0. \end{aligned} \tag{6.41}$$

The benefit of using the explicit counterparts in Eq. (6.41) is that the algorithm terminates in only 96 seconds as opposed to 3,600 seconds, when using the initial linear constraints in Eq. (6.40).

6.3.4 Summary

In this chapter, we discussed the topic of robust optimization for high-dimensional problems. In particular, we addressed simulation-based problems for which an analytically closed form expression is a priori not known. This method is of direct relevance to many engineering design projects, since they often rely on numerical simulations to solve problem, e.g. on PDE solvers. Since in most real-world design problems errors in implementation, modeling, and design are inevitable, an otherwise nominally optimized solution can easily be sub-optimal or, even worse, infeasible. Moreover, we might lose the phenomenon which the design sought to achieve. Therefore, taking errors into account during the optimization process is a first-order effect.

We illustrated the robust local search algorithm in the context of simulation-based problems that are subject to errors. The advantage of this method is that it treats the solver of the actual physical problem as an oracle and does not exploit any internal structure of the problem. Because of this black-box approach, the algorithm is generic and can be applied to most engineering problems.

This algorithm can be summarized as a method which aims to minimize the worst-case scenario. Possible errors for a particular design define a neighborhood whose extension is determined by the actual application (e.g. manufacturing errors). Within the neighborhood, local searches are conducted to identify a set of neighbors with highest costs. The solution to a second-order cone problem determines a direction along which we can update the design and consequently exclude all neighbors with highest costs. Through iterations, the algorithm successively reduces the worst-case cost and, thus, solves the robust problem.

We also illustrated the extension of this robust optimization algorithm to problems with constraints which may be convex or given through simulations. When constraints are encountered within the neighborhood, infeasible

neighbors are collected in a separate set. The new descent direction for the constrained robust problem points to a direction that will exclude both the neighbors with highest cost and the infeasible neighbors. If the constraints have certain structure, such as convexity, we showed significant improvement of the efficiency by using conventional methods, such as robust counterparts.

To demonstrate the performance of this robust optimization method, we discussed three applications in this chapter. First, we illustrated the application of robust optimization to a 100-dimensional electromagnetic scattering problem and showed that the final design was 90% more robust in comparison to the initial configuration. This problem is of direct relevance to nano-photonic design, because it scales with wavelength.

In the second application, we discussed the 208-dimensional robust design of double chirped mirrors as they are used in ultrafast mode-locked lasers. Using additional information about the rare and unphysical events, we were able to transform the search into an unconstrained optimization problem, and thus increased the efficiency. Using this adopted method, we provided a robust solution for the estimated possible errors. Moreover, our final design was, by far, more robust beyond the assumed errors, when compared to the nominal solution. Furthermore, using random sampling, we showed that the robust design has a significantly increased manufacturing yield, which is a crucial aspect of mirror design.

The third example showcased the robust method in the presence of constraints. This application in polynomial optimization was intended to showcase the performance when constraints are encountered as well as to generate intuition. We demonstrated that when these constraints have structure, it can be exploited for a more efficient algorithm.

Overall, we have discussed the robust optimization algorithm that is well suited for engineering design problems and demonstrated its performance based on real-world applications. Nevertheless, there is still a vital need for more advanced robust optimization algorithms. In fact, since engineering design problems increasingly seek to exploit nonlinear (and quantum) effects, the impact of possible errors can have severe outcomes. Moreover, since the spatial extension of objects, whose designs are subject to optimization, will continually shrink in the future, the size of errors becomes increasingly comparable to the actual size of the design. This will cause increased sensitivity and, thus, a heightened need for robust optimization algorithms.

Furthermore, the deeper understanding of the nature of errors will give rise to a more tailored definition of uncertainty sets. If possible, avoiding unphysical and rare events and their parameterization, along with dimension-reduction and separability, may reduce the size of the search space enabling

more efficient algorithms as well as an adjusted conservatism for error estimation.

On the other hand, the search for global robust optima is yet another direction that needs further investigation. Whether known global optimization techniques can be extended to address robust problems, or whether new algorithms have to be developed, will depend on the particular application and the structure of the search space.

6.4 References

1. H. Petroski, *Design Paradigms*, Cambridge University Press, Cambridge, United Kingdom, 1994.
2. A. Ben-Tal and A. Nemirovski, *Robust optimization — methodology and applications*, Mathematical Programming **92**, 453–480 (2002).
3. J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York, New York, 1997.
4. A. Prekopa and A. Ruszczynski, *Stochastic programming*, Optimization Methods and Software **17**, 359–559 (2002).
5. A. Charnes and W. W. Cooper, *Chance-constrained programming*, Management Science **6**, 73–79 (1959).
6. H. Markowitz, *Portfolio selection*, The Journal of Finance **7**, 77–91 (1952).
7. G. J. Park, T. H. Lee, K. H. Lee, and K. H. Hwang, *Robust design — an overview*, American Institute of Aeronautics and Astronautics Journal **44**, 181–191 (2006).
8. B. Ramakrishnan and S. S. Rao, *A robust optimization approach using Taguchi's loss function for solving nonlinear optimization problems*, Advances in Design Automation **32**, 241–248 (1991).
9. A. Ruszczynski and A. Shapiro, *Optimization of Risk Measures*, Springer-Verlag, London, pp. 119–157 (2006).
10. S. Uryasev and R. Rockafellar, *Conditional Value-at-risk: Optimization Approach*, vol. 54 of *Applied Optimization*, Kluwer Academic Publishing, Dordrecht, The Netherlands, 2001.
11. J. Mulvey and A. Ruszczynski, *A new scenario decomposition method for large-scale stochastic optimization*, Operations Research **43**, 477–490 (1995).
12. R. T. Rockafellar and R. J. B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Mathematics of Operations Research **16**, 119–147 (1991).
13. M. Dyer and L. Stougie, *Computational complexity of stochastic programming problems*, Mathematical Programming **106**, 423–432 (2006).
14. A. Nemirovski, *On tractable approximations of randomly perturbed convex constraints*, Proceedings, 42nd IEEE Conference on Decision and Control **3**, 2419–2422 (2003).

15. I. Doltsinis and Z. Kang, *Robust design of structures using optimization methods*, Computational Methods in Applied Mechanical Engineering **193**, 2221–2237 (2004).

16. K. H. Lee and G. J. Park, *Robust optimization considering tolerances of design variables*, Computers and Structures **79**, 77–86 (2001).

17. A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Mathematics of Operations Research **23**, 769–805 (1998).

18. D. Bertsimas and M. Sim, *Robust discrete optimization and network flows*, Mathematical Programming **98**, 49–71 (2003).

19. D. Bertsimas and M. Sim, *Tractable approximations to robust conic optimization problems*, Mathematical Programming **107**, 5–36 (2006).

20. S. Žaković and C. Pantelides, *An interior point algorithm for computing saddle points of constrained continuous minimax*, Annals of Operations Research **99**, 59–77 (2000).

21. M. Diehl, H. G. Bock, and E. Kostina, *An approximation technique for robust nonlinear optimization*, Mathematical Programming, Ser. B **107**, 213–230 (2006).

22. Y. Zhang, *General robust-optimization formulation for nonlinear programming*, Journal of Optimization Theory and Applications **132**, 111–124 (2007).

23. P. G. Ciarlet, *Finite Element Method for Elliptic Problems*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2002.

24. R. D. Cook, D. S. Malkus, M. E. Plesha, and R. J. Witt, *Concepts and Applications of Finite Element Analysis*, John Wiley and Sons, Hoboken, New Jersey, 2007.

25. R. Jin, W. Chen, and T. W. Simpson, *Comparative studies of metamodelling techniques under multiple modelling criteria*, Structural and Multidisciplinary Optimization **23**, 1–13 (2001).

26. T. W. Simpson, J. D. Poplinski, P. N. Koch, and J. K. Allen, *Metamodels for computer-based engineering design: Survey and recommendations.*, Engineering with Computers **17**, 129–150 (2001).

27. E. Stinstra and D. den Hertog, *Robust optimization using computer experiments*, European Journal of Operational Research **191**, 816–83 (2008).

28. D. Bertsimas, O. Nohadani, and K. M. Teo, *Robust optimization for unconstrained simulation-based problems*, Operations Research, accepted for publication (2008).

29. D. Bertsimas, O. Nohadani, and K. M. Teo, *Robust optimization in electromagnetic scattering problems*, Journal of Applied Physics **101**, 074507 1–7 (2007).

30. I. L. Gheorma, S. Haas, and A. F. J. Levi, *Aperiodic nanophotonic design*, Journal of Applied Physics **95**, 1420–1426 (2004).

31. P. Seliger, M. Mahvash, C. Wang, and A. F. J. Levi, *Optimization of aperiodic dielectric structures*, Journal of Applied Physics **100**, 034310 1–6 (2006).

32. D. M. Kingsland, J. Gong, J. L. Volakis, and J. F. Lee, *Performance of an anisotropic artificial absorber for truncating finite-element meshes*, IEEE Transactions on Antennas Propagation **44**, 975–982 (2006).

33. J. P. Berenger, *Three-dimensional perfectly matched layer for the absorption of electromagnetic waves*, Journal of Computational Physics **127**, 363–379 (1996).

34. R. Szipöcs, K. Ferencz, C. Spielmann, and F. Karusz, *Chirped multilayer coatings for broadband dispersion control in femtosecond lasers*, Optics Letters **19**, 201–203 (1994).

35. F. Kärtner, N. Matuschek, T. Schibli, *et al.*, *Design and fabrication of double-chirped mirrors*, Optics Letters **22**, 831–833 (1997).

36. N. Matuschek, F. Kärtner, and U. Keller, *Theory of double-chirped mirrors*, IEEE Journal of Selected Topics in Quantum Electronics **4**, 197–208 (1998).

37. F. Kaertner, U. Morgner, T. Schibli, *et al.*, *Ultrabroadband double-chirped mirror pairs for generation of octave spectra*, Journal of the Optical Society of America B**18**, 882–885 (2001).

38. B. Sullivan and J. Dobrowolski, *Deposition error compensation for optical multilayer coating: I. theoretical description*, Applied Optics **31**, 3821–3835 (1992).

39. V. Pervak, A. Tikhonravov, M. Trubetskov, *et al.*, *1.5-octave chirped mirror for pulse compression down to sub-3 fs*, Applied Optics **87**, 5–12 (2007).

40. V. Yakovlev and G. Tempea, *Optimization of chirped mirrors*, Applied Optics **41**, 6514–6520 (2002).

41. J. Kong, *Electromagnetic Wave Theory*, EMW Publishing, Cambridge, Massachusetts, 2000.

42. J. Birge and F. Kärtner, *Efficient analytic computation of dispersion from multilayer structures*, Applied Optics **45**, 1478–1483 (2006).

43. J. Birge and F. Kärtner, *Efficient optimization of multilayer coatings for ultrafast optics using analytic gradients of dispersion*, Applied Optics **46**, 2656–2662 (2007).

44. O. Mücke, R. Ell, A. Winter, *et al.*, *Self-referenced 200 MHz octave-spanning Ti:saphhire laser with 50 attosecond carrier-envelope phase jitter*, Optics Express **13**, 5163–5169 (2005).

45. P. Verly, *Fourier transform technique with refinement in the frequency domain for the synthesis of optical thin films*, Applied Optics **35**, 5148–5154 (1996).

46. A. Tikhonravov, M. Trubetskov, and G. DeBell, *Applications of the needle optimization technique to the design of optical coatings*, Applied Optics **35**, 5493–5508 (1996).

47. D. Bertsimas, O. Nohadani, and K. M. Teo, *Nonconvex robust optimization for problems with constraints*, INFORMS Journal on Computing, accepted for publication (2008).

48. J. B. Lasserre, *Robust global optimization with polynomials*, Mathematical Programming **107**, 275–293 (2006).

49. D. Henrion and J. B. Lasserre, *Gloptipoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Transactions on Mathematical Software **29**, 165–194 (2003).

50. M. Kojima, *Sums of Squares Relaxations of Polynomial Semidefinite Programs*, Mathematical and Computing Sciences, Tokyo Institute of Technology, Meguro, Tokyo, Japan, pp. 152–8552 (2003).

7 Mathematical framework for optimal design

I.G. Rosen and C. Wang

7.1 Introduction

Fundamental to most optimal design is the replacement of the actual physical device by high fidelity mathematical models. Models are also used in place of sophisticated sensors and performance measuring instrumentation to avoid the costly and time-consuming process of building a large number of prototype devices fitted with sensors and monitored by the associated test equipment. These mathematical models also help to establish an unambiguous relationship between the design parameters and the performance of the device. This relationship is commonly referred to as the optimal design problem's *performance index*. Mathematical optimization theory can then be used to characterize local and global optimal designs. Moreover, by formulating the optimal design problem mathematically, we may employ sophisticated mathematical programming techniques to provide an efficient means to search for these local and global optimal designs. This is especially significant if the design space is high or even infinite dimensional. Thus, mathematics plays a central and essential role in solving optimal design problems.

In this chapter we describe how mathematical systems theory can be used to develop a framework for optimal design problems. In particular, we consider the device which is the basis for the optimal design and the environment it is interacting with as a system. The internal quantities that uniquely characterize the status of the system are referred to as the system states. For example, in the electromagnetic waveguide design problem described previously in Chapter 4, the structure of the waveguide, the dielectric Teflon cylinders, and the surrounding space form a system. The state variables of this system include the intensity and directions of the electric and magnetic fields.

There are at least three categories of input to an optimal design system: *uncontrolled external forces*, *control inputs*, and the *design parameters*. Among the uncontrolled external forces are unknown deviations from nominal system configurations due to manufacturing inaccuracies. The control inputs are excitation signals that the device is subject to during its operation. The design parameters are static variables that define the physical configuration of the device. Once again, in the case of the electromagnetic waveguide design problem, the uncontrolled inputs include deviation from the nominal refractive index of the Teflon cylinders, inaccuracy in the placement of these cylinders, and disturbances in the input microwave signal. The control input is the microwave signal that the antenna is transiting into the waveguide. The design parameters are the locations of the Teflon cylinders.

As in any system theoretical description, the behavior of a system is only revealed through the measurements or observations that we can make. On rare occasions, it is possible to observe or measure all of the system states. More typically, an observation is a finite-dimensional quantity that is a function of the system state and, in some cases, the control inputs and design parameters. In most cases, random observation errors must also be considered in the design process. This is often handled by insisting that an acceptable optimal design be robust with respect to observation disturbances and unmodeled dynamics. The ensemble of relationships between the control inputs and design parameters to the state variables and to the system observations is referred to as the *forward model* for the optimal design problem.

In most cases, the system forward model is represented by a set of algebraic and/or differential equations (with associated boundary and initial conditions) that the internal states of the system must satisfy. These ideas can be described somewhat more formally if we consider a linear vector space V as the space in which the values of the state variables are elements. Let L represent an operator from V into V and let g be a vector in V. An abstract equation for the system state has the form

$$L(p,q)u(p,q) = g(p,q), \tag{7.1}$$

where p represents a vector of model parameters including uncontrolled external forces and control inputs, and q represents the design parameters. Consider an operator H from V into R^n. The observation of the system can then be given by

$$y(p,q) = H(p,q)u(p,q). \tag{7.2}$$

The system forward model consists of the combination of Eq. (7.1) and Eq. (7.2). A system forward model is said to be *well-posed*, if and only if, for specified control inputs and design parameters, and in the absence of

any uncontrolled inputs, the values of the state variable are uniquely determined and depend continuously on the various parameters that appear in the model equations. Thus the demonstration of the existence, uniqueness, and continuous dependence of the solution of the set of algebraic and differential equations which comprise the forward model is an essential aspect of establishing the well-posedness of the system forward model. In addition, in the context of optimal design it is often especially important to ensure that solutions to the forward model equations depend continuously on the control input and the design parameters. This plays a key role in establishing that the optimal design problem itself admits a solution.

Once the well-posedness of the forward model is established for a specified control input, the observations may then be considered to be functions of the design parameters. In this way, we are able to translate the objective of our design problem into an unambiguously defined quantity that depends continuously on the observations, control inputs, and design parameter values. This quantity is referred to as the performance index of the design. It is important to emphasize that the performance index is, in general, only a function of the observations of the system, control inputs, and design variables. Letting F from R^n to R^1 be a given function, a performance index has the form

$$J(q;p) = F(y(p,q)). \tag{7.3}$$

In particular, J is not an explicit function of the internal state variables. This is due to the fact that typically, in an actual system, the internal states are not accessible (i.e. observable or measurable) to the designer and the device application. In many cases the performance index includes terms that measure the difference between the desired system response and the forward model predicted response. In these cases, an optimal design is sought that yields a minimal value for the performance index. A generic form of the performance index which also includes these cost factors takes the form

$$J(q;p) = F(y(p,q)) + G(q,q_0), \tag{7.4}$$

where G is a functional that measures deviation of the design parameters, q, from a nominal value, q_0.

With the performance index simply stated as a function of finitely many design parameters, it may appear that an optimization problem is well defined. However, the relationship between the values of the design parameters and the performance index is usually defined implicitly through the forward model. There are two distinct philosophical approaches that are commonly taken in dealing with this situation. The first approach considers the problem as a state-equation-constrained optimization problem. In this view, the

performance index is a function of the internal state variables and the design variables. The optimization problem is to minimize the performance index under the constraint that the forward model equations are satisfied. In the context of the electromagnetic waveguide design problem described earlier, in this approach, the performance index is viewed as a function of the electric and magnetic fields over the entire space inside and outside the waveguide. This is in addition to it also being viewed as an explicit function of the design parameters as well. Thus, this optimization problem is defined over an infinite dimensional vector space whose elements are functions (describing the electric and magnetic fields). At the same time, the system forward model represents infinitely many constraints on the optimization problem. The well-posedness of the forward model equation provides a guarantee that the admissible set for this constrained minimization problem is not empty.

An alternative approach to dealing with the implicit nature of the performance index is to consider the forward model as an implicit definition of the relationship between the design parameters and the internal state variables and, therefore, between the design parameters and the value of the performance index. In this approach, the only constraints associated with the optimization problem are those explicitly imposed on the values of the design parameters. Those constraints are usually finite dimensional, but the evaluation of the performance index requires the solution of an infinite-dimensional system of equations. Although at this level, the difference between these two approaches seems purely academic, the directions that these two approaches take us in attempting to solve the resulting optimization problem can be dramatically different. For example, to avoid dealing with the constraints explicitly, residuals between the model equations with candidate design parameter values, evaluated at a proposed value for the state variable, are added to the performance index as a *penalty* term. As a result, an optimal solution to the resulting *augmented* performance index may not precisely satisfy the forward model. Thus, in this cases, the state constraints are said to not have been strictly enforced.

In optimal device design problems, the forward model represents our best understanding of the relationship that exists between the system behavior and possible design choices as characterized by the design parameter values. It is therefore important that the forward model equation be satisfied. Consequently, in this treatment, we primarily consider the forward equation as an implicit definition of the relationship between design parameters and the performance index. Thus, in our approach here, the state constraints are strictly enforced. However, in either approach, we cannot avoid the fact that if the model equations are set in an infinite dimensional (e.g. functional)

vector space, they must be approximated by a system of finite dimensional equations so as to permit the use of numerical optimization techniques and high speed computing. As a consequence, numerical sampling must be used to approximate infinite dimensional state variables via finite dimensional vectors (once an appropriate basis has been chosen) and, at the same time, approximate any (ordinary, partial, or functional) differential equations by finite dimensional algebraic equations. More specifically, we consider a finite dimensional subspace V^N of the state space V and an operator L^N from V^N into V^N. The finite dimensional approximation of Eq. (7.1) has the form

$$L^N(p,q)u^N(p,q) + g^N(p,q) = 0, \tag{7.5}$$

where $g^N(p,q)$ is a vector in V^N representing an approximation to $g(p,q)$. Similarly, an approximating observation operator H^N is needed to define an approximating observation

$$y^N(p,q) = H^N(p,q)u(p,q). \tag{7.6}$$

The approximating performance index now takes the form

$$J^N(q;p) = F(y^N(p,q)) + G(q,q_0). \tag{7.7}$$

The significance of numerical approximation and its careful analysis in the context of optimal design problems cannot be over stated. In fact, in virtually any realistic problem, in determining an optimal design, we solve an optimization problem for a somewhat modified performance index J^N derived from the original performance index, J, through the introduction of finite dimensional approximation. Consequently, we must address the issue of convergence of not only the solution to the approximating forward model equations n^N to the solution to the original infinite dimensional forward model equations, for any given values of design variables, but also the convergence of the optimal solutions q^N_{optimal} to the approximating optimal design problems to a solution to the original underlying infinite dimensional optimal design problem as the level of approximation or discretization of the forward model equation and the performance index is further refined.

Mathematical system theory and results for ordinary, partial, and functional differential equations, and their finite dimensional discrete approximation, provide an underlying abstract theoretical framework for reducing an optimal design problem involving infinite dimensional constraints to the computationally tractable problem of minimizing an approximating finite dimensional performance index over a finite dimensional approximation of the space of design parameters. However, it is important to recognize that the optimization of the approximating performance index subject to finitely many constraints over a finite, but typically high dimensional design

parameter space, still presents a number of significant challenges. In particular, the highly implicit nature of the performance index makes it extremely difficult to determine key properties (e.g. convexity) of the underlying optimization problem.

In general, two classes of mathematical programming technique are used to search for optimal solutions: local optimization and global optimization. The local optimization approach assumes that a reasonable sub-optimal solution is available. The goal in this case, is to locate an optimal design in the neighborhood of a sub-optimal solution. However, the size of the neighborhood that a local optimization technique explores is, in general, not explicitly specified. Thus, a local scheme is more accurately characterized by the strategy used in the exploration of the design space. Indeed, in local optimization, the search for an optimal solution is guided by local trends and behavior of the performance index in the neighborhood of a candidate sub-optimal design. These local trends are typically (if the performance index is smooth) identified via calculation of the gradient of the performance index, or, if necessary, via calculation of the gradient of a local approximation of the performance index by polynomial or other readily differentiated smooth functions. The adjoint method provides an efficient means of evaluating the gradient of a performance index.

A global optimization technique, on the other hand, attempts to locate the absolute extremum of the performance index over the entire design space. In the case where the performance index is a continuous function of the design parameters and the design space consists of a compact subset of R^n, the existence of a global optimum is guaranteed.

In the succeeding sections of this chapter, we provide a detailed description of how to mathematically formulate optimal design problems, how to impose finite dimensional approximation when it is required, and how to design and implement both local and global optimization schemes for locating extrema. We also discuss theoretical issues related to the numerical implementation of these techniques, in particular the notions of stability and convergence.

7.2 Constrained local optimal design

We begin by mathematically formulating an abstract, possibly infinite dimensional, optimal design problem. Although it may be possible to say something about the existence and/or uniqueness of either local or global solutions, the resulting optimization problem, being infinite dimensional, typically cannot be solved in any practical manner as it is formally stated. A solution, typically using high speed computing, requires finite dimensional

approximation. Consequently, we then formulate and present an abstract finite dimensional approximation framework that yields a sequence of finite dimensional approximating optimal design problems which are numerically solvable. We next discuss techniques for efficiently locating local optima for the approximating design problems and address the issue of convergence of these approximating optimal designs to a solution to the original infinite dimensional optimal design problem. Finally we illustrate how all of this works in actual practice by considering the optimal design of a layered nanoscale electronic semiconductor device involving the determination of optimal layer potentials that produce a desired ballistic electron transmission probability as a function of applied voltage bias.

Let $\{X, \|\cdot\|_X\}$, $\{U, \|\cdot\|_U\}$, $\{Y, \|\cdot\|_Y\}$, $\{X_0, \|\cdot\|_{X_0}\}$, and $\{W, \|\cdot\|_W\}$ be normed linear spaces over the complex numbers that denote respectively, the state space, the design space, the measurement or observation space, the space of initial conditions, and the external excitation space for our optimal design problem. The variables that describe the state of the system, for example, displacement, velocity, temperature, concentration, etc., will be elements of the state space X. The design parameters, for example, layer potential or thickness, mass or stiffness, Young's Modulus, thermal and other material parameters, etc., will be elements of the design space U. The quantities that can be measured or observed that will form the basis for the optimality in the term optimal design will be elements in the measurement or observation space Y. The space X_0 typically contains elements that describe initial conditions for the state, while the entries in the excitation space, W, represent any external influences acting on the system, for example forces, heat sources or sinks, etc. On occasion, but not often, the quantities that can be measured coincide with the system states. But this is very rare. In general, any or all of these spaces can be infinite dimensional. This will depend upon the nature and character of the optimal design problem and the mathematical elements that it takes to describe them. The elements in these spaces may be scalars, vectors, functions (of time, or location, for example), or even random variables if uncertainty represents a significant modeling issue that must be dealt with in formulating the problem.

We assume that the state of the underlying system may be described or modeled by a state equation which takes the general form

$$f(x, u; x_0, w) = 0, \tag{7.8}$$

where the state function

$$f : X \times U \times X_0 \times W \to X, \tag{7.9}$$

is in general continuous and possibly abstractly differentiable in either a

Frechet or Gateaux sense, $x_0 \in X$ and $w \in W$ denote some sort of exogenous inputs in the form of initial conditions and external excitation, respectively. The variables $x \in X$ and $u \in U$ denote respectively the system state variables and design parameters. In general, the state variables and/or the design parameters are functional in nature, the function f may involve either time or space derivatives or integrals of x and u. For example, if the design problem of interest to us involved the identifying of the mass, stiffness, and damping of a system that can be modeled as a simple harmonic oscillator, then in this case we might take the space X to be $PC[0,T] \times PC[0,T]$, where $PC[0,T]$ denotes the piecewise continuous functions on the interval $[0,T]$, W to be the space $PC[0,T]$, U to be R^3 and the space X_0 to be the space R^2. In this case, the state function $f : X \times U \times X_0 \times W \to X$ would be given by

$$f(x,u;x_0,w) = \mathbf{x}(t) - e^{\mathbf{A}t}\mathbf{x}_0 - \int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}w(s)ds, \tag{7.10}$$

where $\mathbf{A}$ is the 2×2 matrix and $\mathbf{B}$ is the 2×1 matrix respectively given by $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -k/m & -b/m \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 0 \\ 1/m \end{bmatrix}$, $\mathbf{x}_0 \in \mathrm{R}^2$, and the design parameters, $\mathbf{u} \in \mathrm{R}^3$, are given by $\mathbf{u} = [m,k,b]^{\mathrm{T}}$.

We define a performance index for the optimal design problem, $J : U \to \mathrm{R}$, in terms of an objective functional $g : X \times U \times Y \to \mathrm{R}$ as

$$J(u) = g(x,u;y), \tag{7.11}$$

where $y \in Y$ is a given observation or measurement in terms of which the desired optimal performance of the system can be described. So to continue with the simple harmonic oscillator example given above, suppose that $y \in C[0,T]$ is the desired displacement trajectory that you would like the mass to trace out corresponding to the given external excitation $w \in C[0,T]$ and initial displacement and velocity $\mathbf{x}_0 \in \mathrm{R}^2$. In this case we might take $J : U \to \mathrm{R}$ to be given by

$$J(u) = \sum_{i=1}^{\nu} \left| y(t_i) - Cx(t_i;u) \right|^2, \tag{7.12}$$

where $\{t_i\}_{i=1}^{\nu} \subset [0,T]$, $\mathbf{C} = [1,0]$, and $\mathbf{x}(\cdot;u) \in PC[0,T]$ is the solution to

$$f(x,u;x_0,w) = \mathbf{x}(t) - e^{\mathbf{A}t}\mathbf{x}_0 - \int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}w(s)ds = \mathbf{0}, \tag{7.13}$$

corresponding to $\mathbf{u} = [m,k,b]^{\mathrm{T}} \in \mathrm{R}^3$. If we let $\Omega \subset U$ denote the feasible parameter set, our abstract optimal design problem may then be formulated as:

Given $x_0 \in X_0$ and $w \in W$,
find $u* = \arg\min_{u\in\Omega} J(u) = \arg\min_{u\in\Omega} g(x, u; y)$,
subject to $f(x, u; x_0, w) = 0$.

We turn our attention next to finite dimensional approximation. For ease of exposition we will assume only that it is only the state space, X, that is infinite dimensional and requires finite dimensional approximation. This will capture the essence of what is required to make our general approach clear without adding an excess of technical complexity. Moreover, this is often the situation in practice. Toward this end for each $M = 1, 2, \ldots$, let X^M denote a k^M-dimensional subspace of X. In general it should be clear how one could readily introduce finite dimensional approximation in the parameter, initial condition, observation, and external excitation spaces if it were called for in a particular situation. We assume that the remaining spaces, U, X_0, and W are all either finite dimensional or do not require finite dimensional approximation. In addition let $f^M : X^M \times U \times X_0 \times W \to X^M$ and $g^M : X^M \times U \times Y \to \mathrm{R}$, and define the sequence of approximating finite dimensional optimal design problems by:

Given $x_0 \in X_0$ and $w \in W$,
find $u^{M*} = \arg\min_{u\in\Omega} J^M(u) = \arg\min_{u\in\Omega} g^M(x^M, u; y)$,
subject to $f^M(x^M, u; x_0, w) = 0$.

It is these latter approximating problems that will actually be solved for u^{M*}. Of course we would hope that u^{M*} will, in some sense, approximate $u*$ a solution to the original optimal design problem; that is that $\lim_{M\to\infty} u^{M*} = u^*$. We will address the issue of convergence later. First we consider some of the ideas involved in actually solving the approximating finite dimensional optimal design problems.

In this section we are only concerned with identifying local optima. Finding global optima will be discussed elsewhere in this chapter. Assuming that all functions and variables depend smoothly on the design parameters, local optima are typically located via gradient based search. Whether it be steepest descent, Newton's method, or conjugate gradient, constrained or unconstrained, virtually all methods depend upon being able to compute the gradient of the performance index, J^M, with respect to the design parameters $u \in U$. Indeed, the general form of these methods, which are iterative in nature, is

$$u_{k+1} = u_k - H(u_k) \nabla J^M(u_k)^{\mathrm{T}}, \quad k = 0, 1, 2, \ldots \tag{7.14}$$

where the particular form of $H(u)$ depends on which method is being used.

For example, in the case of steepest descent, $H(u)$ is simply a scalar and successive iterates are determined from their predecessor by moving in the direction of the negative gradient. On the other hand, in the case of Newton's method, which is based upon locating a stationary point of J^M in the form of a zero of the gradient, $H(u)$ involves the inverse of the Hessian.

However, in any case, whichever method is used, because these methods are iterative in nature, the topology of the design landscape can be both complex and of high (albeit finite) dimension, and evaluating J^M and ∇J^M involves effectively numerically solving infinite dimensional operator equations in order to solve the state equation for x^M, so that it is essential that gradient calculation be both highly accurate and computationally very efficient.

One approach would be to use a finite-difference-based scheme. But this has two fundamental drawbacks. First it introduces truncation error into the gradient calculation which can make convergence slow, especially near the optimum. It also requires solving the state equation numerous times in each iteration. This is the most time-consuming step in evaluating J^M and consequently it makes the entire approach computationally very slow. The adjoint method, on the other hand, which is commonly used in data assimilation for numerical weather prediction, a problem which shares many of these same challenges, allows for the efficient computation of gradients of dynamically constrained performance indices with zero truncation error.

The adjoint method has its roots in the classical theory of constrained optimization, Lagrange multipliers, and optimal control. Associated with and coupled to the constraints on the system state variables, is another, related, system of equations known as the adjoint. The variables in the adjoint system are known as the co-states, dual variables, or Lagrange multipliers. At optimality, the co-state variables may be interpreted as the marginal benefit derived (with respect to the underlying performance index being optimized) by weakening the constraints on the state variables. Economists sometimes refer to them as either *imputed* or *shadow prices*. However, at nonstationary points, the adjoint formulation may be used to facilitate the efficient computation of the gradient of the performance index with respect to the optimization parameters. It does this by making it unnecessary to compute the derivatives of the state variables with respect to the optimization parameters.

To see how it works, we derive the basic equations in terms of our abstract approximating problems and then see what it actually looks like in the context of our simple harmonic oscillator example. We begin by formally differentiating the performance index, J^M, to obtain

$$\nabla J^M(u) = \frac{\partial g^M(x^M, u; y)}{\partial x^M}\frac{\partial x^M}{\partial u} + \frac{\partial g^M(x^M, u; y)}{\partial u}. \tag{7.15}$$

Similarly, differentiating the state equation we obtain

$$\frac{\partial f^M(x^M, u; x_0, w)}{\partial x^M}\frac{\partial x^M}{\partial u} = -\frac{\partial f^M(x^M, u; x_0, w)}{\partial u}. \tag{7.16}$$

We define the adjoint or co-state variables, $z^M \in X^M$, to be the solution to the linear system given by

$$\left(\frac{\partial f^M(x^M, u; x_0, w)}{\partial x^M}\right)^* z^M = \left(\frac{\partial g^M(x^M, u; y)}{\partial x^M}\right)^*, \tag{7.17}$$

where, for a matrix Λ with complex entries, Λ^* denotes its conjugate transpose. It then follows that

$$\begin{aligned}\nabla J^M(u) &= \frac{\partial g^M(x^M, u; y)}{\partial x^M}\frac{\partial x^M}{\partial u} + \frac{\partial g^M(x^M, u; y)}{\partial u} \\ &= \left(z^M\right)^* \frac{\partial f^M(x^M, u; x_0, w)}{\partial x^M}\frac{\partial x^M}{\partial u} + \frac{\partial g^M(x^M, u; y)}{\partial u} \\ &= -\left(z^M\right)^* \frac{\partial f^M(x^M, u; x_0, w)}{\partial u} + \frac{\partial g^M(x^M, u; y)}{\partial u}.\end{aligned} \tag{7.18}$$

The adjoint method for calculating the gradient may then be stated as follows:

Given $x_0 \in X_0$,$w \in W$,$y \in Y$ and $u \in U$, Step 1: Solve the state equation $f^M(x^M, u; x_0, w) = 0$ for the state variables $x^M \in X^M$. Step 2: Solve the adjoint equation $\left(\frac{\partial f^M\left(x^M, u; x_0, w\right)}{\partial x^M}\right)^* z = \left(\frac{\partial g^M\left(x^M, u; y\right)}{\partial x^M}\right)^*$ for the adjoint variables $z^M \in X^M$. Step 3: Compute the gradient of the performance index from

$$\nabla J^M(u) = -\left(z^M\right)^* \frac{\partial f^M(x^M, u; x_0, w)}{\partial u} + \frac{\partial g^M(x^M, u; y)}{\partial u}.$$

We illustrate the application of the adjoint method on the simple harmonic oscillator example considered above. We begin with the finite dimensional approximation of the state space $X = PC[0,T] \times PC[0,T]$. For each $M = 1, 2, \ldots$ let $\{\varphi_i^M\}_{i=1}^M \subset PC[0,T]$ be a linearly independent set. For $i = 1, 2, \ldots, M$, define elements in X, $\Phi_i^M = \left[\varphi_i^M \; 0\right]^{\mathrm{T}}$ and $\Phi_{i+M}^M = \left[0 \; \varphi_i^M\right]^{\mathrm{T}}$ and set $X^M = \operatorname{span}\{\Phi_i^M\}_{i=1}^{2M}$. We endow $X = PC[0,T] \times PC[0,T]$ with the standard $L_2[0,T] \times L_2[0,T]$ inner product, $\langle \cdot, \cdot \rangle_X$, and let $P^M \colon X \to X^M$

denote the orthogonal projection of X onto X^M. We define the approximating state and objective functions

$$f^M : X^M \times U \times X_0 \times W \to X^M, \tag{7.19}$$

and

$$g^M : X^M \times U \times Y \to \mathrm{R}, \tag{7.20}$$

by

$$f^M\left(\mathbf{x}^M, \mathbf{u}; \mathbf{x}_0, w\right) = P^M f\left(\mathbf{x}^M, \mathbf{u}; \mathbf{x}_0, w\right), \tag{7.21}$$

and

$$g^M\left(\mathbf{x}^M, \mathbf{u}; y\right) = g\left(\mathbf{x}^M, \mathbf{u}; y\right), \tag{7.22}$$

respectively. An easy calculation based on the normal equations reveals that the approximating state equation is given by

$$f^M\left(\mathbf{x}^M, \mathbf{u}; \mathbf{x}_0, w\right) = \mathbf{Q}^M \alpha^M - \mathbf{b}^M\left(\mathbf{u}; \mathbf{x}_0, w\right), \tag{7.23}$$

where $\mathbf{Q}^M_{i,j} = \left\langle \Phi^M_i, \Phi^M_j \right\rangle_X$, $i, j = 1, 2, \ldots, 2M$, $\alpha^M = \left[\, \alpha^M_1 \ \alpha^M_2 \ \ldots \ \alpha^M_{2M} \,\right]^{\mathrm{T}}$, $x^M = \sum_{i=1}^{2M} \alpha^M_i \Phi^M_i$ and $\mathbf{b}^M\left(\mathbf{u}; x_0, w\right) = \int_0^T \Phi^M\left(t\right)^{\mathrm{T}} e^{\mathbf{A}t}\mathbf{x}_0 - \int_0^t \Phi^M\left(t\right)^{\mathrm{T}} e^{\mathbf{A}(t-s)}\mathbf{B}w\left(s\right) dsdt$, with $\Phi^M\left(t\right) = \left[\, \Phi^M_1\left(t\right) \Phi^M_2\left(t\right) \ldots \Phi^M_{2M}\left(t\right) \,\right]^{\mathrm{T}}$. It also follows that the objective function is given by

$$g^M\left(\mathbf{x}^M, \mathbf{u}; y\right) = \sum_{i=1}^{\nu} \left| y\left(t_i\right) - \mathbf{C}\Phi^M\left(t_i\right)\alpha^M \right|^2. \tag{7.24}$$

Recalling that the $2M \times 2M$ matrix $\mathbf{Q}^M$ is symmetric (and positive definite and therefore invertible since the $\{\Phi^M_i\}_{i=1}^{2M}$ are linearly independent), the adjoint system is given by

$$\mathbf{Q}^M \boldsymbol{\beta}^M = -2\sum_{i=1}^{\nu} \left(y\left(t_i\right) - \mathbf{C}\Phi^M\left(t_i\right)\alpha^M \right) \Phi^M\left(t_i\right)^{\mathrm{T}} \mathbf{C}^{\mathrm{T}}. \tag{7.25}$$

The gradient of J^M, ∇J^M, can then be computed with no truncation error from

$$\begin{aligned}
\nabla J^M\left(\mathbf{u}\right) &= -\left(\boldsymbol{\beta}^M\right)^{\mathrm{T}} \frac{\partial}{\partial \mathbf{u}} \left\{ \mathbf{Q}^M \alpha^M - \mathbf{b}^M\left(\mathbf{u}; \mathbf{x}_0, w\right) \right\} \\
&= \left(\boldsymbol{\beta}^M\right)^{\mathrm{T}} \frac{\partial \mathbf{b}^M\left(\mathbf{u}; \mathbf{x}_0, w\right)}{\partial \mathbf{u}} \\
&= \left(\boldsymbol{\beta}^M\right)^{\mathrm{T}} \frac{\partial}{\partial \mathbf{u}} \left\{ \int_0^T \Phi^M\left(t\right)^{\mathrm{T}} e^{\mathbf{A}t}\mathbf{x}_0 - \int_0^t \Phi^M\left(t\right)^{\mathrm{T}} e^{\mathbf{A}(t-s)}\mathbf{B}w\left(s\right) dsdt \right\}
\end{aligned}$$

$$= \left(\boldsymbol{\beta}^M\right)^{\mathrm{T}} \left\{ \int_0^T \Phi^M(t)^{\mathrm{T}} \frac{\partial e^{\mathbf{A}t}\mathbf{x}_0}{\partial \mathbf{u}} - \int_0^t \Phi^M(t)^{\mathrm{T}} \frac{\partial e^{\mathbf{A}(t-s)}\mathbf{B}w(s)}{\partial \mathbf{u}} ds dt \right\} \tag{7.26}$$

where $\mathbf{u} = [m, k, b]^{\mathrm{T}}$.

We refer to the method described above for calculating gradients as the static adjoint. When the state equation is of evolutionary type and the state of the system is specified recursively, there is a dynamic formulation of the adjoint method in which the adjoint variables or costate are also specified recursively. To see how this works, we once again consider our simple harmonic oscillator example but with the added assumption that the times specified in the performance index are equally spaced and with the added restriction that the external excitation is zero-order hold.

Let $t_k = kT/\nu = k\tau$, $k = 0, 1, 2, \ldots, \nu$, and set $y_k = y(t_k) = y(k\tau)$, $\mathbf{x}_k = \mathbf{x}(t_k) = \mathbf{x}(k\tau)$, $w_k = w(t_k) = w(k\tau)$, $k = 0, 1, 2, \ldots, \nu$. We then take the state space $X = \times_{i=0}^{\nu} \mathrm{R}^2$, the observation space $Y = \times_{i=1}^{\nu} \mathrm{R}$, the external excitation space $W = \times_{i=0}^{\nu-1} \mathrm{R}$, and, as before, U to be R^3 and the space X_0 to be the space R^2. The system state equation and objective functions are then given by

$$f\left(\{\mathbf{x}_i\}_{i=0}^{\nu}, u; \mathbf{x}_0, \{w_i\}_{i=0}^{\nu-1}\right) = \left\{ \mathbf{x}_k - \hat{\mathbf{A}}\mathbf{x}_0 - \sum_{j=0}^{k-1} \hat{\mathbf{A}}^{k-1-j} \hat{\mathbf{B}} w_j \right\}_{k=0}^{\nu}, \tag{7.27}$$

and

$$g\left(\{\mathbf{x}_i\}_{i=0}^{\nu}, \mathbf{u}; \{y_i\}_{i=1}^{\nu}\right) = \sum_{k=1}^{\nu} \left| y_k - \mathbf{C}\mathbf{x}_k \right|^2, \tag{7.28}$$

respectively, where $\hat{\mathbf{A}} = e^{\mathbf{A}\tau}$ and $\hat{B} = \int_0^\tau e^{\mathbf{A}t}\mathbf{B}dt$. Note that in this case, no finite-dimensional approximation is required because all of the spaces are already finite dimensional. In addition, in this case, it is not difficult to see that the state variables $\{\mathbf{x}_k\}_{k=0}^{\nu}$ evolve recursively according to the discrete dynamical system

$$\mathbf{x}_{k+1} = \hat{A}\mathbf{x}_k + \hat{B}w_k, k = 0, 1, 2, \ldots, \nu. \tag{7.29}$$

We define the adjoint variables recursively in reverse by

$$\mathbf{z}_{k-1} = \hat{A}^{\mathrm{T}}\mathbf{z}_k - 2\left(y_k - \mathbf{C}\mathbf{x}_k\right)\mathbf{C}^{\mathrm{T}}, \tag{7.30}$$

with $k = \nu, \nu-1, \ldots, 2, 1, z_\nu = 0$. Then noting that $\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{u}} = \frac{\partial \hat{\mathbf{A}}}{\partial \mathbf{u}} x_k + \hat{\mathbf{A}} \frac{\partial \mathbf{x}_k}{\partial \mathbf{u}} + \frac{\partial \hat{\mathbf{B}}}{\partial \mathbf{u}} w_k$, $\frac{\partial \mathbf{x}_0}{\partial \mathbf{u}} = 0$, and $z_\nu = 0$, it follows that

$$\begin{aligned}\nabla J(\mathbf{u}) &= -2\sum_{k=1}^{\nu}(y_k - \mathbf{C}\mathbf{x}_k)\,\mathbf{C}\frac{\partial \mathbf{x}_k}{\partial \mathbf{u}} \\ &= \sum_{k=1}^{\nu}\left\{\mathbf{z}_{k-1} - \hat{\mathbf{A}}\mathbf{z}_k\right\}^{\mathrm{T}}\frac{\partial \mathbf{x}_k}{\partial \mathbf{u}} \\ &= \sum_{k=0}^{\nu-1}\mathbf{z}_k^{\mathrm{T}}\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{u}} - \sum_{k=1}^{\nu}\mathbf{z}_k^{\mathrm{T}}\hat{\mathbf{A}}\frac{\partial \mathbf{x}_k}{\partial \mathbf{u}} \\ &= \sum_{k=0}^{\nu-1}\mathbf{z}_k^{\mathrm{T}}\left\{\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{u}} - \hat{\mathbf{A}}\frac{\partial \mathbf{x}_k}{\partial \mathbf{u}}\right\} \\ &= \sum_{k=0}^{\nu-1}\mathbf{z}_k^{\mathrm{T}}\left\{\frac{\partial \hat{\mathbf{A}}}{\partial \mathbf{u}}\mathbf{x}_k + \frac{\partial \hat{\mathbf{B}}}{\partial \mathbf{u}}w_k\right\}. \end{aligned} \tag{7.31}$$

In this case, the adjoint method for computing the gradient dynamically is given by:

Given $\mathbf{x}_0 \in X_0$,$\{w_i\}_{i=0}^{\nu-1} \in W$, $\{y_i\}_{i=1}^{\nu} \in Y$ and $\mathbf{u} = [m,k,b]^{\mathrm{T}} \in U$,
Step 1: Solve the state equation $\mathbf{x}_{k+1} = \hat{\mathbf{A}}\mathbf{x}_k + \hat{\mathbf{B}}w_k$ for the state variables $\{\mathbf{x}_k\}_{k=0}^{\nu} \in X$.
Step 2: Solve the adjoint equation $\mathbf{z}_{k-1} = \hat{\mathbf{A}}^{\mathrm{T}}\mathbf{z}_k - 2(y_k - \mathbf{C}\mathbf{x}_k)\,\mathbf{C}^{\mathrm{T}}$, $\mathbf{z}_\nu = \mathbf{0}$, for the adjoint variables $\{z_k\}_{k=\nu}^{0} \in X$.
Step 3: Compute the gradient of the performance index from

$$\nabla J(\mathbf{u}) = \sum_{k=0}^{\nu-1} z_k^{\mathrm{T}}\left\{\frac{\partial \hat{\mathbf{A}}}{\partial \mathbf{u}}\mathbf{x}_k + \frac{\partial \hat{\mathbf{B}}}{\partial \mathbf{u}}w_k\right\}.$$

There are a number of theoretical questions that it is natural to ask. In particular, these might include whether or not it is possible to establish the existence and/or uniqueness of solutions to the optimal design problems, and is it possible to demonstrate the convergence of solutions to the approximating optimal design problem to a solution to the original optimal design problem. In considering the first question, the arguments typically used to demonstrate existence and uniqueness of solutions to optimization problems in general do not apply to the kinds of problems of interest to us here. These arguments are typically based on convexity. However, the performance indices that arise in the present context are typically highly nonlinear functions of the optimization parameters. Establishing convexity, if in fact it is even present, is often out of the question. Moreover, it is almost always the case that we are unable to express the performance index as an

explicit function of the design parameters. More typically the performance index is a function of the solution to an ordinary or partial differential equation in which the design variables appear as parameters (coefficients, inputs, boundary or initial conditions, etc.).

However, if we assume that a (approximating) feasible subset $\Omega \subseteq U$ of the design variable space is compact (if U is finite dimensional, this is equivalent to closed and bounded) and we can establish that the (approximating) performance index depends continuously on the design variables, then it is a well-known consequence of elementary analysis that the (approximating) optimal design problems admit a (possibly non-unique) solution.

Establishing the convergence of solutions of the approximating optimal design problems to a solution of the original optimal design problem requires demonstrating that solutions to the approximating state equations converge to the solution of the original state equation uniformly with respect to the design parameters. We make this precise by stating it as a formal proposition.

Proposition 1. Let $x^M(u) \in X^M$ and $x(u) \in X$ denote respectively solutions to the approximating and original state equations corresponding to $u \in U$, and suppose that $\{u^M\}_{M=1}^{\infty} \subset \Omega$ is a convergent sequence in U with $\lim_{M\to\infty} \|u^M - u_0\|_U = 0$ for some $u_0 \in \Omega$. Then $\lim_{M\to\infty} \|x^M(u^M) - x(u_0)\|_X = 0$.

Proving Proposition 1 of course depends on the particular problem at hand. However, if it can be established, then Proposition 2 below follows at once.

Proposition 2. Suppose that the feasible design parameter set $\Omega \subseteq U$ is a compact subset of U, that the objective function $g : X \times U \times Y \to \mathrm{R}$ is continuous, and that Proposition 1 holds. Let $\bar{u}^M \in \Omega$ be a solution to the approximating optimal design problem corresponding to the index M. Then there exist a convergent subsequence $\{\bar{u}^{M_k}\}_{k=1}^{\infty} \subset \{\bar{u}^M\}_{M=1}^{\infty} \subset \Omega$ with $\lim_{k\to\infty} \bar{u}^{M_k} = \bar{u}$ for some $\bar{u} \in \Omega$. Moreover, $\bar{u}$ is a solution to the original optimal design problem.

Proof. The existence of the convergent subsequence follows from the compactness of Ω. Then for every $u \in \Omega$

$$J(\bar{u}) = J\left(\lim_{k\to\infty} \bar{u}^{M_k}\right) = g\left(x\left(\lim_{k\to\infty} \bar{u}^{M_k}\right), \lim_{k\to\infty} \bar{u}^{M_k}, y\right)$$
$$= g\left(\lim_{k\to\infty} x^{M_k}\left(\bar{u}^{M_k}\right), \lim_{k\to\infty} \bar{u}^{M_k}, y\right) = \lim_{k\to\infty} g\left(x^{M_k}\left(\bar{u}^{M_k}\right), \bar{u}^{M_k}, y\right)$$

$$
\begin{aligned}
&= \lim_{k\to\infty} g^{M_k}\left(x^{M_k}\left(\bar{u}^{M_k}\right), \bar{u}^{M_k}, y\right) \\
&= \lim_{k\to\infty} J^{M_k}\left(\bar{u}^{M_k}\right) \leq \lim_{k\to\infty} J^{M_k}(u) = \lim_{k\to\infty} g\left(x^{M_k}(u), u, y\right) \\
&= g\left(\lim_{k\to\infty} x^{M_k}(u), u, y\right) = g(x(u), u, y) = J(u), \qquad (7.32)
\end{aligned}
$$

and the desired result follows.

In some instances it may be possible to establish that the sequence itself converges to a solution of the original optimal design problem. Showing this typically requires uniqueness.

In the next section we show how the ideas we have presented here can be applied in the context of a realistic optimal design problem for a nanoscale layered quantum electronic device.

7.3 Local optimal design of an electronic device

As an example of how we mathematically formulate, and develop efficient computational techniques to solve, the optimal design problem, we consider the design of an electronic semiconductor device whose conduction band potential profile, $V(x)$, can be fabricated with great accuracy in the crystal growth direction, x. The behavior of electronic devices with layers that are a few nm thick may often be characterized by ballistic electron transmission probability as a function of applied voltage bias. The design parameters are the values of potential at each atomic layer and the design criterion is a desired functional relationship between an applied voltage bias, V_{bias}, and electron transmission, T. (For example, in the case of a resistor, by virtue of Ohm's Law, this functional relationship is linear. Here, we are interested in devices that yield much more general functional relationships.) We formulate a design problem in terms of identifying designs for $V(x)$ that result in locally optimal electron transmission characteristics $T = T(V_{\text{bias}})$. The design criterion is formulated in terms of the squared difference between the desired and observed performance of the device. We solve the optimal design problem by seeking local minima via a gradient-based search that makes use of the adjoint method to facilitate efficient and accurate computation of gradients. The underlying constraints involve differential equations. Consequently we will rely on some form of finite-difference approximation to make the optimization problem amenable to numerical solution. We will demonstrate how mathematical analysis can be used to establish the numerical stability and, more significantly, the numerical convergence of the scheme.

7.3.1 The optimal design problem

We consider a layered nanoscale semiconductor electronic device schematically configured as shown in Fig. 7.1. The device is assumed to be of total thickness L and to consist of N layers. For $i = 1, 2, \ldots, N$, the ith layer begins and ends at positions x_{i-1} and x_i, respectively, and is of thickness $L_i = x_i - x_{i-1}$. It follows therefore that $x_N - x_0 = L$. The local potential energy in the ith barrier-layer is assumed to be U_i, with $i = 1, 2, \ldots, N$. For $x < x_0$, the local potential energy is assumed to be U_0 and for $x > x_N$, it is assumed to be U_{N+1}. We assume that a single electron propagating in the right electrode from $-\infty$ is incident upon the left boundary of the device at x_0 and that a voltage bias, V_{bias}, is applied across the device.

Typically, application of the voltage bias illustrated in Fig. 7.1 has the effect of creating an accumulation of charge on the left side of the barrier layers and a depletion region on the far right. Obtaining the precise form of the resulting potential energy profile $V(x)$ requires the solution of an appropriate Poisson equation [1]. However, we assume that the thickness of the depletion and accumulation layers is sufficiently small so as to allow for linear approximation. Consequently, the static potential energy profile takes the form

$$V(x) = V(x; \mathbf{U}, V_{\text{bias}}) = \begin{cases} U_0 & -\infty < x < x_0 \\ \sum_{j=1}^{N} U_j \chi_j(x) - V_{\text{bias}} \frac{x - x_0}{L} & x_0 \le x \le x_N \\ U_{N+1} - V_{\text{bias}} & x_N < x < \infty \end{cases}, \tag{7.33}$$

where $\mathbf{U} = \{U_i\}_{i=1}^{N}$ describes the local layer potentials, for each $j = 1, 2, \ldots, N$, χ_j is the characteristic function corresponding to the jth subinterval, $[x_{j-1}, x_j]$, and we assume unit electron charge. That is

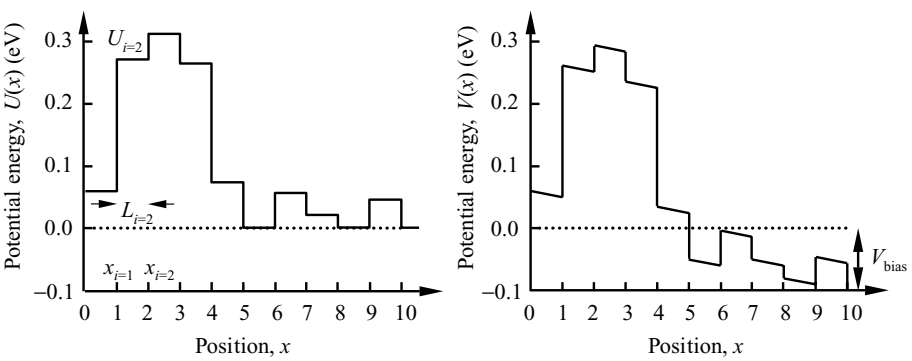

Fig. 7.1. A layered nanoscale semiconductor device (left) and the device with a bias voltage V_{bias} applied across it (right).

$$\chi_j(x) = \begin{cases} 1 & x_{j-1} \le x < x_j \\ 0 & \text{otherwise.} \end{cases} \tag{7.34}$$

The interaction of the electron with the potential $V(x)$ is described by solving for the electron wave function Ψ in the Schrödinger equation

$$-\frac{\hbar^2}{2m_0}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V(x)\Psi(x,t) = i\hbar\frac{\partial\Psi(x,t)}{\partial t} \tag{7.35}$$

where $\hbar = 1.05492 \times 10^{-34}$ J is Planck's constant, $m_0 = 9.10938188 \times 10^{-31}$ kg is the bare electron mass, and $i = \sqrt{-1}$. In a semiconductor the bare electron mass is often replaced by an effective electron mass m^*. The charge density flux, $\partial\rho_e(x,t)/\partial t$, is then given in terms of the solution, Ψ, to the Schrödinger equation, by $\partial\rho_e(x,t)/\partial t = \partial\left|\Psi(x,t)\right|^2/\partial t$ [2]. If unit electron charge is assumed and the potential energy function, $V(x)$, is taken to be real, a straightforward calculation then yields

$$\frac{\partial\rho_e(x,t)}{\partial t} + \frac{\partial}{\partial x}\frac{\hbar}{2m_0 i}\left\{\bar{\Psi}\frac{\partial\Psi}{\partial x} - \Psi\frac{\partial\bar{\Psi}}{\partial x}\right\} = 0. \tag{7.36}$$

Equation (7.36) is in the form of a continuity equation and by analogy to fluid flow suggests conservation of charge. Consequently, we define *current density* $\hat{j} = \hat{j}(x,t)$ by the expression

$$\hat{j}(x,t) = \frac{\hbar}{2m_0 i}\left\{\bar{\Psi}(x,t)\frac{\partial\Psi(x,t)}{\partial x} - \Psi(t,x)\frac{\partial\bar{\Psi}(x,t)}{\partial x}\right\}. \tag{7.37}$$

The transmission coefficient, T, of the device is then defined to be a ratio of current densities as

$$T = T(V_{\text{bias}}, \mathbf{U}) = \frac{\left|\hat{j}_{\text{trans}}\right|}{\left|\hat{j}_{\text{inc}}\right|}, \tag{7.38}$$

where $\hat{j}_{\text{trans}}$ is the current density transmitted from the device at $x = x_N$, and $\hat{j}_{\text{inc}}$ denotes the current density incident upon the device at $x = x_0$.

Since the potential $V(x)$ is independent of time, Eq. (7.35) admits a solution of the form $\Psi(x,t) = \psi(x)\varphi(t)$ which can be found via separation of variables. It follows that $\varphi(t) = \exp(-iEt/\hbar)$, where E, the separation constant, is the sum of electron kinetic and potential energy. For the conservative system we consider, energy E is a constant of the electron motion. The time-independent wave function $\psi(x)$ is then found as the solution to the second-order ordinary differential equation (time-independent Schrödinger equation) given by

$$-\frac{\hbar^2}{2m_0}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x). \tag{7.39}$$

With the potential $V(x)$ as given in Eq. (7.33), on the intervals $-\infty < x < x_0$ and $x_N < x < \infty$, the general solution to Eq. (7.39) is given by

$$\psi_0(x) = A_0 e^{ik_0 x} + B_0 e^{-ik_0 x}, \tag{7.40}$$

and

$$\psi_{N+1}(x) = A_{N+1} e^{ik_{N+1}(x-x_N)} + B_{N+1} e^{-ik_{N+1}(x-x_N)}, \tag{7.41}$$

respectively, where $k_0^2 = \frac{2m_0(E-U_0)}{\hbar^2}$, $k_{N+1}^2 = \frac{2m_0(E-U_{N+1}+V_{\text{bias}})}{\hbar^2}$, and the in general complex coefficients, $A_0,\ B_0,\ A_{N+1},\ B_{N+1}$, are determined by boundary conditions. We assume that $E \neq U_0$ and $E \neq U_{N+1} - V_{\text{bias}}$. The latter two assumptions are made so as to ensure that the time-independent Schrödinger equation (7.39) admits exponential solutions of the form Eq. (7.40) and Eq. (7.41) on the intervals $-\infty < x < x_0$ and $x_N < x < \infty$, respectively, as opposed to polynomial (i.e. linear) solutions. We exclusively treat the exponential case here; our general approach can be modified to handle the polynomial case as well. It is clear that for $-\infty < x < x_0$ and $x_N < x < \infty$, the time-dependent wave function $\Psi(x,t) = \psi(x)\varphi(t)$ is of the form

$$\Psi(x,t) = A_0 e^{ik_0\left(x-\frac{E}{\hbar}t\right)} + B_0 e^{-ik_0\left(x+\frac{E}{\hbar}t\right)}, \tag{7.42}$$

and

$$\Psi(x,t) = A_{N+1} e^{ik_{N+1}\left(x-x_N-\frac{E}{\hbar}t\right)} + B_{N+1} e^{-ik_{N+1}\left(x-x_N+\frac{E}{\hbar}t\right)}, \tag{7.43}$$

respectively. In this way, the wave function can be viewed as the sum of left and right propagating wave amplitudes. Moreover, as a result of the significant likelihood of reflection at interfaces across which there is a significant change in the electron's velocity, it is clear that the second term in Eq. (7.42) and the first term in Eq. (7.43) respectively represent the cumulative sum of the interference effects that result from the reflected and transmitted amplitude at each change in the spatial potential, $V(x)$, of the device.

It is immediately clear that $|A_0|$ is the amplitude of the electron wave function impinging on the left boundary of the device at $x = x_0$. Hence, for an electron incident from the left, $|A_0|^2 = 1$ and since there is neither transmission nor reflection from $x = +\infty$ we require $B_{N+1} = 0$. Furthermore, from Eq. (7.37), (7.42), and (7.43), the transmitted and incident current densities are given by

$$\hat{j}_{\text{trans}} = \frac{\hbar k_{N+1}}{m_0} |A_{N+1}|^2,$$

and

$$\hat{j}_{\text{inc}} = \frac{\hbar k_0}{m_0} |A_0|^2, \tag{7.44}$$

respectively. Combining Eq. (7.38) and Eq. (7.44), we immediately obtain

$$T = T\left(V_{\text{bias}}\right) = T\left(V_{\text{bias}}, \mathbf{U}\right) = \frac{k_{n+1}\left|A_{N+1}\right|^2}{k_0\left|A_0\right|^2}. \tag{7.45}$$

We formulate the optimal design problem mathematically as a constrained least-squares fit to a given desired transmission function $T_0 = T_0\left(V_{\text{bias}}\right)$ defined on the interval $V_{\min} \leq V_{\text{bias}} \leq V_{\max}$. We seek local layer potentials $\mathbf{U} = \{U_i\}_{i=1}^N$ with $U_L \leq U_i \leq U_{\text{H}}$, and $i = 1, 2, \ldots, N$, which minimize the least-squares performance index

$$J\left(\mathbf{U}\right) = \sum_{j=1}^{\nu} \left|T_0\left(V_j\right) - T\left(V_j, \mathbf{U}\right)\right|^2, \tag{7.46}$$

where $T = T\left(V_j; \mathbf{U}\right)$ is given by Eq. (7.45) with $V_{\text{bias}} = V_j$, and $V_{\min} \leq V_j \leq V_{\max}$, $j = 1, 2, \ldots, \nu$, are given arbitrarily spaced bias voltages in the interval $[V_{\min}, V_{\max}]$.

In order to evaluate $J = J\left(\mathbf{U}\right)$ given in Eq. (7.46), we seek solutions to the time-independent Schrödinger equation given in Eq. (7.39) on $(-\infty, \infty)$ which are smooth (i.e. C^1) across the device boundaries at $x = x_0$ and $x = x_N$. That is, we seek solutions $\psi = \psi\left(x\right)$, $-\infty < x < \infty$, to Eq. (7.39) satisfying the boundary conditions

$$\psi\left(x_0\right) = \psi_0\left(x_0\right), \psi'\left(x_0\right) = \psi_0'\left(x_0\right), \tag{7.47}$$

$$\psi\left(x_N\right) = \psi_{N+1}\left(x_N\right), \psi'\left(x_N\right) = \psi_{N+1}'\left(x_N\right), \tag{7.48}$$

where the functions ψ_0 and ψ_{N+1} are given by Eq. (7.40) and Eq. (7.41), respectively. Recalling that $B_{N+1} = 0$, the two conditions given above at $x = x_0$ can be combined to eliminate the constant of integration B_0 and the two conditions given above at $x = x_N$ can be combined to eliminate the constant of integration A_{N+1} to yield the linear second-order two-point boundary value problem on $[x_0,\ x_N]$ parameterized by A_0 given by

$$-\frac{\hbar^2}{2m_0}\frac{d^2\psi\left(x\right)}{dx^2} + V\left(x\right)\psi\left(x\right) = E\psi\left(x\right), x_0 < x < x_N, \tag{7.49}$$

$$ik_0\psi\left(x_0\right) + \psi'\left(x_0\right) = 2ik_0A_0e^{ik_0x_0}, \tag{7.50}$$

$$ik_{N+1}\psi\left(x_N\right) - \psi'\left(x_N\right) = 0. \tag{7.51}$$

Then if ψ is the solution to Eq. (7.49)–(7.51) on the interval $[x_0,\ x_N]$, the desired solution on $(-\infty, \infty)$ can then be obtained by combining ψ on $[x_0,\ x_N]$ with ψ_0 on $(-\infty, x_0]$ and ψ_{N+1} on $[x_N, \infty)$ given by Eq. (7.40) and Eq. (7.41), respectively, with $A_{N+1} = \psi\left(x_N\right)$ and $B_0 = e^{ik_0x_0}\psi\left(x_0\right) - A_0e^{2ik_0x_0}$. It then

follows that $A_{N+1} = \psi(x_N) = \psi(x_N; A_0) = A_0\psi(x_N; 1)$ is linear in A_0, and, moreover, that the value of $T(V_{\text{bias}}, \mathbf{U})$ is independent of the value of A_0. Indeed, in light of Eq. (7.45), it follows that

$$T(V_{\text{bias}}) = T(V_{\text{bias}}, \mathbf{U}) = \frac{|A_{N+1}|^2 k_{N+1}}{|A_0|^2 k_0} = \frac{k_{N+1}|\psi(x_N; A_0)|^2}{k_0 |A_0|^2}$$
$$= \frac{k_{N+1}}{k_0}|\psi(x_N; 1)|^2 = \frac{k_{N+1}}{k_0}|\psi(x_N; V_{\text{bias}}, \mathbf{U})|^2, \tag{7.52}$$

where $\psi(\cdot; V_{\text{bias}}, \mathbf{U})$ denotes the solution to the two-point boundary-value problem Eq. (7.49)–(7.51) corresponding to V_{bias}, $A_0 = 1$, and $\mathbf{U} = \{U_i\}_{i=1}^N$. Then, recalling (7.46), solving the optimal design problem requires the minimization of the least-squares functional

$$J(\mathbf{U}) = \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{k_{N+1,j}}{k_0}|\psi(x_N; V_j, \mathbf{U})|^2 \right|^2, \tag{7.53}$$

where $\psi(\cdot, V_j, \mathbf{U})$ is the solution to the two-point boundary-value problem Eq. (7.49)–(7.51) corresponding to $V_{\text{bias}} = V_j$, $j = 1, 2, \ldots \nu$, $A_0 = 1$, and $\mathbf{U} = \{U_i\}_{i=1}^N$ and we have added the subscript j to k_{N+1} to reflect the fact that $k_{N+1}^2 = \frac{2m_0(E - U_{N+1} + V_{\text{bias}})}{\hbar^2}$ depends on the value of the bias voltage V_{bias}. That is, for $j = 1, 2, \ldots, \nu$, $k_{N+1,j}^2 = \frac{2m_0(E - U_{N+1} + V_j)}{\hbar^2}$.

7.3.2 Approximation

Actually, solving the least-squares minimization problem requires that we be able to numerically solve Eq. (7.49)–(7.51). Toward this end, for each $M = 1, 2, \ldots$ we partition each of the layers, $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, N$ into M equal sub-layers, $[x_{(j-1)M+m-1}^M, x_{(j-1)M+m}^M]$, with $m = 1, 2, \ldots, M$, $x_0^M = x_0$, and $x_{(j-1)M+m}^M = x_{j-1} + mL_j/M$, $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, M$. We then consider the time-independent Schrödinger equation, Eq. (7.39), with the potential function V given by Eq. (7.33) replaced by the piecewise constant approximation V^M given by

$$V^M(x) = V\left(x_{(j-1)M+m-1}^M\right), x_{(j-1)M+m-1}^M \le x < x_{(j-1)M+m}^M, \tag{7.54}$$

$j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, M$. For $\mathbf{U} = \{U_i\}_{i=1}^N$ and V_{bias} given, and $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, M$ we set

$$\left[k_{(j-1)M+m}^M\right]^2 = \frac{2m_0}{\hbar^2}\left(E - U_j + V_{\text{bias}}\frac{x_{(j-1)M+m-1}^M - x_0}{L}\right), \tag{7.55}$$

$k_0^M = k_0$ and $k_{NM+1}^M = k_{N+1}$. Then, for $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots,$

M, on the interval $\left[x^M_{(j-1)M+m-1},\ x^M_{(j-1)M+m}\right]$, the general solution to the time-independent Schrödinger equation with V replaced by V^M is given by

$$\begin{aligned}\psi^M_{(j-1)M+m}(x) &= A^M_{(j-1)M+m} e^{ik^M_{(j-1)M+m}\left(x-x^M_{(j-1)M+m-1}\right)} \\ &\quad + B^M_{(j-1)M+m} e^{-ik^M_{(j-1)M+m}\left(x-x^M_{(j-1)M+m-1}\right)},\end{aligned} \tag{7.56}$$

where $A^M_{(j-1)M+m}$ and $B^M_{(j-1)M+m}$ in the above expressions are arbitrary constants of integration. We also set $\psi^M_0 = \psi_0$ and $\psi^M_{NM+1} = \psi_{N+1}$, where ψ_0 and ψ_{N+1} are given by Eq. (7.40) and Eq. (7.41), respectively. Assuming that E, the sum of electron kinetic and potential energy, the layer potentials, $\mathbf{U} = \{U_i\}_{i=1}^N$, and bias voltages, V_{bias}, of interest are such that the time-independent Schrödinger equation, Eq. (7.39), with V replaced by V^M admits an exponential solution of the form given above on each sub-interval $[x^M_{(j-1)M+m-1},\ x^M_{(j-1)M+m}]$, we seek a smooth (i.e C^1) solution on $[x_0, x_N]$. Consequently, by setting

$$\psi^M_{(j-1)M+m}\left(x^M_{(j-1)M+m}\right) = \psi^M_{(j-1)M+m+1}\left(x^M_{(j-1)M+m}\right), \tag{7.57}$$

$$\frac{d\psi^M_{(j-1)M+m}}{dx}\left(x^M_{(j-1)M+m}\right) = \frac{d\psi^M_{(j-1)M+m+1}}{dx}\left(x^M_{(j-1)M+m}\right), \tag{7.58}$$

$$\psi^M_0\left(x^M_0\right) = \psi^M_1\left(x^M_0\right), \tag{7.59}$$

$$d\psi^M_0/dx\left(x^M_0\right) = d\psi^M_1/dx\left(x^M_0\right), \tag{7.60}$$

$$L^M_{(j-1)M+m} = x^M_{(j-1)M+m} - x^M_{(j-1)M+m-1} = \frac{L_j}{M}, \tag{7.61}$$

for $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, M$, and $L^M_0 = x^M_0$, we obtain the system of equations given in matrix form by

$$\begin{bmatrix} e^{ik^M_n L^M_n} & e^{-ik^M_n L^M_n} \\ ik^M_n e^{ik^M_n L^M_n} & -ik^M_n e^{-ik^M_n L^M_n} \end{bmatrix} \begin{bmatrix} A^M_n \\ B^M_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ ik^M_{n+1} & -ik^M_{n+1} \end{bmatrix} \begin{bmatrix} A^M_{n+1} \\ B^M_{n+1} \end{bmatrix}, \tag{7.62}$$

with $n = 0, 1, 2, \ldots, NM$ or, inverting the two by two matrix on the left-hand side of (7.62), equivalently by

$$\begin{aligned}\begin{bmatrix} A^M_n \\ B^M_n \end{bmatrix} &= \frac{1}{2}\begin{bmatrix} \left(1+\frac{k^M_{n+1}}{k^M_n}\right) e^{-ik^M_n L^M_n} & \left(1-\frac{k^M_{n+1}}{k^M_n}\right) e^{-ik^M_n L^M_n} \\ \left(1-\frac{k^M_{n+1}}{k^M_n}\right) e^{ik^M_n L^M_n} & \left(1+\frac{k^M_{n+1}}{k^M_n}\right) e^{ik^M_n L^M_n} \end{bmatrix} \begin{bmatrix} A^M_{n+1} \\ B^M_{n+1} \end{bmatrix} \\ &\equiv \mathbf{P}^M_n\left(\mathbf{U}, V_{\text{bias}}\right) \begin{bmatrix} A^M_{n+1} \\ B^M_{n+1} \end{bmatrix},\end{aligned} \tag{7.63}$$

$n = 0,\ 1,\ 2, \ldots,\ NM$. With the incident electron of amplitude $|A_0|$ being introduced on the left and no reflective wave propagating to the left from $+\infty$, we also have the two boundary conditions

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} A_0^M \\ B_0^M \end{bmatrix} = A_0 \text{ and } \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} A_{NM+1}^M \\ B_{NM+1}^M \end{bmatrix} = 0. \tag{7.64}$$

The least-squares performance indices for the approximating optimal design problem then become

$$\begin{aligned} J^M(\mathbf{U}) &= \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{\left|A_{NM+1}^M(V_j, \mathbf{U})\right|^2 k_{N+1,j}}{\left|A_0^M\right|^2 k_0} \right|^2 \\ &= \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{\left|A_{NM+1}^M(V_j, \mathbf{U})\right|^2 k_{N+1,j}}{|A_0|^2 k_0} \right|^2, \end{aligned} \tag{7.65}$$

where as before, for $j = 1,\ 2,\ \ldots,\ \nu$, $k_{N+1,j}$ is defined by $k_{N+1,j}^2 = \frac{2m_0(E - U_{N+1} + V_j)}{\hbar^2}$. The scheme we have just outlined is sometimes referred to as the propagation matrix method (see, for example, [3]).

7.3.3 Computing gradients using the static adjoint method

The formulation of the optimal design problem as the minimization of an approximating least-squares performance index Eq. (7.65) subject to the discrete two-point boundary-value problem, Eq. (7.63) and (7.64), use of the adjoint method for efficiently and accurately (in fact, no truncation error) calculating the gradients required by most standard iterative optimization routines. Recall that the adjoint method has its roots in the classical method of Lagrange multipliers and the Maximum Principle from optimal control theory and works its magic by making it unnecessary to directly compute derivatives of the state variables with respect to the optimization parameters.

We begin by combining the $NM + 1$ 2×2 linear systems in $2NM + 4$ unknowns given in Eq. (7.63) and the two boundary conditions given in Eq. (7.64) into a single $2NM + 2$ dimensional linear system of equations in $2NM + 2$ unknowns. Indeed, by making use of the boundary conditions Eq. (7.64) to move the two determined quantities $A_0^M = A_0$ and $B_{NM+1}^M = 0$ to the right-hand side, we obtain the $2NM + 2$-dimensional linear system

$$\mathbf{A}^M(\mathbf{U}, V_{\text{bias}})\, \mathbf{X}^M = \mathbf{b}_0^M, \tag{7.66}$$

where

$$\mathbf{X}^M = \left[B_0^M \; A_1^M \; B_1^M \; A_2^M \; B_2^M \cdots A_{NM}^M \; B_{NM}^M \; A_{NM+1}^M \right]^{\mathrm{T}}, \tag{7.67}$$

$$\mathbf{b}_0^M = \left[-A_0 e^{ik_0 x_0} \; -ik_0 A_0 e^{ik_0 x_0} \cdots 0 \right]^{\mathrm{T}}, \tag{7.68}$$

$$\mathbf{A}^M \left(\mathbf{U}, V_{\mathrm{bias}} \right) = \begin{bmatrix} \mathbf{v}_0 & \mathbf{Q}_0^M & \mathbf{0} & \cdots & & \mathbf{0} \\ \mathbf{0} & I & -\mathbf{P}_1^M & & & \\ & & \mathbf{I} & -\mathbf{P}_2^M & & \vdots \\ \vdots & & & \ddots & \ddots & \\ & & & & \mathbf{I} & -\mathbf{P}_{NM-1}^M \; \mathbf{0} \\ \mathbf{0} & & \cdots & & \mathbf{0} & \mathbf{Q}_N^M \; \mathbf{v}_N \end{bmatrix}, \tag{7.69}$$

with

$$\mathbf{P}_n^M = \mathbf{P}_n^M \left(\mathbf{U}, V_{\mathrm{bias}} \right). \tag{7.70}$$

Note that $n = 1, 2, \ldots, NM - 1$ are as they were defined in Eq. (7.63), $\mathbf{I}$ is the 2×2 identity matrix, $\mathbf{v}_0 = \left[e^{-ik_0 x_0} \; -ik_0 e^{-ik_0 x_0} \right]^{\mathrm{T}}$, $\mathbf{v}_N = \left[-1 \; -ik_{N+1} \right]^{\mathrm{T}}$, $\mathbf{Q}_0^M = \begin{bmatrix} -1 & -1 \\ -ik_1^M & ik_1^M \end{bmatrix}$, and $\mathbf{Q}_N^M = \begin{bmatrix} e^{ik_{NM}^M L_{NM}^M} & e^{-ik_{NM}^M L_{NM}^M} \\ ik_{NM}^M e^{ik_{NM}^M L_{NM}^M} & -ik_{NM}^M e^{-ik_{NM}^M L_{NM}^M} \end{bmatrix}$.

Under rather mild assumptions on the layer potentials $\mathbf{U}$ and the bias voltage V_{bias}, it can be argued (see below) that the matrix given in Eq. (7.69) is nonsingular and consequently that the linear system given in Eq. (7.66) admits a unique solution. If for each $j = 1, 2, \ldots, \nu$, we define the $2NM + 2$ dimensional row vector $\mathbf{c}_j$ by

$$\mathbf{c}_j^M = \left[0 \, 0 \cdots \sqrt{k_{N+1}} \Big/ \left(\sqrt{k_0} A_0 \right) \right], \tag{7.71}$$

the performance index Eq. (7.65) can now be written as

$$J^M \left(\mathbf{U} \right) = \sum_{j=1}^{\nu} \left| T_0 \left(V_j \right) - \left| \mathbf{c}_j^M \mathbf{X}^M \left(\mathbf{U}, V_j \right) \right|^2 \right|^2, \tag{7.72}$$

where $\mathbf{X}^M \left(\mathbf{U}, V_{\mathrm{bias}} \right)$ is the unique solution to the linear system Eq. (7.66) corresponding to the layer potentials U and bias voltage V_{bias}.

We turn our attention next to computing the gradient with respect to $\mathbf{U}$,$\nabla J^M(\mathbf{U})$, of the performance index J^M given in Eq. (7.72) via the application of a static form of the adjoint method. For each $j = 1, 2, \ldots, \nu$, we define the adjoint system by

$$\mathbf{A}^M \left(\mathbf{U}, V_{\mathrm{j}} \right)^* \mathbf{Z}_j^M = 4 \left(\mathbf{c}_j^M \right)^* \mathbf{c}_j^M \mathbf{X}_j^M \left(T_0 \left(V_j \right) - \left| \mathbf{c}_j^M \mathbf{X}_j^M \right|^2 \right), \tag{7.73}$$

where the entries in the $2NM+2$ dimensional vector $\mathbf{Z}_j^M$ are known as the adjoint or co-state variables, $\mathbf{X}_j^M = \mathbf{X}^M(\mathbf{U}, V_j)$ is the unique solution to the linear system Eq. (7.66) corresponding to the layer potentials $\mathbf{U}$ and the bias voltage V_j, and Λ^* denotes the conjugate transpose of a matrix Λ with complex entries. Note that nonsingularity of the matrix $\mathbf{A}^M(\mathbf{U}, V_\mathrm{j})$ given in Eq. (7.69) ensures that the adjoint system admits a unique solution $\mathbf{Z}_j^M$. Then

$$\begin{aligned}\nabla J^M(\mathbf{U}) &= -2\sum_{j=1}^{\nu}\left(T_0(V_j) - \left|\mathbf{c}_j^M\mathbf{X}_j^M\right|^2\right)\left\{2\mathrm{Re}\left(\mathbf{X}_\mathrm{j}^\mathrm{M}\right)^*\left(\mathbf{c}_\mathrm{j}^\mathrm{M}\right)^*\mathbf{c}_\mathrm{j}^\mathrm{M}\partial\mathbf{X}_\mathrm{j}^\mathrm{M}/\partial\mathbf{U}\right\}\\ &= -\sum_{j=1}^{\nu}\mathrm{Re}\left(\mathbf{Z}_\mathrm{j}^\mathrm{M}\right)^*\mathbf{A}^\mathrm{M}(\mathbf{U},\mathrm{V_j})\,\partial\mathbf{X}_\mathrm{j}^\mathrm{M}/\partial\mathbf{U}\\ &= \sum_{j=1}^{\nu}\mathrm{Re}\left(\mathbf{Z}_\mathrm{j}^\mathrm{M}\right)^*\left(\partial\mathbf{A}^\mathrm{M}(\mathbf{U},\mathrm{V_j})/\partial\mathbf{U}\right)\mathbf{X}_\mathrm{j}^\mathrm{M},\end{aligned} \tag{7.74}$$

where in the final expression in Eq. (7.74) we have used the identity

$$\mathbf{A}^M(\mathbf{U}, V_\mathrm{bias})\left(\partial\mathbf{X}_j^M/\partial\mathbf{U}\right) = -\left(\partial\mathbf{A}^M(\mathbf{U}, V_\mathrm{bias})/\partial\mathbf{U}\right)\mathbf{X}_j^M, \tag{7.75}$$

which follows immediately by differentiating Eq. (7.66). We note that it is in fact possible to argue that the matrix $\mathbf{A}^M(\mathbf{U}, V_\mathrm{j})$ and the vector $\mathbf{X}_j^M = \mathbf{X}^M(\mathbf{U}, V_j)$ are both differentiable with respect to $\mathbf{U}$ (see [4]).

It now follows that in each step of an iterative optimization scheme, both the value of the performance index J^M and its gradient $\nabla J^M(\mathbf{U})$ can be computed efficiently by sequentially solving the two linear systems (with only a single LU decomposition required because the two system matrices are the same up to conjugate transpose) Eq. (7.66) and Eq. (7.73)

$$\mathbf{A}^M(\mathbf{U}, V_j)\,\mathbf{X}_j^M = \mathbf{b}_0^M, \tag{7.76}$$

and

$$\mathbf{A}^M(\mathbf{U}, V_\mathrm{j})^*\,\mathbf{Z}_j^M = 4\left(\mathbf{c}_j^M\right)^*\mathbf{c}_j^M\mathbf{X}_j^M\left(T_0(V_j) - \left|\mathbf{c}_j^M\mathbf{X}_j^M\right|^2\right), \tag{7.77}$$

and then computing the sum Eq. (7.72) and inner product in the final expression in Eq. (7.74) given by

$$J^M(\mathbf{U}) = \sum_{j=1}^{\nu}\left|T_0(V_j) - \left|\mathbf{c}_j^M\mathbf{X}_j^M\right|^2\right|^2, \tag{7.78}$$

and

$$\nabla J^M(\mathbf{U}) = \sum_{j=1}^{\nu} \mathrm{Re}\left(\mathbf{Z}_{\mathrm{j}}^{\mathrm{M}}\right)^* \left(\partial \mathbf{A}^{\mathrm{M}}(\mathbf{U}, \mathrm{V}_{\mathrm{j}}) / \partial \mathbf{U}\right) \mathbf{X}_{\mathrm{j}}^{\mathrm{M}}. \tag{7.79}$$

7.3.4 An alternative approach involving the dynamic adjoint

By using the linear relationship that exists between A_{N+1} and A_0 together with the fact that the transmission function involves the quotient of A_{N+1} and A_0, we are able to replace the boundary value problem by the much more easily solved terminal value problem. This in turn permits us to now compute the gradient of J via a dynamic adjoint. As a result we are able to take advantage of the specialized structure of the linear systems given in Eq. (7.63) and develop a highly efficient scheme for calculating the value of J and its gradient.

Noting that A_{NM+1}^M depends linearly on A_0 and consequently on A_0^M as well, it follows that the value of the function $T^M(V_{\mathrm{bias}})$ given by $T^M(V_{\mathrm{bias}}) = T^M(V_{\mathrm{bias}}, \mathbf{U}) = \left|A_{NM+1}^M\right|^2 k_{N+1} \Big/ |A_0^M|^2 k_0$ is independent of the value of A_0. Moreover, it necessarily follows that A_0^M depends linearly on A_{NM+1}^M and that the value of $T^M(V_{\mathrm{bias}}) = T^M(V_{\mathrm{bias}}, \mathbf{U}) = \left|A_{NM+1}^M\right|^2 k_{N+1} \Big/ |A_0^M|^2 k_0$ is independent of the value of A_{NM+1}^M if it is specified instead of A_0^M. Consequently, without affecting the solution to the optimal design problem and without loss of generality, we may set $A_{NM+1}^M = 1$. The boundary conditions given in Eq. (7.64) can now be replaced with the single terminal condition

$$\begin{bmatrix} A_{NM+1}^M \\ B_{NM+1}^M \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{7.80}$$

and the least-squares performance indices given in Eq. (7.65) can be replaced by

$$J^M(\mathbf{U}) = \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{k_{N+1,j}}{\left|A_0^M(V_j, \mathbf{U})\right|^2 k_0} \right|^2. \tag{7.81}$$

Solving the approximating optimal design problem then consists of finding local layer potentials $\mathbf{U}^* = \{U_i^*\}_{i=1}^N$ which minimize the least-squares performance index Eq. (7.81) subject to the system of two linear difference Eq. (7.63) and the terminal condition Eq. (7.80).

To see in this case how the adjoint is derived, we rewrite the underlying system of difference Eq. (7.63) and the terminal condition Eq. (7.80) as

$$\boldsymbol{\alpha}_{i,j} = \mathbf{P}_{i,j} \boldsymbol{\alpha}_{i+1,j} \text{ and } \boldsymbol{\alpha}_{\mathrm{NM+1,j}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{7.82}$$

where for $j = 1,\ 2, \ldots, \nu$ and $i = 0,\ 1,\ 2, \ldots, NM$, $\mathbf{P}_{i,j} = \mathbf{P}_i^M\left(\mathbf{U}, V_j\right)$ and $\boldsymbol{\alpha}_{i,j} = \left[A_i^M,\ B_i^M\right]^T = \left[A_i^M\left(\mathbf{U}, V_j\right),\ B_i^M\left(\mathbf{U}, V_j\right)\right]^T$. The least-squares performance index Eq. (7.81) is then given by

$$J^M(\mathbf{U}) = \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{k_{N+1,j}}{\left|\mathbf{c}\boldsymbol{\alpha}_{0,j}\right|^2 k_0} \right|^2 = \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{k_{N+1,j}}{\bar{\boldsymbol{\alpha}}_{0,j}^T \mathbf{Q}\boldsymbol{\alpha}_{0,j} k_0} \right|^2, \quad (7.83)$$

where $\mathbf{c} = [1, 0]$ and $\mathbf{Q} = \mathbf{c}^T\mathbf{c} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. We define the adjoint or co-state system corresponding to Eq. (7.82) and (7.83) as the initial value problem given by

$\boldsymbol{\beta}_{i+1,j} = \mathbf{P}_{i-1,j}^T \boldsymbol{\beta}_{i,j} + \delta_{i0} 4 \left[T_0(V_j) - \frac{k_{N+1,j}}{\bar{\boldsymbol{\alpha}}_{i,j}^T \mathbf{Q}\boldsymbol{\alpha}_{i,j} k_0} \right] \frac{\mathbf{Q}\bar{\boldsymbol{\alpha}}_{i,j} k_{N+1,j}}{\left(\bar{\boldsymbol{\alpha}}_{i,j}^T \mathbf{Q}\boldsymbol{\alpha}_{i,j}\right)^2 k_0}$ and $\boldsymbol{\beta}_{0,j} = 0$,

where δ_{ij}denotes the Kroeneker delta function. It then follows that

$$\nabla J^M(\mathbf{U}) = \frac{\partial J^M}{\partial \mathbf{U}} = \sum_{j=1}^{\nu} 2 \left[T_0(V_j) - \frac{k_{N+1,j}}{\bar{\boldsymbol{\alpha}}_{0,j}^T \mathbf{Q}\boldsymbol{\alpha}_{0,j} k_0} \right] \frac{2\mathrm{Re}\bar{\boldsymbol{\alpha}}_{0,\mathrm{j}}^{\mathrm{T}} \mathbf{Q} \frac{\partial \boldsymbol{\alpha}_{0,\mathrm{j}}}{\partial \mathbf{U}} \mathrm{k}_{\mathrm{N}+1,\mathrm{j}}}{\left(\bar{\boldsymbol{\alpha}}_{0,j}^T \mathbf{Q}\boldsymbol{\alpha}_{0,j}\right)^2 k_0}$$

$$= \mathrm{Re} \sum_{\mathrm{j}=1}^{\nu} \sum_{\mathrm{i}=0}^{\mathrm{NM}} 4\delta_{\mathrm{i}0} \left[\mathrm{T}_0(\mathrm{V}_\mathrm{j}) - \frac{\mathrm{k}_{\mathrm{N}+1,\mathrm{j}}}{\bar{\boldsymbol{\alpha}}_{\mathrm{i,j}}^{\mathrm{T}} \mathbf{Q}\boldsymbol{\alpha}_{\mathrm{i,j}} \mathrm{k}_0} \right] \frac{\bar{\boldsymbol{\alpha}}_{\mathrm{i,j}}^{\mathrm{T}} \mathbf{Q} \mathrm{k}_{\mathrm{N}+1,\mathrm{j}}}{\left(\bar{\boldsymbol{\alpha}}_{\mathrm{i,j}}^{\mathrm{T}} \mathbf{Q}\boldsymbol{\alpha}_{\mathrm{i,j}}\right)^2 \mathrm{k}_0} \frac{\partial \boldsymbol{\alpha}_{\mathrm{i,j}}}{\partial \mathbf{U}}$$

$$= \mathrm{Re} \sum_{\mathrm{j}=1}^{\nu} \sum_{\mathrm{i}=0}^{\mathrm{NM}} \left(\boldsymbol{\beta}_{\mathrm{i}+1,\mathrm{j}} - \mathbf{P}_{\mathrm{i}-1,\mathrm{j}}^{\mathrm{T}} \boldsymbol{\beta}_{\mathrm{i,j}}\right)^{\mathrm{T}} \frac{\partial \boldsymbol{\alpha}_{\mathrm{i,j}}}{\partial \mathbf{U}}$$

$$= \mathrm{Re} \sum_{\mathrm{j}=1}^{\nu} \left\{ \sum_{\mathrm{i}=0}^{\mathrm{NM}} \boldsymbol{\beta}_{\mathrm{i}+1,\mathrm{j}}^{\mathrm{T}} \frac{\partial \boldsymbol{\alpha}_{\mathrm{i,j}}}{\partial \mathbf{U}} - \sum_{\mathrm{i}=1}^{\mathrm{NM}} \boldsymbol{\beta}_{\mathrm{i,j}}^{\mathrm{T}} \mathbf{P}_{\mathrm{i}-1,\mathrm{j}} \frac{\partial \boldsymbol{\alpha}_{\mathrm{i,j}}}{\partial \mathbf{U}} \right\}$$

$$= \mathrm{Re} \sum_{\mathrm{j}=1}^{\nu} \left\{ \sum_{\mathrm{i}=0}^{\mathrm{NM}} \boldsymbol{\beta}_{\mathrm{i}+1,\mathrm{j}}^{\mathrm{T}} \left(\mathbf{P}_{\mathrm{i,j}} \frac{\partial \boldsymbol{\alpha}_{\mathrm{i}+1,\mathrm{j}}}{\partial \mathbf{U}} + \frac{\partial \mathbf{P}_{\mathrm{i,j}}}{\partial \mathbf{U}} \boldsymbol{\alpha}_{\mathrm{i}+1,\mathrm{j}} \right) - \sum_{\mathrm{i}=1}^{\mathrm{NM}} \boldsymbol{\beta}_{\mathrm{i,j}}^{\mathrm{T}} \mathbf{P}_{\mathrm{i}-1,\mathrm{j}} \frac{\partial \boldsymbol{\alpha}_{\mathrm{i,j}}}{\partial \mathbf{U}} \right\}$$

$$= \mathrm{Re} \sum_{\mathrm{j}=1}^{\nu} \sum_{\mathrm{i}=0}^{\mathrm{NM}} \boldsymbol{\beta}_{\mathrm{i}+1,\mathrm{j}}^{\mathrm{T}} \frac{\partial \mathbf{P}_{\mathrm{i,j}}}{\partial \mathbf{U}} \boldsymbol{\alpha}_{\mathrm{i}+1,\mathrm{j}},$$

where, in light of the terminal condition given in Eq. (7.80), we have used the fact that $\partial \boldsymbol{\alpha}_{NM+1,j} / \partial \mathbf{U} = 0$. Consequently, the gradient of J^Mcan be obtained with no truncation error according to the following steps.

1. For each $j = 1,\ 2, \ldots, \nu$, solve the terminal value problem

$$\begin{bmatrix} A_i^M \\ B_i^M \end{bmatrix} = \mathbf{P}_i^M \left(\mathbf{U}, V_j\right) \begin{bmatrix} A_{i+1}^M \\ B_{i+1}^M \end{bmatrix}, \begin{bmatrix} A_{NM+1}^M \\ B_{NM+1}^M \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{7.84}$$

with $i = NM,\ NM-1,\ \ldots,\ 2,\ 1,\ 0$.

2. For each $j = 1,\ 2, \ldots, \nu$, solve the initial value problem

$$\boldsymbol{\beta}_{i+1,j} = \mathbf{P}_i^M \left(\mathbf{U}, V_j\right)^T \boldsymbol{\beta}_{i,j} \boldsymbol{\beta}_{1,j} = 4 \left[T_0\left(V_j\right) - \frac{k_{N+1,j}}{\left|A_0^M\right|^2 k_0} \right] \frac{k_{N+1,j}}{\left|A_0^M\right|^4 k_0} \begin{bmatrix} \bar{A}_0^M \\ 0 \end{bmatrix}, \tag{7.85}$$

with $i = 1,\ 2,\ \ldots,\ NM$.

3. Compute the gradient of J^M as

$$\nabla J^M\left(\mathbf{U}\right) = \mathrm{Re} \sum_{j=1}^{\nu} \sum_{i=0}^{NM} \boldsymbol{\beta}_{i+1,j}^T \frac{\partial \mathbf{P}_i^M \left(\mathbf{U}, V_j\right)}{\partial \mathbf{U}} \begin{bmatrix} A_{i+1,j}^M \\ B_{i+1,j}^M \end{bmatrix}. \tag{7.86}$$

Central to the approach we have just outlined is the replacement and approximation of the two-point boundary problem given by Eq. (7.63) and (7.64) with the terminal value problem given by Eq. (7.84). The solution of the optimal design problem then requires the iterative sequential solution of Eq. (7.84) for the system *state* and Eq. (7.85) for the system *co-state*. However, in doing this we may have sacrificed the inherent stability of solving a boundary-value problem for the potential instability of integrating a sequence of successively more highly discretized terminal value problems. It turns out that it is in fact possible to demonstrate the numerical stability of the scheme we have just proposed. Indeed, the solutions to the recursions given in Eq. (7.84) and (7.85) take the form

$$\begin{bmatrix} A_k^M \\ B_k^M \end{bmatrix} = \left[\prod_{i=k}^{NM} \mathbf{P}_i^M \left(\mathbf{U}, V_j\right) \right] \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{7.87}$$

where $k = NM,\ NM-1,\ \ldots,\ 2,\ 1,\ 0$ and

$$\boldsymbol{\beta}_{k,j} = \left[\prod_{i=k-2}^{0} \mathbf{P}_i^M \left(\mathbf{U}, V_j\right)^T \right] \boldsymbol{\beta}_{1,j}, \tag{7.88}$$

$$\boldsymbol{\beta}_{1,j} = 4 \left[T_0\left(V_j\right) - \frac{k_{N+1,j}}{\left|A_0^M\right|^2 k_0} \right] \frac{k_{N+1,j}}{\left|A_0^M\right|^4 k_0} \begin{bmatrix} \bar{A}_0^M \\ 0 \end{bmatrix}, \tag{7.89}$$

respectively where $i = 1,\ 2,\ \ldots,\ NM+1$. These recursions will be numerically stable with respect to the discretization if we can demonstrate the

boundedness of the matrix product $\prod_{i=0}^{NM} \mathbf{P}_i^M(\mathbf{U}, V_j)$ uniformly in M in some appropriate matrix norm, where the matrices $\mathbf{P}_i^M(\mathbf{U}, V_j)$ are given in Eq. (7.63). Our stability argument requires an assumption on the device parameters guaranteeing that the time-independent Schrödinger Eq. (7.39) with V replaced by V^M admits exponential solutions on each of the approximating sub-intervals and moreover that these solutions remain bounded away from becoming polynomial on any sub-interval as the discretization level tends to infinity.

Assumption 1. The total energy E, the layer potentials $\mathbf{U} = \{U_i\}_{i=1}^N$, the overall length of the device L, and the bias voltage V_{bias} are such that there exists a constant $\delta > 0$ for which $0 < \delta \leq \left|k_j^M\right|$, $j = 0, 1, 2, \ldots, NM+1$, where the k_j^M, $j = 1, 2, \ldots, NM$, are given by (7.55), and $k_0^M = k_0$ and $k_{NM+1}^M = k_{N+1}$.

Lemma 1. For $j \neq 0, M, 2M, 3M, \ldots, NM$, there exists a positive constant ρ which depends on E, $\mathbf{U} = \{U_i\}_{i=1}^N$, and V_{bias}, for $V_{\min} \leq V_{\text{bias}} \leq V_{\max}$, but which is independent of j and M, such that $\left\|\mathbf{P}_j^M(\mathbf{U}, V_{\text{bias}})\right\| \leq e^{\rho \frac{L}{M}} \left(1 + \frac{2m_0 V_{\text{bias}}}{\hbar^2 \delta^2 M}\right)^{\frac{1}{2}}$, where the constant $\delta > 0$ is as defined in Assumption 1 and where for a complex matrix $\mathbf{A}$, the matrix norm $\|\mathbf{A}\|$ is the spectral norm given in terms of the largest (necessarily real and non-negative) eigenvalue of the symmetric non-negative definite matrix $\mathbf{A}^*\mathbf{A}$ by the expression $\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})}$.

Proof. We begin by writing

$$\begin{aligned}
\mathbf{P}_j^M(\mathbf{U}, V_{\text{bias}}) &= \frac{1}{2}\begin{bmatrix} \left(1 + \frac{k_{j+1}^M}{k_j^M}\right) e^{-ik_j^M L_j^M} & \left(1 - \frac{k_{j+1}^M}{k_j^M}\right) e^{-ik_j^M L_j^M} \\ \left(1 - \frac{k_{j+1}^M}{k_j^M}\right) e^{ik_j^M L_j^M} & \left(1 + \frac{k_{j+1}^M}{k_j^M}\right) e^{ik_j^M L_j^M} \end{bmatrix} \\
&= \begin{bmatrix} e^{-ik_j^M L_j^M} & 0 \\ 0 & e^{ik_j^M L_j^M} \end{bmatrix} \begin{bmatrix} \frac{1}{2}\left(1 + \frac{k_{j+1}^M}{k_j^M}\right) & \frac{1}{2}\left(1 - \frac{k_{j+1}^M}{k_j^M}\right) \\ \frac{1}{2}\left(1 - \frac{k_{j+1}^M}{k_j^M}\right) & \frac{1}{2}\left(1 + \frac{k_{j+1}^M}{k_j^M}\right) \end{bmatrix} \\
&= \begin{bmatrix} e^{-ik_j^M L_j^M} & 0 \\ 0 & e^{ik_j^M L_j^M} \end{bmatrix} \begin{bmatrix} \frac{1+\gamma_j^M}{2} & \frac{1-\gamma_j^M}{2} \\ \frac{1-\gamma_j^M}{2} & \frac{1+\gamma_j^M}{2} \end{bmatrix} \equiv \mathbf{E}_j^M \Gamma_j^M, \qquad (7.90)
\end{aligned}$$

where $\gamma_j^M = k_{j+1}^M / k_j^M$. It follows that $\left\|\mathbf{P}_j^M(\mathbf{U}, V_{\text{bias}})\right\| \leq \left\|\mathbf{E}_j^M\right\| \left\|\Gamma_j^M\right\|$. A straightforward computation immediately reveals that

$$\left\|\mathbf{E}_j^M\right\| \leq e^{L_j^M \left|Im\, k_j^M\right|} \leq e^{\rho \frac{L}{M}}, \qquad (7.91)$$

where $\rho = \sqrt{\frac{2m_0}{\hbar^2}\left|E + V_{\text{bias}} - \|\mathbf{U}\|_\infty\right|}$. To estimate $\left\|\Gamma_j^M\right\|$, the spectral norm of the matrix Γ_j^M, a computation yields the characteristic equation

$$\det\left(\left(\Gamma_j^M\right)^*\Gamma_j^M-\lambda I\right)=\left(\frac{1+\left|\gamma_j^M\right|^2}{2}-\lambda\right)^2-\left(\frac{1-\left|\gamma_j^M\right|^2}{2}\right)^2=0, \quad (7.92)$$

from which we find the two eigenvalues of $\left(\Gamma_j^M\right)^*\Gamma_j^M$ to be $\lambda_+^M=1$ and $\lambda_-^M=\left|\gamma_j^M\right|^2$. It follows that $\left\|\Gamma_j^M\right\|=\max\left\{1,\left|k_{j+1}^M/k_j^M\right|\right\}$. Now, from the definition of k_j^M we have that $\left[k_{j+1}^M\right]^2=\left[k_j^M\right]^2+2m_0V_{\text{bias}}L_j/\hbar^2LM$, for $j\neq 0,\ M,\ 2M,\ 3M,\ldots,NM$, from which it follows that $\left|k_{j+1}^M/k_j^M\right|\leq\sqrt{1+2m_0V_{\text{bias}}/\hbar^2\delta^2M}$, and therefore that

$$\left\|\Gamma_j^M\right\|\leq\sqrt{1+2m_0V_{\text{bias}}/\hbar^2\delta^2M}. \quad (7.93)$$

The result then immediately follows from Eq. (7.91) and Eq. (7.93).

Lemma 2. For $j=0,\ M,\ 2M,\ 3M,\ldots,\ NM$, there exist positive constants ρ and σ which depend on E, $\mathbf{U}=\{U_i\}_{i=1}^N$, and V_{bias}, for $V_{\min}\leq V_{\text{bias}}\leq V_{\max}$, but which are independent of j and M, such that the bounds

$$\left\|\mathbf{P}_j^M\left(\mathbf{U},V_{\text{bias}}\right)\right\|\leq\exp\left(\rho L/M\right)\sqrt{1+\left(2m_0/\hbar^2\sigma^2\right)\left\{2\left\|\mathbf{U}\right\|_\infty+V_{\text{bias}}\right\}},$$

$j\neq 0$ and

$$\left\|\mathbf{P}_0^M\left(\mathbf{U},V_{\text{bias}}\right)\right\|\leq\exp\left(\rho x_0\right)\sqrt{1+\left(2m_0/\hbar^2\sigma^2\right)\left\{2\left\|U\right\|_\infty+V_{\text{bias}}\right\}}$$

obtain, where for a complex matrix $\mathbf{A}$, the matrix norm $\|\mathbf{A}\|$ is the spectral norm given by $\|\mathbf{A}\|=\sqrt{\lambda_{\max}\left(\mathbf{A}^*\mathbf{A}\right)}$.

Proof. For $j=nM,\ n=1,\ 2,\ldots,\ N$, once again from the definition of k_j^M we now have

$$\begin{aligned}\left[k_{j+1}^M\right]^2&=\left[k_j^M\right]^2+\frac{2m_0}{\hbar^2}\left\{U_n-U_{n+1}+V_{bias}\left(\frac{x_{nM}^M-x_{nM-1}^M}{L}\right)\right\}\\&=\left[k_j^M\right]^2+\frac{2m_0}{\hbar^2}\left\{U_n-U_{n+1}+\frac{V_{\text{bias}}L_n}{L}\right\}.\end{aligned} \quad (7.94)$$

In this case it follows from Assumption 1 that $\left|k_{j+1}^M/k_j^M\right|\leq\sqrt{1+2m_0/\hbar^2\delta^2\left\{2\left\|\mathbf{U}\right\|_\infty+V_{\text{bias}}\right\}}$, and therefore that $\|\Gamma_{nM}^M\|\leq\sqrt{1+2m_0/\hbar^2\delta^2\left\{2\left\|\mathbf{U}\right\|_\infty+V_{\text{bias}}\right\}}$, $n=1,2,\ldots,N$. Finally, for $j=0$ we obtain $|k_1^M/k_0|\leq\sqrt{1+4m_0\left\|\mathbf{U}\right\|_\infty/\hbar^2k_0^2}$ and $\|\Gamma_0^M\|\leq\sqrt{1+4m_0\left\|\mathbf{U}\right\|_\infty/\hbar^2k_0^2}$. It continues to remain true that $\left\|\mathbf{E}_j^M\right\|\leq e^{L_j^M\left|Imk_j^M\right|}\leq e^{\rho\frac{L}{M}}$, $j=M,\ 2M,\ 3M,\ldots,\ NM$ and $\left\|\mathbf{E}_0^M\right\|\leq e^{x_0\left|\text{Imk}_0\right|}\leq e^{\rho x_0}$, and consequently the desired result follows at once.

Theorem 1. The linear discrete dynamical systems given in Eq. (7.84) and (7.85) are stable with respect to the approximation index M.
Proof. Lemmas 1 and 2 yield

$$\begin{aligned}
\left\| \prod_{i=0}^{NM} \mathbf{P}_i^M (\mathbf{U}, V_j) \right\| &\leq \prod_{i=0}^{NM} \left\| \mathbf{P}_i^M (\mathbf{U}, V_j) \right\| \\
&= \prod_{i=0}^{N} \left\| \mathbf{P}_{iM}^M (\mathbf{U}, V_j) \right\| \cdot \prod_{\substack{i=1, i \neq nM, \\ n=1,\ldots,N-1}}^{NM-1} \left\| \mathbf{P}_i^M (\mathbf{U}, V_j) \right\| \\
&\leq e^{\rho x_0} \left(1 + \frac{2m_0}{\hbar^2 \sigma^2} \left\{ 2 \|\mathbf{U}\|_\infty + V_{\text{bias}} \right\} \right)^{\frac{1}{2}} \\
&\quad \cdot \left\{ \prod_{i=1}^{N} e^{\frac{\rho L}{M}} \left(1 + \frac{2m_0}{\hbar^2 \sigma^2} \left\{ 2 \|\mathbf{U}\|_\infty + V_{\text{bias}} \right\} \right)^{\frac{1}{2}} \right\} \\
&\quad \cdot \left\{ \prod_{\substack{i=1, i \neq nM, \\ n=1,2,\ldots,N-1}}^{NM-1} e^{\frac{\rho L}{M}} \left(1 + \frac{2m_0 V_{\text{bias}}}{\hbar^2 \delta^2 M} \right)^{\frac{1}{2}} \right\} \\
&\leq \left(1 + \frac{2m_0}{\hbar^2 \sigma^2} \left\{ 2 \|\mathbf{U}\|_\infty + V_{\text{bias}} \right\} \right)^{\frac{N+1}{2}} e^{\rho(x_0 + NL) + \frac{N m_0 V_{\text{bias}}}{\hbar^2 \delta^2}},
\end{aligned} \tag{7.95}$$

and the result follows.

7.3.5 Convergence

We note that in actuality, we are not locating a local minimum of the original performance index Eq. (7.46), but rather we are finding a local minimum of the approximating performance index Eq. (7.81). It would be useful to know that the approximating optimal design located by the method we have described above does in fact approximate a true optimal (in the sense of the performance index Eq. (7.46)) design. In addition there are also a number of other theoretical questions whose answers are of great significance when evaluating the efficacy of our approach. Indeed, there is the question of the existence of a solution to each of the approximating optimal design problems, and since we are using a gradient-based search to locate local minima there is the question of the differentiability of the performance index with

respect to the optimization parameters, the layer potentials. One might also ask if the gradients of the approximating performance indices converge to the gradient of the original performance index as the discretization level grows without bound. It turns out that all of these questions can be answered with complete rigor by using functional analytic techniques involving the reformulation of the time-independent Schrödinger Eq. (7.39) in terms of an abstract bounded and coercive sesquilinear form. In particular, the coercivity property, a form of positive definiteness, is at the heart of all of these arguments. In our treatment here, we illustrate how this works by considering just two of the questions posed above: the existence of solutions to each of the approximating optimal design problems and the convergence of solutions of the approximating optimal design problems to a solution of the original optimal design problems. The arguments establishing differentiability of the performance indices and the convergence of the gradients can be found in [4].

We begin by reformulating the boundary value problem Eq. (7.49)–(7.51) as an abstract elliptic system in weak or variational form. Let H denote the Hilbert space $L_2(x_0,\ x_N)$ and let $W = H^1(x_0,\ x_N)$, each endowed with its standard inner product. It follows that W is densely and continuously embedded in H (see, for example, [5]) and then pivoting (see [6–9]) on H we obtain the well-known dense and continuous embeddings $W \subset H \subset W^*$ where W^* denotes the space of continuous conjugate linear functionals on W. Let Ω be a compact (i.e. closed and bounded) subset of $\mathbf{R}^N$, let $|\cdot|_H$, $\|\cdot\|_W$, and $\|\cdot\|_{W*}$ denote the usual norms on H, W, and W*, respectively, and let κ denote the embedding constant between H and W; that is $|\varphi|_H \leq \kappa \|\varphi\|_W$, for $\varphi \in W$. We let $\|\cdot\|_\infty$ denote the standard max-norm on $\mathbf{R}^N$.

For each $\mathbf{U} = \{U_i\}_{i=1}^N \in \Omega$, $A_0 \in \mathbf{C}$, and V_{bias}, with $V_{\min} \leq V_{\text{bias}} \leq V_{\max}$, we define the sesquilinear form $a(\mathbf{U}, V_{\text{bias}}; \cdot, \cdot) : W \times W \to \mathbf{C}$ and the bounded conjugate linear functional $f \in W^*$ by

$$\begin{aligned} a(\mathbf{U}, V_{\text{bias}}; \varphi, \chi) = &\int_{x_0}^{x_N} D\varphi(x)\, D\bar{\chi}(x)\, dx \\ &-ik_{N+1}\varphi(x_N)\,\bar{\chi}(x_N) - ik_0\varphi(x_0)\,\bar{\chi}(x_0) \\ &+ \frac{2m_0}{\hbar^2}\int_{x_0}^{x_N} \{V(x) - E\}\,\varphi(x)\,\bar{\chi}(x)\, dx, \end{aligned} \tag{7.96}$$

for $\varphi, \chi \in W$ and $\langle f, \chi\rangle = -2ik_0A_0e^{ik_0x_0}\bar{\chi}(x_0)$, for $\chi \in W$, where D denotes the differentiation operator with respect to the variable x. The boundary value problem Eq. (7.49)–(7.51) is then given in abstract form as finding a $\psi \in W$ that satisfies

$$a(\mathbf{U}, V_{\text{bias}}; \psi, \varphi) = \langle f, \varphi\rangle, \text{ for every } \varphi \in \mathrm{W}. \tag{7.97}$$

Straightforward calculations (see, for example, [6]) can be used to establish the following lemma.

Lemma 3. For $\mathbf{U} = \{U_i\}_{i=1}^N \in \Omega$ and V_{bias} with $V_{\min} \leq V_{\text{bias}} \leq V_{\max}$, there exist constants $\lambda \in \mathbf{R}$ and $\alpha, \beta > 0$ which are independent of $\mathbf{U}$ and V_{bias} such that

$$|a(\mathbf{U}, V_{\text{bias}}; \varphi, \chi)| \leq \alpha \|\varphi\|_W \|\chi\|_W, \tag{7.98}$$

for $\varphi, \chi \in W$, and

$$\text{Re} \quad a(\mathbf{U}, \mathrm{V}_{\text{bias}}; \varphi, \varphi) + \lambda |\varphi|_{\mathrm{H}}^2 \geq \beta \|\varphi\|_{\mathrm{W}}^2, \tag{7.99}$$

for every $\varphi \in W$.

Also, there exists a constant $\gamma > 0$ which is independent of V_{bias} such that for every $\mathbf{U}_1 = \{U_i^1\}_{i=1}^N, \mathbf{U}_2 = \{U_i^2\}_{i=1}^N \in \Omega$ we have

$$|a(\mathbf{U}_1, V_{\text{bias}}; \varphi, \chi) - a(\mathbf{U}_2, V_{\text{bias}}; \varphi, \chi)| \leq \gamma \|\mathbf{U}_1 - \mathbf{U}_2\|_\infty |\varphi|_H |\chi|_H, \tag{7.100}$$

for $\varphi, \chi \in W$.

In a similar manner, the sequence of approximating discrete two-point boundary-value problems given by Eq. (7.63) and (7.64) can be reformulated as a sequence of abstract elliptic systems of the form given in Eq. (7.97). Indeed, this is achieved by simply replacing the potential function V in Eq. (7.96) by the piecewise constant approximation V^M given by Eq. (7.54). Toward this end, for each $M = 1, 2, \ldots$, we define the sequence of abstract sesquilinear forms $\{a_M(\mathbf{U}, V_{\text{bias}}; \cdot, \cdot)\}_{M=1}^\infty$ on W, $a_M(\mathbf{U}, V_{\text{bias}}; \cdot, \cdot) : W \times W \to \mathbf{C}$, by

$$\begin{aligned} a_M(\mathbf{U}, V_{\text{bias}}; \varphi, \chi) = & \int_{x_0}^{x_N} D\varphi(x) D\bar{\chi}(x) dx \\ & -ik_{N+1}\varphi(x_N)\bar{\chi}(x_N) - ik_0\varphi(x_0)\bar{\chi}(x_0) \\ & +\frac{2m_0}{\hbar^2}\int_{x_0}^{x_N} \{V^M(x) - E\}\varphi(x)\bar{\chi}(x) dx, \end{aligned} \tag{7.101}$$

where $\varphi, \chi \in W$. The difference between the form defined in Eq. (7.96) and the forms defined in Eq. (7.101) is that the potential function V given by Eq. (7.33) in the form $a(\mathbf{U}, V_{\text{bias}}; \cdot, \cdot) : W \times W \to \mathbf{C}$ has been replaced by its piecewise constant approximation V^M given in Eq. (7.54) by

$$V^M(x) = V\left(x_{(j-1)M+m-1}^M\right), \quad x_{(j-1)M)+m-1}^M \leq x < x_{(j-1)M)+m}^M, \tag{7.102}$$

where $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, M$. Once again, it is not difficult to show that the forms $a_M(\mathbf{U}, V_{\text{bias}}; \cdot, \cdot) : W \times W \to \mathbf{C}$ satisfy the inequalities Eq. (7.98)–(7.100) with the same constants $\lambda \in \mathbf{R}$ and $\alpha, \beta > 0$ which work

for the form $a(\mathbf{U}, V_{\text{bias}}; \cdot, \cdot) : W \times W \to \mathbf{C}$ given by Eq. (7.96). Condition Eq. (7.98) implies that the sesquilinear form is bounded and condition Eq. (7.99) implies that the form is coercive. The condition given in Eq. (7.100) ensures that the form depends continuously on the optimization parameters.

We consider the sequence of abstract elliptic boundary value problems given by

$$a_M\left(\mathbf{U}, V_{\text{bias}}; \psi^M, \varphi\right) = \langle f, \varphi \rangle, \tag{7.103}$$

for every $\varphi \in W$. Now if the constants $\lambda \in \mathbf{R}$ and $\beta > 0$ guaranteed to exist by Lemma 3 are such that $\lambda < \beta/\kappa^2$, a routine application of the Lax–Milgram Theorem (see, for example, [4, 6, 7, 9]) yields the existence of unique solutions to the abstract boundary-value problems Eq. (7.97) and (7.103). This of course also implies, by virtue of its equivalence with the abstract boundary-value problem given in Eq. (7.103), the existence of a unique solution to the $2NM + 2$ dimensional linear system of equations in $2NM + 2$ unknowns given in Section 7.3.3 and the nonsingularity of $2NM + 2$-dimensional square matrix associated with it. It is also immediately clear that a sufficient condition for $\lambda \leq 0 < \beta/\kappa^2$ would be that the design space Ω, V_0, V_{f} and the total energy E are such that there exists a constant $\mu > 0$ for which $(2m_0/\hbar^2)\{V^M(x) - E\} \geq \mu$, for $x_0 \leq x \leq x_N$.

Recalling the approximation framework developed earlier, it follows that the functions $\psi^M \in W$ given by

$$\begin{aligned}
\psi^M(x) &= \psi^M(x; V_{\text{bias}}, \mathbf{U}) \\
&= \psi^M_{(j-1)M+m}(x) \\
&= A^M_{(j-1)M+m} e^{ik^M_{(j-1)M+m}\left(x - x^M_{(j-1)M+m-1}\right)} \\
&+ B^M_{(j-1)M+m} e^{-ik^M_{(j-1)M+m}\left(x - x^M_{(j-1)M+m-1}\right)},
\end{aligned} \tag{7.104}$$

for $x \in \left[x^M_{(j-1)M+m-1}, x^M_{(j-1)M+m}\right]$ with the coefficients $\left\{A^M_{(j-1)M+m}\right\}$ and $\left\{B^M_{(j-1)M+m}\right\}$, for $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, M$, determined via the propagation matrix method are in fact the unique solutions to the abstract elliptic boundary value problems Eq. (7.103). Moreover, from Eq. (7.62), (7.63), and (7.64) it follows that

$$\begin{aligned}
\psi^M(x_N; V_{\text{bias}}, \mathbf{U}) &= \psi^M_{NM}\left(x^M_{NM}\right) \\
&= A^M_{NM} e^{ik^M_{NM}\frac{L_N}{M}} + B^M_{NM} e^{-ik^M_{NM}\frac{L_N}{M}} \\
&= A^M_{NM+1} + B^M_{NM+1} = A^M_{NM+1}.
\end{aligned} \tag{7.105}$$

Consequently, Eq. (7.65) yields

$$
\begin{aligned}
J^M(U) &= \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{\left|A^M_{NM+1}(V_j, \mathbf{U})\right|^2 k_{N+1,j}}{|A_0|^2 k_0} \right|^2 \\
&= \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{k_{N+1,j}}{|A_0|^2 k_0} \left|\psi^M(x_N; V_j, \mathbf{U})\right|^2 \right|^2 \\
&= \sum_{j=1}^{\nu} \left| T_0(V_j) - \frac{k_{N+1,j}}{k_0} \left|\psi^M(x_N; V_j, \mathbf{U})\right|^2 \right|^2, \qquad (7.106)
\end{aligned}
$$

the last expression resulting when A_0 has been arbitrarily set to one.

Theorem 2. If $\lambda < \beta/\kappa^2$, then for each $M = 1, 2, \ldots$, the approximating optimal design problems involving the minimization of the performance indices J^M given in Eq. (7.65) (or equivalently in Eq. (7.106)) over the compact set Ω subject to the boundary value problem given by Eq. (7.63), (7.64) (or equivalently by Eq. (7.103)) have a solution $\hat{\mathbf{U}}^M \in \Omega$.

Proof. The result will follow immediately if for each $M = 1, 2, \ldots$ we can demonstrate the continuous dependence of $\psi^M(x_N; V_j, \mathbf{U})$ on $\mathbf{U} \in \Omega$. Since $W = H^1(x_0, x_N)$, the Sobolev Embedding Theorem (see, for example, [7] or [8]) implies that it is sufficient to demonstrate the continuous dependence of $\psi^M(\cdot; V_j, \mathbf{U}) \in W$ on $\mathbf{U} \in \Omega$ with respect to the W-norm. But this result follows immediately from the bounds given in Lemma 3. Indeed, for M fixed and $\mathbf{U}, \mathbf{U}_0 \in \Omega$, we let $\psi^M = \psi^M(\cdot; V_j, \mathbf{U}) \in W$ and $\psi^M_0 = \psi^M(\cdot; V_j, \mathbf{U}_0) \in W$ denote respectively the unique solutions to the abstract elliptic boundary value problem given in Eq. (7.103) corresponding to $\mathbf{U} \in \Omega$ and $\mathbf{U}_0 \in \Omega$. Then Eq. (7.99), (7.100), and (7.103) yield

$$
\begin{aligned}
&\beta \left\|\psi^M_0 - \psi^M\right\|^2_W \\
&\quad \le \mathrm{Re}\ \ \mathrm{a_M}\left(\mathbf{U}, \mathrm{V_{bias}}, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right) \\
&\quad + \lambda \left|\psi^M_0 - \psi^M\right|^2_H \\
&\quad = \mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}, \mathrm{V_{bias}}, \psi^{\mathrm{M}}_0, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right) - \mathrm{a_M}\left(\mathbf{U}, \mathrm{V_{bias}}, \psi^{\mathrm{M}}, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right)\right\} \\
&\quad + \lambda \left|\psi^M_0 - \psi^M\right|^2_H \\
&\quad = \mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}, \mathrm{V_{bias}}, \psi^{\mathrm{M}}_0, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right) - \left\langle \mathrm{f}, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right\rangle\right\} \\
&\quad + \lambda \left|\psi^M_0 - \psi^M\right|^2_H \\
&\quad = \mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}, \mathrm{V_{bias}}, \psi^{\mathrm{M}}_0, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right) - \mathrm{a_M}\left(\mathrm{U}_0, \mathrm{V_{bias}}, \psi^{\mathrm{M}}_0, \psi^{\mathrm{M}}_0 - \psi^{\mathrm{M}}\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+\lambda \left|\psi_0^M - \psi^M\right|_H^2 \\
&\leq \left|a_M\left(\mathbf{U}, V_{\text{bias}}, \psi_0^M, \psi_0^M - \psi^M\right) - a_M\left(\mathbf{U}_0, V_{\text{bias}}, \psi_0^M, \psi_0^M - \psi^M\right)\right| \\
&+\lambda \left|\psi_0^M - \psi^M\right|_H^2 \\
&\leq \gamma \left\|\mathbf{U} - \mathbf{U}_0\right\|_\infty \left|\psi_0^M\right|_H \left|\psi_0^M - \psi^M\right|_H \\
&+\lambda \left|\psi_0^M - \psi^M\right|_H^2 \\
&\leq \gamma\kappa \left\|\mathbf{U} - \mathbf{U}_0\right\|_\infty \left|\psi_0^M\right|_H \left\|\psi_0^M - \psi^M\right\|_W \\
&+\lambda\kappa^2 \left\|\psi_0^M - \psi^M\right\|_W^2. \qquad (7.107)
\end{aligned}
$$

It then follows that

$$
\left(\beta - \lambda\kappa^2\right) \left\|\psi_0^M - \psi^M\right\|_W^2 \leq \gamma\kappa \left\|\mathbf{U} - \mathbf{U}_0\right\|_\infty \left|\psi_0^M\right|_H \left\|\psi_0^M - \psi^M\right\|_W, \qquad (7.108)
$$

and therefore that

$$
\left\|\psi_0^M - \psi^M\right\|_W \leq \frac{\gamma\kappa}{\hat{\beta}} \left|\psi_0^M\right|_H \left\|\mathbf{U}_0 - \mathbf{U}\right\|_\infty, \qquad (7.109)
$$

where $\hat{\beta} = \beta - \lambda\kappa^2 > 0$, and the result follows.

Our approximation result for the local minima of the approximating optimal design problems takes the form of sub-sequential convergence of the solutions of the approximating optimal design problems, $\hat{\mathbf{U}}^M \in \Omega$, to a solution $\hat{\mathbf{U}} \in \Omega$ of the original optimal design problem.

Theorem 3. Let $\lambda < \beta/\kappa^2$, and for each $M = 1, 2, \ldots$, let $\hat{\mathbf{U}}^M \in \Omega$ be the solution to the Mth approximating optimal design problem. Then the sequence $\left\{\hat{\mathbf{U}}^M\right\}_{M=1}^\infty \subset \Omega$ admits a convergent sub-sequence, $\left\{\hat{\mathbf{U}}^{M_k}\right\}_{k=1}^\infty \subset \left\{\hat{\mathbf{U}}^M\right\}_{M=1}^\infty \subset \Omega$ with $\lim_{k\to\infty} \hat{\mathbf{U}}^{M_k} = \hat{\mathbf{U}} \in \Omega$. Moreover, $\hat{\mathbf{U}} \in \Omega$ is a solution to the optimal design problem given by Eq. (7.49)–(7.53) in the sense that $J\left(\hat{\mathbf{U}}\right) = \min_{U\in\Omega} J\left(\mathbf{U}\right)$.

Proof. The existence of the convergent sub-sequence, $\left\{\hat{\mathbf{U}}^{M_k}\right\}_{k=1}^\infty \subset \left\{\hat{\mathbf{U}}^M\right\}_{M=1}^\infty \subset \Omega$ with $\lim_{k\to\infty} \hat{\mathbf{U}}^{M_k} = \hat{\mathbf{U}} \in \Omega$, follows immediately from the assumption that Ω is a closed and bounded (and therefore compact) subset of $\mathbf{R}^N$. Now let $\left\{\mathbf{U}^M\right\}_{M=1}^\infty \subset \Omega$ be any convergent sequence in Ω with $\lim_{M\to\infty} \mathbf{U}^M = \mathbf{U}_0 \in \Omega$ and for each $M = 1, 2, \ldots$ let ψ^M denote the unique solution to the abstract elliptic boundary-value problem given in Eq. (7.103) with $\mathbf{U} = \mathbf{U}^M$ and let ψ_0 denote the unique solution to the abstract elliptic boundary value problem given in Eq. (7.97) with $\mathbf{U} = \mathbf{U}_0$. Then,

the bounds given in Lemma 3 imply

$$
\begin{aligned}
&\beta \left\|\psi_0-\psi^M\right\|_W^2\\
&\le \mathrm{Re}\{a_M\left(\mathbf{U}^M,V_{\mathrm{bias}},\psi_0-\psi^M,\psi_0-\psi^M\right)\}+\lambda\left|\psi_0-\psi^M\right|_H^2\\
&=\mathrm{Re}\left\{a_M\left(\mathbf{U}^M,V_{\mathrm{bias}},\psi_0,\psi_0-\psi^M\right)-a_M\left(\mathbf{U}^M,V_{\mathrm{bias}},\psi^M,\psi_0-\psi^M\right)\right\}\\
&\quad+\lambda\left|\psi_0-\psi^M\right|_H^2\\
&=\mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}^{\mathrm{M}},\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)-\left\langle \mathrm{f},\psi_0-\psi^{\mathrm{M}}\right\rangle\right\}\\
&\quad+\lambda\left|\psi_0-\psi^M\right|_H^2\\
&=\mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}^{\mathrm{M}},\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)-\mathrm{a}\left(\mathbf{U}_0,\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)\right\}\\
&\quad+\lambda\left|\psi_0-\psi^M\right|_H^2\\
&=\mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}^{\mathrm{M}},\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)-\mathrm{a_M}\left(\mathbf{U}_0,\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)\right\}\\
&\quad+\mathrm{Re}\left\{\mathrm{a_M}\left(\mathbf{U}_0,\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)-\mathrm{a}\left(\mathbf{U}_0,\mathrm{V_{bias}},\psi_0,\psi_0-\psi^{\mathrm{M}}\right)\right\}\\
&\quad+\lambda\left|\psi_0-\psi^M\right|_H^2\\
&\le \gamma\kappa\left\|\mathbf{U}^M-\mathbf{U}_0\right\|_\infty|\psi_0|_H\left\|\psi_0-\psi^M\right\|_W\\
&\quad+\frac{2m_0}{\hbar^2}\int_{x_0}^{x_N}\left\{V^M(x)-V(x)\right\}\psi_0(x)\left(\bar\psi_0(x)-\bar\psi^M(x)\right)dx\\
&\quad+\lambda\kappa^2\left\|\psi_0-\psi^M\right\|_W^2\\
&\le \gamma\kappa\left\|\mathbf{U}^M-\mathbf{U}_0\right\|_\infty|\psi_0|_H\left\|\psi_0-\psi^M\right\|_W\\
&\quad+\frac{2m_0V_{\mathrm{bias}}}{\hbar^2}\sum_{j=1}^{N}\int_{x_{j-1}}^{x_j}\frac{L_j}{ML}|\psi_0(x)|\left|\bar\psi_0(x)-\bar\psi^M(x)\right|dx\\
&\quad+\lambda\kappa^2\left\|\psi_0-\psi^M\right\|_W^2\\
&\le \gamma\kappa\left\|U^M-U_0\right\|_\infty|\psi_0|_H\left\|\psi_0-\psi^M\right\|_W\\
&\quad+\frac{2m_0\kappa V_{\mathrm{bias}}}{\hbar^2M}|\psi_0|_H\left\|\psi_0-\psi^M\right\|_W\\
&\quad+\lambda\kappa^2\left\|\psi_0-\psi^M\right\|_W^2.
\end{aligned}
$$

It then follows that

$$
\begin{aligned}
&\left(\beta-\lambda\kappa^2\right)\left\|\psi_0-\psi^M\right\|_W^2\\
&\le \gamma\kappa\left\|\mathbf{U}^M-\mathbf{U}_0\right\|_\infty|\psi_0|_H\left\|\psi_0-\psi^M\right\|_W+\frac{2m_0\kappa V_{\mathrm{bias}}}{\hbar^2M}|\psi_0|_H\left\|\psi_0-\psi^M\right\|_W,
\end{aligned}
\tag{7.110}
$$

and therefore that

$$
\left\|\psi_0-\psi^M\right\|_W\le\frac{\gamma\kappa}{\hat\beta}\left\|\mathbf{U}^M-\mathbf{U}_0\right\|_\infty|\psi_0|_H+\frac{2m_0\kappa V_{\mathrm{bias}}}{\hat\beta\hbar^2M}|\psi_0|_H,
\tag{7.111}
$$

where once again in Eq. (7.111) we have made use of the fact that $\hat{\beta} = \beta - \lambda\kappa^2 > 0$. Consequently, we have

$$\lim_{M\to\infty} \psi^M\left(\cdot; V_{\text{bias}}, \mathbf{U}^M\right) = \psi_0\left(\cdot; V_{\text{bias}}, \mathbf{U}_0\right), \tag{7.112}$$

in W, and, once again by the Sobolev Embedding Theorem [5], in $C[x_0, x_N]$ as well. Finally, for any $\mathbf{U} \in \Omega$ we find that

$$J\left(\hat{\mathbf{U}}\right) = J\left(\lim_{k\to\infty} \hat{\mathbf{U}}^{M_k}\right) = \lim_{k\to\infty} J^{M_k}\left(\hat{\mathbf{U}}^{M_k}\right) \leq \lim_{k\to\infty} J^{M_k}(\mathbf{U}) = J(\mathbf{U}), \tag{7.113}$$

and the desired result has been established.

It is interesting to note that it is in fact possible to characterize both the gradient of J and the gradient of J^M and the convergence of ∇J^M to ∇J as $M \to \infty$ through the use of an adjoint and a co-state variable. Indeed, for each $\mathbf{U}_0 \in \Omega$ and $j = 1, 2, \ldots, \nu$, a straightforward calculation yields the adjoint and gradient formula $a(\mathbf{U}_0, V_j; \psi_j, \varphi) = \langle f, \varphi\rangle$, for every $\varphi \in W$, $\overline{a(\mathbf{U}_0, V_j; \varphi, \eta_j)} = \langle g_j(\psi_j), \varphi\rangle$, for every $\varphi \in W$

$$\nabla J(\mathbf{U}) = \frac{2m_0}{\hbar^2}\mathrm{Re}\sum_{J=1}^{\nu}\left[\int_{x_0}^{x_1} \psi_j(x)\,\overline{\eta_j(x)}dx, \ldots, \int_{x_{N-1}}^{x_N} \psi_j(x)\,\overline{\eta_j(x)}dx\right], \tag{7.114}$$

where for $\varphi, \psi \in W$, $g_j(\psi) \in W^*$ is given by

$$\langle g_j(\psi), \varphi\rangle = -\frac{4k_{N+1,j}}{k_0}\left[T_0(V_j) - \frac{k_{N+1,j}}{k_0}|\psi_j(x_N)|^2\right]\psi_j(x_N)\,\overline{\varphi(x_N)}. \tag{7.115}$$

Similarly, for each $j = 1, 2, \ldots, \nu$, each $M = 1, 2, \ldots,$ and $\mathbf{U}^M \in \Omega$ we have $a_M\left(\mathbf{U}^M, V_j; \psi_j^M, \varphi\right) = \langle f, \varphi\rangle$, for every $\varphi \in W$, $\overline{a_M\left(\mathbf{U}^M, V_j; \varphi, \eta_j^M\right)} = \left\langle g_j\left(\psi_j^M\right), \varphi\right\rangle$, for every $\varphi \in W$,

$$\nabla J^M(\mathbf{U}) = \frac{2m_0}{\hbar^2}\mathrm{Re}\sum_{J=1}^{\nu}\left[\int_{x_0}^{x_1} \psi_j^M(x)\,\overline{\eta_j^M(x)}dx, \ldots, \int_{x_{N-1}}^{x_N} \psi_j^M(x)\,\overline{\eta_j^M(x)}dx\right]. \tag{7.116}$$

Then if $\lambda < \beta/\kappa^2$ and $\lim_{M\to\infty} \mathbf{U}^M = \mathbf{U}_0 \in \Omega$, as in the proof of Theorem 2, for each $j = 1, 2, \ldots, \nu$, we have $\lim_{M\to\infty} \psi_j^M = \psi_j$ in W, from which it is

straightforward to show $\lim_{M\to\infty} g_j\left(\psi_j^M\right) = g_j(\psi)$ in W^*, or equivalently that $\lim_{M\to\infty} \left\|g_j\left(\psi_j^M\right) - g_j(\psi)\right\|_{W*} = 0$. Once again using the coercivity condition (7.99) co-state convergence can be argued. That is that $\lim_{M\to\infty} \eta_j^M = \eta_j$ in W. The convergence of the gradients, $\lim_{M\to\infty} \nabla J^M(\mathbf{U}_M) = \nabla J(\mathbf{U}_0) \in R^N$ then follows immediately from Eq. (7.114), (7.116), the continuity of the H inner product and the continuous embedding of W in H.

7.3.6 A numerical example

We consider the optimal design of a ten-layer device in which all of the layers have the same thickness of 1 nm. The device is to have a quadratic transmission function, T_0,

$$T_0(V) = 0.05V^2 + 0.015V + 0.001. \tag{7.117}$$

We base our design on 26 equally spaced bias voltages from $V_{\min} = 0$ V to $V_{\max} = 0.25$ V. It follows that we set $N = 10$, $L = 10$, $x_i = i$, $i = 0, 1, 2, \ldots, N$, $L_i = 1$, $i = 1, 2, \ldots, N$, $\nu = 26$, $V_{\min} = 0$, $V_{\max} = 0.25$, and $V_j = V_{\min} + j\frac{V_{\max} - V_{\min}}{\nu}$, $j = 0, 1, 2, \ldots, \nu$.

All computations were carried out on a PC using MATLAB. The resulting approximating optimization problems were solved using the MATLAB constrained optimization routine FMINCON. An initial guess for the layer potential energies had to be supplied. We took it to be constant across all the layers of the device at 0.5 eV. The feasible potential energy levels were constrained to remain between $U_L = 0$ eV and $U_H = 1$ eV. In our calculations we set the incident electron energy to be $E = 0.026$ eV and the effective electron mass $m^* = 0.07 \times m_0$ where $m_0 = 9.10939 \times 10^{-31}$ kg is the bare electron mass. This choice of m^* is appropriate for an electron in the conduction band of $\mathrm{Al}_x Ga_{1-x}$ As.

We determined optimal designs with discretization levels $M = 2, 4, 8, 16, 32, 64, 128$. To evaluate the performance of our scheme we simulated the performance of the optimal designs by evaluating the least-squares performance index, Eq. (7.53), using the propagation matrix method to solve the Schrödinger equation discretized at the level of $M = 512$. To attempt to observe the convergence of the optimal designs, we computed the relative error between the optimal design for discretization level M and the optimal design for discretization level $M = 128$ in both the L_2 and L_∞ norms. Finally we also recorded the number of FMINCON iterations that were required until convergence was achieved and calculated the number of CPU seconds per iteration. We give our results in Table 7.1.

In Fig. 7.2 and 7.3 the potential energy profiles and the values of the

Table 7.1. *Results for optimal potential energy profiles for different discretization levels.*

	$J^M(\widehat{U}^M)$ (10^{-8})	$J^{512}(\widehat{U}^M)$ (10^{-6})	Relative error in L^2	Relative error in L^∞	Iter	CPU sec/iter
2	1.06	7.11	0.22	0.30	26	1.67
4	1.13	1.50	0.12	0.13	38	3.02
8	0.56	0.35	0.04	0.05	24	5.92
16	1.85	0.11	0.15	0.17	38	10.57
32	1.20	0.03	0.06	0.06	34	21.85
64	0.51	0.01	0.09	0.10	55	44.25
128	1.03	0.01	0.00	0.00	107	99.24

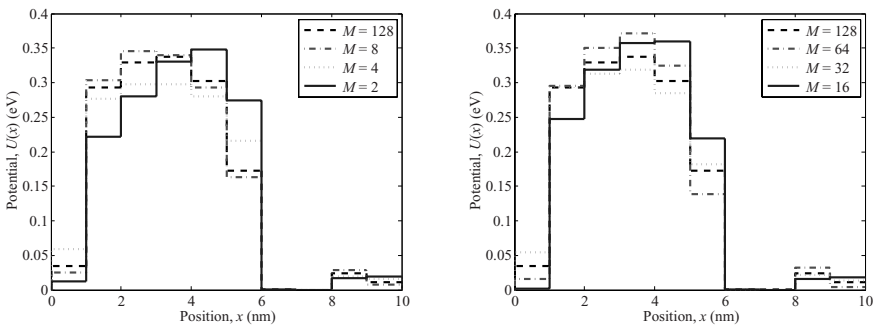

Fig. 7.2. Optimal potential energy profiles for the device given by Eq. (7.117) for discretization levels $M = 2, 4, 8, 128$ (left) and $M = 16, 32, 64, 128$ (right).

transmission function at the bias voltage levels $V_j = V_{\min} + j\frac{V_{\max}-V_{\min}}{\nu}$, $j = 0, 1, 2, \ldots, \nu$, are shown. To simulate the actual performance of the optimal designs, in calculating the transmission functions we discretized the Schrödinger equation using the approach we have described here at discretization level $M = 512$. Inspection of Table 1 reveals that performance improves with increasing level of discretization.

7.4 Techniques for global optimization

A local minimization approach is often successful in finding a design with satisfactory performance when a nearby sub-optimal design in the design space can be pre-determined using a highly simplified intuitive design approach. Global optimization, on the other hand, is necessary for at least two reasons. First, it is always extremely challenging to demonstrate that

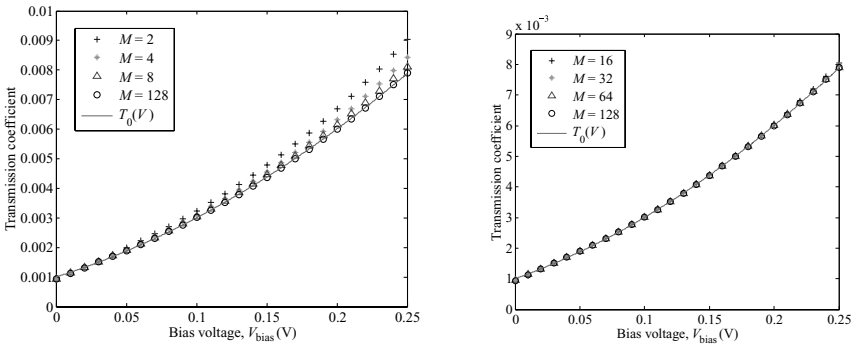

Fig. 7.3. Simulated value of transmission function at bias voltage design values using optimal potential energy profiles for the device given by Eq. (7.117) for discretization levels $M = 2, 4, 8, 128$ (left) and $M = 16, 32, 64, 128$ (right).

the optimal design obtained through local optimization is close to the "best that can be achieved," and secondly, a close to optimal initial design is not always available. The latter is especially true in the case of design problems involving a relatively large number of design parameters. In the examples presented in earlier chapters of this book, we have demonstrated the success of the local optimization approach in designing a high-performance EM waveguide and a nanoscale electronic device with arbitrary functional transmission properties. However, in both of these cases, the local optimization approach does not provide us with any indication as to whether or not the final designs yield the best achievable performance. In fact, numerical experiments in the case of the EM waveguide design problem have shown that substantially different locally optimal designs with similar performance can be identified when an alternative initial configuration of the Teflon cylinders is used.

A global minimization (or maximization) problem in a finite-dimension vector space can generally be stated as follows. Consider a subset K of a finite-dimensional vector space V and a scale function J defined on K. Find an element $x^* \in K$ such that

$$J(x^*) \leq J(x), \forall x \in K, \tag{7.118}$$

for a minimization problem, or, in the case of maximization problem as

$$J(x^*) \geq J(x), \forall x \in K. \tag{7.119}$$

The global optimization of an arbitrary function J is an extremely challenging mathematical problem [10–14]. In fact, even the existence of a global optimum can only be guaranteed in problems where both the function J and the subset K satisfy special, generally difficult to verify, conditions. For

example, one set of conditions is convexity of both the objective function and the design space. A subset K of a vector space is said to be convex if for any two points x_1 and x_2 in K, the line segment linking the two points given by $\{x, x = tx_1 + (1-t)x_2, t \in [0,1]\}$ belongs to K. A function J is said to be convex if the following inequality holds for any $t \in (0,1)$ and x_1 and x_2 in K:

$$J(yx_1 + (1-t)x_2) \leq tJ(x_1) + (1-t)J(x_2). \tag{7.120}$$

A convex function J can have at most one global minimum and any local minimum of J must also be a global minimum. As a result, for any convex function, the search for a local minimum is equivalent to the search for a global minimum. However, for many practical optimal design problems, the performance index, or objective function, is not convex. In fact, the existence of many locally optimal designs in the design space poses a significant challenge for any global optimization technique.

Alternatively, there is a broad class of optimization problems for which the existence of a global optimal solution is guaranteed. We consider a compact subset K of a normed vector space V. A subset K in a finite-dimensional vector space is said to be compact if and only if K is a closed and bounded subset of V. A function J has at least one global minimum and one global maximum in K if J is continuous. In particular, there may be many globally optimum solutions and some of these solutions may be on the boundary of the set K. For most of the optimal design problems occurring in practice, it is not difficult to specify a reasonable bound for the values of the design parameters. Therefore, limiting the search for the optimal design to a bounded and closed subset of the space of design parameters is often quite feasible. On the other hand, the continuity requirement is also readily verified for most design problems of interest. Consequently, we limit our discussion in the remainder of this section to the global optimization of a continuous performance index J over a compact subset K of the design space. We refer to this class of optimization problem as the Continuous Global Optimization Problem over Compact Subset (CGOP-C). In addition to the continuity of the performance index, for many optimal design problems the performance index is in fact smooth or differentiable with respect to the design variables. In the remainder of this chapter, we present a newly developed global optimization algorithm that is specifically designed to handle this class of optimal design problem by combining local optimization techniques with global random search to achieve high efficiency in identifying globally optimal solutions. We refer to this new algorithm as the *Ensemble Global Search* (EGS) algorithm.

As in the case of genetic algorithms (GA) [15–17], the EGS algorithm

makes explicit use of the ability to efficiently evaluate the performance index at multiple locations in the design space in parallel by considering an ensemble of candidate designs at any given step of the algorithm. However, unlike a GA, the selection of the ensemble in the subsequent iteration of the algorithm explicitly depends on all the previously evaluated design candidates. The EGS algorithm considers the size of the ensemble as a measure of available resources for the global search. This is analogous to how modern-day highly-coordinated explorers for valuable minerals search over a vast terrain. Indeed, the EGS deploys finite resources strategically by assessing the potential of each new area of exploration. Two key considerations in our EGS strategy are coverage and efficiency. As in any well-coordinated exploration campaign, the entire targeted region must be interrogated or explored by a survey team. Efficient preliminary assessment of the potential of an area represented by a sample point helps to determine whether or not the region should subsequently be explored in higher resolution and greater detail. The re-evaluation of the region can occur when a finer level of survey is permitted by available resources at a later time. On the other hand, local information such as the gradient can be used to quickly lead the explorer to a potentially highly promising region and thus accelerate the overall exploration process. In addition to following local leads, EGS is also careful to monitor how close its parallel exploration neighborhoods come to each other so as to guard against different local explorations discovering the same locally optimal design. In this way, the EGS algorithm is able to free-up scarce resources so that they may be deployed to explore new previously unexplored regions of the design space.

Another objective in the development of the EGS scheme is that the algorithm not only locate a global optimal solution, but that it also identify as many "next best" sub-optimal solutions as possible. The EGS is designed to be used in conjunction with a highly-efficient local optimization technique so that any potentially promising candidate regions can be refined via the local optimization algorithm. As a result, the primary objective of EGS is to successively place resources in reasonably close proximity to a globally optimal solution so that subsequent refinement via local optimization readily pinpoints the optimal solution. We define the following criterion for a successful EGS search.

Definition An execution of the EGS is said to be successful in finding global optima of a performance index J over a compact subset K of a vector space V if, for any connected subset O_k of global optima and local optima, a sample point x_k^* is placed within $\epsilon_k > 0$ distance to an element $x_k \in O_k$. The radius ϵ_k is determined so that whenever the initial guess for a local

optimization algorithm is within ϵ_k to x_k, the local optimization iteration would converge to an element of O_k.

Given a local optimization scheme we define the region of attraction $A(x_k)$ of a local minimum x_k as the set of all points that if the local optimization is started from $x \in A(x_k)$ the iterative optimization scheme converges to the local optimum x_k. Naturally, the shape of $A(x_k)$ depends on both the performance index function J and the local optimization scheme. In the above definition, we implicitly assumed that $A(x_k)$ contains a sphere $S(x_k)$ centered at x_k with a radius ϵ_k.

Unlike clustering algorithms, EGS does not attempt to determine the shape of the region of attraction $A(x_k)$. Instead, EGS simply considers the subset $S(x_k)$ of $A(x_k)$ as part of its strategy for eliminating duplication in the local search effort. The underlying rationale is that the search space is so large that the exploration of untested regions is of higher priority than the determination of the shape of any particular region of attraction $A(x_k)$.

In initiating a new local search point, the random sampling of the design parameter space selects points at a distance ϵ away from all previously evaluated points. This *exclusion zone* around each previously evaluated point forces the coverage of the design parameter space to grow in size and scope. However, it also introduces the risk of missing high-performing designs, especially when the exclusion radius ϵ is large as it is early in the EGS iteration. The EGS relies on the gradual reduction of ϵ once a survey of the entire design parameter space at a specified resolution has nearly been completd. This provides the capability for capturing missed high-performing designs in earlier, coarser scale exploration. In Fig. 7.4 an example of a selection of initial points is shown.

In order to balance the computational load of computing nodes of a cluster during the execution of the EGS, each node is assigned the same number of candidate designs to evaluate. Recall that evaluating the potential benefit of a candidate design typically involves solving the forward model equations for the current values of the design parameters. In each iteration of the EGS, only a subset of these candidate designs corresponds to the initialization of a local search from random samples of the design space. The remaining portion of the candidates is for the continuation of a local search initiated in a previous iteration. As a result, the population of design candidates evaluated in each EGS iteration is similar to a GA generation. However, the motivation and mechanism for creating the subsequent generation are substantially different from a GA based scheme.

In our initial implementation of the EGS, we employed a gradient descent local search algorithm. This method is simple to implement and it is

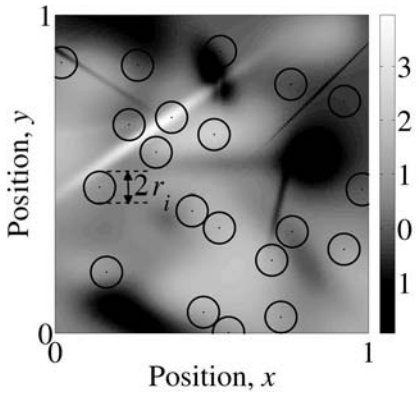

Fig. 7.4. Example of initial point generation for a 2D test-objective function as described below. Initial as well as additional random points must have a minimum distance of r_i at iteration i from all previously generated sample points.

computationally efficient when the gradient of J is computed using the adjoint method. A key element of the EGS algorithm is to *forecast* the potential gain that a continued local search can yield in comparison to already established achievable performance. This forecast is achieved using available local information in the form of first- and second-order tests described in subsections below. Only points that pass both tests successfully are used for continued local search. It is important to note that the points in the population of candidate designs that are generated from local searches are identified. In fact, the exclusion radius does not apply to the selection of these candidate designs. Although points determined by local optimization are allowed to be placed in areas that are closer to previous function evaluations than the exclusion radius ϵ when two local optimization branches lead to points within ϵ distance from each other in the design space, the weaker performing branch is terminated while the other one is allowed to continue. The subsequently freed computational resource is then used for exploring the search space by random placement of a search point. In this manner global search is automatically balanced against local search.

In general the exclusion radius ϵ is reduced once it becomes too difficult to generate new trial points that have a minimum distance ϵ to all prior trial points. In this manner previously explored areas are opened up for more detailed further exploration.

EGS can be terminated in various ways. These include:

(a) A fixed number of iterations or generations is reached;
(b) A minimum exclusion radius $\epsilon_{\min}$ is reached; or
(c) Improvement of highest performing design has stalled.

Detailed descriptions of various aspects of the EGS implementation are presented in the following subsections. These include the first- and second-order tests, considerations for implementation on a parallel computer, archiving of ensemble information, and the handling of the termination of the EGS. The results of our evaluation of the EGS using a family of test problems are also presented.

7.4.1 First-order test

In Section 7.2 we have shown that the performance index for optimal design problems constrained by differential equations often admits an adjoint system that can be used to efficiently evaluate the gradient of the performance index with respect to the design parameters. Therefore, whenever the objective function $J(x_{\text{trial}})$ is evaluated at the coordinate $x_{\text{texttrial}}$ we also compute the gradient and archive it in the database for previously examined designs. The first-order test consists of using a first-order approximation of $J(x)$ in the neighborhood of $x_{\text{mobxtrial}}$ as a predictor of achievable performance. With the limited information given within the sphere $\mathcal{B}_\epsilon(x_{\text{texttrial}})$ a good approximation of the objective function is given by

$$J(x) \approx J(x_{\text{trial}}) + (x - x_{\text{trial}})^T \cdot \nabla J(x_{\text{trial}}). \tag{7.121}$$

With this prediction the EGS tries to determine if a point identified by the local search in the neighborhood of x_{trial} shows enough promise to start or continue local optimization. In the evaluation of the potential of a candidate design, two types of consideration are involved. The first is the absolute performance relative to all previously evaluated designs. That is, if the performance of x_{trial} is already among the best performing designs examined, any improvement is worthwhile to explore. The second consideration is relative improvement. Since the gradient information merely provides the direction and the rate of change in performance, a very slow rate would require massive modification to achieve meaningful performance enhancement. Since in general the performance index is a highly nonlinear function of the design parameters, the larger the required change, the less reliable the performance prediction given by the linear approximation becomes. Consequently, even among the best performing designs, if further improvement requires an unreasonable magnitude change, the local search can still be deemed unpromising. Thus, local optimization is pursued if

- $J(x_{\text{trial}}) < J_{\text{critical}}$, or,
- $J(x_{\text{trial}}) - \|\nabla J(x_{\text{trial}})\| \, \epsilon < J_{\text{critical}}$,

where J_{critical} is a statistical parameter determined from all previous performance index evaluations archived in the database. The cut-off value J_{critical}

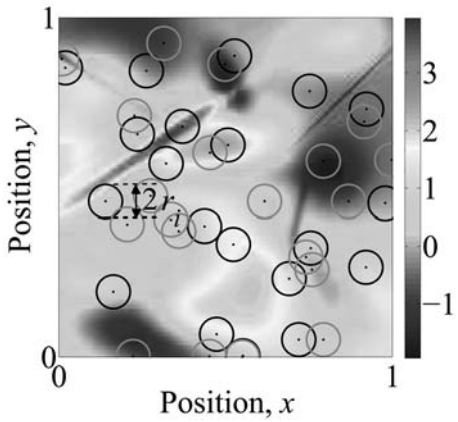

Fig. 7.5. The second generation of points is marked with a lighter shade exclusion radius while the initial generation of trial points is circled in black. The local search procedure does allow the exploration within the exclusion radius of already evaluated points. The new generation of points is strictly outside these exclusion radii.

may, for example, be selected as the βth percentile $P_{\beta(J(x_1),\ldots,J(x_i))}$ of all previously obtained values of performance index evaluations. If a test point "fails" the first-order test, the assigned computing node is freed for further global exploration. If, on the other hand, a test point "passes" the first-order test, it is subjected to the second-order test (see Fig. 7.5) which we describe in the next section.

7.4.2 Second-order test

Once the potential promise of a test point x_{trial} has been established, a more accurate local search method is applied in the form of a second-order test. One additional objective function and gradient evaluation is made at $x_2 = x_{\text{trial}} - \lambda \left\|\nabla J(x_{\text{trial}})\right\|^{-1} \nabla J(x_{\text{trial}})$, where λ is a trust region parameter. The quantities $J(x_{\text{trial}})$, $\|\nabla J(x_{\text{trial}})\|$, $J(x_2)$, and $\left|(\nabla J(x_2))^T \cdot \left\|\nabla J(x_{\text{trial}})\right\|^{-1} \nabla J(x_{\text{trial}})\right|$ are used to construct a quadratic approximation to J along the line connecting x_2 and x_{trial}, as shown in Fig. 7.6. By solving for the unique extremum of this quadratic approximation to the performance index, we obtain the next sample point for the local search. If the fitted parabola is convex, the next iterate x_{new} is the minimum of the quadratic function. If the fitted parabola is concave the maximum allowed step-length of twice the trust region radius 2λ is taken.

The least-squares approximation of $J((1-\theta)x_{\text{trial}} + \theta x_2) = J(x_{\text{trial}} - \theta\lambda \left\|\nabla J(x_{\text{trial}}\right\|^{-1} \nabla J(x_{\text{trial}}))$ along the line linking x_2 and x_{trial} by a quadratic function of the form $p(\theta) = a\theta^2 + b\theta + c$ is obtained by solving the linear

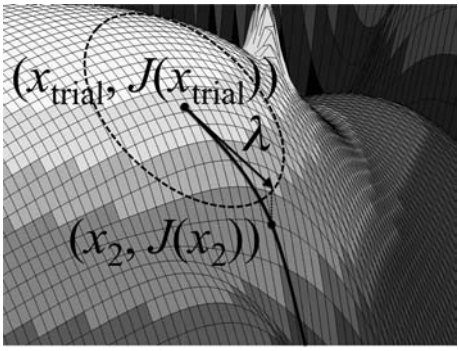

Fig. 7.6. Example of an inverted parabola fitted locally to a trial point x_{trial}. The parabola is fitted along the direction of steepest descent.

system

$$\begin{pmatrix} 0\,0\,1 \\ 1\,1\,1 \\ 0\,1\,0 \\ 2\,1\,0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} J(x_{\text{trial}}) \\ J(x_2) \\ -\lambda \left\| \nabla J(x_{\text{trial}}) \right\| \\ -\lambda \left\| \nabla J(x_{\text{trial}}) \right\|^{-1} (\nabla J(x_{\text{trial}}))^T \cdot \nabla J(x_2) \end{pmatrix}, \tag{7.122}$$

in the least squares sense. When the value b is negative, the function $g(\theta) = J(x_{\text{trial}} - \theta\lambda \left\| \nabla J(x_{\text{trial}} \right\|^{-1} \nabla J(x_{\text{trail}}))$ and its quadratic approximation are both decreasing as a function of θ at 0. This is important since we would like to move in the direction of negative gradient. However, since we solve the above linear system of equations in the least squares sense, we must check whether or not $b < 0$. If $a > 0$, the local quadratic approximation is an upright parabola and the optimal step-length is given by $\theta_{\text{step}} = -\frac{b}{2a}$. If $b < 0$, the minimum value of the function $g(\theta)$ is achieved for a value $\theta_{\text{step}} > 0$. In the case that $a > 0$ but also $b \geq 0$ we choose to omit the gradient information given at $x_{i,2}$ and set $\theta_{\text{step}} = \frac{|\nabla J(x_i)|}{2a}$. If $a < 0$, then the fitted approximation is an inverted parabola and we choose the maximum allowed step-length $\theta_{\text{step}} = 2\epsilon$. Using the computed step-length θ_{step} a new local search point $x_{\text{new}} = x_{\text{trial}} - \theta_{\text{step}} \left\| \nabla J(x_{\text{trial}}) \right\|^{-1} \nabla J(x_{\text{trial}})$ is generated.

A local search can be terminated if either local optimization is successful or the process is stranded on a plateau. More precisely, we terminate a local search and mark the identified local optimum if

$$\left| \frac{J(x_{\text{trial}}) - \epsilon \left\| \nabla J(x_{\text{trial}}) \right\|^{-1} \nabla J(x_{\text{trial}})}{J(x_{\text{trial}})} \right| < \delta, \tag{7.123}$$

where δ is a parameter selected to control the termination process. If the local search process is not terminating, the second-order test compares the

value of $p(\theta_{\text{step}})$, which is a proxy for the lowest value achievable at x_{new} in the next step to the threshold value J_{critical}. If $p(\theta_{\text{step}})$ is above the threshold, the local search is abandoned similarly to the case of the first-order test. Otherwise, the new candidate design is included in the next step of the EGS algorithm.

7.4.3 Parallel computation

Both first-order and second-order tests require the one evaluation of the performance index which includes the solution of the adjoint equation and the computation of the gradient of the performance index. In this manner it is easy to distribute the computation on a cluster in either a synchronized or an asynchronized manner. The default computational architecture for implementation of EGS is an interconnected, nearly homogeneous, computer cluster with Message Passing Interface (MPI) among the nodes. A master node is in charge of dispatching tasks to the rest of the slave nodes. The master node collects test points, function values, and function gradients from the other nodes in a database. From the database the master node extracts the status of the optimization and distributes the computationally intensive first- and second-order tests to the subordinate nodes. These nodes return the function values and gradients evaluated at the designated coordinates to the master node to be inserted into the central database. Since the most computationally intensive portion of the optimization (the evaluation of the performance index) is carried out within a single node and multiple copies of the computation are executed in parallel by different nodes, this parallelization scheme is simply referred to as "trivially parallel." The computational efficiency therefore scales linearly as a function of the size of the cluster. In particular the EGS algorithm tries to keep the assigned nodes at full computational load, minimizing the idle time when subordinate nodes await instructions. At the same time the single master node ensures that all slave nodes work with the most recent information extracted from the central database.

7.5 Database of search iterations

One key element of the EGS is the idea that all decisions on selection of new points to explore or continuation of a local search are based on ensemble information. This is made possible by maintaining a central database of all previously examined candidate designs. In particular, the database archives the values of design variables for the previously examined candidate designs,

the values of the performance index and its gradient, as well as status flags indicating whether or not a point is a part of a chain of local search or a termination point, and the reason for the termination. In addition to the database, the EGS algorithm can be custom configured using the following three parameters:

- the initial exclusion radius r_0, and an optional minimum exclusion radius $r_{\min}$;
- the trust region parameter ϵ (defaults to r_0), and
- the so called "greediness" parameter β that is used in the first-order and second-order tests to compare the performance of a candidate design to the performance of the top β-percentile of previously examined designs.

The initial value of r_0 is selected to force the algorithm to explore the entire search space with a relatively low number of evaluations of the performance index. In addition to the above three parameters, we also need to choose parameters that determine the transitional condition for the reduction of the exclusion radius. In particular, we must select a fixed number of cost function evaluations M as well as a proportion of the design space δ that we consider sufficiently insignificant to leave unexplored. In fact, we typically choose r_0 so that after at most M randomly placed points, a fraction of $1 - \delta$ of the design space is covered by spheres with radius r_0 centered at the already explored points in the design space. The value of M can be chosen as a certain fraction of the maximum allowed iterations. For an example, if the volume of the search space is $V = \text{Volume}(\Omega) \subset \mathrm{R}^n$ a good choice for the initial exclusion radius r_0 might be $r_0 = \sqrt[n]{\delta V/M}$. The well known "curse of dimensionality" may make the above selection of an initial exclusion radius impractical. In fact, the volume covered by a cube with edge-length equal to r_0 in n-dimensional Euclidian space is r_0^n. Thus, it would require an extremely large value for M to allow sufficient coverage of the design space. The exclusion radius r_i is reduced once one is unable to generate further trial points that have a minimum distance of r_i from all previously evaluated sample points. In this case we replace r_i with $r_{i+1} = \alpha r_i$ and proceed to sampling the design space using the new exclusion radius. In an enhanced implementation of EGS the exclusion radii can vary for different regions of the design space based on prior evaluation of the performance index in that region. For example, a low-performing region with small gradient may be assigned a larger exclusion radius than other regions.

The trust region parameter ϵ is primarily chosen to define a neighborhood around a sample point in which the lower-order approximation of the performance index is valid. In particular, this parameter should reflect our estimation of the behavior of the objective function relevant for the local

search and it signifies the radius over which the quadratic approximation adequately models the objective function.

The parameter β determines how aggressively global exploration should be preferred to local optimization. Local optimization is initiated or continued only for points that achieve or are expected to lead to a top-performing point quickly. The smaller the value of β the fewer points fall into the β percentile of all previous function evaluations and the more randomly generated sample points are subjected to the first-order test. The percentile measure is some what self-correcting because the cut-off value J_β is raised with the number of sample points that lie above the cut-off value and shrinks if too many sample points fall below the critical performance measure. In fact, an interrupted local search due to the failure of the first- or the second-order test may appear more attractive to pursue further as the algorithm encounters more and more low-performing designs during the global search. To enable the EGS algorithm to continue an interrupted local search, the gradient of the performance index is also stored in the database.

Another feature of the database is that it tracks the local minima the algorithm has already discovered. If it turns out that after termination of the global search the best found solution lacks some other design requirements such as robustness or sensitivity that may not have been reflected in the objective function, the database offers alternatives in the form of sub-optimal solutions. The user has the option of applying a more advanced local search in the neighborhood of these alternative candidates before making the decision to terminate the search and accept a final design.

7.5.1 Termination of global search

In general, execution of the EGS terminates after the number of function evaluations reaches a specified limit. However, the algorithm can also terminate when nearly complete coverage of the design space is achieved with a sufficiently small exclusion region. For global search in a high-dimensional design space, a complete search at fine scale is almost always too expensive.

7.5.2 Procession flow of EGS

The implementation of the EGS requires the establishing of communication between the master and slave nodes. This communication is illustrated in Fig. 7.7. The slave nodes request a task to be performed from the master node. This is represented by the path (1) in Fig. 7.7. Upon receiving a task request, the master node searches the archive for possible local searches that can be continued by comparing the potential pay-off of the search to the

β-percentile performance. If a point is found, the master node sends the design parameters for the next step of the local search to the slave node. If no local search is worthy of being continued, a randomly generated set of design parameters is sent to the slave node. This communication is represented by the path (2). The search of the archive is represented by path (6). Upon receiving the task instruction, the slave node evaluates the value of the performance index for the design parameter as well as the gradient vector at that point, and then sends the results to the master node. This is represented by path (3). Once the values of the performance index and its gradient vector are available, the first-order test is performed. If a point passes the first-order test, evaluation of the performance index at an additional point is required to enable evaluation of the second-order test. The master node requests the next available slave node to perform this evaluation. After receiving the value of the performance index and its gradient at the new point in the design space, the master node performs the second test. The potential pay-off as well as the coordinates of the next set of design parameters to evaluate are archived in the database along with the value of the performance index and its gradient vector at the current point in the design space. This is represented by path (5). Even though the management of the archive and the coordination of searches are all handled by the master node, in our description we also highlight the interfaces between these two modules. In fact it is possible to designate a special node in the computer cluster to handle archival functions. In such an implementation, it is also useful to identify the interfaces between the search coordination and the archival modules.

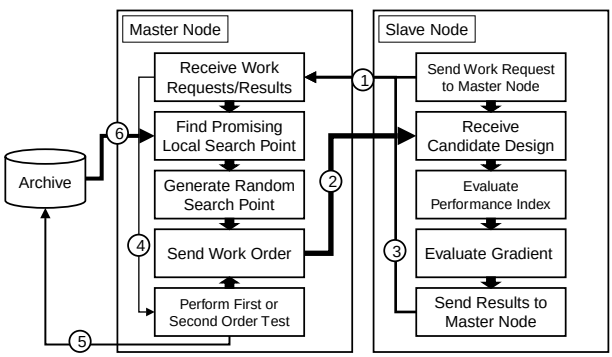

Fig. 7.7. The processing flow and exchange of information in an implementation of EGS.

7.5.3 Evaluation of performance of the GO-algorithm

Before using the EGS on actual engineering design problems, assessment of its performance was conducted on mathematically representative test examples. In particular, since in the EM-scattering problem the parameter space can be over 100-dimensional, the efficiency of an algorithm with an increasing dimensionality of the search space is a crucial criterion for evaluating performance. The assessment of EGS performance was made using a test function that is an explicit function of the design parameters and as such can be evaluated efficiently. The two main performance measures that concern us are runtime, which is represented by the number of evaluations of the performance index, and the performance (i.e the optimality) of the solution. The assessment of solution quality used in our test was the lowest value of the performance index since we formulated the optimal design problem as a minimization problem. In general, in practice the true global minimum is not known, and so the minimum value of the performance index is a good proxy for the quality of the solution. Since we are also interested in finding as many locally optimal solutions as possible we also use the number of identified locally optimal solutions as a measure of performance.

7.5.4 A randomly generated test problem

We use a family of explicitly defined functions parameterized by randomly generated values. The advantage of an explicit test function is that it can be evaluated very quickly and the performance of the EGS can be readily identified and improved. The random selection of the parameters helps make the results of our evaluation more robust and valid for a larger range of design problems. On the other hand, we also required the cost function in our tests to have analytic properties similar to the engineering examples we are interested in.

The family of functions we chose for the performance evaluation is a sum of gaussian functions also considered by Pardalos and Romeijn ([11], p.396). The test function is given by

$$f(x) = \sum_{i=1}^{N} f_i \exp\big(-(x - s_i)^T C_i^{-1} (x - s_i)/2\big), \tag{7.124}$$

where s_i are N *seed points* randomly distributed in and around the search domain. For the test function the domain is the unit cube. The parameters f_i are random values at the seed points s_i and are normally distributed. The matrices C_i are randomly generated positive definite matrices at the seed points. The advantage of this test function is that it is smooth and it

likely has N extrema including maxima and minima separated by saddle points. It also has possible optima on the boundary if the seed points are allowed outside the search domain. Because the approximate locations of the maxima and minima are given by the seed points, the extreme function values are also practically known. The function can be evaluated quickly and efficiently and is in general suitable as a CGOP test function because it easily extends to higher dimensions.

7.5.5 EGS performance on the test function

We first investigate the performance of the EGS on a 2-dimensional test problem in order to understand the benefits and deficiencies of the algorithm. As can be seen in Fig. 7.8, the search space is covered nearly uniformly with trial points. The exclusion radius worked as expected to distribute the randomly generated points across the entire search domain. We can see that the sample point concentration is only slightly higher at the local minima. This is due to the fact that the second-order test generates test points inside the local minima deterministically. The prevention of clustering of local searches succeeds in controlling this increased concentration of trial points near the local optima.

The + signs indicate the local minima that are marked in the database. Almost all local minima are correctly identified. Even though most local minima are identified multiple times, clustering algorithms may be used to reduce the number to a reasonably sized set.

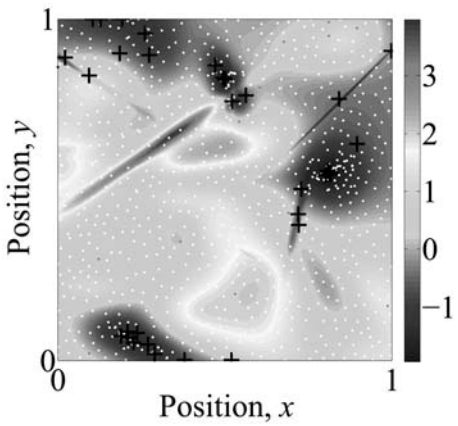

Fig. 7.8. It can be seen that the search space is covered fairly uniformly by the trial points generated during the EGS. Almost all local minima are identified by a + symbol. The concentration of sample points in the local minima is only slightly higher than in the rest of the search domain. The low sampling rate of identified local minima is intended to free computational resources for global exploration.

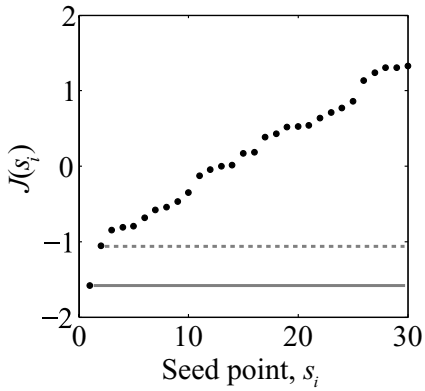

Fig. 7.9. Plot of the test function values $J(s_i)$ at the 30 seed points s_i. The local minima are located very close to the seed points so that $J(s_i)$ is approximately the function value at the local minimum. The horizontal lines are drawn to indicate the gap between the global minimum and the local minimum with second smallest function value. Once the search has found a function value that falls below the upper horizontal line, the global optimum has been identified.

Next we investigate the performance of the EGS on a 5-dimensional test problem. There are a total of 30 seed points s_i placed randomly in the unit hypercube. The randomly distributed function values $J(S_i) = f_i$ at seed points s_i are shown in Fig. 7.9. We observe a large gap between the global minimum and the second lowest local minimum, indicated by the two horizontal lines. Once EGS places a test point with function value below the value of the upper line, the EGS has identified the region of attraction of the global minimum and quickly locates it using a gradient descent method as part of the second-order test.

Figure 7.10 shows the performance of 30 separate instances of EGS optimization for the same test function with function values at the seed points shown in Fig. 7.9. The population size in this case is 20. In all 30 cases the EGS successfully identifies the region of attraction of the global minimum within 10 iterations, i.e. with less than 100 function evaluations. After the neighborhood of the global minimum is identified the local optimization phase of EGS quickly proceeds to find the global minimum and by iteration 30 the EGS has virtually converged to the global minimum in all cases.

The EGS performed extremely well on the given test problem. The performance of EGS for still higher-dimensional problems remains to be investigated.

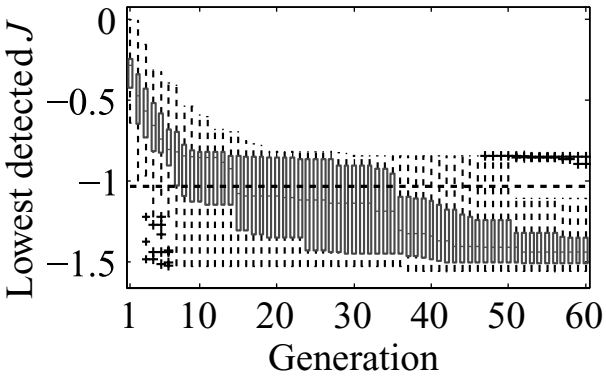

Fig. 7.10. The distribution of 30 instances of EGS minimization procedures is shown. Each generation has 20 members. The horizontal axis shows the index of the iteration. The vertical axis shows the distribution of obtained minimum values for generation i as a box plot. The +s mark outliers and otherwise the upper and lower tails mark the range and quartiles of obtained minima. The box indicates the range from 1st to 3rd quartile. We see that after only 10 iterations all EGS instances place a test point below the second lowest local minimum. This implies that at least one test point is inside the region of attraction of the global minimum.

7.6 Summary

In this chapter we have shown how mathematical system theory may be used to provide a framework for solving the optimal design problem. We emphasized well-posedness of the forward model and convergence of the numerical approximation of the forward model because they are fundamental to ensuring the reliability of the optimal design method. For the class of problems we have presented in this book, the adjoint method is an essential ingredient that enables an effective local search of the design space. The combination of this approach with the ensemble global search algorithm offers a practical way to find globally optimal designs.

7.7 References

1. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice Hall, Englewood Cliffs, New Jersey, 1989.
2. L.I. Schiff, *Quantum Mechanics*, McGraw-Hill, New York, New York, 1968.
3. A.F.J. Levi, *Applied Quantum Mechanics,* Cambridge University Press, Cambridge, United Kingdom, 2006.
4. A.F.J. Levi and I. G. Rosen, *A novel formulation of the adjoint method in the optimal*

design of quantum electronic devices, SIAM Journal of Control and Optimization, submitted, 2009.

5. R.A. Adams, *Sobolev Spaces*, Academic Press, New York, New York, 1975.
6. R.E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, United Kingdom, 1977.
7. H. Tanabe, *Equations of Evolution*, Pitman, London, United Kingdom, 1979.
8. J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, United Kingdom, 1987.
9. T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Germany, 1976.
10. R. Horst and P.M. Pardalos, *Handbook of Global Optimization*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1994.
11. P.M. Pardalos and H.E. Romeijn, *Handbook of Global Optimization*, Volume 2, Kluwer Academic Publishers, Dordrecht, Netherlands, 2002.
12. A. Neumaier, *Complete Search in Continuous Global Optimization and Constraint Satisfaction*, Acta Numerica 2004 (A. Iserles, ed.), Cambridge University Press, Cambridge, United Kingdom, 2004.
13. J.D. Pinterér, *Global Optimization in Action*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1996.
14. C.A. Floudas and P.M. Pardalos, *State of the Art in Global Optimization*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1996.
15. A.H.G.R. Kan and G.T. Timmer, *Stochastic global optimization methods: part 1; clustering methods*, Mathematical Programming **39**, 57–78 (1987).
16. A.H.G.R. Kan and G.T. Timmer, *Stochastic global optimization methods, part II: multi-level methods*, Mathematical Programming **39**, Springer, Berlin, Germany, 1987.
17. Z. Michaelewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd ed., Springer, Berlin, Germany, 1996.

8 Future directions

A.F.J. Levi

8.1 Introduction

The theme of this book is driven by an attempt to exploit system-level complexity that can exist in small devices and quantum systems. The approach adopted involves development of realistic physical models with enough richness in the solutions to allow for discovery of nonintuitive designs. The methodology utilizes a systematic numerical search of solution space to find unexpected behavior. This might include exponential sensitivity, super linearity, polynomial response, or some other desired objective. Obviously, possible future directions of this strategy could be very broad in scope. To narrow the options and help identify a productive path forward it serves to consider some examples that help illustrate the concepts. One way this can be done is by addressing the question of device scaling. That is, what happens when a device is reduced in size.

When electronic and photonic systems are made very small they behave differently. Because we now have access to vast amounts of inexpensive computing power it is possible to find out how these small, but fundamentally complex, systems work, explore differences compared to larger scale systems, and possibly identify opportunities for innovation in the way small systems are designed.

Future directions for research can be guided in part by understanding how small devices and small physical systems differ in their behavior compared to larger systems. One theme that emerges is the increased role fluctuations play in determining the behavior of small complex systems comprising multiple interacting elements. The example of complexity in a small laser diode is discussed in Section 8.2. Another theme is the increased sensitivity scaled devices have to defects and small changes in configuration. Aspects of sensitivity to atomic configuration, reproducibility in manufacturing, and

robustness of design are considered in Section 8.3. Realtime optimization of molecules via high-field chemistry is one approach to atomic control that may eventually become a practical technique and this is discussed in Section 8.4.

A universal challenge for *nanoscience*, and a clear future direction, is to find ways to control and exploit the behavior of these small, yet complex, systems so that they may contribute to *nanotechnology*. As described in Section 8.5, success in this endeavor might lead to a new paradigm in which design and nanoscale manufacture merge into what might best be called *quantum engineering*.

8.2 Example: System complexity in a small laser

As an example of scaling a technologically significant device, consider what happens when a laser diode is made very small. Lasers are critical to modern society, they are a key component that enables the internet, they are used in all optical storage media such as DVDs, and they are the element that makes high-quality printing widely available. The smaller one can make a laser the more uses they can have and the potentially cheaper they can become. Some of the smallest semiconductor lasers include the microdisk device shown in Fig. 8.1(a) [1].

The active elements of a laser are photons and excited electronic states. With increasing drive current around a threshold, photon emission undergoes a nonlinear transition from disordered phase-incoherent thermal light to ordered phase-coherent lasing light. In a conventional edge-emitting semiconductor laser diode with active volume 12 μm^3, the number of excited electronic states n is about 2×10^7 and the number of photons s in the lasing mode under typical operating conditions is around 10^5. In this case continuum mean-field rate equations can be used to describe the average behavior of n and s. The reason for this is that fluctuations about mean electron number $\langle n \rangle$ and mean photon number $\langle s \rangle$ are Poisson distributed and scale as $\sqrt{\langle n \rangle}$ and $\sqrt{\langle s \rangle}$ respectively, and so the fluctuation is only a fraction of a percent of the mean value.

The situation changes when lasers are reduced in size like the device shown in Fig. 8.1(a). The active volume is small, in fact less than 0.12 μm^3, so the number of photons and electrons decreases. For the microdisk laser shown in Fig. 8.1(a) the average number of excited electronic states is about 2×10^5 and the average number of photons in the lasing mode is around 1,000. Now fluctuations about the mean value and the fact that photons and electrons are quantized play an increasingly important role in system performance.

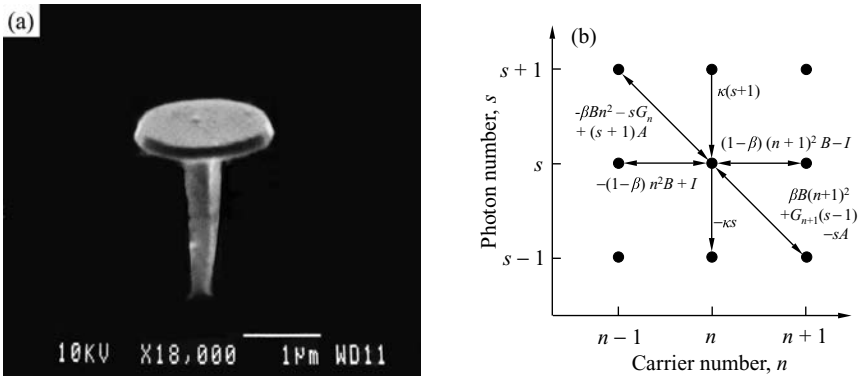

Fig. 8.1. (a) A microdisk laser of radius 0.8 μm [1]. The active volume of the device is less than 0.12 μm^3, the fraction of spontaneous emission feeding the whispering gallery optical mode is $\beta = 0.1$, when optically pumped using $\lambda = 0.98$ μm wavelength radiation the lasing threshold is less than 1 mW and lasing wavelength is near $\lambda_0 = 1.5$ μm. (b) The probability $P_{n,s}$ of a state having n electrons and s photons is determined by the indicated rate of transitions in and out of the state.

The question we seek to answer is what happens if the laser is made even smaller than the microdisk shown in Fig. 8.1(a). Simulation through computation is by far the most efficient way to systematically explore this question.

A master equation model can be used that captures quantization effects and, at the same time, provides information on the statistics of coupled photon and electron particles in the system. These equations are a set of differential equations in continuous functions (probabilities) that can be used to describe the dynamics of the discrete particle system. Figure 8.1(b) shows that the probability $P_{n,s}$ of the system containing n excited electronic states and s photons in the lasing mode can be calculated if all possible transition rates in and out of the state are known [2]. The master equation to be solved is

$$\begin{aligned}\frac{dP_{n,s}}{dt} = &-\kappa(sP_{n,s} - (s+1)P_{n,s+1}) - (sG_nP_{n,s} - (s-1)G_{n+1}P_{n+1,s-1}) \\ &- (sAP_{n,s} - (s+1)AP_{n-1,s+1}) - \beta B(n^2P_{n,s} - (n+1)^2P_{n+1,s-1}) \\ &- (1-\beta)B(n^2P_{n,s} - (n+1)^2P_{n+1,s}) - \frac{I}{e}(P_{n,s} - P_{n-1,s}),\end{aligned} \tag{8.1}$$

where β is the fraction of spontaneous emission feeding into the lasing mode, B is the radiative recombination rate, κ is the optical cavity decay rate, G_n is the stimulated emission coefficient for a system of n excited electronic states, A is stimulated absorption, and I is the pump current.

In both the steady-state and transient response case, there are, in principle, a very large number of coupled equations that must be solved. Even

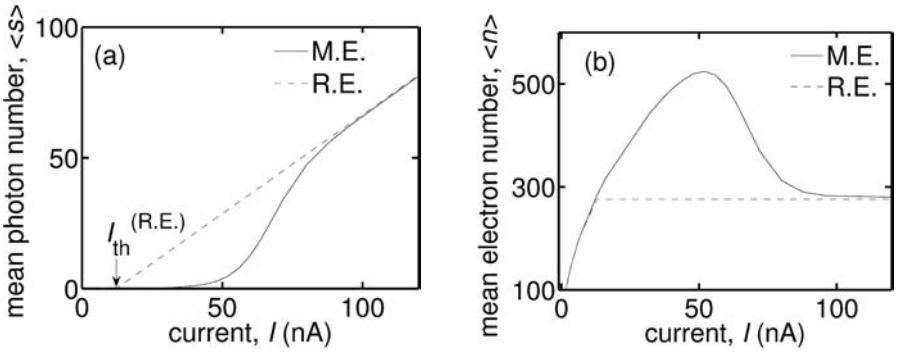

Fig. 8.2. (a) Calculated mean photon number $\langle s \rangle$ as a function of current showing that master equations (M.E.) predict suppression of lasing threshold relative to continuum mean-field rate equation (R.E.) calculations for a very small semiconductor laser with active volume 10^{-4} μm^3. Suppression in lasing is due to quantum fluctuations in particle number n and s. (b) Calculated mean electron number $\langle n \rangle$ as a function of current. For currents above I_{th} the continuum mean-field R.E. predicts carrier pinning while the M.E. shows average carrier number depinning due to quantum fluctuations. Because spontaneous emission is proportional to n^2, the M.E. predicts an enhancement in spontaneous emission around threshold [3].

with appropriate truncation, the system of equations is large, requiring significant compute power to solve and it is only today that such resources are readily available, making it a viable way to explore the consequences of scaling these systems. What makes the effort particularly worthwhile is the fact that these equations are very general and apply to many types of nonlinear systems so that knowledge gained from such a study may very well find application in a broad range of subjects.

For the system we are considering, fluctuations in the value of n and s are correlated such that $\langle ns \rangle$ may not be factorized, i.e., $\langle ns \rangle \neq \langle n \rangle \langle s \rangle$, and normal statistics do not apply. A consequence of this, and the fact that a lowest energy state of the system exists, is that particle number fluctuations suppress lasing and enhance spontaneous emission around the threshold [3]. As illustrated in Fig. 8.2, this remarkable observation is contrary to the predictions of continuum mean-field rate equations in which particle number is not discretized. It is also contrary to the expectations of conventional Landau–Ginzburg theory of phase transitions in which fluctuations enhance lasing below threshold [4].

More insight is gained into the behavior of discrete particle systems by calculating behavior in the time domain. Computationally, the use of Monte Carlo techniques reduces the burden on memory at the expense of increased processor usage making it particularly attractive for large parallel computing machines. Results shown in Fig. 8.3 for the same device parameters used

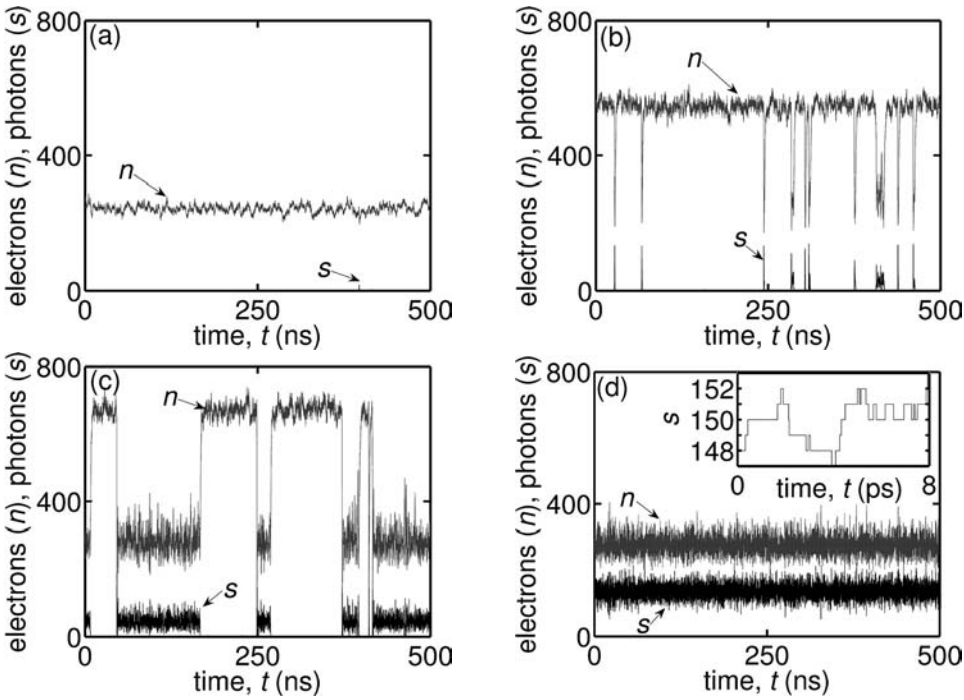

Fig. 8.3. Time evolution of n discrete electrons and s discrete photons calculated by the Monte Carlo method for the laser system of Fig. 8.2. (a) Current, $I = 9.6$ nA with no lasing, (b) $I = 48$ nA showing very short bursts of lasing light, (c) with $I = 72$ nA quantum fluctuations cause the system to switch between lasing and non-lasing states, and (d) when $I = 192$ nA lasing is sustained. The inset illustrates discrete quantum step changes in photon number with time.

in Fig. 8.2 indicate that the system fails to lase continuously and there is switching between two characteristic system states in the region dominated by strong fluctuations.

Because the average electron number in the cavity increases with increasing current, more spontaneous photon emission events become possible. Spontaneous emission in a cavity with no photons initiates the first stimulated processes. So a large number of such events enables the system to lase continuously. A lesser likelihood of such events in a small active-volume laser causes suppression of continuous lasing. If the system loses all its photons, it has to wait for the next spontaneous photon emission event. A direct consequence is non-Poisson carrier and photon statistics in the device.

This example of a small discrete particle system comprising multiple interacting elements shows how one can learn and gain new insights about how such systems work. One is able to use computation to gain a deeper

understanding of a physically small, yet complex, system. From an engineering perspective this presents an opportunity for new system concepts that make use of unusual, and sometimes unexpected, behavior. A future challenge is to find ways to control and exploit the inherent complexity of the system. The potential for unanticipated applications resides in the complexity of the system.

8.3 Sensitivity to atomic configuration

Sometimes solutions that appear obvious for large systems fail to work when the device is scaled to the nano-regime. For example, consider the case of electron transport in a bulk semiconductor containing n randomly distributed substitutional ionized impurities per unit volume. For doping levels of practical interest one expects ionized impurity scattering to play a dominant role in determining carrier mobility.

As a first step to understanding ionized impurity scattering one can write down the potential seen by an electron. For an electron at position $\mathbf{r}$ this is

$$V(\mathbf{r}) = \sum_{j=1}^{n} v(\mathbf{r} - \mathbf{R}_j), \tag{8.2}$$

where $\mathbf{R}_j$ is the location of the ions and

$$v(\mathbf{r}) = \int \frac{d^3q}{(2\pi)^3} \frac{e^2}{\epsilon(\mathbf{q})q^2} e^{i\mathbf{q}\cdot(\mathbf{r})}, \tag{8.3}$$

is the Coulomb potential of an individual ion statically screened by the dielectric function $\epsilon(\mathbf{q})$, and $\mathbf{q}$ is the scattered wave vector.

For elastic scattering one typically considers transitions between a state $\psi_{\mathbf{k}} = Ae^{i(\mathbf{k}\cdot\mathbf{r})} = |\mathbf{k}\rangle$ of energy $E(\mathbf{k})$ and a final state $|\mathbf{k}'\rangle$ with the same energy. Assuming the mean free path l_k between scattering events is such that $kl_k \gg 1$, the elastic scattering rate $1/\tau_{\text{el}}$ may be calculated using Fermi's golden rule. This involves evaluating the matrix element $\langle\mathbf{k}'|v(\mathbf{r})|\mathbf{k}\rangle$. Since $|\mathbf{k}\rangle$ and $\langle\mathbf{k}'|$ are assumed to be plane-wave states of the form $e^{-i\mathbf{k}\cdot\mathbf{r}}$, the matrix element is

$$\begin{aligned}\langle\mathbf{k}'|v(\mathbf{r})|\mathbf{k}\rangle &= \int d^3r e^{i\mathbf{k}'\cdot\mathbf{r}} v(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} = \int d^3r v(\mathbf{r}) e^{-i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} \\ &= \int d^3r v(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} = v(\mathbf{q}),\end{aligned} \tag{8.4}$$

where $v(\mathbf{q})$ is just the Fourier transform of the Coulomb potential in real space. In this expression $\mathbf{q} = \mathbf{k} - \mathbf{k}'$, since momentum conservation requires

$\mathbf{k} = \mathbf{k}' + \mathbf{q}$. The scattering angle θ is such that $k \sin(\theta/2) = q/2$.

Using the first term in the Born series (Fermi's golden rule) to calculate the elastic scattering rate in a bulk doped n-type semiconductor gives

$$\frac{1}{\tau_{\text{el}}} = \frac{2\pi}{\hbar} \int \frac{d^3q}{(2\pi)^3} |V(\mathbf{q})|^2 \delta(E(\mathbf{k}) - E(\mathbf{k} - \mathbf{q})), \tag{8.5}$$

where the delta function ensures that no energy is exchanged and $V(\mathbf{q})$, the Fourier transform of $V(\mathbf{r})$, is

$$V(\mathbf{q}) = \sum_{j=1}^{n} \int d^3r \;\; v(\mathbf{r} - \mathbf{R}_j) e^{-i\mathbf{q}\cdot(\mathbf{r}-\mathbf{R}_j)} e^{i\mathbf{q}\cdot\mathbf{R}_j}. \tag{8.6}$$

The expression for $|V(\mathbf{q})|^2$ in Eq. (8.5) may be written

$$|V(\mathbf{q})|^2 = |v(\mathbf{q})|^2 s(\mathbf{q}), \tag{8.7}$$

where $v(\mathbf{q})$ is the Fourier transform of Eq. (8.3) and

$$s(\mathbf{q}) = \sum_{j=1}^{n} e^{-i\mathbf{q}\cdot\mathbf{R}_j} \sum_{k=1}^{n} e^{i\mathbf{q}\cdot\mathbf{R}_k} = \sum_{j=1}^{n} 1 + \sum_{j\neq k}^{n} e^{-i\mathbf{q}\cdot(\mathbf{R}_j - \mathbf{R}_k)}, \tag{8.8}$$

is a structure factor that contains phase information on the scattered electron wave from sites $\mathbf{R}_j$. The second term on the right-hand side is a pair correlation that can be written in terms of sine and cosine functions to give

$$s(\mathbf{q}) = n + \sum_{j\neq k}^{n} (\cos(\mathbf{q} \cdot (\mathbf{R}_j - \mathbf{R}_k)) + i \sin(\mathbf{q} \cdot (\mathbf{R}_j - \mathbf{R}_k))). \tag{8.9}$$

For n large and random $\mathbf{R}_j$, the sum in Eq. (8.9) is zero. It follows that if there are n spatially uncorrelated scattering sites corresponding to random impurity positions, we expect $s(\mathbf{q}) = n$ [5] and the magnitude of the matrix element squared in Eq. (8.7) becomes

$$|V(\mathbf{q})|^2 = n|v(\mathbf{q})|^2. \tag{8.10}$$

Hence, the total elastic scattering rate from n impurities per unit volume is

$$\frac{1}{\tau_{\text{el}}} = \frac{2\pi}{\hbar} \;\; n \int \frac{d^3q}{(2\pi)^3} \left| \frac{e^2}{\varepsilon(\mathbf{q})q^2} \right|^2 \delta(E(\mathbf{k}) - E(\mathbf{k} - \mathbf{q})), \tag{8.11}$$

where the integral over d^3q is the final density of plane-wave electron states.

Physically, each impurity is viewed as contributing independently, so that the scattering rate is n times the scattering rate from a single impurity atom. This approach to explaining the contribution of elastic ionized impurity scattering to the measured mobility of bulk semiconductors has been remarkably successful [6].

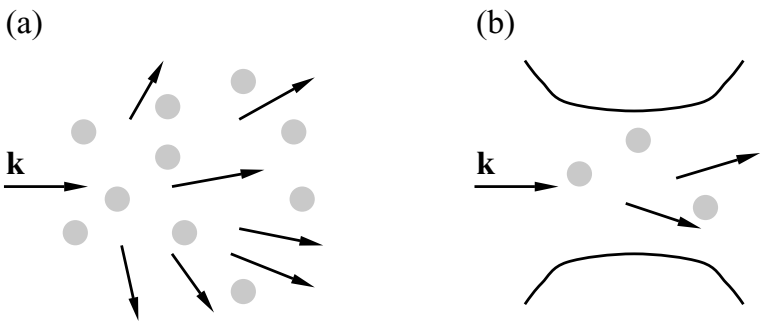

Fig. 8.4. (a) Showing that an electron of wave vector **k** can explore many different paths when elastically scattering from ionized impurities in a bulk semiconductor. (b) In the presence of a small number of impurities and a constriction such as occurs in a semiconductor quantum wire, the electron has fewer paths through the system and is more sensitive to small changes in scattering potential.

Thus far we have assumed that each substitutional dopant atom occupies a random crystal lattice site. This, however, is not strictly correct because impurity sites cannot be chosen randomly. The constraint that substitutional impurity atoms in a crystal occupy crystal lattice sites gives rise to a correlation effect because double occupancy of a site is not allowed. Suppose a fraction f of sites is occupied. In this case, we no longer have a truly random distribution, and, for small f, the scattering rate will be reduced by

$$s(\mathbf{q}) = n(1 - f). \tag{8.12}$$

The term $(1 - f)$ reflects the fact that not allowing double occupancy of a site is a correlation effect [7].

Other spatial correlation effects are possible and can, in principle, dramatically alter scattering rates. A periodic array of impurity sites can be arranged such that elastic ionized impurity scattering becomes insignificant. This can be true even in the presence of some disorder [7]. The reason why such a strategy works so well is that one performs an ensemble average over many scattering sites and Coulomb scattering favors transitions involving small values of $\mathbf{q}$.

Figure 8.4 illustrates another way of looking at the effect. Figure 8.4(a) shows schematically that an electron with initial wave vector $\mathbf{k}$ in a bulk doped semiconductor can explore many different paths by elastically scattering. The ensemble average over many scattering sites causes the sum in Eq. (8.9) to asymptotically approach zero for random ion positions. As a consequence the overall response, in this case the scattering rate $1/\tau_{\text{in}}$, is robust with respect to small changes in local impurity potential. However, when the number of impurity sites contributing to the potential seen by an

electron is small the sum in Eq. (8.9) is almost never zero and so contributes to $s(\mathbf{q})$. Hence, $1/\tau_{\text{in}}$ becomes sensitive to even small changes in position of the ionized impurities. Figure 8.4(b) shows schematically that an electron with initial wave vector $\mathbf{k}$ in the presence of a few scattering sites near a constriction can explore only a limited number of different paths and so is sensitive to small changes in local impurity potential.

8.3.1 Reproducibility in manufacturing

There is no doubt that reproducibility in manufacturing is one of the outstanding challenges facing any future transition of nanoscience to nanotechnology. Ideally, one would like to be able to guarantee that each nanostructure produced is exactly the same. The precision required might be at the atomic level. In nature, this is readily achieved at the molecular level. The structures of small molecules in free-space are identical. However, larger, more complex molecules might have a number of different configurations.

Manmade structures such as semiconductor quantum dots, quantum wires, and quantum wells can only be fabricated to an accuracy of a few atomic spacings. The sensitivity to small variations in device dimensions depends on material parameters and characteristic device size, L.

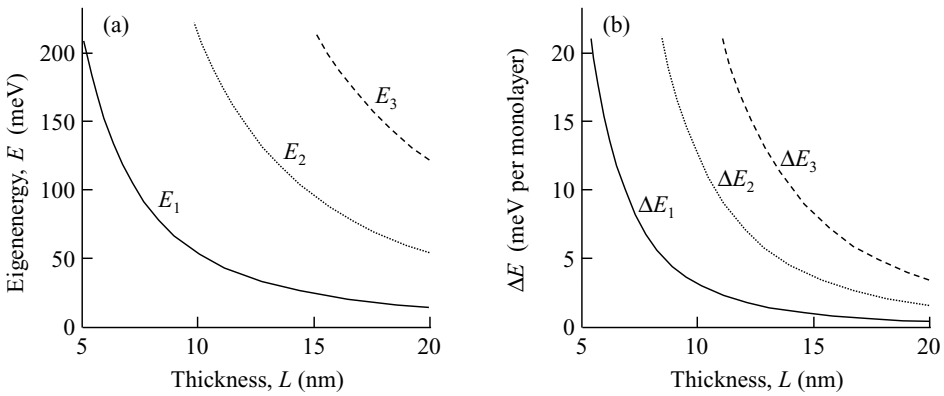

Fig. 8.5. (a) Eigenenergy for the first three eigenstates of an electron in the conduction band of GaAs with effective mass $0.07 \times \text{m}_0$ confined in one-dimension by an infinite potential well of thickness, L. (b) Change in eigenenergy for one monolayer change in thickness (0.2825 nm) as a function of L.

By way of example, consider an electron in the conduction band of GaAs confined in one dimension by an infinite potential well of thickness, L. As is well known and illustrated in Fig. 8.5(a), electron eigenenergy scales as

$E \propto 1/L^2$. The lattice constant of GaAs is approximately 0.565 nm so the thickness of one monolayer in the (100) direction is 0.2825 nm. Figure 8.5(b) shows the change in eigenenergy, ΔE, for one monolayer change in thickness (0.2825 nm) as a function of L. The eigenenergies of the first three eigenstates have increasing sensitivity to even one monolayer change in characteristic size, L. This fact becomes especially important if the device requires response from many identical nanostructures. The resulting ensemble average broadens response in the composite system and can limit functionality. As illustrated in Fig. 8.5(b), if it is possible to control the value of L to one monolayer, then the eigenenergy, and hence the response of the device, is inhomogeneously broadened by ΔE.

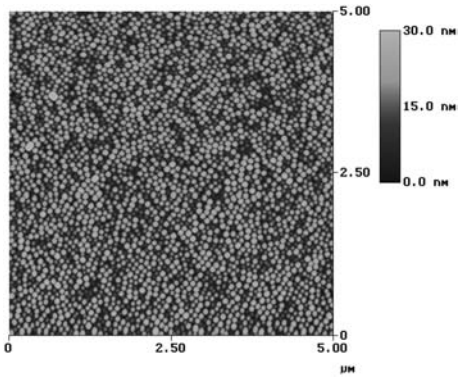

Fig. 8.6. Area view of InP self-assembled quantum dots grown using low-pressure MOCVD on an InAlP matrix layer lattice-matched to a GaAs substrate. As measured from Atomic Force Microscope (AFM) images, areal density of quantum dots is 1.5×10^{10} cm^{-2} and dominant size is in the range of 15–20 nm for a 15-monolayer planar-growth-equivalent deposition time at a growth temperature of 650 °C. Dominant sizes are controllable by changing the deposition time. Image courtesy of R. Dupuis.

Even the best experiments [8, 9] using self-assembly techniques to grow semiconductor layers containing a high density of quantum dots have resulted in structures with inhomogeneously broadened response due to quantum dot size variations. Figure 8.6 is an area view of InP self-assembled quantum dots grown using low-pressure MOCVD on an InAlP matrix layer lattice-matched to a GaAs substrate. The areal density of quantum dots is 1.5×10^{10} cm^{-2} and dominant size is in the range of 15–20 nm for a 15-monolayer planar-growth-equivalent deposition time at a growth temperature of 650 °C. The high density of quantum dots makes this particular structure suitable for inclusion in the optically active region of a quantum dot laser or infrared focal plane array photodetector.

For example, vertical cavity surface emitting laser diodes that make use of the properties of high-density quantum dots have demonstrated optical

modulation bandwidths in excess of 35 GHz [10]. In another example, by carefully engineering the electron wave functions of quantum dots in quantum wells, infrared focal plane arrays with voltage bias wavelength tunability and multi-color operation have been successfully demonstrated [11].

Clearly, the application and nanoscale control of quantum dots has significant challenges that are directly related to reproducibility and manufacturing. The opportunity for enhanced device performance and function is increased if the size distribution of quantum dots, each of which contain a few tens of thousands of atoms, can be controlled with greater precision.

At another extreme, individual atoms can be placed on a crystal surface with great accuracy using a Scanning Tunneling Microscope (STM). For example, gold atoms have been arranged on an AlNi surface and a large number of interesting physics experiments performed that study the behavior of various atomic configurations [12–15]. However, so far, no plausible way to manufacture the structures while maintaining the atomic precision has been identified.

8.3.2 Robustness

What this all means is that design at the nanoscale must either rely on atomically precise manufacturing techniques or be inherently robust against small variations in configuration. The latter requires consideration of robust optimization techniques similar to those discussed in Chapter 6 or, at a minimum, sensitivity analysis of designs.

Robustness in quantum systems can be fundamentally different from classical systems. In electronics, the wave nature of the electron, the existence of superposition states, and many-body phenomena can lead to increased sensitivities to design parameters compared to the corresponding classical system. Some aspects of this, including a discussion of the classical-quantum boundary in system response, has been explored in Chapter 5.

Clearly, robustness of device design is a major future direction for any strategy that employs optimization.

It is worth noting that sometimes quantum systems can have a response that is *naturally robust*. For example, the use of broad resonant electron states in the multi-barrier semiconductor diode structures of Section 1.3.2 resulted in robustness to small manufacturing defects. This occurred as a consequence of constructing device response from low-lying eigenstates of the quantum system. Another example is quantized conductance that occurs in a one-dimensional channel. For a given variation in device configuration over an otherwise perfect one-dimensional conductance region and over a specific energy range, resistance to electron transport per independent channel per

spin is almost exactly $2\pi\hbar/e^2 = 25,812.8075\ \Omega$. Understanding *natural robustness* is both a challenge and an opportunity for optimal device design.

8.4 Realtime optimal design of molecules

As illustrated in previous sections of this chapter, when electronic and photonic systems are made very small they behave differently. Electronic nanostructure devices become sensitive to the exact position of defects and the performance of scaled photonic devices becomes dominated by fluctuations in particle number. These facts are indicative of the challenges facing those who wish to find technological applications of nanoscience. Nanotechnology will require practical solutions to these and other related issues.

For example, either nanoelectronic devices will be fabricated to atomic precision or device designs and system architectures will be adopted that are robust against large variations in individual component performance. Increased sensitivity to the exact position of impurities or defects in nanoscale structures appears to be limiting applications of nanoscience. One approach to fabricating device elements with precise atomic configuration involves the synthesis of molecules.

It is possible to apply closed-loop control to strong-field photo-chemistry [16–18]. As illustrated in Fig. 8.7, laser pulse sequences can be generated to excite specific molecular states and thereby produce particular molecular products. For example, a large molecule might dissociate into a number of specific fragments. The laser pulse sequences control the exact number and type of fragments.

Because optical pulse generation can produce many different very intense electric field pulse shapes, chemical reaction can be driven by high-field chemistry. A search algorithm may be used to find the optimal pulse shape and sequence that most efficiently creates the desired molecular species. This is a realtime version of optimal design in which the system acts as an analog computer to solve Schrödinger's equation and discover a pulse sequence or chemical reaction pathway to obtain the objective molecule or molecules. Since the intense high-field optical pulse shape defines the interaction Hamiltonian of the system, one is, in effect, searching for the Hamiltonian that best obtains the objective.

Similar ideas may be applied to chemical sensing in which interfering signals are suppressed relative to the signal sought. In this case, the interaction Hamiltonian of the high-field chemistry performed is modified to maximize the sensing signal relative to interfering (noise) signals [19]. This is analogous to the idea of changing the physical model to discover an optimal

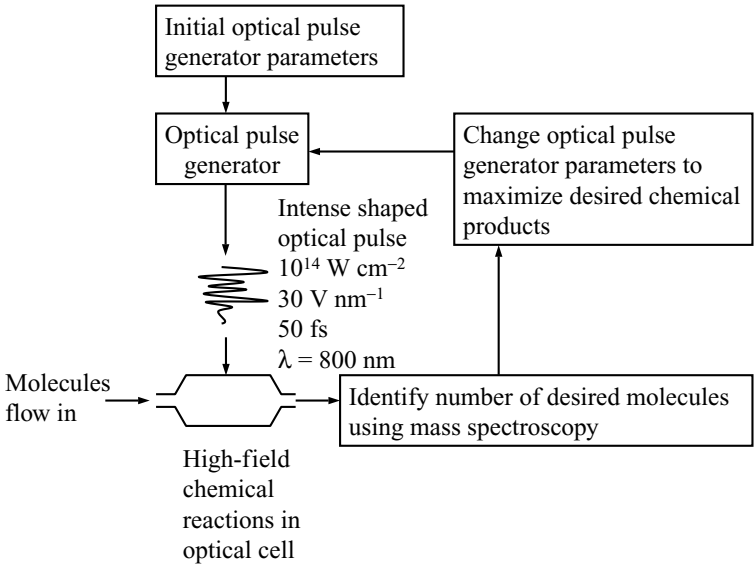

Fig. 8.7. System designed to find chemical reaction pathways using adaptive high-field chemistry. An optical pulse generator is capable of creating intense shaped optical pulses on very short timescales. Intensity can be as high as 10^{14} W cm^{-2}, peak electric fields are in the 30 V nm^{-1} range, and pulse duration is about 50 fs centered at $\lambda = 800$ nm.

device design [20] that was mentioned at the end of Chapter 1.

8.5 The path to quantum engineering

Availability of compute resources, improved realistic physical models, and the use of optimal design point to a future in which quantum engineering emerges as a mainstream activity in technology development [20]. However, the best way to incorporate optimal design of small quantum systems into a viable technology is unknown. In one vision, the machinery needed for manufacture and component testing is located in a factory. This might be called the *nanofactory*. Of course, nanoscale optimal device design does not have to be performed at the nanofactory, it could take place at a remote location.

Figure 8.8 illustrates the concept of an integrated approach to design and manufacture at the nanoscale. The figure shows information flow starting with device and system needs being specified by engineers. These specifications are translated into physical constraints followed by search for optimal configuration. Because it is likely that many nano-devices will have unusual functionality, it is essential that they be modeled correctly using physically

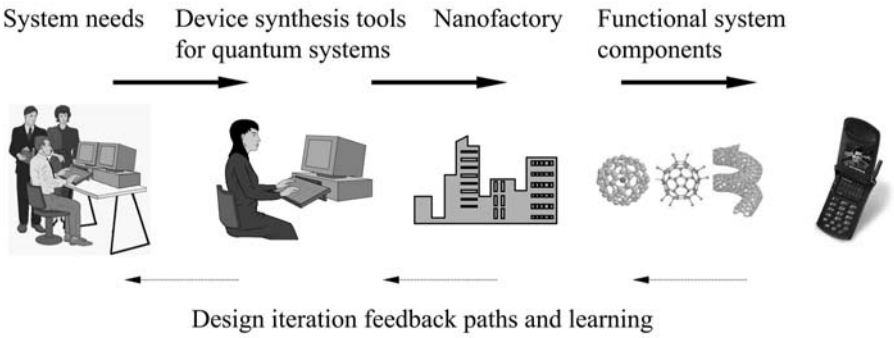

Fig. 8.8. The nanofactory concept in which system needs are specified by engineers. These specifications are translated into physical constraints followed by search for optimal configuration. Optimized nonintuitive design fabricated in nanofactory and shipped for use in systems.

realistic models and appropriate choice of physical variables. This important point has been emphasized by Chua [21].

Optimized, and often nonintuitive, device designs with nanoscale and possibly atomic-scale tolerance are then fabricated in a centralized nanofactory. Tested functional system components and subsystems are subsequently delivered to the system user. In addition to device manufacture, the nanofactory accumulates knowledge and expertise, including process development, that can be fed back in the form of constraints in the design process. In a foundry model of operation, the nanofactory could manufacture components submitted by a large number of different design groups each offering their own variant of specialized system function made to order.

8.6 Summary

The transition of nanoscience to nanotechnology could benefit from the use of optimal device design techniques that are the subject of this book. However, the challenges seem to be significant at every level.

Physical models capable of capturing the richness and complexity of nanoscale device response are a prerequisite. Often it will be necessary to describe the transition from classical to quantum behavior and, of course, making these models computationally efficient is essential. Methods to explore nonconvex solution space for globally optimal device designs is another major challenge. A further difficult requirement is that often these designs will have to guarantee robustness. If it were possible to manufacture nanoscale devices with atomic precision, some aspects of optimal design would become

easier to implement.

Despite these challenges, there are important positive aspects to pursuing an optimal device design strategy. For example, the merging of discovery and design in small complex quantum systems could be a more efficient way to drive technology forward. The inevitable technological surprises, with potentially enormous value to the owner of the technology, might also be significant.

8.7 References

1. A.F.J. Levi, S.L. McCall, S.J. Pearton, and R.A. Logan, *Room temperature operation of submicrometre radius disc laser*, Electronics Letters **29**, 1666–1667 (1993).
2. P.R. Rice and H.J. Carmichael, *Photon statistics of a cavity-QED laser: A comment on the laser-phase-transition analogy*, Physical Review A **50**, 4318–4328 (1994).
3. K. Roy-Choudhury, S. Haas, and A.F.J. Levi, *Quantum fluctuations in small lasers*, Physical Review Letters **102**, 053902 1–4 (2009).
4. J. O'Gorman, A.F.J. Levi, S. Schmitt-Rink, *et al.*, *On the temperature sensitivity of semiconductor lasers*, Applied Physics Letters **60**, 157–159 (1992).
5. $s(\mathbf{q}) = n$ is the standard result when calculating elastic ionized impurity scattering rates from a large number of random impurity sites. This situation was first analyzed by W. Kohn and J.M. Luttinger, *Quantum theory of electrical transport phenomena*, Physical Review **108**, 590–611 (1957).
6. P. Chattopadyay and H.J. Queisser, *Electron scattering by ionized impurities in semiconductors*, Reviews of Modern Physics **53**, 745–768 (1981).
7. A.F.J. Levi, S.L. McCall, and P.M. Platzman, *Nonrandom doping and elastic scattering of carriers in semiconductors*, Applied Physics Letters **54**, 940–942 (1989).
8. J.H. Ryou, R.D. Dupuis, D.T. Mathes, *et al.*, *High-density InP self-assembled quantum dots embedded in $In_{0.5}Al_{0.5}P$ grown by metalorganic chemical vapor deposition*, Applied Physics Letters **78**, 3526–3528 (2001).
9. J.H. Ryou, R.D. Dupuis, G. Walter, *et al.*, *Photopumped red-emitting $InP/In_{0.5}Al_{0.3}Ga_{0.2}P$ self-assembled quantum dot heterostructure lasers grown by metalorganic chemical vapor deposition*, Applied Physics Letters **78**, 4091–4093 (2001).
10. N.N. Ledentsov, F. Hopfer, and D. Bimberg, *High-speed quantum-dot vertical-cavity surface-emitting lasers*, Proceedings of the IEEE **95**, 1741–1756 (2007).
11. S. Krishna, S.D. Gunapala, S.V. Bandara, C. Hill, and D.Z. Ting, *Quantum dot based infrared focal plane arrays*, Proceedings of the IEEE **95**, 838–1852 (2007).
12. T.M. Wallis, N. Nilius, and W. Ho, *Electronic density oscillations in gold atomic chains assembled atom by atom*, Physical Review Letters **89**, 236802 1–4 (2002).
13. N. Nilius, T.M. Wallis, and W. Ho, *Building alloys from single atoms: Au-Pd chains*

on NiAl(110), Journal of Physical Chemistry B **108**, 14616–14619 (2004).

14. S.W. Wu, G.V. Nazin, X. Chen, X.H. Qiu, and W. Ho, *Control of relative tunneling rates in single molecule bipolar electron transport*, Physical Review Letters **93**, 236802 1–4 (2004).

15. N. Nilius, T.M. Wallis, and W. Ho, *Tailoring electronic properties of atomic chains assembled by STM*, Applied Physics A **80**, 951–956 (2005).

16. R.S. Judson and H. Rabitz, *Teaching lasers to control molecules*, Physical Review Letters **68**, 1500–1503 (1992).

17. R.J. Levis and H. Rabitz, *Closing the loop on bond selective chemistry using tailored strong field laser pulses*, Journal of Physical Chemistry A **106**, 6427–6444 (2002).

18. M. Wollenhaupt, A. Präkelt, C. Sarpe-Tudoran, *et al.*, *Femtosecond strong-field quantum control with sinusoidally phase-modulated pulses*, Physical Review A **73**, 063409 1–15 (2006).

19. D.A. Romanov, D.M. Healy, J.J. Brady, and R.J. Levis, *Adaptive reshaping of objects in (multiparameter) Hilbert space for enhanced detection and classification: an application of receiver operating curve statistics to laser-based mass spectroscopy*, Journal of the Optical Society of America A **25**, 1039–1050 (2008).

20. A.F.J. Levi, *Towards quantum engineering*, Proceedings of the IEEE **96**, 335–342 (2008).

21. L.O. Chua, *Nonlinear circuit foundations for nanodevices, Part I: The four-element torus*, Proceedings of the IEEE **91**, 1830–1859 (2003).

Appendix A
Global optimization algorithms

A.1 Introduction

There are many approaches to global optimization (for a review, see [1]). The popular genetic algorithm is discussed in Chapter 1 and a novel ensemble stochastic method is described in Chapter 7. This appendix summarizes some other global search algorithms.

A.2 Tabu search

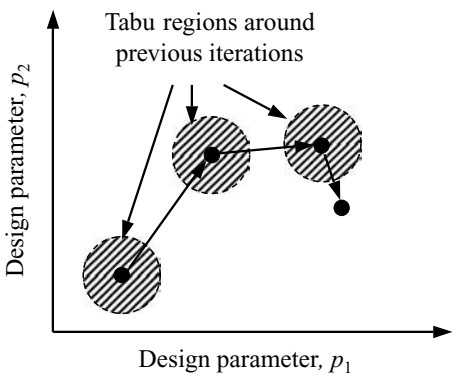

Fig. A.1. Tabu search is an iterative algorithm that maintains a short-term memory of previous iterations. Trial points in short-term memory are not revisited during the next iteration of the optimization search. For continuous search space an exclusion region around trial points can be used.

Tabu search is a method for global optimization that was originally proposed by Glover for combinatorial optimization problems [2, 3]. The two main concepts of a tabu search are *adaptive memory* and *responsive*

exploration. Tabu search uses adaptive memory in contrast to memoryless algorithms like GA and simulated annealing (described in Section A.4), which have no explicit memory of previously evaluated design parameter settings. Rigid memory algorithms like branch and bound (not discussed here) retain all information about prior function evaluations and treat all points in memory equally. In contrast to rigid memory, adaptive memory actively manages the information it saves in memory. Usually only the most recent function evaluations are retained.

The memory serves two purposes. First, and as illustrated in Fig. A.1, already explored regions are not revisited. Second, previously unexplored areas that should be explored are determined in response to already good solutions as well as regions that were determined as less promising. Responsive exploration can be implemented in a stochastic or a deterministic manner but it uses the best as well as the worst memories to dynamically explore the most promising regions where the greatest improvement of the objective function is expected or the maximum information can be gained.

Not all combinatorial aspects of tabu search translate directly to continuous global optimization problems but, in conjunction with for instance clustering methods, it is reasonable to avoid generating candidates near solutions that are known to perform significantly worse than the best solutions found up to that point. For continuous design parameters the search domain can be discretized and combinatorial methods applied directly, or a tabu region around a trial point is used instead of excluding a single point on a discrete map. The tabu search algorithm is a heuristic search framework and, depending on the particular implementation, it is at best asymptotically complete. It is designed to prevent entrapment in a local minimum and oscillation between multiple solutions.

A.3 Particle swarm algorithm

An asymptotically incomplete but successfully implemented search algorithm is the particle swarm search. The particle swarm algorithm resembles particles in a hyperspace and is a strictly heuristic global optimization method. A particle represents a parameter setting and the path of a particle represents a sequence of test points in the parameter space. Each particle i has a location x_i and a velocity vector v_i. The particle's initial location might be completely random but the evolution in space of a particle is determined by its velocity vector. Each particle moves in the direction of its velocity. The manner in which the velocity vector of each particle is updated reflects the swarm characteristic of the algorithm. Initially a random

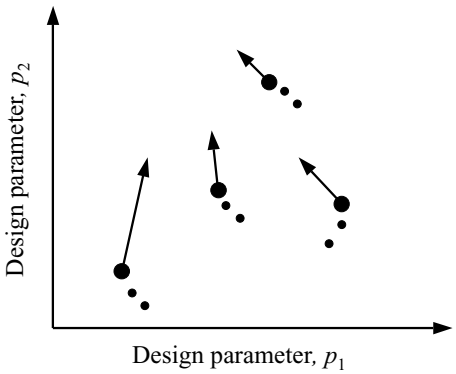

Fig. A.2. Particle swarm search is an algorithm designed for cluster computation. A family of trial points creates new trial points in a coordinated manner, with all samples approximately following the best performing trial point.

set of particles with random velocities are generated on Ω. The objective function is evaluated at each point and the best test point with lowest objective function value is determined. In the next iteration both the location as well as the velocity must be updated. The position is simply updated in the direction of the velocity $x_{i+1} = x_i + \lambda_i v_i$ with some scaled step-length λ_i. Let $\hat{g}$ be the best parameter setting found up to iteration i and $\hat{x}_j$ be the best parameter setting encountered by particle j at iteration i. To add a random component let r_1 and r_2 be randomly generated numbers, usually distributed uniformly on $[0, 1]$. The velocity update is then some variation of $v_{i+1} = \omega v_i + c_1 r_1(\hat{x}_i - x_i) + c_2 r_2(\hat{g} - x_i)$. The velocity is influenced by the velocity in the previous iteration, and moves in general towards the minimum value encountered by the particle itself as well as the minimum value found up to that point by any particle. Let $\hat{x}_i$ be the test point in sequence i that has the lowest cost function value $J(\hat{x}_i)$. Note that $\hat{g}$ and $\hat{x}_i$ must be updated after the objective function has been evaluated at iteration $i + 1$ and the next iteration is constructed. Because gradient information is available in many cases, including the gradient information into the velocity vector makes sense. There are suggested choices of ω, c_1, and c_2 which can be fixed as well as adaptive. The particle swarm algorithm has been analyzed using dynamical system techniques as its iteration scheme resembles the discretized solution of a dynamical system in time.

The intuition of the particle swarm method is that the sample points move through the sample space in a coordinated manner following the best sample point of the collection of points, see Fig. A.2. In this manner a certain minimum distance between sample points is maintained while allowing some degree of individual movement.

A.4 Simulated annealing

A popular stochastic search algorithm is simulated annealing (SA). The algorithm was originally motivated by the process of annealing in material science, a technique that is used to allow a system with many degrees of freedom (e.g. a liquid) to reach a low-energy state (e.g. a crystal). The theory of statistical mechanics is used for analysis of the process [4]. Unlike GA, SA is not per se an evolutionary algorithm because it does not rely on a population of trial points that evolves from generation to generation. Rather it is aptly described as a filtered random walk [5].

The SA algorithm computes the cost function $J(x_i)$ at a sequence of stochastic trial points $\{x_i\}$. At each point x_i in the sequence a new random trial point x_{trial} is generated and $J(x_{\text{trial}})$ is evaluated. Under certain conditions x_{trial} is accepted as the next point in the search sequence x_{i+1} or it is discarded and a different x_{trial} is generated. The decision on whether to accept a new trial point in the sequence of points is based on an *acceptance function*. We begin the detailed discussion of SA with the acceptance function. The acceptance function depends on the comparison $J(x_{\text{trial}})$ with $J(x_i)$ as well as an abstract temperature T. The temperature T controls the rate at which trial points are accepted. The temperature changes with the number of iterations that have been performed in the optimization algorithm and this behavior is called the *cooling schedule*. The cooling schedule is directly related to the efficiency of the algorithm.

The decision of accepting a candidate parameter setting x_{trial} as the next point in the search sequence depends on a certain probability given by the *acceptance function*. The standard acceptance function is based on the original Metropolis *et al.* algorithm [6] and is given by

$$p_{\text{accept}}(x_i, x_{\text{trial}}, T) = \min\left(1, \exp\left(\frac{J(x_i) - J(x_{\text{trial}})}{T}\right)\right). \tag{A.1}$$

If $J(x_{\text{trial}}) \leq J(x_i)$ the trial point is accepted and $x_{i+1} = x_{\text{trial}}$. If $J(x_{\text{trial}}) > J(x_i)$ the candidate x_{trial} can still be accepted with a probability p_{accept}. This feature makes it possible that the sequence of cost function values $\{J(x_i)\}$ is not monotonically decreasing and therefore, as illustrated in Fig. A.3, can escape local minima and continue to search for a global minimum. The probability of accepting x_{trial} when $J(x_{\text{trial}}) > J(x_i)$ is controlled by the temperature T at step i. The probability of accepting an x_{trial} inferior to x_i is given by $\exp(J(x_{\text{trial}}) - J(x_i))/T$. For large T the probability of accepting an increase in J is close to one, but as T is cooled with iteration i this probability diminishes. The rule governing the behavior of $T(i)$ is called

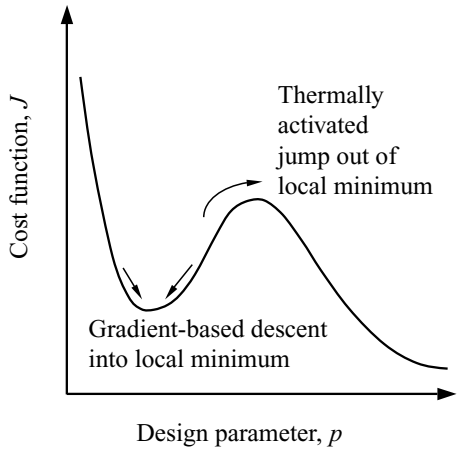

Fig. A.3. Showing how simulated annealing can be used to escape a local minimum.

the *cooling schedule*. Alternative acceptance functions have been suggested but most alternatives can be achieved by using the acceptance function Eq. (A.1) and transforming the cooling schedule, see [1], p. 183.

A cooling schedule works in combination with the generation of the next candidate x_{trial}. Given a certain probability distribution of candidates, sufficiently slow cooling guarantees finding the location of the global optimum. The cooling schedule *de facto* balances local against global exploration. Lowering T too quickly with the iteration count results in *simulated quenching* and can result in terminating the search in a local minimum prematurely. Simple, iteration-based cooling schedules as well as more complicated ones have been proposed and implemented. An adaptive cooling schedule for example controls T depending on the progress of the optimization which is reflected in the change of the objective function. Originally SA was implemented for problems with a discrete set of states and most theoretical results are for combinatorial optimization problems. Nevertheless, SA was quickly adapted to continuous global optimization. It is significant that for continuous search spaces the perturbations that generate the candidate parameters are assumed to be of global reach. A generated trial point can be located anywhere in the feasible set Ω with a certain probability independent of the current state x_i. The probability density controlling the candidate generation only depends on the temperature T. For large T global perturbations are more likely than for small T.

An example of random point generation with corresponding cooling schedule was developed by Ingber for the *Adaptive Simulated Annealing* algorithm (ASA) [7]. The trial point is given by

$$x_{\text{trial}} = x_i + \Delta x_i, \tag{A.2}$$

where Δx_i is a random variable. The probability distribution $g_i(\Delta x)$ of the perturbation Δx_i is given by

$$g_i(\Delta x_i) = (2\pi T(i))^{-\frac{n}{2}} \exp\left(-\frac{||\Delta x_i||^2}{2T(i)}\right), \tag{A.3}$$

where Δx_i is the deviation from the current point x_i and n is the dimension of Ω [1], p. 193. Again convergence to the global optimum for this algorithm is guaranteed for a sufficiently slow cooling schedule. If $T(i)$ does not decrease faster than

$$T(i) \geq \frac{T_0}{\log(i)}, \tag{A.4}$$

for sufficiently large and constant T_0, the algorithm will find the global optimum, i.e. ASA is rigorous. This very slow criterion is derived from the probability that the sampled points visit every set of positive Lebesgue-measure infinitely many times. Of course this logarithmic cooling schedule is useless for practical purposes and speed-ups were suggested. By sampling the perturbation from a Cauchy distribution given by

$$g_i(\Delta x_i) = \frac{T(i)}{(||\Delta x_i||^2 + T(i)^2)^{\frac{n+1}{2}}}, \tag{A.5}$$

the convergence to a global optimum is now guaranteed if

$$T(i) \geq \frac{T_0}{i}, \tag{A.6}$$

for sufficiently large T_0. This is called fast annealing and the speed-up stems from the fact that global exploration is improved because the Cauchy distribution has fatter tails than the distribution given in Eq. (A.3). Many more re-annealing and cooling schedules have been developed but are beyond the scope of the discussion here. The self-adaptation dynamically adjusts $T(i)$ depending on the progress of the optimization. If the optimization progress stagnates, for instance an increase in temperature or so called *re-annealing* can improve the chances of escaping from a local minimum or plateau. For more details on re-annealing and alternative perturbation methods see [1], p. 195. After considering the convergence properties of a few basic varieties of SA we have seen that in practice the convergence is guaranteed for unrealistically long times only. This is especially true when the objective function evaluations introduced in our design examples take in excess of one minute. Alternative termination conditions must be considered and can be replaced by standard heuristic stopping rules. See Section A.9.

Besides ease of implementation there are two main attractive features of SA. First, SA does not require or use the derivative information at a point

x_i in the search sequence. This is useful when the function to be optimized is not differentiable or the derivative is not accessible. Also, the sequence of values that is produced by an SA search is not necessarily decreasing. This climbing feature is necessary to avoid terminating the search prematurely in a local minimum. Parallel implementations of for instance ASA are available and might enjoy a revival with the increased availability of large cluster computers. The details of parallel ASA are omitted here. There are a few disadvantages that make SA not the best choice for the type of problems we consider. First of all, standard SA as discussed above does not use all the available information, i.e. the gradient $\nabla J(x_i)$. In this form random perturbations can result in spending too many cost function evaluations on local optimization. The choice of cooling schedule is critical. In order to have theoretical results that guarantee convergence to a global minimum a very slow logarithmic cooling schedule is required [8]. Rapid cooling without intermittent increases in T is called simulated quenching and will result in local optimization only, without necessarily climbing out of local minima. Sluggish cooling on the other hand will result in extensive global exploration but lack of local convergence. Note that SA does not utilize the sequence of previous function evaluations effectively to generate search points. High cost function values actually do contain useful information which remains unused in SA. Also, redundant function evaluations are not explicitly suppressed. The result is a waste of computational resources that makes SA less fitting for the type of optimization we are using. Next, a class of two-phase algorithms that combine global and local search techniques is presented.

A.5 Two-phased algorithms

This subsection suggests a classification of sampling methods that is different from the categories of rigor described above. Such classification does not focus on convergence properties but presents search heuristics that have been shown to work in practice but belong to the incomplete category. Many algorithms have two distinct phases. A global phase that ensures the exploration of the entire search domain and a local phase that takes advantage of fast and efficient local optimization techniques.

A.5.1 Multiple starting points

The simplest global optimization heuristic based on local search is the multiple start point search. Initial search points are generated in some stochastic manner in Ω and efficient local searches are started from each. If the initial

points are generated uniformly in Ω the global component of this algorithm guarantees the convergence to a global optimum, albeit on an unrealistic time scale. The local optimization component is supposed to provide significant speed-up.

A.6 Clustering algorithms

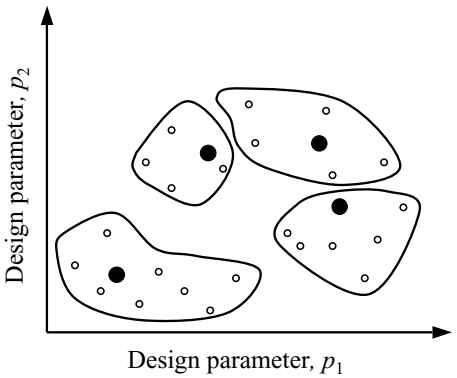

Fig. A.4. Clusters of sample points (○) are formed and a single cluster representative (●) of each cluster is determined. A computationally intensive local search procedure is started from all cluster representatives.

Clustering, a version of a two-phased algorithm, can be traced back to a paper by Becker and Lago [9] as well as Törn [10]. All cluster algorithms are based on principles of efficient local search as well as elimination of redundant cost function evaluations within a single cluster. First, clusters are formed from an initial set of sample points. From each cluster a single representative is determined as the initial point of a local search procedure. Figure A.4 is an illustration of the basic idea. A local search L is started for only the best performer in each cluster in order to avoid duplicating local searches within the region of attraction of the same local minimum. Here the region of attraction of a local minimizer $\hat{x}_i$ is defined to be the subset $S_i \subset \Omega$ such that if L is applied to $x \in S_i$ it will converge to $\hat{x}_i$. The efficiency of the algorithm largely depends on how well the cluster formation approximates the region of attraction of a local minimum. The variations of cluster formation for a sample of trial points include density clustering, single-linkage clustering, multi-level single-linkage clustering, simple-linkage clustering, and topographical linkage ([5], p. 833). Once the local optimization procedures are completed a new set of sample points is generated, clusters are formed again and the next set of local optimizations is started.

At iteration k the global phase consists of generating N_k sample points

$x_1, \ldots, x_{N_k}$ usually uniformly in Ω. Note that J is evaluated at each point and a *sample concentration* method is applied to increase the concentration of sample points into clusters around local minima. The resulting concentrated set of points is Z_c, which is combined with the sequence of all previous objective function evaluations $Z = Z \cup Z_c$. At this point a clustering algorithm is used to sort Z into a finite number of clusters $Z_1, \ldots, Z_h$. From each cluster $1, \ldots, h$ a single representative $p_1, \ldots, p_h$ is determined, usually the best performing point, and handed to the local phase of the algorithm. For cluster algorithms *without recall* local searches are started strictly at points from the set of new sample points. Cluster algorithms *with recall* on the other hand use the entire history of cost function evaluations to form all clusters anew and as possible starting points for the local search L. From each cluster representative in X a local search is started unless it previously served as the initial point for a local search. Most algorithms only store the starting point p_i and the endpoint q_i of the local optimization phase to conserve memory. After the local optimization phase is complete the global termination conditions are checked and either the algorithm terminates or a new global phase begins for further exploration. In the following the *sample concentration* and *clustering* methods are presented in more detail. For the local optimization phase any efficient local optimization technique is acceptable.

The purpose of the sample concentration phase is to make it easier to identify clusters. Clusters can be thought of as a collection of points in the basins of convergence near local minima. As the superior method to achieve this, Schoen suggests *sample reduction* ([1], p. 160). This is used to separate neighboring clusters more clearly and consists of reducing the entire set of samples Z by a fixed proportion γ, $|Z_c| = \gamma |Z|$ where $|\cdot|$ denotes the number of elements in a set. To ensure sample concentration we require $J(y) \geq J(x), \forall y \in Z \setminus Z_c, x \in Z_c$, [5]. For greater γ the separation between clusters increases. Following Boender and Romeijns ([in, 5] p. 832), several methods that sort the sample points into clusters can be used.

Density clustering (with recall) is based on approximating the basin of convergence around local minima by ellipsoids. In the global phase random points are generated uniformly over Ω. Sample reduction with recall is used to create a reduced set Z_c. Let Y be the set of local minima $\{q_1, \ldots, q_h\}$ found so far. If $Y = \emptyset$, start L from the point with lowest cost function value in Z_c and insert the resulting local minimum as the first point into Y, the set of local minima. If $Y \neq \emptyset$ we assign the point $x_i \in Z_c$ to the cluster with seed point $q_j \in Y$ if $||x_i - q_j||_2 \leq r_i(q_j)$ unless it is already assigned to another cluster. The radius r_i is given by Kan and Timmer [11] as

$$r_i(x) = \frac{1}{\sqrt{\pi}} \left(i\Gamma(1 + n/2) \cdot \sqrt{-\det H(x)} \cdot m(S) \cdot \frac{\zeta \ln(kN)}{kN} \right)^{1/n}. \tag{A.7}$$

Here $H(x)$ denotes the Hessian at x and ζ is a positive constant. Also, n is the dimension of Ω, N is the population size, and S is the level set $\{x \in \Omega : J(x) \leq f_\gamma\}$. Note that f_γ is chosen in a manner such that S has measure $m(S) = \gamma$. For all points in Z_c that are still not assigned to a cluster, L is applied and the local minimum q_j is determined. If the endpoint q_j of L is a cluster representative in Y, then the starting point is also assigned to cluster j. If q_j is not yet in the set of cluster representatives Y, the local minimum q_j is new and it is added to $Y_{\text{new}} = Y_{\text{old}} \cup q_{h+1}$. This method works in particular for level sets that are approximated well by an ellipsoid and the clustering process based on r_i is faster than clustering based on local minimization. Multi-level single linkage clustering does not rely on any particular shape of the level sets.

We give a brief pseudo algorithm for multi-level single linkage clustering (MLSL) as described by Pardalos and Romeijn ([1], p. 163). During the initialization let $X = \emptyset, Y = \emptyset, Z = \emptyset, k = 0$ and choose $\nu \in (0, 1)$ as well as $\sigma > 0$. As before N_k points are generated in Ω during the kth generation and the objective function is evaluated $J_i = J(x_i), i = 1, \ldots, N_k$. As this algorithm is performed with recall, a sample concentration method is applied to the entire search history Z to create the concentrated set Z_c. To form the cluster each point $x \in Z_c$ is marked as belonging to a cluster if and only if there exists another point $y \in Z_c$ such that

$$||x - y|| \leq \frac{1}{\sqrt{2\pi}} \left(\frac{\log(\sum_k N_k)}{(\sum_k N_k)} \sigma \Gamma \left(1 + \frac{n}{2}\right) \right)^{1/n}, \tag{A.8}$$

and

$$J(y) \leq J(x), \tag{A.9}$$

where n is the dimension of Ω and $\Gamma(\cdot)$ is the factorial function. The clustering radius in multi-level single linkage clustering is based on probabilistic considerations of covering the entire search space with sample points. After the clustering phase the local search procedure L is applied to all points in S_c that are not marked clustered and that have not been the initial point for a local search before. Complications might arise with the boundary of Ω so the local search L should be able to handle these constraints if they are present. The set of local optima Y is augmented by the results of the local search phase

$$Y_{\text{new}} = Y_{\text{old}} \cup \{(x, y, J(y))\}, \tag{A.10}$$

where x is the initial point of the local search and y is the local minimizer found. After the local search phase the global termination conditions are checked and the global search either terminates or another global phase is initiated. The recall feature is crucial for the convergence of this algorithm. As the size of the search history grows, condition Eq. (A.8) is tightened and points that are assigned to clusters early on might be pursued at later iterations. Under some analytical assumptions MLSL has attractive analytical properties and it is asymptotically complete. If $\sigma > 2$ the probability of local searches being started from a sample point decreases to 0. If $\sigma > 4$ the number of local searches is finite with probability 1 even if the algorithm is not explicitly stopped. This is true only if no local searches are started too close to the boundary. Unfortunately, experience shows that the performance of this algorithm fails for dimensions larger than around twelve [1].

Topological clustering is another simple clustering method, independent of the shape of the region of attraction around local minima. In this type of algorithm the clustering phase is performed on a fixed number of closest neighbors of a sample point. If any of the g closest neighbors has a lower cost function value the point is marked clustered. Local searches are started from those points whose g nearest neighbors have larger cost function values and have not been the seed of L before.

In addition to the global optimization schemes presented so far there are heuristics, essentially a bag of tricks, that help to cope with some of the challenges of global optimization stemming from local characteristics. These often address the difficulty of terminating in local minima prematurely and incorporating constraints.

A.7 Global optimization based on local techniques

Various global heuristics have been developed based on local optimization techniques. Some of these fit into the framework of two-phased algorithms more than others. Cluster algorithms for instance can be considered both two-phased and based on local techniques. In the following a collection of tricks is listed that may be useful in practice ([1], p. 87).

A.7.1 Direct elimination of local minimizers

The direct elimination of local minima is based purely on local optimization routines ([1], p. 95). After a local minimum x_k^* is reached, the objective function J is modified in order to remove the local minimum from J and the local minimization is continued. A positive function is added to the original

J in order to create a local maximum at x_k^*. The modification is simple and it should not add significantly to the computational burden of the objective function evaluation or its gradient. The modification only has a local impact in a neighborhood of x_k^* so that the other local minimizers of J are preserved. The modification should be removed once an x_l is found such that $J(x_l) < J(x_k)$. The size of the local perturbation is difficult to determine but should be based on the scale of the search domain Ω and the cost function J. If the neighborhood that is significantly impacted by the perturbation is too small new local minima may be created. If the perturbation is too large, neighboring minima might be eliminated because the original J is modified too heavily. Another difficulty is the first step after adding a local elimination function to J. Because x_k^* is a local minimizer its gradient is zero and computationally costly second-order information would have to be considered.

A similar method to escape local minima is *tunneling*. To tunnel out of a local minimum a *tunneling function* must be minimized, starting from the last known local minimum. An example of a tunneling function T is given by

$$T(x, \lambda, x_k^*, J(x_k^*)) = \frac{J(x) - J(x_k^*)}{||x - x_k^*||_2^{2\lambda}}, \tag{A.11}$$

and it depends on λ, a control parameter for the strength of the singularity as well as the last found minimum x_k^* and $J(_k^*)$. Initially a small λ is chosen. If x can be found such that $T(x, \lambda, x_k^*, J(x_k^*)) < 0$ the local minimization can continue at x. If no such x can be found, λ is increased and the search for an x such that $T(x, \lambda, x_k^*, J(x_k^*)) < 0$ is continued. The disadvantage here is, as before, that it is not clear how to search for x. If no improvement can be made this algorithm offers no alternative course of action.

A.8 Global smoothing

Global smoothing can be used to eliminate secondary or shallow local minima. In global smoothing techniques the objective function J is modified with an additional convex, global term to yield

$$\tilde{J}(x, \lambda) = J(x) + \lambda\Phi(x). \tag{A.12}$$

The function Φ should be at least as smooth as J and for a given λ the modified objective function $\tilde{J}$ should be more readily optimized using local optimization algorithms. The obtained minimum is then used as a starting point for a local optimization with a smaller λ. The algorithm consists of

a sequence of local optimization problems that converge to the unmodified problem not unlike interior point methods. Because Φ is known, so is usually its minimizer x^*_Φ and therefore the solution for sufficiently large λ is basically known and independent of the initial starting point. According to Murray and Ng the key to successfully using global smoothing is to pick Φ such that x^*_Φ does not coincide with any of the minima of J ([1], p. 97). Depending on the given problem this may be possible or not. Engineering intuition can sometimes be used to focus the search on different regions in the search space but this is contrary to the notion of automating the entire design process.

A.9 Stopping rules

Stopping rules are necessary for terminating a global search. In [5], p. 853, Horst states that unfortunately for

> a broad class of global optimization problems, it can never be verified in finite time that the global optimum is identified with certainty. Therefore a need emerges for stopping rules which decide if the expected benefit of further searching outweighs the required computational effort.

In the case of a purely random search with uniform sample placement over the feasible set Ω the probability of placing a random point in a neighborhood $B_\epsilon(x^*)$ of the global optimum x^* is given by $p_\epsilon = |B_\epsilon|/|\Omega|$, the ratio of the Lebesgue measures of $B_\epsilon(x^*)$ to that of the entire search space. After generating n points the probability of placing a point in the neighborhood $B_\epsilon(x^*)$ increases as $1-(1-p_\epsilon)^n$ with the number of sample points, n. To be $1-\delta$ certain of having explored the space, the simplest and most conservative of all stopping criteria is then given by termination for sufficiently many function evaluations $n \geq \frac{\log(\delta)}{\log(1-p_\epsilon)}$. Of course this stopping rule is usually too expensive, and independent of achieved function value or necessary or sufficient conditions for global or local optimality. The challenge for high dimensional problems is of course p_ϵ which shrinks exponentially with the dimension of the search space m, also called the curse of dimensionality. All algorithms face this challenge and it becomes crucial to exploit all possible speed-ups and accelerations available. This includes analytic as well as computational tools. For practical purposes a realistic stopping rule must be chosen. Possible criteria are listed in [5], p. 853, such as sample dependent, problem dependent, method dependent, loss dependent, and resource dependent.

Sample dependent stopping rules take into account the distribution of evaluated cost function values and their location in the feasible set. This can be a purely statistical analysis and if a probabilistic model of the distribution of cost function values is formulated, *optimal stopping* rules can be applied.

Problem dependent stopping rules make use of prior information, for instance if the number of local minima is known a priori the algorithm can be terminated once all of them have been located. This type of information is usually unknown in the engineering tasks we consider. Another aspect may be the minimum size of the region of attraction. If an estimate of for instance the minimum manufacturing accuracy is known, Ω only needs to be searched to a certain minimum scale.

Method dependent stopping rules incorporate known sufficient or necessary conditions if the applied method is rigorous. This is in particular true for the well-known first- and second-order conditions for local optima or the duality gap conditions for convex problems.

More esoteric termination conditions are given by *loss dependent* stopping rules. The cost of continuing the search is weighed against the benefit of doing so versus the cost involved with stopping prematurely. This type of analysis can be of benefit if possible cost savings can be weighed against the cost of waiting before implementing an improved design.

Resource dependent stopping rules take into account the computational cost and what is realistically possible. For example using the given computational resources only a finite number of cost function evaluations are possible in the allotted time.

Many stopping rules combine the different varieties of stopping rules mentioned here. If a minimal degree of functionality is required for a device design to operate in an acceptable manner there may be no option to continue the search. When this minimal requirement can be quantified as a numerical value of the objective function it can be used as a termination condition for global optimization. For the engineering problems considered here a complete search of the parameter space is most desirable but unrealistic. A simple and practical stopping rule is a fixed number of function evaluations, with the option of continuing the search if the best solution found does not meet the engineering requirements.

A.10 References

1. P.M. Pardalos and H.E. Romeijn, *Handbook of Global Optimization,* Volume 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

2. F. Glover, *Tabu search - Part 1*, ORSA Journal on Computing **1**, 190–206 (1989).

3. F. Glover, *Tabu search - Part 2*, ORSA Journal on Computing **2**, 4–32 (1990).

4. S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, *Optimization by simulated annealing*, Science **220**, 671–680 (1983).

5. R. Horst and P.M. Pardalos, *Handbook of global optimization*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

6. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, Journal of Chemical Physics **21**, 1087–1092 (1953).

7. L. Ingber, *Adaptive simulated annealing (ASA): Lessons learned*, Control and Cybernetics **25**, 33–54 (1996).

8. L. Ingber, *Simulated annealing: Practice versus theory*, Mathematical and Computer Modelling **18**, 29–57 (1993).

9. R. Becker and G. Lago, *A global optimization algorithm*, Proceedings of the 8th Allerton Conference on Circuits and Systems Theory, Monticello, Illinois, 1970, pp. 3–12.

10. A. Törn, *Global optimization as a combination of global and local search*, Doctoral thesis, Abo Akademi University, Turku, Finland, 1974.

11. A.H.G. Rinnooy Kan and G.T. Timmer, *Stochastic global optimization methods: Part 1; Clustering methods*, Mathematical Programming **39**, 57–78 (1987).

About the authors

(in alphabetic order)

Dimitris Bertsimas, *Massachusetts Institute of Technology*

Dimitris Bertsimas is the Boeing professor of Operations Research and the co-director of the Operations Research Center at MIT. He has received a BS in Electrical Engineering and Computer Science at the National Technical University of Athens, Greece, in 1985, an MS in Operations Research at MIT in 1987, and a PhD in Applied Mathematics and Operations Research at MIT in 1988. He has been in the faculty at MIT ever since. His research interests include optimization, stochastic systems, data mining and their applications. He has published widely and has written three graduate-level textbooks. He is a member of the National Academy of Engineering and has received several awards for his work including the Erlang prize, the Farkas prize, the SIAM optimization prize, the Bodossaki prize, and NSF's presidential young investigator award.

Stephan Haas, *University of Southern California*

Stephan Haas is a professor of theoretical condensed matter physics at the University of Southern California. He received his undergraduate education at the Technical University of Berlin and his PhD in Physics from the Florida State University. His research interests are in the area of strongly correlated electronic systems, including quantum magnets and unconventional superconductors. Recently, he has focused on quantum-to-classical crossover phenomena in nanoscale systems and the development of adaptive design techniques to efficiently explore and control the collective response properties of multi-parameter complex systems. He is a specialist in numerical techniques applied to many-body systems, including Quantum Monte Carlo, numerical diagonalization, and genetic algorithms.

A.F.J. Levi, *University of Southern California*

Tony Levi joined the USC faculty as professor of Electrical Engineering in mid 1993 after working for 10 years at AT&T Bell Laboratories, Murray Hill,

New Jersey. He invented hot electron spectroscopy, was the first to measure ballistic transport in unipolar transistors and heterostructure bipolar transistors, created the first unipolar ballistic electron transistor to operate at room temperature, created the first microdisk laser, and carried out work in parallel fiber optic interconnect components in computer and switching systems. To date he has published over 200 scientific papers, several book chapters, is author of the book "Applied Quantum Mechanics", and holds 17 US patents. His current research interests include scaling of ultra-fast electronic and photonic devices, system-level integration of advanced optoelectronic and RF technologies, manufacturing at the nanoscale, and optimal design.

Kelly Magruder, *University of Southern California*

Kelly Magruder received his BS in Electrical Engineering from Arizona State University in 2006, MS in Electrical Engineering from the University of Southern California in 2008, and is currently a PhD student at USC. He was named the Distinguished Senior upon graduating from ASU and is a USC Viterbi Fellow. His current research focuses on the effects of inelastic scattering on current flowing through nanowires and bulk semiconductors. His research is aimed at bridging the gap between perturbative methods and more exact solutions to the inelastic scattering problem to understand why perturbative methods such as Fermi's golden rule appear to work so well.

Rodrigo A. Muniz, *University of Southern California*

Rodrigo Muniz obtained his BS degree in Physics from the Fluminense Federal University, Brazil, in 2005 and currently he is a physics PhD student at the University of Southern California. He has been performing computer simulations of the dielectric response of nanostructures in order to analyze the properties of plasmonic modes such as space localization and scalability with system properties. His research aims to understand the fundamental aspects of collective excitations in broken symmetry systems and how to exploit the new degrees of freedom absent in more restrictive bulk structures.

Omid Nohadani, *Massachusetts Institute of Technology*

Omid Nohadani is a post-doctoral researcher at the Operations Research

Center at MIT. He received a Diploma degree in mathematical physics from the University of Bonn, Germany, and a PhD in physics from the University of Southern California in 2005. His work on quantum magnets and optimized quantum Monte Carlo algorithms was awarded by the US Council for Graduate Schools as the top-five dissertation of the year. He then became a postdoctoral researcher at USC's Quantum Engineering lab, working on optimizing nano-photonic materials. Since 2006 at MIT, he has developed the first robust optimization algorithm that can handle simulation-based problems in the presence of errors. He has designed robust optimization methods to a wide range of applications: nano-photonics, vehicle routing, statistics, ultrafast optics, and radiation oncology. Currently, he also holds a research fellowship at Harvard, developing novel techniques in radiation therapy. He created robust algorithms for motion management in lung cancer at the Massachusetts General Hospital in conjunction with Harvard Medical School clinicians.

Gary Rosen, *University of Southern California*

Gary Rosen is professor and chair of the Department of Mathematics at the University of Southern California. He received his ScB in Mathematics in 1975, his ScM in Applied Mathematics in 1976, and his PhD in Applied Mathematics in 1980 all from Brown University. He was an Assistant Professor of Mathematics at Bowdoin College from 1980–1984, was a member of the Technical Staff at the Charles Stark Draper Laboratory in Cambridge, Massachusetts, and was a consultant at the Institute for Computer Applications in Science and Engineering (ICASE) at the NASA Langley Research Center in Hampton, Virginia. His research is focused on topics related to the modeling, estimation, control, and optimal design of systems governed by infinite dimensional dynamical systems, and in particular partial differential equations. He has developed and analyzed numerical approximation methods and efficient computational algorithms for both local and global optimization. Most recently he has been studying the application of these techniques to the development of a data analysis system for a passive and non-invasive alcohol biosensor, to the adjoint method based optimal design of nanoscale layered quantum electronic devices, to the development of data assimilation techniques for a global assimilative model for the ionosphere, and to the development of intelligent methods for the manufacture of

advanced semiconductor devices. His research has been supported by grants from AFOSR, ONR, DARPA, NASA, NSF, and NIH, and he has served as a member of the editorial board of the IEEE Transactions on Control Technology.

Philip Seliger, *University of Southern California*

Philip Seliger received a BS in Applied Mathematics and Physics from the University of California, San Diego, in 2001. Afterwards, he worked for one year as a research assistant at the Institute for Non-linear Science at UCSD before joining the Mathematics graduate program at the University of Southern California. After receiving an MS in Statistics and an MS in Applied Mathematics, he received a PhD in Applied Mathematics from the University of Southern California in 2008. His thesis work focused on streamlining the design process of electromagnetic devices and similar state-constrained optimization problems. He is currently working for TM Tech on signal processing of broadband radar signals and electronics firmware development.

Chunming Wang, *University of Southern California*

Chunming Wang received his undergraduate degree, Diplôme d'ingénieur in Applied Mathematics and Computer Science, in 1984 from Université de Compiègne in France. He received his PhD degree in Applied Mathematics in 1988 from Brown University. Since his graduation he joined the faculty at the University of Southern California as an assistant professor and was subsequently promoted to associate professor and professor of mathematics. His main research area is control and optimization of distributed parameter systems governed by partial differential equations. He has been involved extensively in multidisciplinary research in intelligent manufacturing of nanoscale semiconductor devices, optimal design of nano-photonic devices, and ionospheric data assimilation. In addition to his academic research, he also has extensive collaborations with scientists in government research laboratories and industry. In particular, he has worked with Northrop Grumman Space Technology Inc. in the development of the National Polar Orbiting Environmental Sensing Satellite (NPOESS) since 2000.

Index