School of Mathematics and Statistics

Computational Studies of Some Static and Dynamic Optimisation Problems

Wei Rong Lee

This thesis is presented as part of the
requirements for the award of the
Degree of Doctor of Philosophy of the
Curtin University of Technology

June 1999

Contents

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of my knowledge, no material is contained herein which has been previously published or written by any other person, except where due reference is made in the text.

Acknowledgments

It is my pleasure to express my sincere gratitude and appreciation to my supervisors Dr. S. Wang of the School of Mathematics and Statistics, Curtin University of Technology, and Professor K.L. Teo of the Department of Applied Mathematics, Hong Kong Polytechnic University for their guidance, patience and encouragement throughout the period of my doctoral studies. Without these the completion of this thesis would be impossible.

I would like to thank Professor L.S. Jennings of the Center for Applied Dynamics and Optimization, the University of Western Australia for generously giving his time to discussing theoretical and computational aspects of the materials in the thesis.

I would also like to thank my associate supervisor, Dr. V. Rehbock for his helpful discussions.

I am very grateful to Curtin University for the financial support during the period of my study.

I also wish to thank the School of Mathematics and Statistics for providing me with the necessary facilities.

Acknowledgment is also made to all the staff at the School for their assistance during my study at Curtin University.

Finally I thank my parents who made it possible, and my wife Mei Ying Huang and son Simon for their understanding and support.

Publications

1. Lee W.R and Wang S, " Finding Unknown parameters in Semiconductor Device Design by an optimization Approach ", in Caccetta L, Teo K.L, Siew P.F, Leung Y.H, Jennings L.S and Rehbock V (eds.), *Proceeding of ICOTA98, July 1998, Perth*, Curtin University of Technology (1998) 503–509.

2. Lee W.R, Wang S and Teo K.L, " An Optimization Approach to Numerical Integration in Two Dimensions", *Applied Mathematics and Computation*, to appear.

3. Lee W.R, Wang S and Teo K.L, "Optimal Recharge and Driving Strategies for a Battery - Powered Electric Vehicle", *Mathematical Problems in Engineering*, **Vol** 3, (1999) 1–31.

4. Lee W.R, Wang S and Teo K.L, " Solving a Class of Nonlinear Optimal Feedback Control Problems Using 3rd order B-Splines with Optimal Partition Points", *Proceeding of DYCONS99, August 1999, Canada*, to appear.

5. Teo K.L, Lee W.R, Jennings L.S, Wang S and Liu Y, "Numerical Solution of an Optimal Control Problem with Variable Time Points in the Objective Function", *J. Aust. Math. Soc.*, Series B, to appear.

6. Wang S, Lee W.R and Teo K.L, "On Optimal Grid Construction in Numerical Integration", *Engineering Optimization*, to appear.

7. Lee W.R, Wang S and Teo K.L, "An Optimization Approach to a Finite Dimensional Parameter Estimation Problem in Semiconductor Device Design", *J. Comp. Phys.*, to appear.

Summary

In this thesis we shall investigate the numerical solutions to several important practical static and dynamic optimization problems in engineering and physics. The thesis is organized as follows.

In Chapter 1 a general literature review is presented, including motivation and development of the problems, and existing results. Furthermore, some existing computational methods for optimal control problems are also discussed.

In Chapter 2 the design of a semiconductor device is posed as an optimization problem: given an ideal voltage-current ($V - I$) characteristic, find one or more physical and geometrical parameters so that the V-I characteristic of the device matches the ideal one optimally with respect to a prescribed performance criterion. The voltage–current characteristic of a semiconductor device is governed by a set of nonlinear partial differential equations (PDE), and thus a black-box approach is taken for the numerical solution to the PDEs. Various existing numerical methods are proposed for the solution of the nonlinear optimization problem. The Jacobian of the cost function is ill-conditioned and a scaling technique is thus proposed to stabilize the resulting linear system. Numerical experiments, performed to show the usefulness of this approach, demonstrate that the approach always gives optimal or near-optimal solutions to the test problems in both two and three dimensions.

In Chapter 3 we propose an efficient approach to numerical integration in one and two dimensions, where a grid set with a fixed number of vertices is to be chosen so that the error between the numerical integral and the exact integral is minimized. For one dimensional problem two schemes are developed for sufficiently smooth functions based on the mid–point rectangular quadrature rule and the trapezoidal rule respectively, and another method is also developed for integrands which are not sufficiently smooth. For two dimensional problems two schemes are first developed for sufficiently smooth functions. One is based on the barycenter rule on a rectangular partition, while the other is on a triangular partition. A scheme for insufficiently smooth functions is also developed. For

illustration, several examples are solved using the proposed schemes, and the numerical results show the effectiveness of the approach.

Chapter 4 deals with optimal recharge and driving plans for a battery–powered electric vehicle. A major problem facing battery–powered electric vehicles is in their batteries: weight and charge capacity. Thus a battery–powered electric vehicle only has a short driving range. To travel for a longer distance, the batteries are required to be recharged frequently. In this chapter we construct a model for a battery–powered electric vehicle, in which driving strategy is to be obtained so that the total traveling time between two locations is minimized. The problem is formulated as an unconventional optimization problem. However, by using the control parameterization enhancing transformation (CPET)(see [100]) it is shown that this unconventional optimization is equivalent to a conventional optimal parameter selection problem. Numerical examples are solved using the proposed method.

In Chapter 5 we consider the numerical solution to a class of optimal control problems involving variable time points in their cost functions. The CPET is first used to convert the optimal control problem with variable time points into an equivalent optimal control problem with fixed multiple characteristic times (MCT). Using the control parameterization technique, the time horizon is partitioned into several subintervals. Let the partition points also be taken as decision variables. The control functions are approximated by piecewise constant or piecewise linear functions in accordance with these variable partition points. We thus obtain a finite dimensional optimization problem. The CPET transform is again used to convert approximate optimal control problems with variable partition points into equivalent standard optimal control problems with MCT, where the control functions are piecewise constant or piecewise linear functions with pre-fixed partition points. The transformed problems are essentially optimal parameter selection problems with MCT. The gradient formulae are obtained for the objective function as well as the constraint functions with respect to relevant decision variables. Numerical examples are solved using the proposed method.

A numerical approach is proposed in Chapter 6 for constructing an approximate optimal feedback control law of a class of nonlinear optimal control problems. In this approach, the state space is partitioned into subdivisions, and the controllers are approximated by a linear combination of the 3rd order B-spline basis functions. Furthermore, the partition points are also taken as decision variables in this formulation. To show the effectiveness of the proposed approach, a two dimensional and a three dimensional examples are solved by the approach. The numerical results demonstrate that the method is superior to the existing methods with fixed partition points.

Chapter 1

Introduction

Static and dynamic optimization problems arise in many disciplines such as engineering, economics, physics and the biomedical sciences. Although many theoretical results are available in the literature for a large range of model optimization problems (see, for example, [114, 112]), most practical problems are too complicated to be solved analytically. Thus numerical methods are essential for solving these kind of practical problems. There are numerous computational methods for solving various practical optimization problems. For details, see [31, 85, 72, 77, 9, 115, 36, 48, 18, 26, 35, 39, 59, 79, 80, 87, 91, 99, 102, 100, 86, 103, 104, 105].

Some earlier computational methods for solving dynamic optimization or optimal control problem are based on the solution of nonlinear two-point boundary-value-problems. One important family of this approach which is comprised of several variations is the neighboring extremal methods (cf [11]). However, there is no guarantee of convergence even for relatively simple problems, though it has been used successfully in solving a number of rather complex problems.

Another earlier method is called the extremal field method (cf. [11, 5, 6]). It requires the solution of the Hamilton-Jacob-Bellman Partial differential equation in a domain of the state space containing the optimal solution. This method can be used for solving relatively simple problems since it demands an exorbitant amount of memory storage. Most of recent algorithms are found in the family of gradient methods. In [19, 27, 52, 57, 93, 94], the differential equations are discretized into difference equations, while in [38], the differential equation

is handled with a Lagrange multiplier while other constraints are treated explicitly. A straightforward transformation method is used in [40, 84], and in [35, 58, 62, 113], sequential gradient-restoration algorithms have been developed for the solution of different classes of optimal control problems. The latter is further enhanced by the dual version (cf. [60, 61]).

More recently, the classical control technique is introduced in [98, 101, 99] as a basis for solving various constrained optimal control problems in a unified fashion. The technique is a flexible and efficient approach for a large class of optimal control problems. The central idea of the method relies on a simple approximation scheme, i.e., to approximate an optimal control problem by a sequence of finite dimensional optimal parameter selection problems. Each of these optimal parameter selection problem can be viewed as a mathematical programming problem.

Despite the flexibility and the efficiency of the approach, there are several numerical difficulties associated with it. The accuracy of method depends greatly on the choice of knot distribution. For example, if the true optimal control belongs to the class of piecewise continuous functions, and if one has no insight of how the switching times of the true optimal control are distributed, a set of dense and evenly distributed knots of the approximating control is usually chosen in the hope that there would be a knot placed at each switching time. Thus, the number of parameters in the approximate optimal parameter selection problem is usually very large. As the number of parameters increases, the optimization process becomes computational more expensive. It appears that one may be able to reduce the overall number of parameters used if the switching times are treated as decision variables. However, it is known that the gradients of the cost functional and constraint functionals with respect to these switching times are discontinuous. See Chapter 5 of [39] for details. Furthermore, in order to obtain an accurate solution, the differential equation solver will need to perform the integration with respect to time over the consequent subintervals in the time horizon according to the particular partition. These subintervals would be varying from one iteration to the next during the optimization process.

Finally, the number of the decision variables would change when two or more switching times collapse. Thus, the task of integrating the differential equations accurately can be very involved. For these reasons, the gradient formulae presented in Chapter 5 of [39] were never implemented. In [100, 50], a novel problem transform, to be referred to as the control parameterization enhancing transform (CPET), was first introduced to overcome these numerical difficulties for time optimal control problems. The transformed optimal control problem can be solved readily and accurately by the classical control parameterization technique. In particular, the control parameterization technique, which can be used to approximate the optimal control problem by a constrained non-linear programming problem, has been developed from some recent research (for detail, see. [98, 101, 99]), while the control parameterization enhancing transform is developed in [100, 50].

Software packages implementing computational algorithms for both static and dynamic optimization problems are also available now. For example, FFSQP [116] was developed for solving general nonlinear programming problems, and MINOS5.4 [73] was developed for solving large-scale optimization problems expressed in the following standard form:

$$\begin{aligned} \min_{x,y} \quad & F(x) + c^T x + d^T y \\ \text{subject to} \quad & f(x) + A_1 y = b_1, \\ & A_2 x + A_3 y = b_2, \\ & l \leq \begin{pmatrix} x \\ y \end{pmatrix} \leq u, \end{aligned}$$

where the vectors c, d, b_1, b_2, l, u and the matrixes A_1, A_2, A_3 are constant. A general optimal control software package MISER3.2 [41, 43] was also developed based on the method of the control parameterization and the control parameterization enhancing transform.

In the following subsections we present a general literature review of several important static and dynamic optimization problems arising in different areas.

1.1 Semiconductor Device Design

The fundamental behavior of semiconductor devices is governed by a coupled system of nonlinear second-order partial differential equations (PDEs) ([107]). This system consists of a Poisson equation and two current continuity equations. Early work in the numerical solution of these PDEs includes Gummel ([37]) and De Mari ([23, 24]). Since the coefficient functions and the unknowns of the equations vary by several orders of magnitude in some small subregions of a device, the application of classical discretisation methods may cause unacceptable errors unless impractically, fine meshes which require an enormous amount of computer resources are used. Numerical solution of the PDEs has become practical since Scharfetter and Gummel (1969) proposed a one-dimensional discretisation method for these equations (cf.[88]). The method was also proposed independently by Allen and Southwell (cf. [2]). The existence of the solution to these PDEs is proved by Mock *et al* (cf. [70, 71, 55]) under some restrictions. The proofs are based on Schauder's fixed principle (cf. [30]). The uniqueness of the solution to these PDEs is proved when the applied biases V are sufficiently small (cf. for example, [55]). To solve these nonlinear PDEs, a typical way is to apply the Newton-like method (cf. [76]), or a modified Newton-like method called the damped Newton method to this system. A popular decoupling method for solving these PDEs is Gummel's method, introduced by Gummel (cf. [37]), which also has a modification (cf. [90]). However, when the recombination term R is sufficiently large Gummel's method may fail to converge. To overcome this difficulty, Seidman and Choo proposed their algorithm (cf. [89]). Using Gummel's method [37] and Newton's method [76] we can decouple and linearize these PDEs so that at each iteration step we have to solve a set of three linear equations of the form

$$-\nabla \cdot (a(x)\nabla u) + G(x)u = F(x) \qquad \text{in} \quad \Omega$$

$$u|_{\partial\Omega_D} = \gamma(x), \qquad \nabla u \cdot \boldsymbol{n}|_{\partial\Omega_N} = 0$$

where $\Omega \subset \mathbb{R}^{\overline{m}}$ ($\overline{m} = 1, 2, 3$), $\partial\Omega = \partial\Omega_D \cup \partial_N$ is the boundary of Ω. $\partial\Omega_D \cap \partial\Omega_N = \emptyset$, $\boldsymbol{n}$ denotes the unit outward normal vector on $\partial\Omega$, $G \in C^0(\overline{\Omega}_{\cup} H^1(\Omega)$, $F \in L^2(\Omega)$. Over the last twenty years, some efficient and stable discretisation schemes (cf. [63, 65, 66, 67, 68, 69, 109, 110, 29]) and meshing techniques (cf. [64]) have been developed for the above equations.

The conventional computer-aided design cycle of a semiconductor device is as follows: given physical parameters and dimensions of a device, find the voltage–current $(V-I)$ characteristic curve of the device by numerically solving the above set of nonlinear partial differential equations [110]. If the result does not match the required characteristic curve well enough, modify the parameters and resolve the problem. This design approach is time-consuming because the choice of the parameters is empirical. In this thesis, we will address this problem in a different way. More specifically the problem is posed as the following optimization problem: Given an ideal $V - I$ characteristic curve $I_g(v)$, find physical and geometric parameters such that the $V - I$ characteristic curve of the device matches the ideal characteristic curve optimally with reference to a specified performance criterion.

1.2 Numerical Integration

Integration plays a key role in many areas of science, engineering and economics. Since most of the integrals encounted in real–world problems cannot be evaluated exactly, numerical approximations of these integrations by some quadrature or cubature rules are normally sought in practice. Thus, efficient numerical methods are crucial. Numerical integration is one of the oldest problems in mathematics which can be traced back to the era of Archimedes. It is used to obtain approximate solutions of many real-world problems. Many results on the construction of quadrature and cubature rules are now available in the literature. For example, see [20] and [28]. Some quadrature rules with minimal error norm have also been developed using interpolation (cf. [111]). In [21] and [78], adaptive schemes are obtained for numerical integration, where a typical

approach is to initially evaluate the integral numerically by a quadrature rule on a uniform partition. Then, estimate the error bound for the numerical integral on each subinterval, use the maximum error to refine the partition uniformly, and subsequently evaluate the integral numerically on the corresponding new partition. This process is repeated until the maximum error is smaller than a given tolerance. Although this approach is efficient, neither the initial mesh nor each of the subsequent refined meshes is optimal. In this thesis, the partition points of the numerical integration are taken to be decision variables, a grid set with a fixed number of vertices is to be obtained such that the error between the numerical integral using a given quadrature rule and the exact integral is minimized.

1.3 Multiple Characteristic Time Constraints

Consider a process described by the following system of nonlinear differential equations defined on [0,T].

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{z}(t)), \tag{1.3.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}^0 \tag{1.3.2}$$

where T is a fixed terminal time, $\boldsymbol{x} = [x_1, \cdots, x_n]^T \in \mathbb{R}^n$, $\boldsymbol{u} = [u_1, \cdots, u_m]^T \in \mathbb{R}^m$ and $\boldsymbol{z} = [z_1, \cdots, z_p]^T \in \mathbb{R}^p$ are , respectively, state, control, and system parameter, while $\boldsymbol{f} = [f_1, \cdots, f_n]^T \in \mathbb{R}^n$ is a continuously differentiable function with respect to all its arguments, and $\boldsymbol{x}^0$ is a given vector.
Let a_i and b_i, $i = 1, \cdots, s$, c_i and d_i, $i = 1, \cdots, m$, be fixed constants. Define

$$\boldsymbol{Z} = \{\boldsymbol{z} = [z_1, \cdots, z_r]^T \in \mathbb{R}^r : a_i \leq z_i \leq b_i,\ i = 1, \cdots, r\}$$

$$\boldsymbol{U} = \{\boldsymbol{u} = [u_1, \cdots, u_m]^T \in \mathbb{R}^m : c_i \leq u_i \leq d_i,\ i = 1, \cdots, m\}$$

Any Borel measurable function $\boldsymbol{u} : [0, T] \longrightarrow \boldsymbol{U}$ is called an admissible control. Let $\mathcal{U}$ be the class of all admissible controls. For each $(\boldsymbol{u}, \boldsymbol{z}) \in \mathcal{U} \times \boldsymbol{Z}$, let$\boldsymbol{x}(\cdot|\boldsymbol{u}, \boldsymbol{z})$ denote the corresponding solution of the system (1.3.1) –(1.3.2).

We now state the canonical optimal control problem as follows:
Given the system (1.3.1) –(1.3.2), find a $(\boldsymbol{u}, \boldsymbol{z}) \in \mathcal{U} \times \boldsymbol{Z}$ such that the cost function

$$g_0(\boldsymbol{u}, \boldsymbol{z}) = \phi_0(\boldsymbol{x}(T|\boldsymbol{u}, \boldsymbol{z})) + \int_0^T \mathcal{L}_0(t, \boldsymbol{x}(t|\boldsymbol{u}, \boldsymbol{z}), \boldsymbol{u}(t))dt, \tag{1.3.3}$$

is minimized subject to the equality constraints:

$$g_i(\boldsymbol{u}, \boldsymbol{z}) = \phi_i(\boldsymbol{x}(\tau_i|\boldsymbol{u}, \boldsymbol{z})) + \int_0^{\tau_i} \mathcal{L}_0(t, \boldsymbol{x}(t|\boldsymbol{u}, \boldsymbol{z}), \boldsymbol{u}(t))dt = 0, \quad i = 1, \cdots, N_e, \tag{1.3.4}$$

and the inequality constraints:

$$g_i(\boldsymbol{u}, \boldsymbol{z}) = \phi_i(\boldsymbol{x}(\tau_i|\boldsymbol{u}, \boldsymbol{z})) + \int_0^{\tau_i} \mathcal{L}_0(t, \boldsymbol{x}(t|\boldsymbol{u}, \boldsymbol{z}), \boldsymbol{u}(t))dt =\geq 0, \quad i = N_e, \cdots, N, \tag{1.3.5}$$

where ϕ_i, $i = 0, 1, \cdots, N$, and $\mathcal{L}_i, i = 0, 1, \cdots, N$, are given real valued functions; and $\tau_i \leq T$ is referred to as the characteristic time for the ith constraint with $\tau_0 = T$ by convention. The terminal terms of the objective function (1.3.3) and the constraints (1.3.4) and (1.3.5) depend only upon the state vector evaluated at single points τ_i.

The concept of multiple characteristic time (MCT) constraints, which are constraints that depend upon the state vector specified at two or more discrete time–points, is introduced in [56].

MCT constraints are encounted in many real-world problems (see, [56]). For example, the constraints $g_i = x(\tau_1) - x(\tau_2) \geq 0$, where $0 < \tau_1 < \tau_2$, and τ_1 and τ_2 are characteristic time, is a simple MCT constraint.

MCT constraints also arise through a technique called constraint transcription (cf. Section 3.7 of [56], and [44]). This is a method for reducing a large number of constraints to a single equivalent constraint.

In [56] the control parameterization technique is extended to problems that have MCT constraints. Some convergence results are established, and the gradient formulae are constructed for the MCT constraints. It is proved that the constraint transcription technique can be applied to optimal control problems that have a set of MCT constraints.

1.4 Optimal Feedback Control

For many optimal control problems in engineering, management and finance, feedback controls are much preferred solutions in comparison with open–loop solutions. This is because an optimal feedback control provides the optimal trajectory from an arbitrary initial condition in a set, and is robust in the presence of system noise and parameter variation. Because of its importance, construction of feedback controls has been discussed for many years. For problems with a linear dynamical systems and a quadratic performance index, a genuine optimal control law can be derived [3, 46]. Techniques for constructing optimal controls are also available in the literature for nonlinear problems with a special structure (cf. [7, 97]). However, finding the optimal feedback control for a general non-linear problem is extremely difficult.

Recently, several methods based on neural networks have been proposed in [74, 33, 34] for approximating optimal feedback controllers. Although the methods are promising, the performance of a feedback control obtained from these method is very much dependent on the structure and the size of the neural network used. The choice of network structure is by large through trial and error experience, which can be rather time-consuming. More recently, an iterative multivariate interpolation method is proposed in [49, 83]. In this approach, several open-loop optimal control trajectories using different initial conditions are first computed, and then a spline interpolation is constructed using these scattered data to provide the optimal feedback control. One problem associated with this approach is that although the initial conditions are chosen uniformly in the state space, the trajectories of the open-loop solutions may shrink up as time steps forward. Thus, extra open-loop control problems need to be solved to cover the regions which have no or unpractically few data from the previous iteration. The efficiency of this approach depends strongly on the problem solved. Furthermore, the number of the partition points grows exponentially as the number of dimensions of the state variables increases. Thus, this approach is only applicable to problems of low dimensions.

In this thesis the optimal feedback control is approximated by a linear combination of the 3rd order B-spline basis functions constructed on a partition in state and time spaces. Furthermore, the partition points are also taken to be decision variables. Once an initial condition is selected, the resulting optimal control problem is an optimal parameter selection problem. The optimal control obtained is represented by a linear combination of the 3rd order B-spline with optimal coefficients and optimal partition points. This control is a function of state and time.

Chapter 2

A Finite Dimensional Parameter Estimation Problem in Semiconductor Device Design

2.1 Introduction

The electrical behavior of a semiconductor device is governed by the following system of nonlinear second-order elliptic equations ([90], [107], Chapter 2 of [63], and Chapter 2 of [55])

$$\nabla \cdot (\varepsilon \nabla \psi) = q(n - p - N), \tag{2.1.1}$$

$$\nabla \cdot \boldsymbol{J}_n - q\frac{\partial n}{\partial t} = qR(\psi, n, p), \tag{2.1.2}$$

$$\nabla \cdot \boldsymbol{J}_p + q\frac{\partial p}{\partial t} = -qR(\psi, n, p), \tag{2.1.3}$$

in $\Omega \subset \mathbb{R}^m$ $(m = 1, 2, 3)$ with appropriate boundary conditions, where $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ are, respectively, the electron and hole current densities, defined by

$$\boldsymbol{J}_n = q(D_n \nabla n - \mu_n n \nabla \psi), \tag{2.1.4}$$

$$\boldsymbol{J}_p = -q(D_p \nabla p + \mu_p p \nabla \psi). \tag{2.1.5}$$

Here ψ is the electrostatic potential, n is the electron concentration, p is the hole concentration, $N = N_D - N_A$ denotes the doping function where N_D and N_A are the donor and acceptor concentrations respectively, R denotes the recombination/generation rate which is assumed to be monotone with respect to n and p, q is the electronic charge, $\varepsilon = \varepsilon_0 \varepsilon_\delta$ is the permittivity of the medium

which is positive and bounded away from zero, and ε_δ is the relative permittivity depending on materials. In (2.1.4)–(2.1.5), μ_n μ_p are electron and hole mobilities respectively, and D_n and D_p are diffusion coefficients of electron and hole respectively. For these quantities, we assume that the Einstein's relationship holds:

$$D_n = \frac{kT}{q}\mu_n, \quad D_p = \frac{kT}{q}\mu_p$$

where T is the absolute temperature and k is the Boltzmann's constant.

In this thesis, we are only concerned with the stationary problem, *i.e.* $\frac{\partial n}{\partial t} = 0$ in (2.1.2–2.1.3).

We now derive the boundary conditions for (2.1.1–2.1.5). Assume that the boundary $\partial\Omega$ of the device region Ω is polygonal or polyhedral, and let $\partial\Omega_D \subset \partial\Omega$ denote the union of all contacts (i.e. terminals of a device) and $\partial\Omega_N$ the part of boundary such that

$$\overline{\partial\Omega}_D \cup \overline{\partial\Omega}_N = \partial\Omega, \quad \partial\Omega_D \cap \partial\Omega_N = \emptyset.$$

On the boundary $\partial\Omega_N$, we assume that all the outward normal derivatives vanish, *i.e.*

$$\nabla\psi \cdot \boldsymbol{n} = \nabla n \cdot \boldsymbol{n} = \nabla p \cdot \boldsymbol{n} = 0 \quad \forall x \in \partial\Omega_N$$

where $\boldsymbol{n}$ denotes the unit outward normal vector along the boundary $\partial\Omega_N$. Under this assumption, there is no current flowing out of $\partial\Omega_N$. The current flowing in or out of a terminal $c \in \partial\Omega_D$ is given by

$$I = \int_c (\boldsymbol{J}_n + \boldsymbol{J}_p) \cdot \boldsymbol{n} ds \tag{2.1.6}$$

where $\boldsymbol{n}$ denotes the outward unit normal vector of $\partial\Omega_D$.

The portion $\partial\Omega_D$ is usually composed of three kinds of contacts: ohmic contact, Schottky contact and interface to insulation material.

For the sake of convenience, we only derive the boundary condition on the ohmic contacts. For the boundary conditions of Schottky contacts and the interfaces to insulation material, we refer to [90] and [55].

An ohmic contact has a negligible contact resistance relative to the bulk or spreading resistance of the semiconductor. It does not perturb the device performance significantly ([95]). Therefore, thermal equilibrium and charge neutrality are usually assumed on an ohmic contact; *i.e.* for any ohmic contact $C \subset \partial\Omega_D$, we have

$$np = n_i^2 \tag{2.1.7}$$

$$n - p - N = 0 \tag{2.1.8}$$

for all $x \in C$, where n_i is the intrinsic density.

One formula of n_i given by Alder *et al* (see. [1]) is

$$n_i = 3.88 \times 10^{16} T^{1.5} \exp\Big(- \frac{7.00 \times 10^3}{T} \Big).$$

There are some other formulas for n_i such as Slotboom and De Graaff (see. [92]). Solving (2.1.7–2.1.8) we get

$$n|_C = \frac{N}{2} + \Big[\Big(\frac{N}{2} \Big)^2 + n_i^2 \Big]^{\frac{1}{2}}, \qquad p|C = \frac{n_i^2}{n} \quad \text{if } N > 0 \tag{2.1.9}$$

$$p|_C = -\frac{N}{2} + \Big[\Big(\frac{N}{2} \Big)^2 + n_i^2 \Big]^{\frac{1}{2}}, \qquad n|C = \frac{n_i^2}{p} \quad \text{if } N < 0. \tag{2.1.10}$$

If we introduce the quasi–Fermi potentials ϕ_n and ϕ_p such that

$$n = n_i \exp\Big(\frac{q(\psi - \phi_n)}{kT} \Big)$$

$$p = n_i \exp\Big(\frac{q(\phi_p - \psi)}{kT} \Big)$$

then we have

$$\psi = \phi_n + \frac{kT}{q} \ln\Big(\frac{n}{n_i} \Big) = \phi_p - \frac{kT}{q} \ln\Big(\frac{p}{n_i} \Big). \tag{2.1.11}$$

Since an ohmic contact supplies the required current with a voltage drop that is sufficiently small compared with the drop across the active region of the device (cf. p304, [95]), the quasi–Fermi potentials are assumed to be constant at an ohmic contact and are equal to the bias applied on the contact (cf. p21 of [45]). Therefore from (2.1.9), (2.1.10) and (2.1.11) we know that ψ is also constant at any contact. Thus, at an ohmic contact $C \subset \partial\Omega_D$ we obtain

$$\psi|_C = V_C + \frac{kT}{q} \ln\Big(\frac{n|_C}{n_i} \Big) = V_C - \frac{kT}{q} \ln\Big(\frac{p|_C}{n_i} \Big)$$

where V_C denotes the bias applied on C.

To summarize, the stationary behavior of semiconductor devices is governed by the following system of PDEs:

$$\nabla \cdot (\varepsilon \nabla \psi) = q(n - p - N), \tag{2.1.12}$$

$$\nabla \cdot (D_n \nabla n - \mu_n n \nabla \psi = R(\psi, n, p), \tag{2.1.13}$$

$$\nabla \cdot (D_p \nabla p + \mu_p p \nabla \psi) = R(\psi, n, p) = 0, \tag{2.1.14}$$

with boundary conditions

$$\psi|_{\partial\Omega_D} = \gamma_1(x), \quad n|_{\partial\Omega_D} = \gamma_2(x), \quad p|_{\partial\Omega_D} = \gamma_3(x), \tag{2.1.15}$$

and

$$\nabla\psi \cdot \boldsymbol{n}|_{\partial\Omega_N} = \nabla n \cdot \boldsymbol{n}|_{\partial\Omega_N} = \nabla p \cdot \boldsymbol{n}|_{\partial\Omega_N} = 0 \tag{2.1.16}$$

where $\gamma_i(x)$ $(i = 1, 2, 3)$ are known functions defined on $\partial\Omega_D$ which are usually piecewise constant.

2.1.1 The Physical Parameters

The mobilities μ_n and μ_p are usually functions of n, p, N_A, N_D and the electric field $\boldsymbol{E}$. For the sake of simplicity, we assume that the mobilities are constant in Ω given by

$$\begin{aligned} \mu &= \mu_n^0 \Big(\frac{T}{300K}\Big)^{-\gamma_n} \\ \mu_p &= \mu_p^0 \Big(\frac{T}{300K}\Big)^{-\gamma_p} \end{aligned}$$

where μ_n^0, μ_p^0, γ_n and γ_p are constants determined by experiment, and T is the absolute temperature.

The recombination–generation R usually consists of three parts (cf [45, 71, 90]), *i.e.*

$$R = R_{SRH} + R_{Aug} - G$$

The first term is Shockley–Read–Hall type recombination given by

$$R_{SRH} = \frac{np - n_i^2}{\tau_n(p + n_i)(np - n_i^2)}$$

where τ_n and τ_p are the lifetimes for electrons and holes respectively, which typically lie in the range from $100ns$ to $5\mu s$. The term R_{Aug} is the net Auger recombination rate given by

$$R_{Aug} = (C_n n + C_p p)(np - n_i^2)$$

where C_n and C_p are constants in Ω. The last term G is the generation by the impact ionisation and is of the form:

$$G = \frac{1}{q}(\alpha_n |\boldsymbol{J}_n| + \alpha_p |\boldsymbol{J}_n|)$$

where α_n, α_p are ionisation rates given by (cf. [45, 81]):

$$\begin{aligned} \alpha_n &= A_n \exp\Big[- \Big(\frac{E_n^{crit}}{|\boldsymbol{E}|} \Big)^{\beta_n} \Big] \\ \alpha_p &= A_p \exp\Big[- \Big(\frac{E_p^{crit}}{|\boldsymbol{E}|} \Big)^{\beta_p} \Big] \end{aligned}$$

Here A_n, A_p, E_n^{crit}, E_p^{crit}, β_n and β_p are constants. Obviously α_n and α_p strongly depend on the electric field $\boldsymbol{E}$.

Among the three types of recombination–generation rates, R_{SRH} is the most fundamental one. R_{Aug} becomes dominant as the carrier concentrations increase. G is assumed to be present only in a high electric field. In practice, we take account of G if at least one junction is reversed biased. When the condition of thermal equilibrium holds, we have $R = 0$. Therefore, we assume $R = 0$ when conditions are close to thermal equilibrium.

The constants mentioned above are listed in Table 2.1.1

2.1.2 Scaling

The ranges of value of the solutions to (2.1.12–2.1.16) lie in a very large interval. The solution n and p may change by many orders of magnitudes in a small subregion of Ω. Therefore, we usually scale (2.1.12–2.1.16) before solving them numerically so that the solutions of the scaled equations behave more moderately in Ω. This procedure can be fulfilled by the following transformations:

1. $x \longrightarrow \frac{x}{l_0}$, $\qquad x \in \Omega$

Constant	Value	Constant	Value
q	1.6×10^{-10}	C_n	$7 \times 10^{-32}\text{cm}^6 s^{-1}$
ε_0	$8.86 \times 10^{-14} F/\text{cm}$	C_p	$10^{-32} cm^6 s^{-1}$
k	$1.38 \times 10^{-23} \boldsymbol{J} \ / \ K$	A_n	10^6cm^{-1}
μ_n^0	$1448\ \text{cm}^2 V^{-1} s^{-1}$	A_p	$1.5 \times 10^6 \text{cm}^{-1}$
μ_p^0	$437\ \text{cm}^2 V^{-1} s^{-1}$	E_n^{crit}	$1.4 \times 10^6 V \ / \ \text{cm}$
γ_n	2.33	E_p^{crit}	$1.8 \times 10^6 V \ / \ \text{cm}$
γ_p	2.33	β_n	1
τ_n	$4 \times 10^{-5} s$	β_p	1
τ_p	$2 \times 10^{-5} s$		

Table 2.1.1: Typical values of physical constant

Quality	Symbol	Value	Quality	Symbol	Value
x	l_0	$\sqrt{\varepsilon kT/(q^2 n_i)}$	ψ	ψ_0	kT/q
$n,\ p,\ N,\ n_i$	N_0	n_i	$D_n,\ D_p$	D_0	$1\text{cm}^2 s^{-1}$
$\mu_n,\ \mu_p$	μ_0	D_0/ψ_0	R		$D_0 N_0/l_0^2$
$\boldsymbol{J}_n,\ \boldsymbol{J}_p$		$qD_0 N_0/l_0^2$		λ^2	1

Table 2.1.2: De Mari scaling factors

2. $\psi \longrightarrow \frac{\psi}{\psi_0}$

3. $n \longrightarrow \frac{n}{N_0}, \qquad p \longrightarrow \frac{p}{N_0}, \qquad N \longrightarrow \frac{N}{N_0}$

4. $D_n \longrightarrow \frac{D_n}{D_0}, \qquad D_p \longrightarrow \frac{D_p}{D_0}, \qquad \mu_n \longrightarrow \frac{\mu_n \psi_0}{D_0}, \qquad \mu_p \longrightarrow \frac{\mu_p \psi_0}{D_0}$

where l_0, ψ_0, D_0 and μ_0 are scaling factors. De Mari (cf. [23]) presented a set of scaling factors as listed in Table 2.1.2. Another choice of the scaling factors, called singular perturbation scaling factors, was introduced by Vasileva and Butuzov (cf. [108]) and further developed by Markowich *et al* (cf. [54]). Under this scaling the equations are singularly perturbed. These scaling factors are listed in Table 2.1.3.

After scaling, (2.1.12)–(2.1.16) become the following set of equations:

$$\lambda^2 \nabla^2 \psi = n - p - N, \tag{2.1.17}$$

$$\nabla \cdot (\nabla n - n \nabla \psi) = R(\psi, n, p), \tag{2.1.18}$$

$$\nabla \cdot (\nabla p + p \nabla \psi) = R(\psi, n, p) \tag{2.1.19}$$

Quality	Symbol	Value
x	l_0	diameter (Ω)
ψ	ψ_0	kT/q
$n,\ p,\ N,\ n_i$	N_0	$\max_{x\in\Omega}\lvert N(x)\rvert$
$D_n,\ D_p$	D_0	$\max_{x\in\Omega}(D_n(x),\ D_p(x))$
$\mu_n,\ \mu_p$	μ_0	D_0/ψ_0
R		D_0N_0/l_0^2
$\boldsymbol{J}_n,\ \boldsymbol{J}_p$		qD_0N_0/l_0^2
	λ^2	$\varepsilon\psi_0/(l_0^2qN_0)$

Table 2.1.3: Singular perturbation scaling factors

with boundary conditions

$$\psi|_{\partial\Omega_D} = \gamma_1'(x), \quad n|_{\partial\Omega_D} = \gamma_2'(x), \quad p|_{\partial\Omega_D} = \gamma_3'(x), \tag{2.1.20}$$

and

$$\nabla\psi\cdot\boldsymbol{n}|_{\partial\Omega_N} = \nabla n\cdot\boldsymbol{n}|_{\partial\Omega_N} = \nabla p\cdot\boldsymbol{n}|_{\partial\Omega_N} = 0 \tag{2.1.21}$$

where $\gamma_i'(x)$ $(i=1,2,3)$ are scaled forms of $\gamma_i(x)$ $(i=1,2,3)$ defined on $\partial\Omega_D$.

2.1.3 The Finite Dimensional Parameter Estimation Problem

A typical 2D p-n diode with two ohmic contacts is depicted in Figure 2.1.1 where the interior curve, called the p-n junction, represents the interface of the p and n regions. We denote this curve by C. In the rest of this chapter, we assume that the doping function N is a step function of the form

$$N = \begin{cases} a & \text{in } p\text{-region} \\ -b & \text{in } n\text{-region} \end{cases} \tag{2.1.22}$$

where a and b are constants in the range from 10^{10} to 10^{20}. Note that N is also a function of the p-n junction C. Given N, the dimensions of a device and the applied bias V, we can solve (2.1.17)–(2.1.21) numerically using an appropriate numerical method and then evaluate the terminal current $I(v)$ flowing in or out of the device (cf. [65, 66]). The problem can be solved for various biases V to obtain the V-I characteristic for a given device.

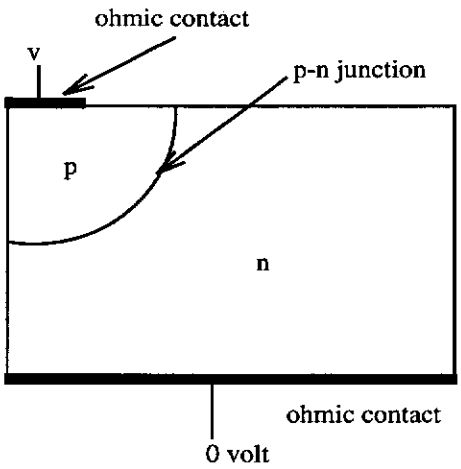


Figure 2.1.1: A typical 2 dimensional diode

In the conventional design cycle of a device, the above process will be repeated using different doping functions N and different geometries until the V-I characteristic matches the required one within a given error range. This approach is time-consuming, because after each iteration in the design cycle, the new values of the parameters to be used in the next iteration are chosen empirically from previous experience which may be far away from the optimal choice. In this chapter the problem is posed as the following optimization problem: given an ideal V-I characteristic curve $I_g(v)$, find some physical and/or geometric parameters such that the V-I characteristic of the device matches the ideal one optimally with respect to a specified performance criterion. This problem is formulated as a nonlinear optimization problem, and the cost function of the problem consists of two competing quadratic terms with penalty parameters. By a judicious choice of these parameters, one can balance the competing costs. Various existing efficient numerical methods are proposed for the numerical solution of this nonlinear optimization problem. Due to the large variations in the parameters and in the solutions to the semiconductor device equations, the Jacobian of the optimization problem is ill-conditioned. To overcome this difficulty, a scaling technique is proposed to balance the entries of the Jacobian so that the problem is numerically stable. This approach is applied to

some test problems and all the numerical results confirm the usefulness of the approach. To our best knowledge, this approach has not been used in semiconductor device design, though, in practice, it will dramatically reduce the time required in the design cycle of a semiconductor device.

2.2 The Formulation of the Problem

We consider the formulation of the semiconductor device parameter design. For simplicity we assume hereafter that the geometry of a device is rectangular or brick, and thus the device region is $\overline{\Omega} = [0, L_x] \times [0, L_y]$ in two dimensions or $\overline{\Omega} = [0, L_x] \times [0, L_y] \times [0, L_z]$ in three dimensions. We also assume that the p-n junction of the device consists of line segments or facets parallel to one of the axes. In this case, the p-n junction is uniquely determined by its intercept on each of the axes in the interval $[0, L_x]$, $[0, L_y]$ or $[0, L_z]$. Let c_x denote the ratio of the intercept on the x-axis and L_x, c_y the ratio of the intercept on the y-axis and L_y, and c_z the ratio of the intercept on the z-axis and L_z. Then, the p-n junction is uniquely determined by c_x, c_y and c_z. Using this notation, we formulate the above parameter selection problem in semiconductor design as the following optimization problem:

Problem 2.1: *given an ideal V-I characteristic function $I_g(v)$ on $[0, V_{\max}]$, find the doping parameters a, b, and the geometrical parameters L_x, L_y , L_z, c_x, c_y and c_z such that*

$$F(a, b, c_x, c_y, c_z, L_x, L_y, L_z) = \alpha \int_0^{V_{\max}} (I(v) - I_g(v))^2 dv + \beta(L_x^2 + L_y^2 + L_z^2)$$

is minimized subject to the bound constraints

$$10^{10} \leq a \leq 10^{20}, \tag{2.2.1}$$

$$10^{10} \leq b \leq 10^{20}, \tag{2.2.2}$$

$$c_x^{\min} \leq c_x \leq c_x^{\max}, \tag{2.2.3}$$

$$c_y^{\min} \leq c_y \leq c_y^{\max}, \tag{2.2.4}$$

$$c_z^{\min} \leq c_z \leq c_z^{\max}, \tag{2.2.5}$$

$$L_x^{\min} \leq L_x \leq L_x^{\max}, \tag{2.2.6}$$

$$L_y^{\min} \leq L_y \leq L_y^{\max}, \tag{2.2.7}$$

$$L_z^{\min} \leq L_z \leq L_z^{\max}, \tag{2.2.8}$$

where L_x, L_y and L_z are the length, width and height of the device, c_x, c_y and c_z are the ratios defined before, α, β are two positive constants, and I is the terminal current.

This is a continuous least squares problem and the cost function F contains two competing performance criteria. This is because $L_x = L_y = L_z = 0$ is the obvious optimal solution for the second term of the cost. Thus, we need to choose α and β properly to balance the two terms in F. In practice, Problem 2.1 can be approximated by taking a set of appropriate sampling points $v_i, i = 1, 2, \cdots, m$, in $[0, V_{\max}]$, leading to

Problem 2.2: *Given an ideal V-I characteristic function $I_g(v)$ on $[0, V_{\max}]$, find the doping parameters a, b, and the geometrical parameters c_x, c_u, c_z, L_x, L_y and L_z such that*

$$E(\theta) = (\boldsymbol{I} - \boldsymbol{I}_g)^T A (\boldsymbol{I} - \boldsymbol{I}_g) + \beta(L_x^2 + L_y^2 + L_z^2) \tag{2.2.9}$$

is minimized subject to (2.2.1)-(2.2.8), where

$$\begin{aligned} \boldsymbol{I} &= (I(v_1, \theta), I(v_2, \theta), \ldots, I(v_m, \theta)), \\ \boldsymbol{I}_g &= (I_g(v_1), I_g(v_2), \ldots, I_g(v_m)), \\ \theta &= (a, b, c_x, c_y, c_z, L_x, L_y, L_z) \end{aligned}$$

and $A = \mathrm{diag}(\alpha_i)$ is a diagonal matrix.

Here $\alpha_i > 0$ $(i = 1, 2, \cdots, m)$ and $\beta > 0$ are weights to be chosen later.

Problem 2.2 is a nonlinear optimization problem with only bound constraints. The nonlinear differential equations (2.1.17)–(2.1.19) do not appear explicitly in the formulation, but the current $\boldsymbol{I}$ depends on these equations through the expression (2.1.6). The dependence of the cost function on the parameter θ and the applied bias v is complicated, and thus the solvability of Problem 2.2 is theoretically difficult. (Even the solvability of the nonlinear PDE system (2.1.17)–(2.1.19) is still an open problem unless under some restrictive assumptions (cf., for example, [54])). However, from our computational experience, Problem 2.2 is computable, though local minima may exist, as will be seen in Section 2.4.

2.3 The Numerical Methods

We now consider the numerical solution of Problem 2.2.

Starting from an initial guess θ_0, Problem 2.2 can be solved iteratively. At each step an increment $\delta\theta_i$ is calculated such that

$$E(\theta_i + \delta\theta_i)$$

is minimized with respect to $\delta\theta_i$, where θ_i and $\delta\theta_i$ are the *i-th* approximation and *i-th* increment of θ respectively. The iterative procedure continues until the relative error $\frac{\|\boldsymbol{I}-\boldsymbol{I}_g\|_2}{\|\boldsymbol{I}_g\|_2}$ and the change in (L_x, L_y, L_z) in the Euclidean norm $\| L \|_2$ between two consecutive iterations are smaller than a given tolerance.

The Gauss-Newton Method

To calculate the increment $\delta\theta_i$ at each step, the Gauss-Newton Method is used. Let

$$\hat{\boldsymbol{I}} = (I(v_1, \theta), I(v_2, \theta), \ldots, I(v_m, \theta), L_x, L_y, L_z)^T,$$
$$\hat{\boldsymbol{I}}_g = (I_g(v_1), I_g(v_2), \ldots, I_g(v_m), 0, 0, 0)^T,$$
$$B = \mathrm{diag}(\alpha_1, \cdots, \alpha_m, \beta, \beta, \beta).$$

Then

$$E(\theta) = (\hat{\boldsymbol{I}} - \hat{\boldsymbol{I}}_g)^T B(\hat{\boldsymbol{I}} - \hat{\boldsymbol{I}}_g).$$

Let $\hat{\boldsymbol{I}}^i = \hat{\boldsymbol{I}}(\theta_i)$, then Taylor's formula for vector valued functions gives

$$\hat{\boldsymbol{I}} = \hat{\boldsymbol{I}}^i + J_i\delta\theta_i + \frac{1}{2}\{\delta\theta_i^T G_1 \delta\theta_i, \cdots, \delta\theta_i^T G_{m+3}\delta\theta_i\}^T \tag{2.3.1}$$

where

$$J_i = \begin{pmatrix} \frac{\partial \boldsymbol{I}(v_1)}{\partial a_i} & \frac{\partial \boldsymbol{I}(v_1)}{\partial b_i} & \frac{\partial \boldsymbol{I}(v_1)}{\partial c_{x_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial c_{y_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial c_{z_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial L_{x_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial L_{y_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial L_{z_i}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{I}(v_m)}{\partial a_i} & \frac{\partial \boldsymbol{I}(v_m)}{\partial b_i} & \frac{\partial \boldsymbol{I}(v_m)}{\partial c_{x_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial c_{y_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial c_{z_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial L_{x_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial L_{y_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial L_{z_i}} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and the matrix G_j denotes the Hessian of $\hat{I}_j^i$ evaluated at $\theta_i + r\delta\theta_i$ with $0 \leq r \leq 1$. Omitting the second order term in (2.3.1), we have

$$\hat{\boldsymbol{I}} = \hat{\boldsymbol{I}}^i + J_i\delta\theta_i$$

when $\delta\theta_i$ is small. Using this, $E(\theta_i + \delta\theta_i)$ can be approximated by

$$\begin{aligned} E(\theta_i + \delta\theta_i) &= (\hat{\boldsymbol{I}}^i + J_i\delta\theta_i - \hat{\boldsymbol{I}}_g)^T B(\hat{\boldsymbol{I}}^i + J_i\delta\theta_i - \hat{\boldsymbol{I}}_g) \\ &= (\hat{\boldsymbol{I}}^i - \hat{\boldsymbol{I}}_g)^T B(\hat{\boldsymbol{I}}^i - \hat{\boldsymbol{I}}_g) + (\hat{\boldsymbol{I}}^i - \hat{\boldsymbol{I}}_g)^T B J_i\delta\theta_i \\ &+ (J_i\delta\theta_i)^T B(\hat{\boldsymbol{I}}^i - \hat{\boldsymbol{I}}_g) + (J_i\delta\theta_i)^T B(J_i\delta\theta_i). \end{aligned}$$

This is a quadratic form in $\delta\theta_i$, and the minimum point $\delta\theta_i^*$ of this quadratic function satisfies

$$\nabla E(\theta_i + \delta\theta_i^*) = 0,$$

which leads to

$$(J_i^T B J_i)\delta\theta_i^* = -J_i^T B(\hat{\boldsymbol{I}} - \hat{\boldsymbol{I}}_g). \tag{2.3.2}$$

The solution to (2.3.2) defines the *i-th* search direction called the Gauss-Newton direction.

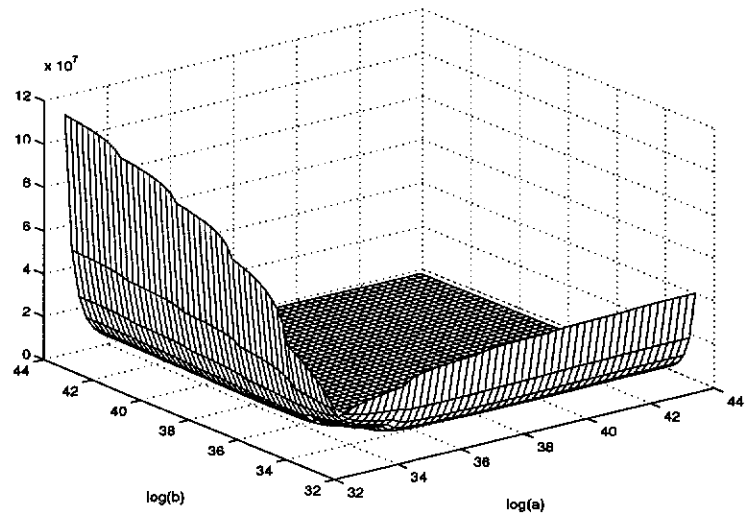


Figure 2.3.1: The cost function against doping concentrations a and b.

The Levenberg-Marquardt Method

From computational studies we see that the partial derivatives of the cost function E with respect to the parameters a and b in the doping function are large when a and b are close to their lower bounds and small when a and b are close to their upper bound (see. Figure 2.3.1). Thus, initial guesses for a and b are always chosen to be their lower bound 10^{10}, and at the first few iterations, we also need to restrict the step size to avoid oscillations. Mathematically, this can be formulated as

$$\begin{aligned} \min \quad & E(\theta_i + \delta\theta_i) \\ \text{subject to} \quad & (\delta\theta_i)^T(\delta\theta_i) \leq \Delta, \end{aligned}$$

where Δ is a positive constant. The above inequality constraint can be added to the cost function to form

$$\min E(\theta_i + \delta\theta_i) - \lambda\{\Delta - (\delta\theta_i)^T(\delta\theta_i)\}$$

where $\lambda > 0$ is a penalty parameter called the Marquardt parameter. The optimal point $\delta\theta_i^*$ of this problem is given by

$$(J_i^T B J_i + \lambda I)\delta\theta_i^* = -J_i^T B(\hat{\boldsymbol{I}} - \hat{\boldsymbol{I}}_g),$$

where I denotes the unit matrix. This method is called the Levenberg-Marquardt method (cf. [51, 53]) which is a combination of the Gauss-Newton Method and the Steepest Descent Method (cf. [82]). The Marquardt parameter λ can be chosen properly to avoid unbounded oscillation in the re-estimation procedure. The same technique can also be applied to the decision variables a and b in the original cost function (2.2.9). This yields a cost function

$$\hat{E}(\theta) = (\boldsymbol{I} - \boldsymbol{I}_g)^T A(\boldsymbol{I} - \boldsymbol{I}_g) + \beta(L_x^2 + L_y^2 + L_z^2) + \gamma(a^2 + b^2)$$

corresponding to (2.2.9).

Let

$$\begin{aligned}
\hat{\hat{\boldsymbol{I}}} &= (I(v_1,\theta), I(v_2,\theta), \ldots, I(v_m,\theta), L_x, L_y, L_z, a, b)^T, \\
\hat{\hat{\boldsymbol{I}}}_g &= (I_g(v_1), I_g(v_2), \ldots, I_g(v_m), 0, 0, 0, 0, 0)^T, \\
\hat{B} &= \mathrm{diag}(\alpha_1, \cdots, \alpha_m, \beta, \beta, \beta, \gamma, \gamma).
\end{aligned}$$

Then $\hat{E}(\theta)$ can be rewritten as

$$\hat{E}(\theta) = (\hat{\hat{\boldsymbol{I}}} - \hat{\hat{\boldsymbol{I}}}_g)^T \hat{B} (\hat{\hat{\boldsymbol{I}}} - \hat{\hat{\boldsymbol{I}}}_g).$$

The corresponding Levenberg-Marquardt correction $\delta\theta^*$ satisfies

$$(\hat{J}_i^T \hat{B} \hat{J}_i + \lambda I) \cdot \delta\theta_i^* = -\hat{J}_i^T \hat{B} (\hat{\hat{\boldsymbol{I}}} - \hat{\hat{\boldsymbol{I}}}_g), \tag{2.3.3}$$

where

$$\hat{J}_i = \begin{pmatrix}
\frac{\partial \boldsymbol{I}(v_1)}{\partial a_i} & \frac{\partial \boldsymbol{I}(v_1)}{\partial b_i} & \frac{\partial \boldsymbol{I}(v_1)}{\partial c_{x_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial c_{y_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial c_{z_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial L_{x_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial L_{y_i}} & \frac{\partial \boldsymbol{I}(v_1)}{\partial L_{z_i}} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{\partial \boldsymbol{I}(v_m)}{\partial a_i} & \frac{\partial \boldsymbol{I}(v_m)}{\partial b_i} & \frac{\partial \boldsymbol{I}(v_m)}{\partial c_{x_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial c_{y_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial c_{z_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial L_{x_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial L_{y_i}} & \frac{\partial \boldsymbol{I}(v_m)}{\partial L_{z_i}} \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$

Using the correction $\delta\theta_i^*$, we update $\hat{\hat{\boldsymbol{I}}}$ by

$$\hat{\hat{\boldsymbol{I}}} = \hat{\hat{\boldsymbol{I}}}^i + \hat{J}_i \delta\theta_i^* \tag{2.3.4}$$

The values of the elements in $\hat{J}_i$ may differ from each other by several orders of magnitudes. This may cause some stability problems in real computation. To avoid this, we scale $\hat{J}_i$ by a diagonal matrix M with positive diagonal entries so that $\hat{J}_i$ has more balanced entries. The scaled equation corresponding to (2.3.3) is

$$(\hat{\hat{J}}_i^T \hat{B} \hat{\hat{J}}_i + \lambda I)\delta\hat{\theta}_i^* = -\hat{\hat{J}}_i^T \hat{B}(\hat{\hat{I}} - \hat{\hat{I}}_g),$$

where

$$\hat{\hat{J}}_i = \hat{J}_i M \quad \text{and} \quad \delta\hat{\theta}_i^* = M^{-1}\delta\theta_i^*. \tag{2.3.5}$$

The updating formula corresponding to (2.3.4) then becomes

$$\hat{\hat{I}} = \hat{\hat{I}}^i + \hat{\hat{J}}_i \delta\hat{\theta}_i^*$$

We comment that all the partial derivatives in $\hat{J}_i$ are approximated by forward finite differences. More specifically,

$$\frac{\partial \boldsymbol{I}(v_i, \theta)}{\partial \theta_j} \simeq \frac{\boldsymbol{I}(v_i, \theta + h \cdot \boldsymbol{e}_j) - \boldsymbol{I}(v_i, \theta)}{h}, \tag{2.3.6}$$

where $\boldsymbol{e}_j$ is a unit vector and h is a small positive increment.

We also comment that the above method is based on the assumption that the independent variables, $\boldsymbol{v} = \{v_i, 1 \leq i \leq m\}$, do not contain any observation errors. This is because these variables are normally not obtained from an experimental observation. In the case that $\boldsymbol{v} = \{v_i, 1 \leq i \leq m\}$ does contain observation errors, the Orthogonal Distance Regression method may be used, instead of the above Gauss-Newton or Levenberg-Marquardt method. For details of the Orthogonal Distance Regression method, we refer to [8].

Gummel's Method

The solvability of the nonlinear system (2.1.17)–(2.1.19) in general is a long-standing open problem, but in the case that the $R = 0$ and the applied bias is close to zero, it can be shown that the system is uniquely solvable (cf., for example, [54]). In practice, this nonlinear system can be solved iteratively by Gummel's method [37] defined as follows:

1. Given an initial value (ψ^0, n^0, p^0) let $k = 0$.

2. Solve the following system sequentially for $(\psi^{k+1}, n^{k+1}, p^{k+1})$

$$\nabla^2\psi^{k+1} = n^k - p^k - N,$$
$$\nabla \cdot (\nabla n^{k+1} - n^{k+1}\nabla\psi^{k+1}) - R(\psi^{k+1}, n^k, p^k) = 0,$$
$$\nabla \cdot (\nabla p^{k+1} + p^{k+1}\nabla\psi^{k+1}) - R(\psi^{k+1}, n^{k+1}, p^k) = 0,$$

   with appropriate boundary conditions.

3. Test for convergence. If failed, increase k and repeat step 2.

In step 2 of the above algorithm we deal with equations of the form

$$-\nabla \cdot (\nabla u - \boldsymbol{c}u) + Gu = F \qquad \text{in } \Omega \tag{2.3.7}$$
$$u\,|_{\partial\Omega_D} = 0, \quad (\nabla u - \boldsymbol{c}u) \cdot \boldsymbol{n}\,|_{\partial\Omega_N} = 0. \tag{2.3.8}$$

This problem can be solved effectively by the exponentially fitted finite volume method proposed in [66] and [67]. We now give a brief account of this method in two dimensions.

The Exponentially Fitted Finite Volume Scheme

To discuss the exponential fitted finite volume scheme (cf. [65], [67]), we first define some meshes on Ω. Let T be any partition of $\overline{\Omega}$ by a set of triangles. Let $X = \{x_i\}_1^N$ be the set of all vertices of T and $E = \{e_i\}_1^M$ the set of edges of T. Without loss of generality we assume that the nodes in X and the edges in E are numbered such that $X' = \{x_i\}_1^{N'}$ and $E' = \{e_i\}_1^{M'}$ are respectively the set of nodes in X not on $\partial\Omega_D$ and the set of edges in E not on $\partial\Omega_D$.

DEFINITION 2.3.1. T is a Delaunay mesh if, for every $t \in T$, the circumcircle of the element contains no other vertices in X (cf. [22]).
We assume henceforth that T is a Delaunay triangulation.

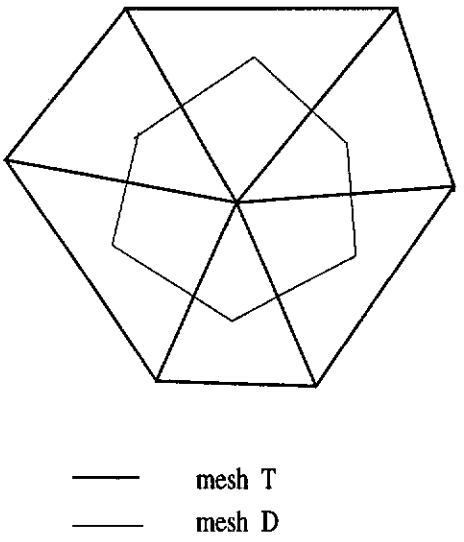


Figure 2.3.2: Part of a Delaunay mesh T and dual Dirichlet tessellation D.

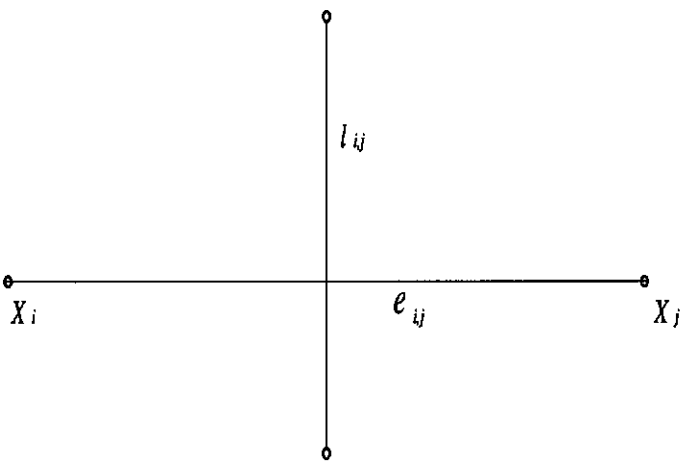


Figure 2.3.3: Notation for edges and nodes.

DEFINITION 2.3.2. The Dirichlet tessellation D, corresponding to the triangulation T is defined by $D = \{d_i\}_1^N$ where the tile

$$d_i = \{x \in \Omega : |x - x_i| < |x - x_j|, x_j \in X, j \neq i\}$$

for all $x_i \in X$ (cf. [25]).

We remark that for each $x_i \in X$, the boundary ∂d_i of the tile d_i is the polygon having as its vertices the circumcentres of all triangles with common vertex x_i. Each segment of ∂d_i is perpendicular to one of the edges sharing the vertex x_i (see Figure 2.3.2). The Dirichlet tessellation D is a polygonal mesh dual to the Delaunay mesh T. For each $i = 1, 2, \cdots, N'$, integrating (2.3.7) – (2.3.8) over d_i and applying Green's formula to the first term we have

$$-\int_{\partial d_i} (\nabla u - \boldsymbol{c}u) \cdot \boldsymbol{n} ds + \int_{d_i} G u d\Omega = \int_{d_i} F d\Omega.$$

For $i = 1, 2, \cdots, N'$, let u_i be the approximate value of $u(x)$ at x_i. Using the one-point quadrature rule we have from the above

$$-\int_{\partial d_i} (\nabla u - \boldsymbol{c}u) \cdot \boldsymbol{n} ds + G_i u_i |d_i| = F_i |d_i| \tag{2.3.9}$$

where $G_i = G(x_i)$ and $F_i = F(x_i)$. We now consider the approximation of the first term in (2.3.9). Let $I_i = \{j : e_{i,j} \in E\}$ denote the index set of neighboring nodes of x_i, where $e_{i,j}$ denotes the edge joining x_i and x_j, as shown in Figure 2.3.3. Since ∂d_i is polygonal and each of its sides is perpendicular to one of the edges joining x_i, we have

$$\int_{\partial d_i} (\nabla u - \boldsymbol{c}u) \cdot \boldsymbol{n} ds = \sum_{j \in I_i} \left(\int_{l_{i,j}} (\nabla u - \boldsymbol{c}u) \cdot \boldsymbol{e}_{i,j} ds \right). \tag{2.3.10}$$

where $l_{i,j}$ denotes the segment of ∂d_i perpendicular to the edge $e_{i,j}$ and is oriented counterclockwise with respect to x_i (see Figure 2.3.3) and $\boldsymbol{e}_{i,j}$ denotes the unit vector from x_i to x_j. For any $j \in I_i$ we now consider the two-point boundary value problem

$$\nabla (\nabla u \cdot \boldsymbol{e}_{i,j} - c_{i,j} u) \cdot \boldsymbol{e}_{i,j} = 0 \qquad \text{on} \quad e_{i,j} \tag{2.3.11}$$

$$u(x_i) = u_i, \quad u(x_j) = u_j \tag{2.3.12}$$

where $c_{i,j}$ is a constant approximation to $\boldsymbol{c} \cdot \boldsymbol{e}_{i,j}$ on $e_{i,j}$. Solving this equation analytically we obtain

$$f_{i,j} \equiv \nabla u \cdot \boldsymbol{e}_{i,j} - c_{i,j} u = \frac{1}{|e_{i,j}|} \left(B(c_{i,j}|e_{i,j}|)u_j - B(-c_{i,j} \mid e_{i,j} \mid)u_i \right) \qquad (2.3.13)$$

where $B(x)$ is the Bernoulli function defined by

$$B(x) = \begin{cases} \frac{x}{e^x - 1} & \text{if } x \neq 0 \,, \\ 1 & \text{if } x = 0. \end{cases} \qquad (2.3.14)$$

Obviously $f_{i,j}$ defines a constant approximation to the integrand on the right side of (2.3.11). Furthermore the solution of (2.3.10) also defines a piecewise exponential approximation to the solution of (2.3.7) on $e_{i,j}$. Substituting (2.3.13) into (2.3.10) and the result into (2.3.9) we obtain

$$\sum_{j \in I_i} \frac{|l_{i,j}|}{|e_{i,j}||d_{i,j}|} (B(-c_{i,j}|e_{i,j}|)u_i - B(c_{i,j}|e_{i,j}|)u_j) + G_i u_i = F_i \qquad (2.3.15)$$

for all $i = 1, 2, \cdots, N'$. In matrix form, we have

$$(E + D)\boldsymbol{U} = \boldsymbol{F}$$

where E and D denote the matrices corresponding to, respectively, the first and second terms of (2.3.15), $\boldsymbol{U} = (u_1, u_2, \cdots, u_{N'})^T$ and $\boldsymbol{F} = (f_1, f_2, \cdots, f_{N'})^T$. The matrix $E + D$ is unsymmetric unless $c_{i,j} = 0$ for all $i = 1, 2, \cdots, N'$ and all $j \in I_i$. However, it is easy to verify that E is an M-matrix (cf. [66]).

We comment that the above method can be extended to three dimensions easily. A detailed discussion of this can be found, for example, in [29]. We also comment that using the numerical solution from the above discretization scheme, we can evaluate the currents flowing in or out of a device, based on (2.1.6). For detailed discussions of this, we refer to [65, 66, 29].

2.4 Numerical Experiments

The numerical methods described in the previous section are applied to some two and three dimensional test problems. All computations were performed in Fortran double precision on a Unix Workstation. In what follows, we use

objective characteristic to denote the discrete I-V characteristic generated by directly solving the equations (2.1.17)–(2.1.18) using a given parameter set. This given parameter set is referred to as the **ideal solution**, and the solution to Problem 2.2 using the objective characteristic is referred to as the **optimal solution**. In all the examples below, the bounds in the constraints (2.2.6)–(2.2.8) are chosen to be

$$L_x^{\min} = L_y^{\min} = L_z^{\min} = 2\mu m, \quad L_x^{\max} = L_y^{\max} = L_z^{\max} = 15\mu m.$$

Also, the increment h in the finite difference approximation (2.3.6) is chosen to be 10^5 for the doping parameters a and b and $10^{-9}\mu m$ for L_x, L_y and L_z. The units for the doping concentration parameters a and b and the geometric parameters L_x, L_y and L_z are respectively $1/cm^2$ (or $1/cm^3$ in 3D) and μm. However, these are omitted below for brevity.

Example 2.1. A two dimensional diode

A two dimensional rectangular diode with width L_x and height L_y, depicted in Figure 2.4.1, is chosen to be our first test problem. The p-region of the device is rectangular with width $\frac{3}{10}L_x$ and height $\frac{3}{10}L_y$, as shown in Figure 2.4.1. (Correspondingly, $c_x = c_y = \frac{7}{10}$ in (2.2.3) and (2.2.4).) The parameters α_i, $(i = 1, \cdots, m)$, β and γ are chosen to be 10^{26}, 10^{10}, and 10^{-18}, respectively, and the scaling matrix M in (2.3.5) is chosen to be diag(10^5, 10^5, 10^{-9}, 10^{-9}). This example is studied numerically in the following two different situations.

Case 1: Decision variables: a, b and $L_x = L_y$.

In this case, we assume that the device is a square and the doping function is piecewise constant. Thus, we are to determine three parameters: a, b in the doping function, and the dimension of the device $L_x = L_y$. Two different discrete V-I characteristics $I_{g,1}$ and $I_{g,2}$, listed in Tables 2.4.1 and 2.4.2 respectively, are used as the objective characteristics. These objective characteristics are generated by solving the equations 2.1.17-2.1.19 using the parameter sets

$$(a = 10^{14}, b = 10^{14}, L_x = 10, L_y = 10) \qquad \text{and}$$
$$(a = 10^{16}, b = 10^{16}, L_x = 10, L_y = 10)$$

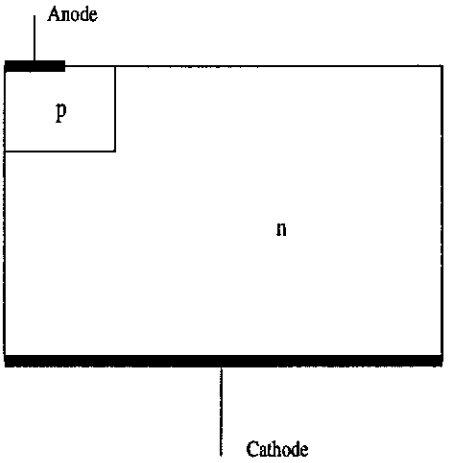


Figure 2.4.1: A 2 dimensional rectangular diode

V	0.2143	0.2857	0.3571	0.4286	0.5
I	$6.490 \cdot 10^{-11}$	$8.771 \cdot 10^{-10}$	$1.139 \cdot 10^{-8}$	$1.073 \cdot 10^{-7}$	$4.837 \cdot 10^{-7}$

Table 2.4.1: The V-I Characteristic $I_{g,1}$

respectively for the applied forward biases listed in Tables 2.4.1 and 2.4.2.

To solve the optimization problems, we choose the following stopping criteria:

1. The relative error $\frac{\|\boldsymbol{I}-\boldsymbol{I}_g\|_2}{\|\boldsymbol{I}_g\|_2} \leq 10^{-4}$ (10^{-11} for Case 1, using $I_{g,1}$ and 10^{-6} for Case 2, using $I_{g,1}$), and

2. The difference in $(L_x^2 + L_y^2)^{1/2}$ between two consecutive iterates is smaller than 10^{-3}.

The solution procedure stops when both of the above criteria are satisfied. The results obtained for various initial values are listed in Tables 2.4.3 and 2.4.4. From the tables we see that the relative error is always smaller than 10^{-4}. From the table we also see that the optimal solutions are very accurate for $I_{g,1}$, and are reasonably close to the ideal solution for $I_{g,2}$.

Case 2: Decision variables a, b, L_x and L_y.

This case differs from Case 1 in the way that we do not assume that $L_x = L_y$.

V	0.2143	0.2857	0.3571	0.4286	0.5
I	$4.196 \cdot 10^{-13}$	$6.618 \cdot 10^{-12}$	$1.042 \cdot 10^{-10}$	$1.640 \cdot 10^{-9}$	$2.575 \cdot 10^{-8}$

Table 2.4.2: The V-I Characteristic $I_{g,2}$

Initial Values				Optimal Solution				No.iter.	Rel.err.
a	b	L_x	L_y	a	b	L_x	L_y		
1e10	1e10	15	15	1e14	1e14	10	10	23	5.9e-13
1e11	1e11	15	15	1e14	1e14	10	10	22	1.7e-13
1e12	1e12	15	15	1e14	1e14	10	10	24	8.8e-14
1e13	1e13	15	15	1e14	1e14	10	10	21	7.9e-13
1e12	1e12	8	8	1e14	1e14	10	10	19	8.7e-12

Table 2.4.3: The results for Example 2.1, Case 1, using V-I characteristic $I_{g,1}$

Thus, there are four independent parameters to be determined. The problem is solved using the objective characteristics listed in Tables 2.4.1 and 2.4.2, the same criterion as in Case 1 is used in this case, and the results are listed in Tables 2.4.5 and 2.4.6.

Example 2.2. A three dimensional diode

The second test problem is chosen to be a three dimensional rectangular prismatic diode with width L_x, depth L_y and height L_z. The p-region is also chosen to be a rectangular prism with width $\frac{3}{10}L_x$, depth $\frac{3}{10}L_y$ and height $\frac{3}{10}L_z$. (Correspondingly, the constraints (2.2.3)–(2.2.5) are replaced by $c_x = c_y = c_z = 7/10$.) For simplicity, we assume that $L_x = L_y$. Thus the decision variables are a, b in the doping function, and the dimensions of the device $L_x = L_y$ and L_z. The parameters α_i, $(i = 1, \cdots, m)$, β and γ are chosen

Initial Values				Optimal Solution				No.iter.	Rel.err.
a	b	L_x	L_y	a	b	L_x	L_y		
1e12	1e12	15	15	9.8e15	1.0e16	9.4	9.4	57	5.5e-5
1e13	1e13	15	15	1.1e16	3.1e16	6.5	6.5	40	2.9e-5
1e11	1e11	15	15	5.3e15	7.1e16	12	12	65	1.5e-5
1e10	1e10	15	15	8.1e15	1.3e16	10.3	10.3	100	2.4e-5

Table 2.4.4: The results for Example 2.1, Case 1, using V-I characteristic $I_{g,2}$

Initial Values				Optimal Solution				No.iter.	Rel.err.
a	b	L_x	L_y	a	b	L_x	L_y		
1e12	1e12	15	15	1e14	1e14	10	10	22	1.6e-8
1e10	1e10	15	15	1e14	1e14	10	10	19	8.2e-7
1e10	1e10	15	5	1e14	1e14	10	10	40	2.6e-8
1e12	1e12	15	5	1e14	1e14	10	10	84	4.0e-8

Table 2.4.5: The results for Example 2.1, Case 2, using V-I characteristic $I_{g,1}$

Initial Values				Optimal Solution				No.iter.	Rel.err.
a	b	L_x	L_y	a	b	L_x	L_y		
1e11	1e11	15	15	1.3e16	1.4e16	12.3	6.6	52	3.3e-5
1e12	1e12	15	15	1.4e16	1.4e16	12.2	6.4	53	5.5e-5
1e13	1e13	15	15	6.0e15	1.7e16	11.3	13.8	43	5.7e-5
1e12	1e12	5	5	1.8e16	2.3e16	12.5	3.8	40	9.6e-5
1e14	1e14	15	5	1.6e16	1.4e16	12.3	6.1	48	9.2e-5
1e10	1e10	15	5	1.9e16	1.1e16	13.7	7.4	46	6.9e-5

Table 2.4.6: The results for Example 2.1, Case 2, using V-I characteristic $I_{g,2}$

to be 10^{26}, 10^{8}, and 10^{-18}, respectively, and the scaling matrix M in (2.3.5) is chosen to be diag(10^5, 10^5, 10^{-9}, 10^{-9}, 10^{-9}). Two different V-I characteristics $I_{g,3}$ and $I_{g,4}$, listed in Tables 2.4.7 and 2.4.8 respectively, are used as the objective characteristics. These objective characteristics are generated by solving (2.1.17)–(2.1.19) using the parameter sets

$$(a = 10^{14}, b = 10^{14}, L_x = 10, L_y = 10, L_z = 10) \qquad \text{and}$$
$$(a = 10^{16}, b = 10^{16}, L_x = 10, L_y = 10, L_z = 10)$$

respectively for various applied forward biases. The solution procedure stops if the following items are satisfied:

1. relative error $\frac{\|\boldsymbol{I}-\boldsymbol{I}_g\|_2}{\|\boldsymbol{I}_g\|_2} \leq 1.2 \times 10^{-4}$ (10^{-6} for the case using $I_{g,3}$) and
2. the difference in $(L_x^2 + L_y^2 + L_z^2)^{1/2}$ between two consecutive iterates is smaller than 10^{-3}.

Various initial values are used, and the results are listed in Tables 2.4.9 and 2.4.10 respectively. From the tables we see that the relative errors are always

V	0.2143	0.2857	0.3571	0.4286	0.5
I	$3.205 \cdot 10^{-11}$	$4.127 \cdot 10^{-10}$	$5.220 \cdot 10^{-9}$	$4.575 \cdot 10^{-8}$	$1.894 \cdot 10^{-7}$

Table 2.4.7: The V-I Characteristic $I_{g,3}$

V	0.2143	0.2857	0.3571	0.4286	0.5
I	$2.116 \cdot 10^{-13}$	$3.341 \cdot 10^{-12}$	$5.266 \cdot 10^{-11}$	$8.284 \cdot 10^{-10}$	$1.299 \cdot 10^{-8}$

Table 2.4.8: The V-I Characteristic $I_{g,4}$

smaller than 1.2×10^{-4}. It is also seen that the optimal solutions are very accurate for $I_{g,3}$, and are reasonably close to the ideal solution for $I_{g,4}$.

We remark that the optimal solutions for $I_{g,2}$ and $I_{g,4}$ are harder to obtain than for $I_{g,1}$ and $I_{g,3}$, since the partial derivatives of the cost function E with respect to the doping parameter a and b are very small when the doping parameters are close to their upper bounds (cf. Figure 2.3.1).

Example 2.3. A two dimensional diode with a variable p-n junction

The third test problem is chosen to be a two dimensional rectangular diode with a variable p-n junction. The configuration of the device is the same as that in Figure 2.4.1, but the p-region has the width $(1 - c_x)L_x$ and the height $(1 - c_y)L_y$ where c_x and c_y are the ratios defined in Section 2. For simplicity, we assume that $L_x = L_y$ and $c_x = c_y$. Thus, the decision variables are a, b in the doping function, the dimension parameter $L_x = L_y$ and the parameter for the p-n junction $c_x = c_y$. The penalty parameters α_i, $(i = 1, \cdots, m)$, β and γ

Initial Values				Optimal Solution				No.	Rel.
a	b	L_x= L_y	L_z	a	b	L_x= L_y	L_z	iter	err
1e10	1e10	15	15	9.9e13	9.7e13	10.2	12	19	4.3e-7
1e12	1e12	15	15	1.0e14	1.0e14	10.0	10	24	6.6e-8
1e11	1e11	15	12	9.9e13	9.7e13	10.2	12	23	3.6e-7
1e12	1e12	15	10	1.0e14	1.0e14	10.0	10	19	2.2e-7
1e11	1e13	14	14	9.9e13	9.7e13	10.2	12	27	6.0e-7

Table 2.4.9: The results for Example 2.2, using objective V-I characteristic $I_{g,3}$

Initial Values				Optimal Solution				No. iter	Rel. err
a	b	L_x= L_y	L_z	a	b	L_x= L_y	L_z		
1e10	1e10	15	15	1.1e16	9.9e15	11.5	15	31	1.0e-4
1e10	1e10	15	10	1.1e16	7.7e15	10.0	9.8	30	7.9e-5
1e11	1e11	15	9	1.6e16	6.0e15	10.7	11	30	6.7e-5
1e11	1e11	14	14	1.2e16	6.3e15	10.2	15	30	3.6e-5
1e13	1e11	14	14	1.2e16	8.8e15	11.4	15	30	8.9e-5

Table 2.4.10: The results for Example 2.2, using objective V-I characteristic $I_{g,4}$

are chosen to be 10^{26}, 10^{10}, and 10^{-18}, respectively, and the scaling matrix M in (2.3.5) is chosen to be diag(10^5, 10^5, 10^{-9}, 4.0×10^{-9}). The bounds in (2.2.3) and (2.2.4) are chosen to be

$$c_x^{\min} = c_y^{\min} = 0.6 \quad \text{and} \quad c_x^{\max} = c_y^{\max} = 0.75.$$

The increment h in (2.3.6) for c_x is chosen to be 4×10^{-3}.

To solve the problem, we choose the V-I characteristic $I_{g,1}$, listed in Table 2.4.7 as the objective characteristic, which is generated by solving (2.1.17)–(2.1.19) using the parameter set

$$(a = 10^{14}, b = 10^{14}, L_x = 10, L_y = 10, c_x = 0.7, c_y = 0.7).$$

The solution procedure stops when both of the conditions

1. the relative error $\frac{\|\boldsymbol{I}-\boldsymbol{I}_g\|_2}{\|\boldsymbol{I}_g\|_2} \leq 10^{-5}$, and
2. the difference in $(L_x^2 + L_y^2)^{1/2}$ between two consecutive iterates is smaller than 10^{-3}

are satisfied. Two initial guesses are used, and the results are listed in Table 2.4.11. From the table we see that the relative errors are always smaller than 6.0×10^{-6}.

2.5 Conclusion

In this chapter, we posed the parameter selection in semiconductor device design as an optimization problem with competing costs. Various existing efficient

Initial Values			Optimal Solution				No. iter	Rel. err
a= b	L_x= L_y	c_x= c_y	a	b	L_x= L_y	c_x= c_y		
1e10	15	0.72	9.98e13	1.0e14	10.002	0.7008	25	5.6e-6
1e10	15	0.68	1.07e14	9.6e13	10.129	0.6773	22	5.5e-6

Table 2.4.11: The results for Example 2.3

methods for nonlinear optimization, nonlinear partial differential equations were discussed for the numerical solution of this problem. A scaling technique was also proposed to avoid numerical instability in computation. Numerical experiments for various model devices were performed and the numerical results showed the effectiveness of the optimization approach to the semiconductor device design.

Chapter 3

Optimization Approach to Numerical Integration

3.1 Introduction

In this chapter, we propose an approach to numerical integration of functions of one and two variables. In this approach, a grid set with a fixed number of vertices is to be obtained such that the error between the numerical integral using a given quadrature rule and the exact integral is minimized. A similar approach was used in finding the numerical solution for a differential equation in [16]. In this approach, the numerical integration is performed by using the quadrature rule on variable partition nodes rather than on a uniform partition as in the classical numerical integration schemes. These variable nodes are decision variables which are to be chosen such that the error between the numerical integral and the exact integral is minimized. The rest of this Chapter is organized as follows:

In Section 3.2, we shall discuss optimal grid construction for numerical integration in one dimension. The method is based on the mid-point rectangular and trapezoidal quadrature rules for sufficiently smooth integrand in Subsection 3.2.1, and trapezoidal rule with quadrature approximation of the integrand functions in Subsection 3.2.2. In Subsection 3.2.3, the method developed in Subsections 3.2.1 and 3.2.2 are used to solve several examples.

In Section 3.3, we consider optimal grid construction for numerical integration in two dimension. Two schemes for sufficiently smooth integrand will be

developed in Subsection 3.3.1: one is based on the barycenter rule on a rectangular partitions while the other is on a triangular partition. A scheme for non-sufficiently smooth functions is developed in Subsection 3.3.2. In Subsection 3.3.3, the methods developed in Subsections 3.3.1 and 3.3.2 are used to solve several examples.

3.2 Optimal Grid Construction in Numerical Integration of One Variable

Notation used in this section :

$\boldsymbol{I}(f)$: integral of function $f(x)$ on [a, b], where a, b $\in \mathbb{R}$ with $a < b$,

P_N : partition of $[a, b]$ with N sub-intervals,

$\boldsymbol{I}_R(f)$: mid-point rectangular quadrature rule on the partition P_N,

$\boldsymbol{I}_T(f)$: the trapezoidal quadrature rule on the partition P_N,

R : the Lagrange remainder of Taylor's expansion,

h_i : step length $x_{i+1} - x_i$,

3.2.1 Method for sufficiently smooth integrands

Consider an integral of the form

$$I(f) = \int_a^b f(x)dx, \tag{3.2.1}$$

where $a, b \in \mathbb{R}$ with $a < b$. We assume that f is sufficiently smooth. More specifically, we assume that at least $f'''(x)$ is continuous on (a, b). (This assumption will be relaxed in the next subsection.) Let P_N be a partition with N sub-divisions defined by

$$P_N : a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b. \tag{3.2.2}$$

Let $h_i = x_{i+1} - x_i$ for $i = 0, 1, \ldots, N-1$. In what follows we discuss the mid-point rectangular quadrature rule and the trapezoidal quadrature rule separately.

Mid-point rectangular quadrature rule

The mid-point rectangular quadrature rule for (3.2.1) is defined by

$$I_R(f) := \sum_{i=0}^{N-1} f(\xi_i)h_i \tag{3.2.3}$$

where $\xi_i = (x_i + x_{i+1})/2$. The following theorem gives a representation for $I(f) - I_R(f)$.

Theorem 3.2.1 *If $f'''(x)$ is continuous on (a, b), then*

$$I(f) - I_R(f) = \sum_{i=0}^{N-1} \left(\frac{f''(\xi_i)}{24} h_i^3 + \frac{1}{6} \int_{x_i}^{x_{i+1}} f'''(\eta_i(x))(x - \xi_i)^3 dx \right) \tag{3.2.4}$$

where $x < \eta_i(x) < \xi_i$ or $\xi_i < \eta_i(x) < x$. Furthermore, if $f^{(4)}$ is continuous on (a, b), then

$$I(f) - I_R(f) = \sum_{i=0}^{N-1} \left(\frac{f''(\xi_i)}{24} h_i^3 + \frac{1}{24} \int_{x_i}^{x_{i+1}} f^{(4)}(\mu_i(x))(x - \xi_i)^4 dx \right) \tag{3.2.5}$$

where $x < \mu_i(x) < \xi_i$ or $\xi_i < \mu_i(x) < x$.

PROOF. The proof is based on Taylor's expansion. We consider the first case. Since f''' is continuous on (a, b), for each $i = 0, 1, ..., N-1$, $f(x)$ can be expanded into

$$f(x) = f(\xi_i) + f'(\xi_i)(x - \xi_i) + \frac{1}{2} f''(\xi_i)(x - \xi_i)^2 + \frac{1}{6} f'''(\eta_i(x))(x - \xi_i)^3,$$

where $x < \eta_i(x) < \xi_i$ or $\xi_i < \eta_i(x) < x$. The last term is the Lagrange remainder of the Taylor's expansion. Thus,

$$\begin{aligned} \int_{x_i}^{x_{i+1}} f(x)dx - f(\xi_i)h_i &= \int_{x_i}^{x_{i+1}} \left(f(\xi_i) + f'(\xi_i)(x - \xi_i) + \frac{1}{2} f''(\xi_i)(x - \xi_i)^2 \right. \\ &\qquad \left. + \frac{1}{6} f'''(\eta_i(x))(x - \xi_i)^3 \right) dx - f(\xi_i)h_i \\ &= \frac{f''(\xi_i)}{24} h_i^3 + \frac{1}{6} \int_{x_i}^{x_{i+1}} f'''(\eta_i(x))(x - \xi_i)^3 dx, \end{aligned}$$

since $x - \xi_i$ is an odd function with respect to ξ_i. Summing both sides of the above from $i = 0$ to $N - 1$, we obtain (3.2.4).

We now prove (3.2.5). Since $f^{(4)}(x)$ is continuous on (a,b), $f(x)$ can be expanded as

$$f(x) = \sum_{i=0}^{3} \frac{1}{i!} f^{(i)}(\xi_i)(x-\xi_i)^i + \frac{1}{24} f^{(4)}(\mu_i(x))(x-\xi_i)^4$$

for $i = 0, 1, \ldots, N-1$. Notice that both $x - \xi_i$ and $(x-\xi_i)^3$ are odd functions with respect to ξ_i and the integrals of these on (x_i, x_{i+1}) vanish. We have

$$\int_{x_i}^{x_{i+1}} f(x)dx - f(\xi_i)h_i = \frac{f''(\xi_i)}{24} h_i^3 + \frac{1}{24} \int_{x_i}^{x_{i+1}} f^(4)(\mu_i(x))(x-\xi_i)^4 dx.$$

Summing this from $i = 0$ to $N-1$ gives (3.2.5). □

From Theorem 3.2.1, we see that the error in the approximate integral from the mid-point rule is dominated by

$$E_R(x_1, x_2, \ldots, x_{N-1}) := \frac{1}{24} \sum_{i=0}^{N-1} f''(\xi_i) h_i^3 \tag{3.2.6}$$

when all $h_i < 1$ (or $f''(x)$ is almost constant if $h_i > 1$). Thus, the partition P_N should be chosen such that $|E_R|$ is minimized. Based on this, we propose an optimization problem as follows:

$$\min \quad g(x_1, x_2, \ldots, x_{N-1}) = E_R^2(x_1, x_2, \ldots, x_{N-1})$$

subject to

$$a - x_1 < 0, \tag{3.2.7}$$

$$x_i - x_{i+1} < 0, \qquad i = 1, \ldots, N-2, \tag{3.2.8}$$

$$x_{N-1} - b < 0, \tag{3.2.9}$$

where E_R is defined by (3.2.6). This problem is referred to as Problem $P3.2.1$. This is a nonlinear mathematical programming problem with N linear inequality constraints. The gradient of the cost function g is given by

$$\frac{\partial g}{\partial x_j} = \frac{1}{12} E_R \left[\frac{1}{2} f'''(\xi_{j-1}) h_{j-1}^3 + 3f''(\xi_{j-1}) h_{j-1}^2 + \frac{1}{2} f'''(\xi_j) h_j^3 - 3f''(\xi_j) h_j^2 \right]$$

for $j = 1, 2, \ldots, N-1$, where h_j and ξ_j are defined above. The solution to this problem will yield the optimal partition P_N.

The trapezoidal quadrature rule

The trapezoidal quadrature rule on the partition P_N defined by (3.2.2) is:

$$I_T(f) = \sum_{i=0}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2} h_i. \tag{3.2.10}$$

This is obtained from the approximation of $f(x)$ on (x_i, x_{i+1}) by the linear function $f(x_i) + (x - x_i)(f(x_{i+1}) - f(x_i))/h_i$. The error for the trapezoidal rule is given in the following theorem.

Theorem 3.2.2 *If $f'''(x)$ is continuous in (a, b), then*

$$I(f) - I_T(f) = \sum_{i=0}^{N-1} \left[-\frac{1}{12} f''(x_i) h_i^3 - \frac{h_i}{2} R(x_{i+1}) + \int_{x_i}^{x_{i+1}} R(x) dx \right] \tag{3.2.11}$$

where $R(x)$ is defined by

$$R(x) = \frac{1}{6} f'''(\sigma_i(x))(x - x_i)^3 \tag{3.2.12}$$

with $x_i < \sigma_i(x) < x$.

PROOF. Since f''' is continuous, we expand f on $[x_i, x_{i+1}]$ as the following Taylor's series with a remainder:

$$f(x) = \sum_{n=0}^{2} \frac{f^{(n)}(x_i)}{n!} (x - x_i)^n + R(x),$$

where the remainder is given by (3.2.12). From this we have

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} f(x) dx - \frac{f(x_i) + f(x_{i+1})}{2} h_i &= \sum_{n=0}^{2} \frac{f^{(n)}(x_i)}{(n+1)!} h_i^{n+1} \\
&\quad + \int_{x_i}^{x_{i+1}} R(x) dx - \frac{f(x_i) + f(x_{i+1})}{2} h_i \\
&= \frac{1}{2} f'(x_i) h_i^2 - \frac{h_i}{2} [f(x_{i+1}) - f(x_i)] \\
&\quad + \frac{1}{6} f''(x_i) h_i^3 + \int_{x_i}^{x_{i+1}} R(x) dx \\
&= -\frac{1}{12} f''(x_i) h_i^3 - \frac{h_i}{2} R(x_{i+1}) \\
&\quad + \int_{x_i}^{x_{i+1}} R(x) dx.
\end{aligned}$$

In the above, we used the Taylor's expansion of $f(x_{i+1})$ at x_i. Finally, (3.2.11) follows from summing the above from $i = 0$ to $N - 1$. □

Obviously the above result is similar to that of Theorem 3.2.1. Again, from Theorem 3.2.2, we see that the error in I_T is dominated by

$$E_T(x_1, x_2, \ldots, x_{N-1}) := \frac{1}{12} \sum_{i=0}^{N-1} f''(x_i) h_i^3 \tag{3.2.13}$$

when $h_i < 1$. This motivates us to define the following optimization problem.

$$\min \quad g(x_1, x_2, \ldots, x_{N-1}) = E_T^2(x_1, x_2, \ldots, x_{N-1})$$

subject to the constraints (3.2.7), (3.2.8) and (3.2.9), where E_T is the function defined by (3.2.13). This problem is refereed to as Problem P3.2.2. Similar to the case of mid-point rectangular rule, Problem P3.2.2 is also a nonlinear mathematical programming problem with N linear constraints and the gradient of the cost function is given by

$$\frac{\partial g}{\partial x_j} = \frac{1}{6} E_T \left[3f''(x_{j-1}) h_{j-1}^2 + f'''(x_j) h_j^3 - 3f''(x_j) h_j^2 \right]$$

for $j = 1, 2, ..., N - 1$. The solution to the problem will give the optimal distribution of vertices for the trapezoidal quadrature rule.

3.2.2 Method for not sufficiently smooth integrands

The techniques presented in the previous subsection require that at least $f'''(x)$ is continuous on (a, b). Following the notation defined in subsection 3.2.1, we now discuss a technique for the case that only f' is continuous on (a, b). This is based on the idea of approximating the error in a quadrature rule by the difference between the the quadrature rule and a higher order quadrature rule. We will demonstrate it using the trapezoidal rule defined in (3.2.10) and a quadrature rule obtained by interpolating $f(x)$ on any $[x_i, x_{i+1}]$ by a quadratic function using the values $f(x_i)$, $f(\xi_i)$ and $f(x_{i+1})$ where ξ_i is the mid-point of $[x_i, x_{i+1}]$ as defined previously.

For any $i = 0, 1, \ldots, N - 1$, the trapezoidal rule on $[x_i, x_{i+1}]$ is defined as

$$\int_{x_i}^{x_{i+1}} f(x) dx \approx \frac{1}{2}(f(x_i) + f(x_{i+1})) h_i.$$

As mentioned in the previous section, this is obtained by approximating $f(x)$ by the linear function $L(x) = f(x_i) + (x - x_i)(f(x_{i+1}) - f(x_i))/h_i$ on $[x_i, x_{i+1}]$. Now we fit $f(x)$ on $[x_i, x_{i+1}]$ by a quadratic function

$$Q(x) = A + Bx + Cx^2$$

such that

$$Q(x_i) = f(x_i), \quad Q(\xi_i) = f(\xi_i) \quad \text{and} \quad Q(x_{i+1}) = f(x_{i+1}).$$

These yield three equations for A, B and C. To simplify the problem, we translate the graph of $Q(x)$ to the position such that $(x_i, f(x_i))$ coincides with the origin $(0, 0)$. In this case $A = 0$ and B and C satisfy

$$\begin{aligned} \frac{h_i}{2}B + \frac{h_i^2}{4}C &= f(\xi_i) - f(x_i), \\ h_iB + h_i^2C &= f(x_{i+1}) - f(x_i). \end{aligned}$$

Solving this system we obtain

$$Q_0(x) = \frac{-3f(x_i) + 4f(\xi_i) - f(x_{i+1})}{h_i}x + \frac{2f(x_i) - 4f(\xi_i) + 2f(x_{i+1})}{h_i^2}x^2,$$

where $x \in [0, h_i]$ and Q_0 denotes the translated quadratic form of Q. Similarly, the linear approximation $L(x)$ can also be translated into $L_0 = (f(x_{i+1}) - f(x_i))x/h_i$ for $x \in [0, h_i]$. Since this translation does not change the value of $Q(x) - L(x)$, we have

$$\begin{aligned} \int_{x_i}^{x_{i+1}} (Q(x) - L(x))dx &= \int_0^{h_i} (Q_0(x) - L_0(x))dx \\ &= \int_0^{h_i} \left(\frac{-3f(x_i) + 4f(\xi_i) - f(x_{i+1})}{h_i} \right. \\ &\quad \left. - \frac{f(x_{i+1}) - f(x_i)}{h_i} \right) xdx \\ &\quad + \int_0^{h_i} \frac{2f(x_i) - 4f(\xi_i) + 2f(x_{i+1})}{h_i^2} x^2 dx \\ &= (-2f(x_i) + 4f(\xi_i) - 2f(x_{i+1}))\frac{h_i}{2} \\ &\quad + (2f(x_i) - 4f(\xi_i) + 2f(x_{i+1}))\frac{h_i}{3} \\ &= \frac{h_i}{6}(-2f(x_i) + 4f(\xi_i) - 2f(x_{i+1})). \end{aligned}$$

From this we define

$$E_{QT}(x_1, x_2, \ldots, x_{N-1}) := \sum_{i=0}^{N-1} \frac{h_i}{6}(-2f(x_i) + 4f(\xi_i) - 2f(x_{i+1})). \qquad (3.2.14)$$

Obviously this dominates the true error $I(f) - I_T(f)$. Therefore we define the following optimization problem.

$$\min \quad g(x_1, x_2, \ldots, x_{N-1}) = E_{QT}^2(x_1, x_2, \ldots, x_{N-1})$$

subject to the constraints (3.2.7), (3.2.8) and (3.2.9), where E_{QT} is the function defined by (3.2.14). This is a nonlinear mathematical programming problem and is referred to as Problem $P3.2.3$.

We comment that in the case that f''' is continuous on (a, b), it is easy to show that E_{QT} reduces to E_T defined by (3.2.13) if all terms on the right side of (3.2.14) are expanded as (truncated) Taylor's series.

3.2.3 Numerical Experiments

To demonstrate the effectiveness and usefulness of the above techniques, some numerical experiments have been carried out. The package FFSQP [116] for solving general nonlinear mathematical programming problems is used as an optimizer for Problems $P3.2.1$, $P3.2.2$ and $P3.2.3$. All computations are performed in Fortran double precision on an Pentium PC under LINUX 2.0 environment. For all the test problems solved, we use the uniform mesh (for a given N) as the initial mesh and the numerical integral obtained from the mid-point rectangular quadrature rule on the uniform mesh with 1000 subintervals as the exact integral $I(f)$.

Example 3.2.1.

$$I(f) = \int_a^b \gamma \exp\left(-\alpha(x-\beta)^2\right) dx.$$

This function often appears in statistics and it is not analytically integrable on a finite interval $[a, b]$. We now choose $a = 0$, $b = 5$, $\alpha = \beta = 1$ and $\gamma = 50$. The number of interior vertices is chosen to be 18, i.e., $N = 19$. Problems

Grid	$g(\boldsymbol{x}_{init})$	$g(\boldsymbol{x}_{final})$	Error	Error on U. Grid
MP on G1	1.1398e-1	7.2416e-12	1.3668e-3	1.0643e-1
TR on G2	5.6546e-2	7.1032e-16	4.6544e-3	2.1295e-1
TR on G3	4.5332e-2	3.6541e-8	2.0017e-3	2.1295e-1

Table 3.2.1: Results for Example 3.2.1 using different quadrature rules and grids

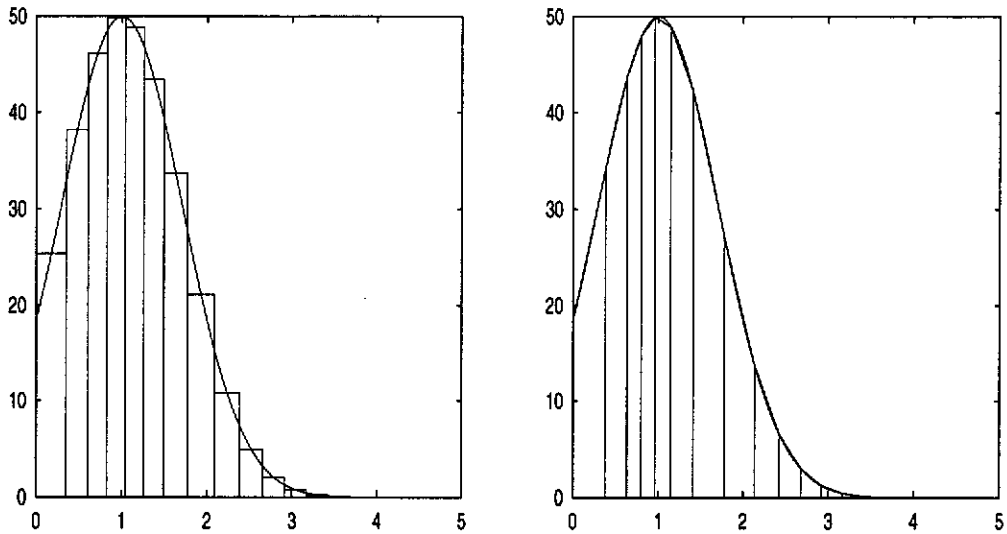
$P3.2.1$, $P3.2.2$ and $P3.2.3$ were used to construct three integral grids label with G1, G2 and G3 respectively. Table 3.2.1 is a list of the initial and final values ($g(\boldsymbol{x}_{init})$ and $g(\boldsymbol{x}_{final})$) of the cost functions and the absolute errors in the numerical integrals computed respectively by the mid-point rule (MP) on G1, the trapezoidal rule (TR) on G2 and G3. For comparison, the errors in the numerical integrals from the mid-point and trapezoidal rules on the uniform mesh are listed in the last column. From this table, we see that the results on the optimal grids are about two orders of magnitudes more accurate than that on the uniform mesh. Also, Problems $P3.2.1$ and $P3.2.3$ produce more accurate results than that from Problem $P3.2.2$. To visualize the the grids obtained from the solutions to Problems $P3.2.1$, $P3.2.2$ and $P3.2.3$, we plot the results in Figure 3.2.1.

In what follows, we concentrate on Problems $P3.2.1$ and$P3.2.3$, as E_T defined by (3.2.13) can be obtained from E_{QT} defined by expanding the functions on the right side of (3.2.14) at each x_i.

Example 3.2.2.

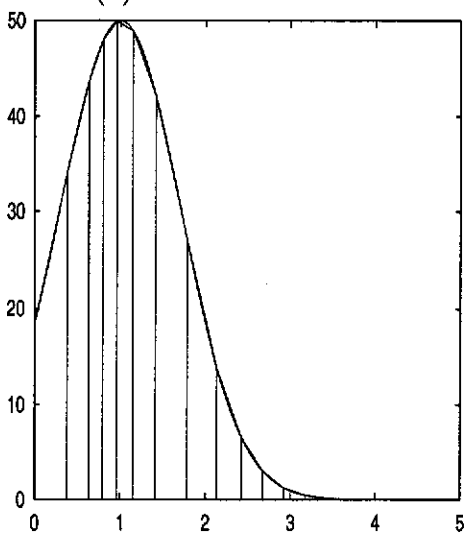
$$I(f) = \int_a^b \frac{1}{\alpha + \beta x^2} \frac{\sin x}{x} dx.$$

The integrand is even and not analytically integrable. We choose $a = 0$, $b = 1$, $\alpha = 0.05$ and $\beta = 50$. The number of subintervals is chosen to be 10, i.e., $N = 10$. Two optimal grids G1 and G3 were obtained by solving Problems $P3.2.1$ and $P3.2.3$ respectively and Table 3.2.2 is a list of the initial and final values of the cost functions and the computed errors on G1, G3 and the uniform mesh. From the table, we see that the result from Problems $P3.2.1$ and $P3.2.2$ are, respectively, two orders and one order of magnitudes more accurate than those

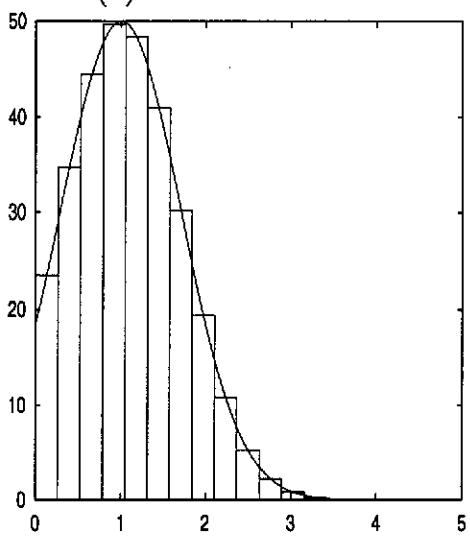


(a) MP rule on G1

(b) TR rule on G2



(c) TR rule on G3



(d) MP rule on the U. Grid

Figure 3.2.1: Graphs of the numerical solutions of Example 3.2.1 using various meshes

Grid	$g(\boldsymbol{x}_{init})$	$g(\boldsymbol{x}_{final})$	Error	Error on U. Grid
MP on G1	6.9184e-2	2.9428e-6	1.0471e-3	2.4693e-1
TR on G3	1.3708e-1	1.6163e-3	3.6292e-2	3.1575e-1

Table 3.2.2: Results for Example 3.2.2 using different quadrature rules and grids

Grid	$g(\boldsymbol{x}_{init})$	$g(\boldsymbol{x}_{final})$	Error	Error on U. Grid
MP on G1	1.1430e-7	5.4524e-21	2.6178e-6	3.3470e-4
TR on G3	4.4224e-7	8.0773e-19	2.3104e-6	6.6281e-4

Table 3.2.3: Results for Example 3.2.3 using different quadrature rules and grids

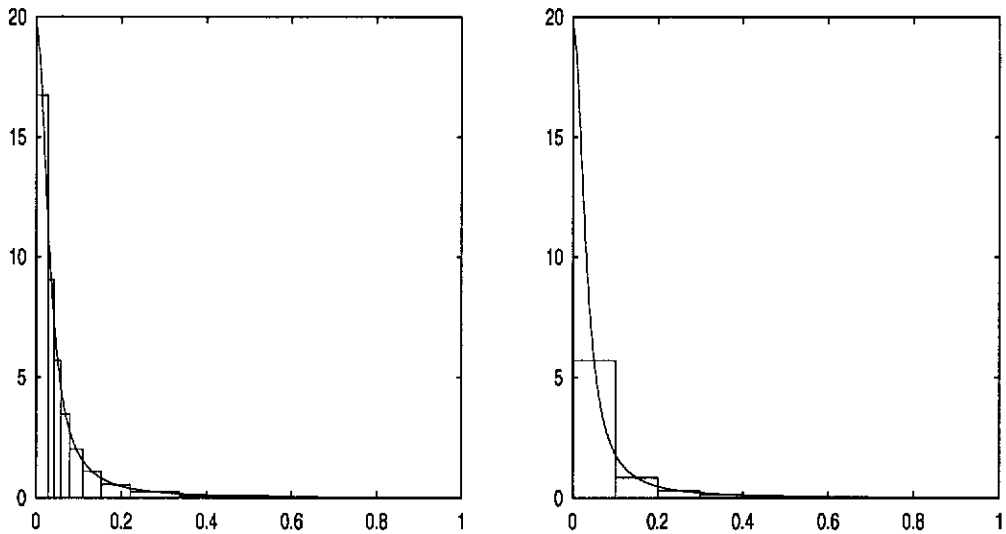
on the uniform grid. The numerical results using the mid-point rectangular and trapezoidal rules on different grids are also plotted in Figure 3.2.2.

Example 3.2.3.

$$I(f) = \int_0^1 \frac{\sin(10x)}{10x} dx.$$

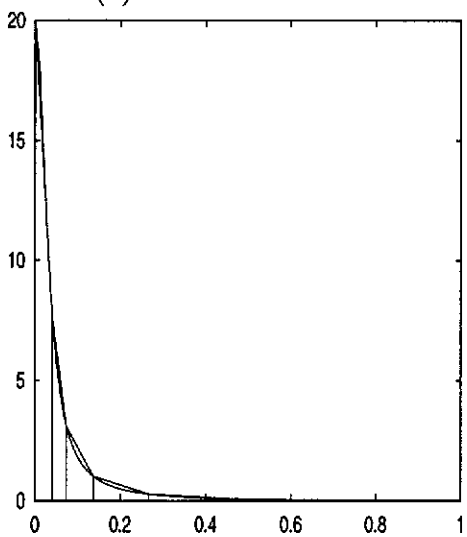
The problem is the same as Example 3.2.2, if we choose $\alpha = 1$, $\beta = 0$, $a = 0$ and $b = 10$, and then scale the integral interval by 10. This integral often appears in signal processing and is not analytically integrable. The integrand of this problem oscillates in the integral interval rather than a sharp pick as in Example 3.2.2. We now choose $N = 10$. The computed costs and errors corresponding to Problems $P3.2.1$ and $P3.2.3$ are listed in Table 3.2.3. From the figure, it is seen that the results for both of the two methods are two orders of magnitudes more accurate than that from the uniform mesh.

The numerical results are plotted in Figure 3.2.3. From Figure 3.2.3 and Table 3.2.3, we see that although the grids G1 and G3 are close to the uniform one, the numerical integrals on G1 and G3 are much more accurate than those on the uniform grid.

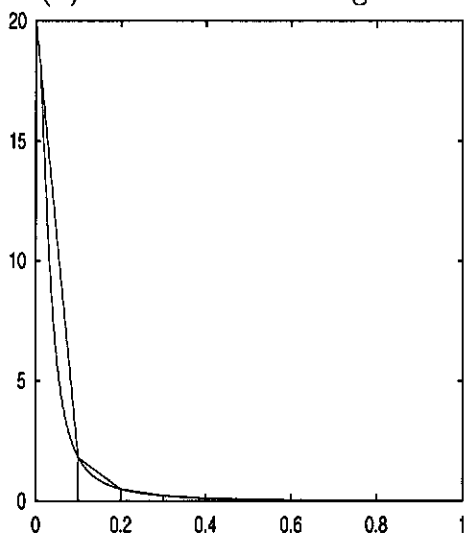


(a) MP rule on G1

(b) MP rule on the u. grid

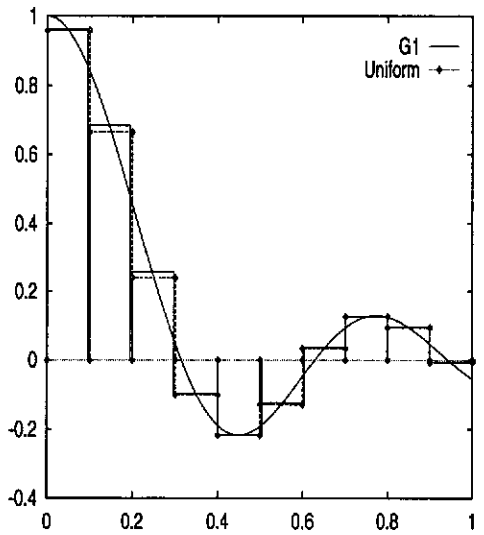


(c) TR rule on G3

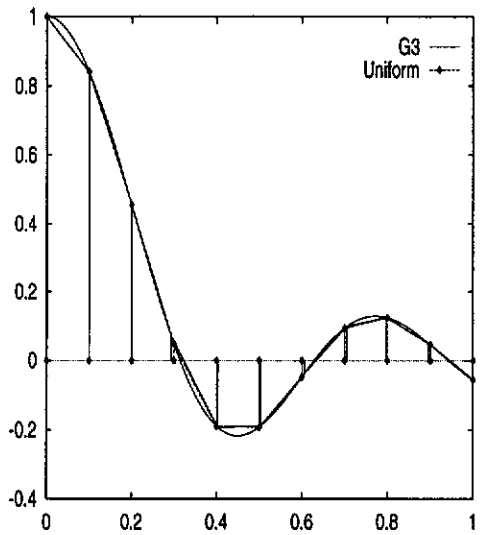


(d) TR rule on the u. grid

Figure 3.2.2: Graphs of the numerical solutions of Example 3.2.2 using various meshes



(a) MP rule on G1 and the u grid



(b) TR rule on G1 and the u. grid

Figure 3.2.3: Graphs of the numerical solutions of Example 3.2.3 using various meshes

3.3 Optimal Grid Construction in Numerical Integration of Two variables

Notation used in this section :

Ω : domain, on which function $f(x, y)$ is integrated,

$\boldsymbol{I}(f)$: integral of function $f(x, y)$ on Ω,

ω_{ij} : rectangle with vertices (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}), and (x_{i+1}, y_{j+1}),

ω_{ij}^1 : triangle with vertices (x_i, y_j), (x_{i+1}, y_j), and (x_i, y_{j+1}),

ω_{ij}^2 : triangle with vertices (x_{i+1}, y_j), (x_{i+1}, y_{j+1}), and (x_i, y_{j+1}),

$P_{m\times n}$: partition of Ω with $m \times n$ rectangular elements ω_{ij},

$P_{m\times n}^*$: partition of Ω with $2 \times m \times n$ triangular elements ω_{ij}^1 or ω_{ij}^2,

$\boldsymbol{I}_R(f)$: the barycenter quadrature rule on the partition $P_{m\times n}$,

$\boldsymbol{I}_T(f)$: the barycenter quadrature rule on the partition $P_{m\times n}^*$,

$\boldsymbol{I}_{TR}(f)$: the trapezoidal rule on the partition $P_{m\times n}^*$,

R : the Lagrange remainder of Taylor's expansion,

h_i : step length $x_{i+1} - x_i$,

k_j : step length $y_{j+1} - y_j$.

3.3.1 Methods for sufficiently smooth integrands

Consider an integral of the form

$$I(f) = \int_{\Omega} f(x,y)dxdy$$

where $\Omega = (a,b) \times (c,d)$ with $a,b,c,d \in I\!R$. In this subsection we assume that at least the third order partial derivatives of the function $f(x,y)$ are continuous on Ω. Let $P_{m\times n}$ be a partition of Ω with $m \times n$ sub-domains defined by

$$P_{m\times n} = \{\omega_{ij} : \ i = 0,1,2,\cdots,m-1;\ j = 0,1,2,\cdots,n-1\} \tag{3.3.1}$$

where

$$\omega_{ij} = (x_i, x_{i+1}) \times (y_j, y_{j+1}), \tag{3.3.2}$$

$$a = x_0 < x_1 < x_2 < \cdots < x_m = b, \tag{3.3.3}$$

and

$$c = y_0 < y_1 < y_2 < \cdots < y_n = d. \tag{3.3.4}$$

Let $h_i = x_{i+1} - x_i$ for $i = 0,1,\cdots,m-1$, and $k_j = y_{j+1} - y_j$ for $j = 0,1,\cdots,n-1$. Numerical schemes using the barycenter quadrature rule on rectangular partition and on triangular partition will be developed in the next two subsections, respectively.

Barycenter quadrature rule on a rectangular partition

The barycenter quadrature rule on the partition $P_{m\times n}$ defined in (3.3.1) - (3.3.2) is:

$$I_R(f) = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} f(\xi_i, \eta_j)h_i k_j \tag{3.3.5}$$

where $\xi_i = \frac{x_i + x_{i+1}}{2}$, and $\eta_j = \frac{y_j + y_{j+1}}{2}$. The following theorem establishes the error $I(f) - I_R(f)$.

Theorem 3.3.1 *Let the third order partial derivatives of the function $f(x,y)$ be continuous on Ω. If the absolute values of these third order partial derivatives are bounded by $M_{ij} > 0$ in ω_{ij} for $i = 0, 1, \cdots, m-1$; and $j = 0, 1, \cdots, n-1$, where ω_{ij} is defined by (3.3.2), then there exist $\theta_{ij} \in [-M_{ij}, M_{ij}]$ such that*

$$\begin{aligned} I(f) - I_R(f) &= \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\Big\{\frac{1}{24}[f_{xx}(\xi_i,\eta_j)h_i^3k_j + f_{yy}(\xi_i,\eta_j)h_ik_j^3] \\ &\quad + \frac{1}{96}\theta_{i,j}[\frac{1}{2}h_i^4k_j + h_i^3k_j^2 + h_i^2k_j^3 + \frac{1}{2}h_ik_j^4]\Big\} \end{aligned} \tag{3.3.6}$$

PROOF. Using Taylor's formula for functions of two variables, we obtain

$$\begin{aligned} f(x,y) &= f(\xi_i,\eta_j) + f_x(\xi_i,\eta_j)(x-\xi_i) + f_y(\xi_i,\eta_j)(y-\eta_j) \\ &+ \frac{1}{2}\{f_{xx}(\xi_i,\eta_j)(x-\xi_i)^2 + 2f_{xy}(\xi_i,\eta_j)(x-\xi_i)(y-\eta_j) \\ &\quad + f_{yy}(\xi_i,\eta_j)(y-\eta_j)^2\} \\ &+ \frac{1}{6}\Big\{(x-\xi_i)\frac{\partial}{\partial x} + (y-\eta_j)\frac{\partial}{\partial y}\Big\}^3 f(\xi_i + r(x-\xi_i), \eta_j + r(y-\eta_j)) \end{aligned}$$

where $0 \le r \le 1$, $i = 0, 1, \cdots, m-1$, and $j = 0, 1, \cdots, n-1$. The last term in the above is the Lagrange remainder of Taylor's expansion. Clearly,

$$\begin{aligned} &\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}}\{f(x,y) - f(\xi_i,\eta_j)\}dxdy \\ =\ &\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}}\big\{f_x(\xi_i,\eta_j)(x-\xi_i) + f_y(\xi_i,\eta_j)(y-\eta_j)\big\}dxdy \\ +\ &\frac{1}{2}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}}\{f_{xx}(\xi_i,\eta_j)(x-\xi_i)^2 + 2f_{xy}(\xi_i,\eta_j)(x-\xi_i)(y-\eta_j) \\ &\qquad + f_{yy}(\xi_i,\eta_j)(y-\eta_j)^2\}dxdy \\ +\ &\frac{1}{6}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}}\Big\{(x-\xi_i)\frac{\partial}{\partial x} + (y-\eta_j)\frac{\partial}{\partial y}\Big\}^3 \\ &\qquad f(\xi_i + r(x-\xi_i), \eta_j + r(y-\eta_j))dxdy. \end{aligned} \tag{3.3.7}$$

Since $x - \xi_i$ and $y - \eta_j$ are, respectively, odd functions with respect to ξ_i and η_j, (3.3.7) can be written as:

$$\begin{aligned} &\int_{y_j}^{y_{j+1}}\int_{x_i}^{xi+1}\{f(x,y) - f(\xi_i,\eta_j)\}dxdy \\ =\ &\frac{1}{24}\{f_{xx}(\xi_i,\eta_j)h_i^3k_j + f_{yy}(\xi_i,\eta_j)h_ik_j^3\} + R_1 + R_2 + R_3 + R_4 \end{aligned} \tag{3.3.8}$$

where

$$
\begin{aligned}
R_1 &= \frac{1}{6}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} f_{xxx}(\xi_i + r(x-\xi_i), \eta_j + r(y-\eta_j))(x-\xi_i)^3 dxdy, \\
R_2 &= \frac{1}{2}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} f_{xxy}(\xi_i + r(x-\xi_i), \eta_j \\
&\qquad + r(y-\eta_j))(x-\xi_i)^2(y-\eta_j) dxdy, \\
R_3 &= \frac{1}{2}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} f_{xyy}(\xi_i + r(x-\xi_i), \eta_j \\
&\qquad + r(y-\eta_j))(x-\xi_i)(y-\eta_j)^2 dxdy, \\
R_4 &= \frac{1}{6}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} f_{yyy}(\xi_i + r(x-\xi_i), \eta_j + r(y-\eta_j)(y-\eta_j)^3 dxdy.
\end{aligned}
$$

We now estimate each of R_1, R_2, R_3 and R_4. For R_1, we have

$$
\begin{aligned}
|R_1| &\leq \frac{1}{6}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} \left| f_{xxx}(\xi_i + r(x-\xi_i), \eta_j \right. \\
&\qquad \left. + r(y-\eta_j))(x-\xi_i)^3 \right| dxdy \\
&\leq \frac{M_{i,j}}{6}\int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} \left|(x-\xi_i)^3\right| dxdy \\
&= \frac{1}{6}\cdot\frac{1}{32} M_{ij} h_i^4 k_j.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
|R_2| &\leq \frac{1}{2}\cdot\frac{1}{48} M_{ij} h_i^3 k_j^2, \\
|R_3| &\leq \frac{1}{2}\cdot\frac{1}{48} M_{ij} h_i^2 k_j^3, \\
|R_4| &\leq \frac{1}{6}\cdot\frac{1}{32} M_{ij} h_i k_j^4.
\end{aligned}
$$

Combining these four inequalities, we obtain

$$
|R_1 + R_2 + R_3 + R_4| \leq \frac{1}{96}\left\{\frac{1}{2}h_i^4 k_j + h_i^3 k_j^2 + h_i^2 k_j^3 + \frac{1}{2} h_i k_j^4\right\} M_{ij}.
$$

Thus, there exists a $\theta_{ij} \in [-M_{ij}, M_{ij}]$ such that

$$
R_1 + R_2 + R_3 + R_4 = \frac{1}{96}\left\{\frac{1}{2}h_i^4 k_j + h_i^3 k_j^2 + h_i^2 k_j^3 + \frac{1}{2} h_i k_j^4\right\}\theta_{ij}. \tag{3.3.9}
$$

Substituting (3.3.9) into (3.3.8) and summing both sides of the resulting equation, we obtain (3.3.6). This completes the proof. □

From Theorem 3.3.1, it is clear that the error between the numerical integral using the barycenter quadrature rule on a rectangular partition and the exact integral is domained by

$$E_R(\boldsymbol{x},\boldsymbol{y}) := \frac{1}{24}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\Big\{f_{xx}(\xi_i,\eta_j)h_i^3k_j + f_{yy}(\xi_i,\eta_j)h_ik_j^3\Big\}$$

where $\boldsymbol{x} = (x_1,\cdots,x_{m-1})$, $\boldsymbol{y} = (y_1,\cdots,y_{n-1})$, and h_i and k_j are such that $h_i < 1$, $k_j < 1$. These conditions are satisfied unless $f_{xx}(x,y)$, and $f_{yy}(x,y)$ are nearly constants. The corresponding numerical scheme may now be posed as an optimization problem:

$$\min g(\boldsymbol{x},\boldsymbol{y}) = \Big\{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[f_{xx}(\xi_i,\eta_j)h_i^3k_j + f_{yy}(\xi_i,\eta_j)h_ik_j^3]\Big\}^2 \tag{3.3.10}$$

subject to

$$x_i - x_{i+1} < 0, \qquad i = 0,1,\cdots,m-1, \tag{3.3.11}$$

$$y_j - y_{j+1} < 0, \qquad j = 0,1,\cdots,n-1. \tag{3.3.12}$$

Let this optimization problem be referred to as Problem $P3.3.1$. The gradient of the cost function g is given by

$$\begin{aligned}\frac{\partial g}{\partial x_i} \;=\; 2g\sum_{j=0}^{n-1}\Big\{&\frac{1}{2}f_{xxx}(\xi_i,\eta_j)h_i^3k_j - \frac{3}{2}f_{xx}(\xi_i,\eta_j)h_i^2k_j\\ &+\frac{1}{2}f_{xxx}(\xi_{i-1},\eta_j)h_{i-1}^3k_j + \frac{3}{2}f_{xx}(\xi_{i-1},\eta_j)h_{i-1}^2k_j\\ &+\frac{1}{2}f_{yyx}(\xi_i,\eta_j)h_ik_j^3 - \frac{1}{2}f_{yy}(\xi_i,\eta_j)k_j^3\\ &+\frac{1}{2}f_{yyx}(\xi_{i-1},\eta_j)h_{i-1}k_j^3 + \frac{1}{2}f_{yy}(\xi_{i-1},\eta_j)k_j^3\Big\}\end{aligned}$$

$$\begin{aligned}\frac{\partial g}{\partial y_j} &= 2g\sum_{i=0}^{m-1}\Big\{\frac{1}{2}f_{xxy}(\xi_i,\eta_j)h_i^3k_j - \frac{1}{2}f_{xx}(\xi_i,\eta_j)h_i^3 \\ &\quad +\frac{1}{2}f_{xxy}(\xi_i,\eta_{j-1})h_i^3k_{j-1} + \frac{1}{2}f_{xx}(\xi_i,\eta_{j-1})h_i^3 \\ &\quad +\frac{1}{2}f_{yyy}(\xi_i,\eta_j)h_ik_j^3 - \frac{3}{2}f_{yy}(\xi_i,\eta_j)h_ik_j^2 \\ &\quad +\frac{1}{2}f_{yyy}(\xi_{i-1},\eta_{j-1})h_ik_{j-1}^3 \\ &\quad +\frac{3}{2}f_{yy}(\xi_{i-1},\eta_{j-1})h_ik_{j-1}^2\Big\}\end{aligned}$$

for $i = 1, 2, \cdots, m-1$, and $j = 1, 2, \cdots, n-1$.

Barycenter quadrature rule on a triangular partition

Dividing each rectangle $\omega_{i,j} = (x_i, x_{i+1}) \times (y_j, y_{j+1})$ in the partition $P_{m\times n}$ into two triangles $\omega_{i,j}^1$ with vertices (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}), and $\omega_{i,j}^2$ with vertices (x_{i+1}, y_j), (x_{i+1}, y_{j+1}), (x_i, y_{j+1}), we obtain a triangular partition $P_{m\times n}^*$ defined by

$$P_{m\times n}^* = \{\omega_{i,j}^1, \omega_{i,j}^2 :\ i = 0, 1, 2, \cdots, m-1\ ,\ j = 0, 1, 2, \cdots, n-1\} \qquad (3.3.13)$$

The barycenter quadrature rule on the partition $P_{m\times n}^*$ is defined as (cf. Chapter 4 of [17].)

$$I_T(f) = \frac{1}{2}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\Big\{f(\xi_{i1},\eta_{j1}) + f(\xi_{i2},\eta_{j2})\Big\}h_ik_j$$

where $\xi_{i1} = \frac{2x_i+x_{i+1}}{3}$, $\eta_{j1} = \frac{2y_j+y_{j+1}}{3}$, $\xi_{i2} = \frac{x_i+2x_{i+1}}{3}$, and $\eta_{j2} = \frac{y_j+2y_{j+1}}{3}$. For this quadrature rule, we have the following theorem.

Theorem 3.3.2 *If all the third partial derivatives of the function $f(x, y)$ are continuous on Ω, and their absolute values are bounded by M_{ij} in ω_{ij} for $i = 0, 1, \cdots, m-1$, and $j = 0, 1, \cdots, n-1$, then there exist $\theta_{ij} \in [-M_{ij}, M_{ij}]$ such*

that

$$
\begin{aligned}
I(f) - I_T(f) &= \frac{1}{72}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} h_i k_j \Big\{ h_i^2[f_{xx}(\xi_{i1},\eta_{j1}) + f_{xx}(\xi_{i2},\eta_{j2})] \\
&\qquad -h_i k_j[f_{xy}(\xi_{i1},\eta_{j1}) + f_{xy}(\xi_{i2},\eta_{j2})] \\
&\qquad +k_j^2[f_{yy}(\xi_{i1},\eta_{j1}) + f_{yy}(\xi_{i2},\eta_{j2})]\Big\} \\
&+ \frac{1}{3^5\times 10}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\theta_{ij}\left\{8h_i^4k_j + \frac{91}{6}h_i^3k_j^2 + \frac{91}{6}h_i^2k_j^3 + 8h_ik_j^4\right\}
\end{aligned}
$$

PROOF. Using Taylor's series expansion, we have

$$
\begin{aligned}
f(x,y) - f(\xi_{ik},\eta_{jk}) &= f_x(\xi_{ik},\eta_{jk})(x-\xi_{ik}) + f_y(\xi_{ik},\eta_{jk})(y-\eta_{jk}) \\
&\quad +\frac{1}{2}\Big\{ f_{xx}(\xi_{ik},\eta_{jk})(x-\xi_{ik})^2 \\
&\quad +2f_{xy}(\xi_{ik},\eta_{jk})(x-\xi_{ik})(y-\eta_{jk}) \\
&\quad +f_{yy}(\xi_{ik},\eta_{jk})(y-\eta_{jk})^2\Big\} + R(\xi_{ik},\eta_{jk},r_{ijk}),
\end{aligned}
$$

where $(x,y) \in \omega_{i,j}^k$ and $R(\xi_{ik},\eta_{jk},r_{ijk})$ is the Lagrange remainder given by:

$$
\begin{aligned}
R(\xi_{ik},\eta_{jk},r_{ijk}) &= \frac{1}{6}\Big\{(x-\xi_{ik})\frac{\partial}{\partial x} + (y-\eta_{jk})\frac{\partial}{\partial y}\Big\}^3 \\
&\quad \cdot f(\xi_{ik}+r_{ijk}(x-\xi_{ik}),\eta_{jk}+r_{ijk}(y-\eta_{jk}))
\end{aligned}
$$

with $0 \le r_{ijk} \le 1$, $i = 1,2,\cdots,m$, $j = 1,2,\cdots,n$, and $k = 1,2$. Integrating over $\omega_{ij}^1 \cup \omega_{ij}^2$, we get

$$
\begin{aligned}
&\int_{x_i}^{x_{i+1}}\int_{y_j}^{\delta(x)} /\{f(x,y) - f(\xi_{i1},\eta_{j1})\}dydx \\
&\qquad + \int_{x_i}^{x_{i+1}}\int_{\delta(x)}^{y_{j+1}} \{f(x,y) - f(\xi_{i2},\eta_{j2})\}dydx \\
&= \frac{1}{72}h_ik_j\Big\{h_i^2f_{xx}(\xi_{i1},\eta_{j1}) - h_ik_jf_{xy}(\xi_{i1},\eta_{j1}) + k_j^2f_{yy}(\xi_{i1},\eta_{j1}) \\
&\qquad +h_i^2f_{xx}(\xi_{i2},\eta_{j2}) - h_ik_jf_{xy}(\xi_i2,\eta_j2) + k_j^2f_{yy}(\xi_{i2},\eta_{j2})\Big\} \\
&\quad + \int_{x_i}^{x_{i+1}}\int_{y_j}^{\delta(x)} R(\xi_{i1},\eta_{j1},r_{ij1})dydx \\
&\quad + \int_{x_i}^{x_{i+1}}\int_{\delta(x)}^{y_{j+1}} R(\xi_{i2},\eta_{j2},r_{ij2})dydx \qquad (3.3.14)
\end{aligned}
$$

where $\delta(x) = (1 - \frac{x-x_i}{h_i})k_j + y_j$.

We now estimate the integrals of each of these four terms in the remainders R over $\omega_{ij}^1 \cup \omega_{ij}^2$. For the first term, we have

$$\begin{aligned}
&\left| \int_{x_i}^{x_{i+1}} \int_{y_j}^{\delta(x)} f_{xxx}(\xi_{i1} + r_{ij1}(x-\xi_{i1}), \eta_{j1} + r_{ij1}(y-y_{j1}))(x-\xi_{i1})^3 dydx \right. \\
&\left. + \int_{x_i}^{x_{i+1}} \int_{\delta(x)}^{y_{j+1}} f_{xxx}(\xi_{i2} + r_{ij2}(x-\xi_{i2}), \eta_{j1} + r_{ij2}(y-y_{j2}))(x-\xi_{i2})^3 dydx \right| \\
\leq\ & M_{ij} \int_{x_i}^{x_{i+1}} \int_{y_j}^{\delta(x)} \left|(x-\xi_{i1})^3\right| dydx + M_{ij} \int_{x_i}^{x_{i+1}} \int_{\delta(x)}^{y_{j+1}} \left|(x-\xi_{i2})^3\right| dydx \\
=\ & \frac{8}{5 \times 3^4} M_{ij} h_i^4 k_j. \qquad (3.3.15)
\end{aligned}$$

For the other three terms, we have

$$\begin{aligned}
&\left| \int_{x_i}^{x_{i+1}} \int_{y_j}^{\delta(x)} \{f_{xxy}(\xi_{i1} + r_{ij1}(x-\xi_{i1}), \eta_{j1} \right. \\
&\qquad + r_{ij1}(y-y_{j1}))(x-\xi_{i1})^2 (y-\eta_{j1})\} dydx \\
&+ \int_{x_i}^{x_{i+1}} \int_{\delta(x)}^{y_{j+1}} \{f_{xxy}(\xi_{i2} + r_{ij2}(x-\xi_{i2}), \eta_{j1} \\
&\qquad \left. + r_{ij2}(y-y_{j2}))(x-\xi_{i2})^2 (y-\eta_{j2})\} dydx \right| \\
\leq\ & \frac{91}{10 \times 3^6} M_{ij} h_i^3 k_j^2, \qquad (3.3.16) \\
&\left| \int_{x_i}^{x_{i+1}} \int_{y_j}^{\delta(x)} \{f_{xyy}(\xi_{i1} + r_{ij1}(x-\xi_{i1}), \eta_{j1} \right. \\
&\qquad + r_{ij1}(y-y_{j1}))(x-\xi_{i1})(y-\eta_{j1})^2\} dydx \\
&+ \int_{x_i}^{x_{i+1}} \int_{\delta(x)}^{y_{j+1}} \{f_{xyy}(\xi_{i2}) + r_{ij2}(x-\xi_{i2}), \eta_{j1} \\
&\qquad \left. + r_{ij2}(y-y_{j2}))(x-\xi_{i2})(y-y_{j2})^2\} dydx \right| \\
\leq\ & \frac{91}{10 \times 3^6} M_{ij} h_i^2 k_j^3, \qquad (3.3.17)
\end{aligned}$$

and

$$\begin{aligned}
&\left| \int_{x_i}^{x_{i+1}} \int_{y_j}^{\delta(x)} \{f_{yyy}(\xi_{i1} + r_{ij1}(x-\xi_{i1}), \eta_{j1} + r_{ij1}(y-y_{j1}))(y-\eta_{j1})^3\} dydx \right. \\
&\left. + \int_{x_i}^{x_{i+1}} \int_{\delta(x)}^{y_{j+1}} \{f_{yyy}(\xi_{i2} + r_{ij2}(x-\xi_{i2}), \eta_{j1} + r_{ij2}(y-y_{j2}))(y-\eta_{j2})^3\} dydx \right| \\
\leq\ & \frac{8}{5 \times 3^4} M_{ij} h_i k_j^4 \qquad (3.3.18)
\end{aligned}$$

From (3.3.15) - (3.3.18), we obtain

$$\left|\int_{x_i}^{x_{i+1}}\int_{y_j}^{\delta(x)} R(\xi_{i1},\eta_{j1},r_{ij1})dydx + \int_{x_i}^{x_{i+1}}\int_{\delta(x)}^{y_{j+1}} R(\xi_{i2},\eta_{j2},r_{ij2})dydx\right|$$

$$\leq \quad \frac{1}{3^5\times 10}\Big\{8h_i^4k_j + \frac{91}{6}h_i^3k_j^2 + \frac{91}{6}h_i^2k_j^3 + 8h_ik_j^4\Big\}M_{ij} \qquad (3.3.19)$$

Thus, there exists a $\theta_{ij} \in [-M_{ij}, M_{ij}]$ such that

$$\int_{x_i}^{x_{i+1}}\int_{y_j}^{\delta(x)} R(\xi_{i1},\eta_{j1},r_{ij1})dydx + \int_{x_i}^{x_{i+1}}\int_{\delta(x)}^{y_{j+1}} R(\xi_{i2},\eta_{j2},r_{ij2})dydx$$

$$= \quad \frac{1}{3^5\times 10}\Big\{8h_i^4k_j + \frac{91}{6}h_i^3k_j^2 + \frac{91}{6}h_i^2k_j^3 + 8h_ik_j^4\Big\}\theta_{ij} \qquad (3.3.20)$$

The main result of the theorem follows by combining (3.3.14) and (3.3.20). □

From Theorem 3.3.2, we see that the error between the numerical integral using the barycenter quadrature rule on triangular partition and the exact integral is dominated by

$$\begin{aligned} E_T(\boldsymbol{x},\boldsymbol{y}) \;:=\; \frac{1}{72}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} h_ik_j\Big\{ & h_i^2[f_{xx}(\xi_{i1},\eta_{j1}) + f_{xx}(\xi_{i2},\eta_{j2})] \\ & -h_ik_j[f_{xy}(\xi_{i1},\eta_{j1}) + f_{xy}(\xi_{i2},\eta_{j2})] \\ & +k_j^2[f_{yy}(\xi_{i1},\eta_{j1}) + f_{yy}(\xi_{i2},\eta_{j2})]\Big\} \end{aligned}$$

where all $h_i < 1$, $k_j < 1$ (or otherwise $f_{xx}(x,y)$, and $f_{yy}(x,y)$ are nearly constant). Based on this, we can pose the corresponding numerical scheme as an optimization problem.

$$\begin{aligned} \text{Min } g(\boldsymbol{x},\boldsymbol{y}) \;=\; \Big\{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} h_ik_j\Big\{ & h_i^2[f_{xx}(\xi_{i1},\eta_{j1}) + f_{xx}(\xi_{i2},\eta_{j2})] \\ & -h_ik_j[f_{xy}(\xi_{i1},\eta_{j1}) + f_{xy}(\xi_{i2},\eta_{j2})] \\ & +k_j^2[f_{yy}(\xi_{i1},\eta_{j1}) + f_{yy}(\xi_{i2},\eta_{j2})]\Big\}\Big\}^2 \end{aligned}$$

subject to constraints (3.3.11) - (3.3.12). This optimization problem is referred to as Problem $P3.3.2$. The gradient of the cost function g is given by

$$
\begin{aligned}
\frac{\partial g}{\partial x_i} \;=\; 2g\sum_{j=0}^{n-1}\Big\{ & -3h_i^2k_j[f_{xx}(\xi_{i1},\eta_{j1})+f_{xx}(\xi_{i2},\eta_{j2})] \\
& +2h_ik_j^2[f_{xy}(\xi_{i1},\eta_{j1})+f_{xy}(\xi_{i2},\eta_{j2})] \\
& -k_j^3[f_{yy}(\xi_{i1},\eta_{j1})+f_{yy}(\xi_{i2},\eta_{j2})] \\
& +h_i^3k_j[\frac{1}{3}f_{xxx}(\xi_{i2},\eta_{j2})+\frac{2}{3}f_{xxx}(\xi_{i1},\eta_{j1})] \\
& -h_i^2k_j^2[\frac{1}{3}f_{xyx}(\xi_{i2},\eta_{j2})+\frac{2}{3}f_{xyx}(\xi_{i1},\eta_{j1})] \\
& +h_ik_j^3[\frac{1}{3}f_{yyx}(\xi_{i2},\eta_{j2})+\frac{2}{3}f_{yyx}(\xi_{i1},\eta_{j1})] \\
& +3h_{i-1}^2k_j[f_{xx}(\xi_{(i-1)1},\eta_{j1})+f_{xx}(\xi_{(i-1)2},\eta_{j2})] \\
& -2h_{i-1}k_j^2[f_{xy}(\xi_{(i-1)1},\eta_{j1})+f_{xy}(\xi_{(i-1)2},\eta_{j2})] \\
& +k_j^3[f_{yy}(\xi_{(i-1)1},\eta_{j1})+f_{yy}(\xi_{(i-1)2},\eta_{j2})] \\
& +h_{i-1}^3k_j[\frac{1}{3}f_{xxx}(\xi_{(i-1)1},\eta_{j1})+\frac{2}{3}f_{xxx}(\xi_{(i-1)2},\eta_{j2})] \\
& -h_{i-1}^2k_j^2[\frac{1}{3}f_{xyx}(\xi_{(i-1)1},\eta_{j1})+\frac{2}{3}f_{xyx}(\xi_{(i-1)2},\eta_{j2})] \\
& +h_{i-1}k_j^3[\frac{1}{3}f_{yyx}(\xi_{(i-1)1},\eta_{j1})+\frac{2}{3}f_{yyx}(\xi_{(i-1)2},\eta_{j2})]\Big\}
\end{aligned}
$$

and

$$\begin{aligned}\frac{\partial g}{\partial y_j} = 2g\sum_{j=0}^{n-1}\Big\{ & -h_i^3[f_{xx}(\xi_{i1},\eta_{j1})+f_{xx}(\xi_{i2},\eta_{j2})] \\ & +2h_i^2k_j[f_{xy}(\xi_{i1},\eta_{j1})+f_{xy}(\xi_{i2},\eta_{j2})] \\ & -3h_ik_j^2[f_{yy}(\xi_{i1},\eta_{j1})+f_{yy}(\xi_{i2},\eta_{j2})] \\ & +h_i^3k_j[\frac{1}{3}f_{xxy}(\xi_{i2},\eta_{j2})+\frac{2}{3}f_{xxy}(\xi_{i1},\eta_{j1})] \\ & -h_i^2k_j^2[\frac{1}{3}f_{xyy}(\xi_{i2},\eta_{j2})+\frac{2}{3}f_{xyy}(\xi_{i1},\eta_{j1})] \\ & +h_ik_j^3[\frac{1}{3}f_{yyy}(\xi_{i2},\eta_{j2})+\frac{2}{3}f_{yyy}(\xi_{i1},\eta_{j1})] \\ & +h_i^3[f_{xx}(\xi_{i1},\eta_{(j-1)1})+f_{xx}(\xi_{i2},\eta_{(j-1)2})] \\ & -2h_i^2k_{j-1}[f_{xy}(\xi_{i1},\eta_{(j-1)1})+f_{xy}(\xi_{i2},\eta_{(j-1)2})] \\ & +3h_ik_{j-1}^2[f_{yy}(\xi_{i1},\eta_{(j-1)1})+f_{yy}(\xi_{i2},\eta_{(j-1)2})] \\ & +h_i^3k_j[\frac{1}{3}f_{xxy}(\xi_{i1},\eta_{(j-1)1})+\frac{2}{3}f_{xxy}(\xi_{i2},\eta_{(j-1)2})] \\ & -h_i^2k_{j-1}^2[\frac{1}{3}f_{xyy}(\xi_{i1},\eta_{(j-1)1})+\frac{2}{3}f_{xyy}(\xi_{i2},\eta_{(j-1)2})] \\ & +h_ik_{j-1}^3[\frac{1}{3}f_{yyy}(\xi_{i1},\eta_{(j-1)1})+\frac{2}{3}f_{yyy}(\xi_{i2},\eta_{(j-1)2})]\Big\}\end{aligned}$$

for $i = 0, 1, \cdots, m-1$, and $j = 0, 1, \cdots, n-1$, where ξ_{ij}, η_{ij}, h_i, and k_j are as defined previously.

3.3.2 Method for not sufficiently smooth integrands

The technique presented in the previous section requires that the third order partial derivatives of $f(x, y)$ on Ω are continuous. Following the notation defined in Subsection 3.3.1, we now develop a numerical scheme for the case in which only first order partial derivatives of $f(x, y)$ are continuous on Ω. This is based on the idea of approximating the error in a quadrature rule by the difference between the numerical integrals by the quadrature rule and a higher order quadrature rule.

The trapezoidal rule on the triangle

$$\omega_{\xi,\eta}^1 = \{(x, y) | 0 < x < \xi,\ 0 < y < (1 - \frac{x}{\xi})\eta\}$$

is defined by

$$\int_{\omega^1_{\xi,\eta}} f(x,y)d\omega \approx \frac{1}{6}\{(f(0,0)+f(\xi,0)+f(0,\eta))\}\xi\eta,$$

which is obtained by approximating $f(x,y)$ by the linear function

$$L(x,y) = f(0,0) + \frac{f(\xi,0)-f(0,0)}{\xi}x + \frac{f(\xi,\eta)-f(\xi)}{\eta}y \qquad \text{on } \omega^1_{\xi,\eta}.$$

Now we fit $f(x,y)$ on $\omega^1_{\xi,\eta}$ by a quadratic function

$$Q(x,y) = a_1 + a_2x + a_3y + a_4x^2 + a_5xy + a_6y^2$$

such that

$$Q(0,0) = f(0,0) = f_1, \quad Q(\frac{\xi}{2},0) = f(\frac{\xi}{2},0) = f_2, \quad Q(\xi,0) = f(\xi,0) = f_3,$$

$$Q(\frac{\xi}{2},\frac{\eta}{2}) = f(\frac{\xi}{2},\frac{\eta}{2}) = f_4, \quad Q(0,\eta) = f(0,\eta) = f_5, \quad Q(0,\frac{\eta}{2}) = f(0,\frac{\eta}{2}) = f_6.$$

Let

$$Q_0(x,y) = b_2x + b_3y + b_4x^2 + b_5xy + b_6y^2 = Q(x,y) - f(0,0) \text{ on } \omega^1_{i,j}.$$

Then, we have

$$f_2 - f_1 = b_2\frac{\xi}{2} + b_4\frac{\xi^2}{4} \tag{3.3.21}$$

$$f_3 - f_1 = b_2\xi + b_4\xi^2 \tag{3.3.22}$$

$$f_4 - f_1 = b_2\frac{\xi}{2} + b_3\frac{\eta}{2} + b_4\frac{\xi^2}{4} + b_5\frac{\xi\eta}{4} + b_6\frac{\eta^2}{4} \tag{3.3.23}$$

$$f_5 - f_1 = b_3\eta + b_6\eta^2 \tag{3.3.24}$$

$$f_6 - f_1 = b_3\frac{\eta}{2} + b_6\frac{\eta^2}{4} \tag{3.3.25}$$

Solving the system (3.3.21) - (3.3.25), we have

$$b_2 = \frac{-f_3 + 4f_2 - 3f_1}{\xi}$$

$$b_3 = \frac{-f_5 + 4f_6 - 3f_1}{\eta}$$

$$b_4 = \frac{2(f_3 - 2f_2 + f_1)}{\xi^2}$$

$$b_5 = \frac{4(f_1 - f_2 + f_4 - f_6)}{\xi\eta}$$

$$b_6 = \frac{2(f_5 - 2f_6 + f_1)}{\eta^2}$$

Therefore

$$\begin{aligned} Q_0(x,y) &= \frac{-f_3+4f_2-3f_1}{\xi}x+\frac{-f_5+4f_6-3f_1}{\eta}y+\frac{2(f_3-2f_2+f_1)}{\xi^2}x^2 \\ &\quad +\frac{4(f_1-f_2+f_4-f_6)}{\xi\eta}xy+\frac{2(f_5-2f_6+f_1)}{\eta^2}y^2 \end{aligned}$$

and

$$\begin{aligned} \int_0^{\xi}\int_0^{(1-\frac{x}{\xi})\eta} Q_0(x,y)dydx &= \frac{-3f_1+f_2+f_4+f_6}{6}\xi\eta \\ \int_0^{\xi}\int_0^{(1-\frac{x}{\xi})\eta} (Q(x,y)-L(x,y))dydx &= \frac{f_2+f_4+f_6-f_1-f_3-f_5}{6}\xi\eta \end{aligned}$$

Similarly, on

$$\omega^2_{\xi\eta}=\{(x,y)\ :\ 0\le x\le\xi,\ (1-\frac{x}{\xi})\eta\le y\le\eta\},$$

we have

$$\int_0^{\xi}\int_{(1-\frac{x}{\xi})\eta}^{\eta}(Q(x,y)-L(x,y))dydx=\frac{f_7+f_4+f_9-f_8-f_3-f_5}{6}\xi\eta,$$

where

$$\begin{aligned} &L(x,y)=f(\xi,\eta)+\frac{f(\xi,0)-f(\xi,\eta)}{\xi}x+\frac{f(\xi,\eta)-f(\xi)}{\eta}y, \\ &f_7=f(\frac{\xi}{2},\eta),\qquad f_8=f(\xi,\eta),\qquad f_9=f(\xi,\frac{\eta}{2}). \end{aligned}$$

Transforming $\omega^1_{\xi\eta}$ and $\omega^2_{\xi\eta}$ into the triangular elements ω^1_{ij} and ω^2_{ij} of the partition $P^*_{m\times n}$, respectively , we have

$$\begin{aligned} &\int_{x_i}^{x_{i+1}}\int_{y_j}^{y_{j+1}}(Q(x,y)-L(x,y))dydx \\ =\ &\frac{1}{6}h_ik_j\big\{f(\xi_i,y_j)+f(x_i,\eta_j)+f(x_{i+1},\eta_j)+2f(\xi_i,\eta_j) \\ &\qquad -f(x_i,y_j)-2f(x_{i+1},y_j)-2f(x_i,y_{j+1})-f(x_{i+1},y_{j+1})\big\} \end{aligned}$$

Summing over $i=0,1,2,\cdots,m-1$, and $j=0,1,2,\cdots,n-1$, we obtain

$$\begin{aligned} &\int_{\Omega}\big\{(Q(x,y)-L(x,y)\big\}dxdy \\ =\ &\frac{1}{6}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}h_ik_j\Big\{f(\xi_i,y_j)+f(x_i,\eta_j)+f(\xi_i,y_{j+1})+f(x_{i+1},\eta_j)+2f(\xi_i,\eta_j) \\ &\qquad -f(x_i,y_j)-2f(x_{i+1},y_j)-2f(x_i,y_{j+1})-f(x_{i+1},y_{j+1})\Big\} \end{aligned}$$

Let $I_{TR}(f)$ denote the trapezoidal rule on the partition $P^*_{m\times n}$. Then, the error $I(f) - I_{TR}(f)$ is domained by

$$\begin{aligned} E_{QTR}(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{6}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} h_i k_j \Big\{ f(\xi_i, y_j) + f(x_i, \eta_j) + f(\xi_i, y_{j+1}) + f(x_{i+1}, \eta_j) \\ &\qquad +2f(\xi_i, \eta_j) - f(x_i, y_j) - 2f(x_{i+1}, y_j) \\ &\qquad -2f(x_i, y_{j+1}) - f(x_{i+1}, y_{j+1}) \Big\} \end{aligned}$$

This motivates us to define the following optimization problem:

$$\begin{aligned} \min g(\boldsymbol{x},\boldsymbol{y}) &= \Big\{ \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} h_i k_j \{ f(\xi_i, y_j) + f(x_i, \eta_j) + f(\xi_i, y_{j+1}) + f(x_{i+1}, \eta_j) \\ &\qquad +2f(\xi_i, \eta_j) \qquad - f(x_i, y_j) - 2f(x_{i+1}, y_j) \\ &\qquad -2f(x_i, y_{j+1}) - f(x_{i+1}, y_{j+1}) \} \Big\}^2 \end{aligned}$$

subject to (3.3.11) - (3.3.12). This problem is referred to as Problem $P3.3.3$. The gradient of g is given by

$$\begin{aligned} \frac{\partial g}{\partial x_i} &= 2g \sum_{j=0}^{n-1} \Big\{ - k_j \{ f(\xi_i, y_j) + f(x_i, \eta_j) + f(\xi_i, y_{j+1}) \\ &\qquad +f(x_{i+1}, \eta_j) + 2f(\xi_i, \eta_j) - f(x_i, y_j) \\ &\qquad -2f(x_{i+1}, y_j) - 2f(x_i, y_{j+1}) - f(x_{i+1}, y_{j+1}) \} \\ &\quad +k_j \{ f(\xi_{i-1}, y_j) + f(x_{i-1}, \eta_j) + f(\xi_{i-1}, y_{j+1}) \\ &\qquad +f(x_i, \eta_j) + 2f(\xi_{i-1}, \eta_j) - f(x_{i-1}, y_j) \\ &\qquad -2f(x_i, y_j) - 2f(x_{i-1}, y_{j+1}) - f(x_i, y_{j+1}) \} \\ &\quad +h_i k_j \{ \frac{1}{2} f_x(\xi_i, y_j) + f_x(x_i, \eta_j) + \frac{1}{2} f'_x(\xi_i, y_{j+1}) \\ &\qquad +f_x(\xi_i, \eta_j) - f_x(x_i, y_j) - 2f_x(x_i, y_{j+1}) \} \\ &\quad -h_{i-1} k_j \{ \frac{1}{2} f_x(\xi_{i-1}, y_j) + \frac{1}{2} f_x(\xi_{i-1}, y_{j+1}) + f_x(x_i, \eta_j) \\ &\qquad +f_x(\xi_{i-1}, \eta_j) - 2f_x(x_i, y_j) - f_x(x_i, y_{j+1}) \} \Big\} \end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial g}{\partial y_j} \;=\; 2g\sum_{j=0}^{n-1}\Big\{ & -h_i\{f(\xi_i,y_j)+f(x_i,\eta_j)+f(\xi_i,y_{j+1}) \\
& +f(x_{i+1},\eta_j)+2f(\xi_i,\eta_j)-f(x_i,y_j) \\
& -2f(x_{i+1},y_j)-2f(x_i,y_{j+1})-f(x_{i+1},y_{j+1})\} \\
& +h_i\{f(\xi_i,y_{j-1})+f(x_i,\eta_{j-1})+f(\xi_i,y_j) \\
& +f(x_{i+1},\eta_{j-1})+2f(\xi_i,\eta_{j-1})-f(x_i,y_{j-1}) \\
& -2f(x_{i+1},y_{j-1})-2f(x_i,y_j)-f(x_{i+1},y_j)\} \\
& +h_ik_j\{f_y(\xi_i,y_j)+\frac{1}{2}f_y(x_i,\eta_j)+\frac{1}{2}f_y(x_{i+1},\eta_j) \\
& +f_y(\xi_i,\eta_j)-f_y(x_i,y_j)-2f_x(x_{i+1},y_j)\} \\
& -h_ik_{j-1}\{\frac{1}{2}f_y(\xi_i,\eta_{j-1})+f_y(\xi_i,y_j)+\frac{1}{2}f_y(x_{i+1},\eta_{j-1}) \\
& +f_y(\xi_i,\eta_{j-1})-2f_y(x_i,y_j)-f_x(x_{i_1},y_j)\}\Big\}
\end{aligned}
$$

3.3.3 Numerical Experiments

To demonstrate the effectiveness and usefulness of the above methods, some numerical experiments are carried out. The package FFSQP [116] for solving general nonlinear programming problems is used as the minimizer for Problems $P3.3.1$, $P3.3.2$, and $P3.3.3$. All computations are performed in Fortran double precision on an SGI workstation. For all the test problems solved, we use the uniform mesh (for given m and n) as the initial mesh and numerical integral obtained from the quadrature rule on the uniform triangle partition with $2 \times 1000 \times 1000$ sub-triangles as the exact integral $I(f)$. For simplicity, we use R, T, QTR denote the barycenter rule with rectangular partition, the barycenter rule with triangular partition and the trapezoidal rule with quadrature approximation of function f(x,y) on triangular partition, respectively.

Example 3.3.1.

$$I(f)=\int_\Omega \frac{100\sin(10x+9y)}{xy}d\Omega$$

This function is not analytically integrable on the region

$$\Omega=\{(x,y)\ :\ 0.1<x<0.6,\ 0.1<y<0.6\}.$$

Grid	$g(\boldsymbol{x}_{init})$	$g(\boldsymbol{x}_{final})$	*Error*	*Err on Unif.Grid*
R on G1	-46.29904	-45.82994	1.54e-2	4.85e-1
T on G2	-46.04443	-45.81842	3.91e-3	2.30e-1
QTR on G3	-45.48281	-45.81509	5.88e-4	3.31e-1

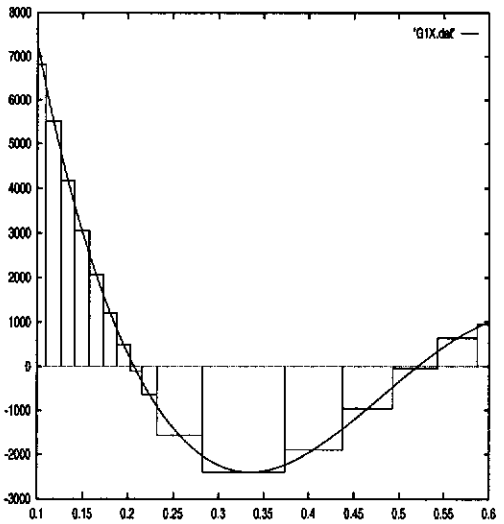
Table 3.3.1: Results for Example 3.3.1 using different quadrature rules and grids

We now divide Ω into 16×16 uniform subregions and use Problem $P3.3.1$, $P3.3.2$, $P3.3.3$ to construct three integral grids labeled with G1, G2 and G3, respectively. Table 3.3.1 contain a list of the initial and final values of the cost function ($g(\boldsymbol{x}_{init})$ and $g(\boldsymbol{x}_{final})$) and the absolute errors in the numerical integrals computed, respectively, by the barycenter quadrature rule with rectangular partition, the barycenter quadrature rule with triangular partition and the trapezoidal rule with quadrature approximation of function $f(x, y)$ on triangular partition. For comparison, the errors in the numerical integrals from the above methods on the uniform mesh are listed in the last column. From this table we see that the results obtained by using the optimal grids are at least one order of magnitude more accurate than those based on the uniform grids. Also Problems $P3.3.2$ and $P3.3.3$ produce more accurate results than those obtained from Problem $P3.3.1$. To visualize the grids obtained from the solutions of Problems $P3.3.1$, $P3.3.2$ and $P3.3.3$, see Figures 3.3.1 - 3.3.4

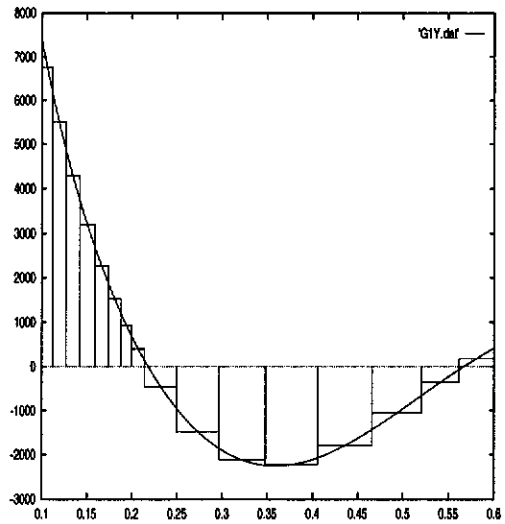
Example 3.3.2.

$$I(f) = 10 \int_{\Omega} \exp\{-0.2(x-2.0)^2(y-3.0)^2\} d\Omega$$

where $\Omega = \{(x, y) : 0 < x < 5,\ 1 < y < 6\}$, which is partitioned into 16×16 subregions. Problems $P3.3.1$, $P3.3.2$, $P3.3.3$ are used to construct three integral grids, labeled by G1, G2, G3 respectively. Table 3.3.2 contains a list of the initial and final values of the cost function. From this table we see that the results obtained by using the optimal grid is at least two orders of magnitudes more accurate than that based on the uniform grid. Figures 3.3.5 and 3.3.6 show the difference between optimal grid and uniform grid for trapezoidal rule on partition $P^*_{m \times n}$.

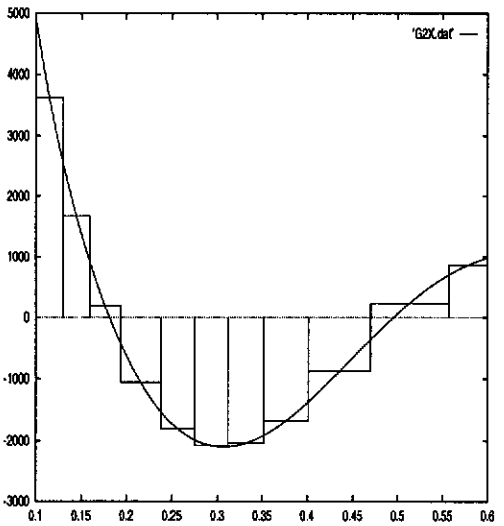


(a) A section perpendicular to y axis

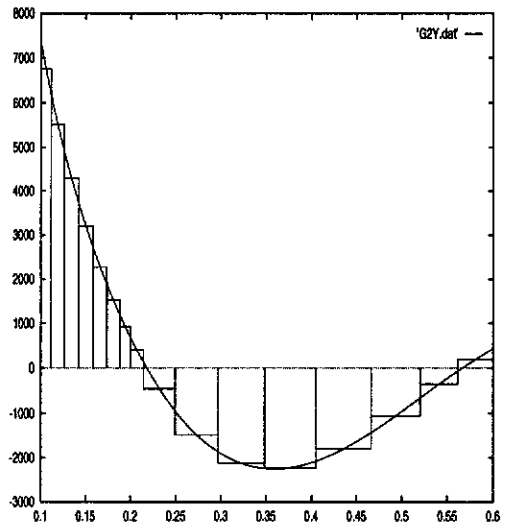


(b) A section perpendicular to x axis

Figure 3.3.1: Graphs of the numerical solution of Example 3.3.1 using R rule on G1



(a) A section perpendicular to y axis

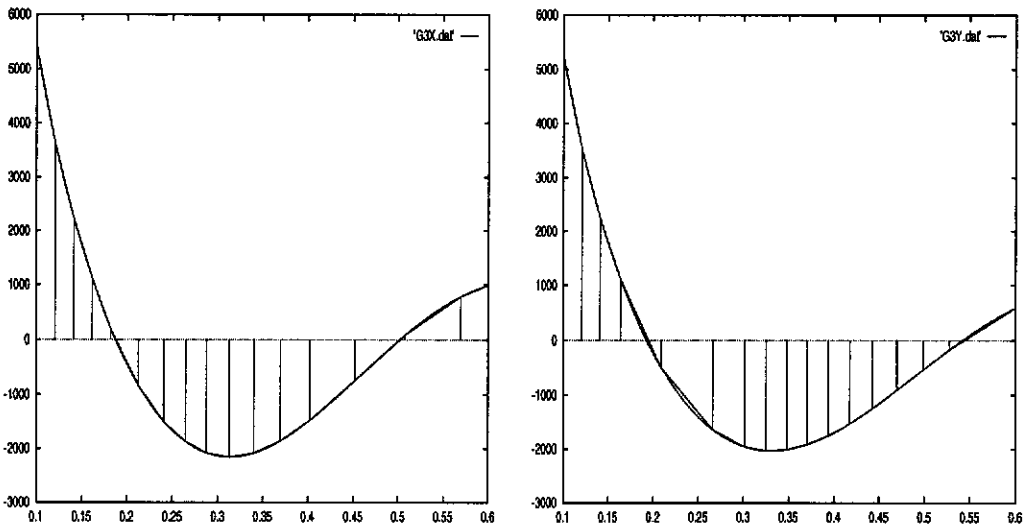


(b) A section perpendicular to x axis

Figure 3.3.2: Graphs of the numerical solution of Example 3.3.1 using T rule on G2

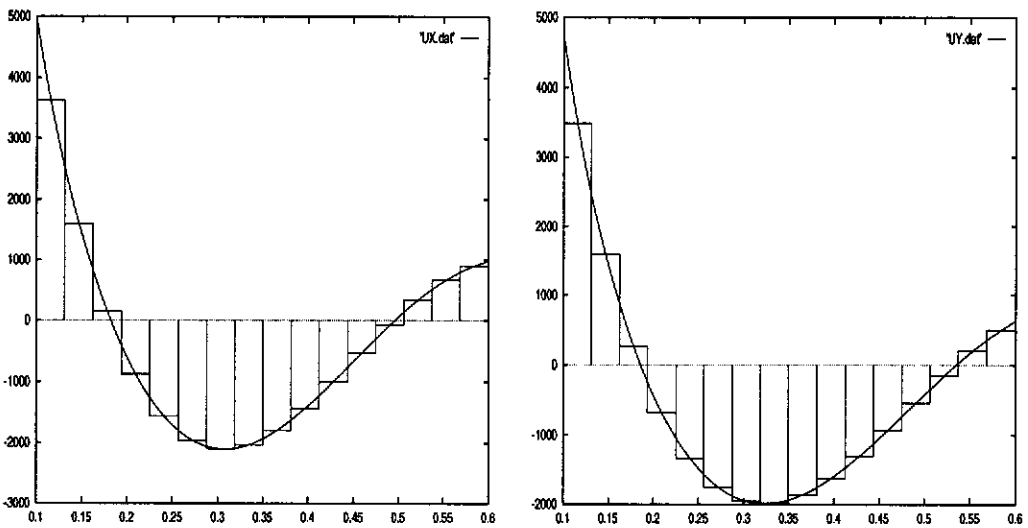
Grid	$g(\boldsymbol{x}_{init})$	$g(\boldsymbol{x}_{final})$	*Error*	*Err on Unif.Grid*
R on G1	156.1982	156.08733	1.52e-3	1.12e-1
T on G2	156.2460	156.08538	4.38e-4	1.60e-1
QTR on G3	155.8580	156.08246	3.35e-3	2.20e-1

Table 3.3.2: Results for test2 using different quadrature rules and grids



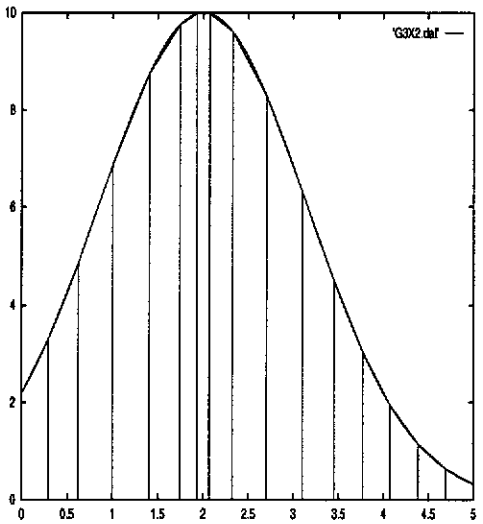
(a) A section perpendicular to y axis (b) A section perpendicular to x axis

Figure 3.3.3: Graphs of the numerical solution of Example 3.3.1 using QTR rule on G3

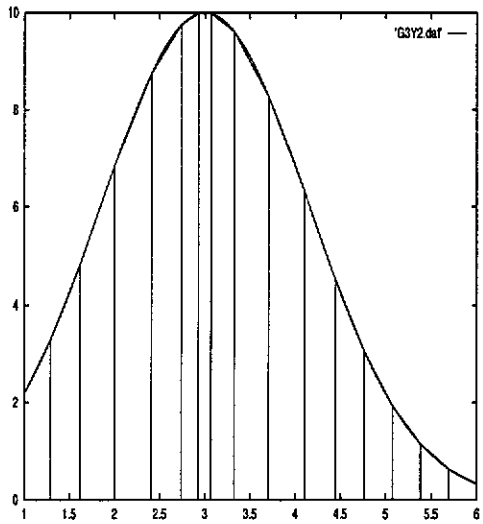


(a) A section perpendicular to y axis (b) A section perpendicular to x axis

Figure 3.3.4: R rule on Uniform Grid

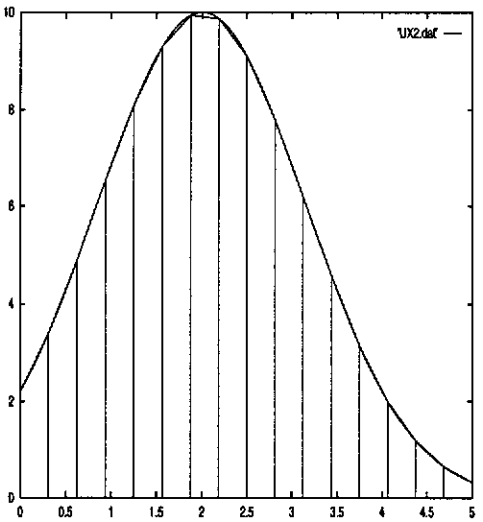


(a) A section perpendicular to y axis

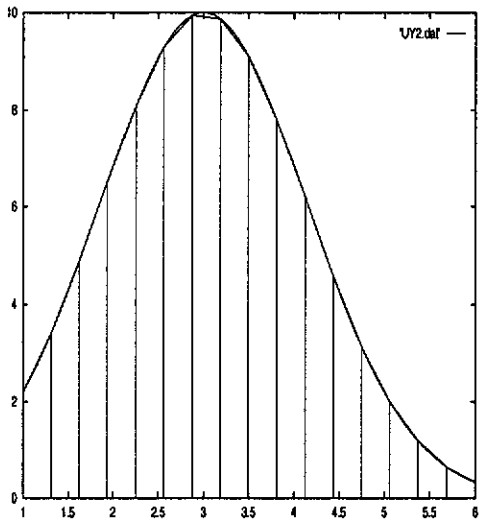


(b) A section perpendicular to x axis

Figure 3.3.5: Graphs of the numerical solution of Example 3.3.2 using QTR rule on G3



(a) A section perpendicular to y axis



(b) A section perpendicular to x axis

Figure 3.3.6: trapezoidal rule on Uniform Grid.

3.4 Conclusion

In this chapter, we present some numerical methods for constructing optimal grids for numerical integrations in one and two dimensions by some well–known quadrature rules. In this approach, a conventional numerical integration problem is posed as an optimization problem so that the solution of the latter yields the optimal grid. Numerical experiments are performed to verify the effectiveness of the approach. The numerical results show that, for a fixed number of mesh nodes, the numerical integrals on a grid from the present method is at least one order of magnitudes more accurate than that on the uniform grid.

We comment that an optimal solution of the problems in this chapter may not exist, since the feasible region constrained by $x_i - x_{i+1} < 0,\ i = 0, 1, \cdots, m-1$ and/or $y_j - y_{j+1} < 0,\ j = 0, 1, \cdots, n-1$ is an open set. To overcome this problem, in practice the above constraints are replaced by $x_i - x_{i+1} \leq \epsilon_i$ and $y_j - y_{j+1} \leq \delta_j$, where ϵ_i and δ_j are user's chosen small positive numbers. If the integrand is continuous in $\overline{\Omega}$, then the above problem has at least one solution, though multiple solutions may exist.

Chapter 4

Optimal Recharge and Driving Strategies for a Battery – Powered Electric Vehicle

4.1 Introduction

Due to the advancement in the battery technology, there is a great revival of interest among the researchers to return to work on the development of battery-powered electric vehicles over the past decade. For details, see [12], [75], [96] and [106].

The main components of an electric vehicle are: (i) battery cells which are energy storage system; (ii) a traction system which consists of an electric motor, a steeling wheel and motor controller; and (iii) a friction system which is a mechanical braking device. The driver can control the power applied to the traction system as well as to the mechanical braking system. The major problem associated with a battery- powered electric vehicle is in its batteries. The total weight of the batteries is directly related to the charging capacity of the electric vehicle. Due to the limited charging capacity, the vehicle can only travel for a relatively short distance when compared with a conventional gasoline-powered vehicle. It needs to be recharged much more frequently.

In this chapter, we consider a situation in which an electric vehicle is to be driven on an even road or an undulating road connecting two given cities. The distance between the two cities is beyond the driving range of the vehicle

without recharging its batteries. In other words, the batteries are required to be recharged before completing the journal. The rest of this chapter is organized as follows:

We construct a model for the battery - powered electric vehicle in Section 4.2. In Section 4.3, we formulate an optimization problem in which the vehicle is to be driven on an even road. The recharge points, and the speed of the vehicle are considered as decision variables. The objective is to minimize the completion time of the journey with respect to these decision-variables. The distance between the two cities, the maximum allowable power to be applied to the motor, and the number of battery cells are assumed fixed. In this optimization problem, the capacity of each of the battery cells is regarded as the state. Clearly, the state will exhibit a jump at each switching time. We further assume that the maximum allowable number of recharge points is fixed. A challenging task is to remove this assumption in that theoretical results are obtained for determining this maximum number of recharge points. To solve this optimization problem, we first fix a number of recharge points . This leads to a simpler problem in which only the switching times and the vehicle speed are to be chosen optimally. The control parameterization enhancing transform (CPET) (see [100] for details) is then used to transform the problem into a standard optimal control problem solvable by MISER3.2 (cf. [41] and [42]). We then increase the number of the recharge points and solve the corresponding standard optimal control problem. This process is repeated by gradually increasing the number of the recharge points until the maximum allowable number is reached. In Section 4.4, we formulate an optimization problem in which the recharge points are to be chosen from several fixed points rather than free as in Section 4.3. A computational method similar to that given in Section 4.3 is developed for solving this optimization problem. Note that in both Sections 4.3 and 4.4 the vehicle is driven on an even road, and we do not consider acceleration and deceleration effects in each of the time intervals. Thus, we construct a new model in Section 4.5. This new model takes into account of acceleration effects. In Sections 4.6 and 4.7, we consider the case in which the vehicle is driven on an

undulating road. In Section 4.6, we formulate an optimization problem in which the recharge points are fixed, and the switching times and the corresponding speed of the vehicle are considered as decision variables. In Section 4.7, recharge points as well as the switching times and the corresponding speed of the vehicle are taken as decision variables. In Section 4.8, five examples are constructed and solved by using the methods developed in the previous sections.

4.2 Battery-Powered Electric Car Model

Let t denote the time, and p the power being applied to the motor. The power p, which is a function of t, is a control variable. It is assumed that

$$0 \leq p(t) \leq P, \qquad \text{for all } t \in [0, t_{final}]$$

where t_{final} is the specified time for completing the journey, and P is the maximum allowable power to be applied to the motor. The electric power is directly applied to the motor. We assume that the control force at the wheel of the vehicle is:

$$F = \frac{p}{v}, \tag{4.2.1}$$

where $v = v(t)$ is the speed of the vehicle. We also assume that the resistance on the vehicle due to the friction and air is:

$$r(v) = m(b_1 + b_2 v), \tag{4.2.2}$$

where b_1 and b_2 are known positive constants, and m is the mass of the vehicle. The dynamics of the vehicle is described by the following system of differential equations.

$$\frac{dx}{dt} = v, \tag{4.2.3}$$

$$m\frac{dv}{dt} = \frac{p}{v} - r(v) - mg\sin(\theta) \tag{4.2.4}$$

with initial conditions

$$v(0) = 0, \tag{4.2.5}$$

$$x(0) = 0 \tag{4.2.6}$$

where $\theta = \theta(x)$ is the angle of slope of the road at location $x = x(t)$, and g is the acceleration due to gravity. The speed of the vehicle is assumed to satisfy the following obvious constraint.

$$v(t) \geq 0.$$

For each battery cell, the recharge-discharge rate at time t is:

$$\frac{dq}{dt} = \begin{cases} -c_1 p_1, & \text{discharge} \\ \frac{c_2}{q+c_3} - \exp\{c_4 - c_5 t\}, & \text{recharge} \end{cases} \tag{4.2.7}$$

with initial condition :

$$q(0) = q_0,$$

where c_i, $i = 1, \cdots, 5$, are given positive constants, and p_1 is the power flowing out from each battery cell. The function q is required to satisfy the following constraint.

$$q_{\min} \leq q(t) \leq q_{\max}.$$

4.3 The Problem on An Even Road

In this section, the road connecting two given cities is assumed to be even. The distance between the two cities, the maximum allowable power to be applied to the motor, and the number of battery cells are assumed to be fixed.

4.3.1 Problem Formulation

Let $[0, t_{final}]$ be the total driving time, and let $\{t_i\}$, $i = 0, 1, \cdots, 2l$, be switching times, which satisfy

$$0 = t_0 < t_1 < \cdots < t_{2l} < t_{2l+1} = t_{final}$$

where l is the number of recharge points , $[t_{2i-1},\ t_{2i}]$, $i = 1, \cdots, l$, are the time intervals at which the vehicle stops for recharging its batteries, while $[t_{2i-2},\ t_{2i-1}]$, $i = 1, \cdots, l+1$, are the time intervals at which the vehicle is driven at a constant speed.

We ignore the times taken for acceleration (respectively, deceleration) at the beginning (respectively, ending) of each of the time intervals. Thus, in each time interval, we obtain

$$\frac{dv}{dt} = 0 \tag{4.3.1}$$

From (4.2.2), (4.2.4), and (4.3.1), we obtain

$$\begin{aligned} p &= v \times r(v) \\ &= v \times m(b_1 + b_2 v) \end{aligned} \tag{4.3.2}$$

and the power flowing out from each of the n parallel battery cells is:

$$p_1 = v \times \frac{m(b_1 + b_2 v)}{n}. \tag{4.3.3}$$

From (4.2.7) and (4.3.3), we have

$$\frac{dq}{dt} = \begin{cases} -c_1 \frac{(b_1 v + b_2 v^2)\cdot m}{n}, & t \in [t_{2i-2}, t_{2i-1}) \\ \frac{c_2}{q+c_3} - \exp\{c_4 - c_5(t - t_{2i-1})\}, & t \in [t_{2i-1}, t_{2i}) \\ -c_1 \frac{(b_1 v + b_2 v^2)\cdot m}{n}, & t \in [t_{2l}, t_{2l+1}] \end{cases}$$
$$\text{here} \qquad i = 1, \cdots, l \tag{4.3.4}$$

with initial condition

$$q(0) = q_0. \tag{4.3.5}$$

Define

$$\mathcal{T} = \{t = (t_1, \cdots, t_{2l}) :\ t_i \in \mathbb{R},\ t_{i-1} < t_i,\ i = 1, \cdots, 2l,\ l \in \mathbb{N}\}$$

where $\mathbb{R}$, and $\mathbb{N}$ are the set of all real numbers and the set of all natural numbers, respectively.

A Borel measurable function $v : [0,T] \longrightarrow \mathbb{R}$ is called an admissible control. Let $\mathcal{V}$ be the set of all such admissible controls. For each $(\boldsymbol{t}, v, l) \in \mathcal{T} \times \mathcal{V} \times \mathbb{N}$, let $q(\cdot|\boldsymbol{t}, v, l)$ denote the corresponding solution of system (4.3.4)–(4.3.5).

We may now specify the corresponding optimization problem formally as follows:
given the dynamical system (4.3.4)–(4.3.5), find a $(\boldsymbol{t}, v, l) \in \mathcal{T} \times \mathcal{V} \times \mathbb{N}$ such that the cost function

$$g_1(\boldsymbol{t}, v, l) = t_{final}(\boldsymbol{t}, v, l) \tag{4.3.6}$$

is minimized subject to the constraints

$$q_{\min} \leq q(t) \leq q_{\max}, \tag{4.3.7}$$

$$0 < p = m(av + bv^2) < P, \tag{4.3.8}$$

$$0 < t_i - t_{i-1}, \qquad i = 1, \cdots, 2l+1, \tag{4.3.9}$$

$$\int_0^{t_{final}} v dt = S, \tag{4.3.10}$$

$$l \leq L, \tag{4.3.11}$$

where n, l and v are, respectively, the number of battery cells, number of recharge points, and the corresponding speed. $t_{final} = t_{2l+1}$ is the required time for completing the journey, S is the total distance traveled, and L is the upper bound of the number of the recharge points. Let this problem be referred to as Problem P_1.

4.3.2 Transformation

To solve Problem P_1, we need to determine the optimal number of recharge points l, the optimal switching time points $\boldsymbol{t} = (t_1, \cdots, t_{2l})$, and the optimal speed v. Let us initially set the number of recharge points to be k. Then only the optimal switching time points with a fixed number of recharge points, and the corresponding optimal speed are to be determined. This simplified problem is referred to as Problem P_1^k.

We now apply the Control Parameterization Enhancing Transform (CPET) [100, 50] to Problem P_1^k. Let $s \in [0, 2k+1]$ be a new time variable, and define

$$\eta(s) = \sum_{i=1}^{2k+1} \eta_i \chi_{[i-1,i)}(s)$$

where $\chi_{[i-1,i)}(s)$ is the characteristic function on the interval [i-1,i), and the $\eta_i's$ are nonnegative constants. Clearly, $\eta(s)$, which is called an enhancing control, is a nonnegative piecewise constant function defined on [0,$2k$+1] with fixed switching points $\{1, 2, \cdots, 2k\}$. The CPET:

$$\begin{aligned} \frac{dt}{ds} &= \eta(s) \\ t(0) &= 0 \end{aligned}$$

maps $t \in [0, t_{final}]$ to $s \in [0, 2k+1]$, where

$$\eta(s) = t_i - t_{i-1}, \qquad s \in [i-1, i), \quad i = 1, 2, \cdots, 2k+1,$$

satisfying

$$\int_0^{2k+1} \eta(s) ds = t_{final}.$$

Let Θ denote the class of all such enhancing controls. Under the CPET, the system dynamics (4.3.4) and (4.3.5) becomes

$$\frac{d}{ds}\begin{pmatrix} \hat{q}(s) \\ t(s) \end{pmatrix} = \begin{pmatrix} \eta \begin{cases} -c_1 \frac{(b_1\hat{v}+b_2\hat{v}^2)\cdot m}{n}, & s \in [2i-2, 2i-1) \\ \frac{c_2}{\hat{q}+c_3} - \exp\{c_4 - c_5(t-2i+1)\}, & s \in [2i-1, 2i) \\ -c_1 \frac{(b_1\hat{v}+b_2\hat{v}^2)\cdot m}{n}, & s \in [2k, 2k+1] \end{cases} \\ \eta(s) \end{pmatrix} \tag{4.3.12}$$

with initial condition

$$\begin{pmatrix} \hat{q}(0) \\ t(0) \end{pmatrix} = \begin{pmatrix} q_0 \\ 0 \end{pmatrix} \tag{4.3.13}$$

where $\hat{q}(s) = q(t(s))$, $\hat{v}(s) = v(t(s))$, $s \in [0, 2k+1]$ and $i = 1, \cdots, k$, while

the constraints (4.3.7) - (4.3.11) reduce to

$$q_{min} \leq \hat{q}(s) \leq q_{max}, \tag{4.3.14}$$

$$0 < p = m(b_1\hat{v} + b_2\hat{v}^2) < P, \tag{4.3.15}$$

$$0 < \eta_i \qquad i = 1, \cdots, 2k+1, \tag{4.3.16}$$

$$\int_0^{2k+1} \hat{v}ds = S, \tag{4.3.17}$$

$$l \leq L. \tag{4.3.18}$$

Lemma 4.3.1 *The all-time state constraint (4.3.14) is equivalent to the following terminal state inequality constraints:*

$$q_{min} \leq \hat{q}(2i-1) \qquad i = 1, \cdots, k+1, \tag{4.3.19}$$

$$\hat{q}(2i) \leq q_{max} \qquad i = 1, \cdots, k. \tag{4.3.20}$$

PROOF. Obviously: (4.3.14) $\Longrightarrow$ (4.3.19) and (4.3.20).

Conversely, suppose that (4.3.19) and (4.3.20) exist. Then:
For each i=1, $\cdots, k+1$, $\hat{q}(s)$, s $\in$[2i-2, 2i-1], is monotonically decreasing, since the vehicle is driven at a constant speed and no battery recharge has taken place in this sub-interval. Thus

$$q_{min} \leq \hat{q}(2i-1) \leq \hat{q}(s) \leq \hat{q}(2i-2) \leq q_{max}.$$

For each i=1, $\cdots, k$, $\hat{q}(s)$, s $\in$[2i-1, 2i], is monotonically increasing, since the vehicle stops for recharging the batteries in this sub-interval. Thus

$$q_{min} \leq \hat{q}(2i-1) \leq \hat{q}(s) \leq \hat{q}(2i) \leq q_{max}.$$

Therefore, the condition (4.3.14) is satisfied, and the results follows. □

Let $\mathcal{V}_1$ be the set of all those feasible $\hat{v}(s)$, such that the conditions (4.3.15)–(4.3.20) are satisfied. Problem P_1^k is now transformed into the following optimal control problem:
given the dynamical system (4.3.12)–(4.3.13), find a $(\boldsymbol{\eta}, \hat{v}) \in \Theta \times \mathcal{V}_1$ such that the cost function

$$\hat{g}_1(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{2k+1} \eta_i \tag{4.3.21}$$

is minimized subject to the constraints (4.3.15)–(4.3.20).
This problem is referred to as Problem $P_1^{\tilde{k}}$.

Definition 4.3.1 $(\boldsymbol{\eta}, \hat{v}) \in \Theta \times \mathcal{V}_1$ (respectively, $(\boldsymbol{t}, v)$) is said to be a feasible element if the constraints (4.3.15)–(4.3.20) (respectively, constraints (4.3.7)–(4.3.11)) are satisfied.

Theorem 4.3.1 *Problem* $(P_1^{\tilde{k}})$ *is equivalent to Problem* (P_1^k) *in the sense that* $(\boldsymbol{\eta}^*, \hat{v}^*)$ *is a solution of Problem* $(P_1^{\tilde{k}})$ *if and only if* $(\boldsymbol{t}^*, v^*)$ *is a solution of problem* (P_1^k), *and* $\hat{g}_1(\boldsymbol{\eta}^*, \hat{v}^*) = g_1(\boldsymbol{t}^*, v^*)$, *where* $\eta^*(s) = t_i^* - t_{i-1}^*$, $s \in [i-1, i)$, $i = 1, 2, \cdots, 2k+1$, *and* $\hat{v}^*(s) = v(t^*(s))$, $s \in [0, 2k+1]$.

PROOF. Let $(\boldsymbol{t}^1, v^1) \in \mathcal{T} \times \mathcal{V}$ be a feasible element of Problem (P_1^k), and let $(\boldsymbol{\eta}, \hat{v}) \in \Theta \times \mathcal{V}_1$ be the corresponding feasible element of Problem $(P_1^{\tilde{k}})$. Then it is easy to check that $q(t)$ is the solution of (4.3.4)–(4.3.5)(with $l = k$) if and only if $\hat{q}(s)$ is the solution of (4.3.12)–(4.3.13), and

$$g_1(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v) = \sum_{i=1}^{2k+1} (t_i - t_{i-1}) = \sum_{i=1}^{2k+1} \eta_i = \hat{g}_1(\boldsymbol{\eta}, \hat{v}).$$

Hence, the results follow readily. □

In Problem (P_1^k) the cost function (4.3.6) is minimized with respect to $(\boldsymbol{t}, v) \in \mathcal{T} \times \mathcal{V}$ where $\boldsymbol{t} = (t_1, t_2, \cdots, t_{2k})$ is the vector of switching time points. On the other hand, the cost function (4.3.21) in Problem $(P_1^{\tilde{k}})$ is minimized with respect to $(\boldsymbol{\eta}, \hat{v}) \in \Theta \times \mathcal{V}_1$ where $\boldsymbol{\eta}(t)$ is a nonnegative piecewise constant function. Problem $(P_1^{\tilde{k}})$ is equivalent to Problem (P_1^k). However, Problem $(P_1^{\tilde{k}})$ is numerically much more tractable, because it does not involve variable switching times.

Let us now address the question of finding the optimal number l of the recharge points. Let $\boldsymbol{t} = (t_1, \cdots, t_{2k})$ be the optimal solution of Problem (P_1^k) corresponding to the positive integer k. We propose the following algorithm.

Algorithm A.

1. Choose an initial guess $k = k_0$.

2. Let $t_0 = 0$, $t_{2k+1} = t_{final}$, and solve Problem (P_1^k) to obtain $\boldsymbol{t} = (t_1, \cdots, t_{2k})$. Let $\hat{g}_1^k$ be the corresponding optimal cost.

3. If $k = L$, stop, find k^* such that $\hat{g}_1^{k^*} \leq \hat{g}_1^l$ for all $l \in \{k_0, \cdots, L\}$, k^* is the optimal number of the recharge points. Otherwise, let $k = k + 1$ and go to step 2.

4.4 Fixed Recharge Locations

Following the notation defined before consider again the situation in which the road is even. However, We assume that the vehicle can only be recharged at a given set of locations $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_h\}$, satisfying $0 < \tau_1 < \tau_2 < \cdots < \tau_h < S$, where S is the maximum distance the vehicle is expected to travel as defined in the previous section. We assume that h is fixed. We also assume that the number of recharge points $l = h$. The problem can then be formulated as the following optimization problem:
given the dynamical system (4.3.4)–(4.3.5) (with $l = h$), find a $(\boldsymbol{t}, v) \in \mathcal{T} \times \mathcal{V}$ such that the cost function

$$g_2(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v) \tag{4.4.1}$$

is minimized subject to the constraints

$$q_{min} \leq q(t) \leq q_{max}, \tag{4.4.2}$$

$$0 < p = m(av + bv^2) < P, \tag{4.4.3}$$

$$0 < t_i - t_{i-1} \qquad i = 1, \cdots, 2h+1, \tag{4.4.4}$$

$$\int_0^{t_{final}} v dt = S, \tag{4.4.5}$$

$$\int_0^{t_{2i-1}} v(t) dt = \tau_i \qquad i = 1, \cdots, h. \tag{4.4.6}$$

This problem is referred to as Problem P_2.

To begin, we choose the number of recharge points to be h. However, the exact optimal number of recharge points is equal to h minus the number of those recharge points corresponding to the occurrence of redundancies.

Under the CPET, the constraints (4.4.2)–(4.4.6) reduce to

$$q_{min} \leq \hat{q}(2i-1) \qquad i = 1, \cdots, h+1, \tag{4.4.7}$$

$$\hat{q}(2i) \leq q_{max} \qquad i = 1, \cdots, h, \tag{4.4.8}$$

$$0 < p = m(b_1\hat{v} + b_2\hat{v}^2) < P, \tag{4.4.9}$$

$$0 < \eta_i \qquad i = 1, \cdots, 2h+1, \tag{4.4.10}$$

$$\int_0^{2h+1} \hat{v}ds = S, \tag{4.4.11}$$

$$\int_0^{2i-1} v(s)ds = \tau_i \qquad i = 0, \cdots, h. \tag{4.4.12}$$

Let $\mathcal{V}_2$ be the set of all those $\hat{v}(s)$, such that the conditions (4.4.7)–(4.4.12) are satisfied. Problem P_2 is now transformed into the following optimal control problem:
given the dynamical system (4.3.12)–(4.3.13) (with $k = h$), find an admissible element $(\boldsymbol{\eta}, \hat{v}) \in \Theta \times \mathcal{V}_2$ such that the cost function

$$\hat{g}_2(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{2h+1} \eta_i \tag{4.4.13}$$

is minimized subject to the constraints (4.4.7)–(4.4.12).
This problem is referred to as Problem P_2^0.

Theorem 4.4.1 *Problem (P_2^0) is equivalent to Problem (P_2) in the sense that $(\boldsymbol{\eta}^*, \hat{v}^*)$ is a solution of Problem (P_2^0) if and only if $(\boldsymbol{t}^*, v^*)$ is a solution of problem (P_2), and $\hat{g}_2(\boldsymbol{\eta}^*, \hat{v}^*) = g_2(\boldsymbol{t}^*, v^*)$, where $\eta^*(s) = t_i^* - t_{i-1}^*$, $s \in [i-1, i)$, $i = 1, 2, \cdots, 2h+1$, and $\hat{v}^*(s) = v(t^*(s))$, $s \in [0, 2h+1]$.*

PROOF. The proof is similar to the proof for Theorem 4.3.1. □

4.5 With Acceleration

In this section, we construct a model, involving the acceleration at each interval at which the vehicle is driven. The distance between two cities, the maximum allowable power to be applied to the motor, the number of the battery cells, the number of recharge points, and the recharge locations are assumed to be fixed.

4.5.1 Problem Formulation

Since the force for acceleration is

$$F_1 = ma, \tag{4.5.1}$$

where a is the acceleration $a = a(t)$.

From 4.2.1, 4.2.2, and 4.5.1, we obtain

$$\begin{aligned} p &= v \times (r(v) + F_1) \\ &= v \times m(b_1 + b_2 v + a) \end{aligned}$$

and for each battery cells, the power flowing out is

$$p_1 = v \times \frac{m}{n}(b_1 + b_2 v + a). \tag{4.5.2}$$

The charge stored in each battery cell is governed by the following differential equation:

$$\frac{dq}{dt} = \begin{cases} -c_1 \frac{(b_1 v + b_2 v^2 + a)\cdot m}{n}, & t \in [t_{2i-2}, \zeta_{2i-2}) \\ -c_1 \frac{(b_1 v + b_2 v^2)\cdot m}{n}, & t \in [\zeta_{2i-2}, t_{2i-1}) \\ \frac{c_2}{q+c_3} - \exp\{c_4 - c_5(t - t_{2i-1})\}, & t \in [t_{2i-1}, t_{2i}) \\ t \in [t_{2l}, \zeta_{2l}) & \\ -c_1 \frac{(b_1 v + b_2 v^2)\cdot m}{n}, & t \in [\zeta_{2l}, t_{2l+1}] \\ \text{here} & i = 1, \cdots, l \end{cases} \tag{4.5.3}$$

with initial condition

$$q(0) = q_0, \tag{4.5.4}$$

where ζ_{2i-2} are switching times at which the acceleration varies from a positive constant to zero, and l is the number of recharge points, which is given.

Define

$$\mathcal{T}^* = \left\{ \boldsymbol{t} = (t_0, \zeta_0, t_1, t_2, \zeta_2, \cdots, \zeta_{2l}) : \begin{array}{l} t_i, \zeta_i \in \mathbb{R}, \\ t_{i-1} < \zeta_{i-1} < t_i < t_{i+1}, \\ \text{here } i = 1, \cdots, 2l,\ l \in \mathbf{N} \end{array} \right\}.$$

The problem can be formulated as:

given the dynamical system (4.5.3) - (4.5.4), find a $(\boldsymbol{t}, v) \in \mathcal{T}^* \times \mathcal{V}$ such that the cost function

$$g_3(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v) \tag{4.5.5}$$

is minimized subject to the constraints

$$q_{min} \leq q(t) \leq q_{max}, \tag{4.5.6}$$

$$0 < p = m(b_1 v + b_2 v^2 + a) < P, \tag{4.5.7}$$

$$0 < t_i - t_{i-1} \qquad i = 1, \cdots, 2l+1, \tag{4.5.8}$$

$$\int_0^{t_{final}} v dt = S, \tag{4.5.9}$$

$$\int_0^{t_{2i-1}} v(t) dt = \tau_i \qquad i = 1, \cdots, l, \tag{4.5.10}$$

where τ_i, $i = 1, \cdots, l$, are recharge points, which are given.
Let this problem be referred to as Problem P_3.

4.5.2 Transformation

To find the optimal solution of Problem P_3, we need to find the optimal switching times $\boldsymbol{t}$, and the corresponding optimal speed v.

We now apply the Control Parameterization Enhancing Transform (CPET) [100, 50] to Problem P_3. Let $s \in [0, 3l+2]$ be a new time variable, and let $\eta(s)$ be defined by

$$\eta(s) = \sum_{i=1}^{3l+2} \eta_i \chi_{[i-1,i)}(s)$$

where $\chi_{[i-1,i)}(s)$ is the indicate function of the interval [i-1,i), and the $\eta_i's$ are nonnegative constants. The enhancing control $\eta(s)$ is defined on [0, $3l+2$] with fixed switching times, located at $\{1, 2, \cdots, 3l+1\}$. The CPET maps $t \in [0, t_{final}]$ to $s \in [0, 3l+2]$ as follows:

$$\begin{aligned} \frac{dt}{ds} &= \eta(s) \\ t(0) &= 0 \end{aligned}$$

where

$$\eta(s) = \begin{cases} t_{2i-2} - \zeta_{2i-2} & s \in [3i-3, 3i-2), \\ \zeta_{2i-2} - t_{2i-1} & s \in [3i-2, 3i-1) \\ t_{2i-1} - t_{2i} & s \in [3i-1, 3i) \\ t_{2l} - \zeta_{2l} & s \in [3l, 3l+1) \\ t_{3l+1} - t_{3l+2} & s \in [3l+1, 3l+2] \end{cases} \qquad i = 1, 2, \cdots, l$$

which satisfies

$$\int_0^{3l+2} \eta(s)ds = t_{final},$$

$$and \quad \int_0^{3i-1} \eta(s)ds = \tau_i, \qquad i = 1, \cdots, l.$$

Let Θ^* denote the class of all such enhancing controls. Under the CPET, the system dynamics changes to

$$\frac{d}{ds}\begin{pmatrix} \hat{q}(s) \\ t(s) \end{pmatrix} = \begin{pmatrix} \eta \begin{cases} -c_1 \frac{(b_1\hat{v}+b_2\hat{v}^2+a)\cdot m}{n}, & s \in [3i-3, 3i-2) \\ -c_1 \frac{(b_1\hat{v}+b_2\hat{v}^2)\cdot m}{n}, & s \in [3i-2, 3i-1) \\ \frac{c_2}{\hat{q}+c_3} - \exp\{c_4 - c_5(t-3i+1)\}, & s \in [3i-1, 3i) \\ -c_1 \frac{(b_1\hat{v}+b_2\hat{v}^2+a)\cdot m}{n}, & s \in [3l, 3l+1) \\ -c_1 \frac{(b_1\hat{v}+b_2\hat{v}^2)\cdot m}{n}, & s \in [3l+1, 3l+2] \end{cases} \\ \eta(s) \end{pmatrix} \tag{4.5.11}$$

with initial condition

$$\begin{pmatrix} \hat{q}(0) \\ t(0) \end{pmatrix} = \begin{pmatrix} q_0 \\ 0 \end{pmatrix} \tag{4.5.12}$$

(4.5.13)

where $\hat{q}(s) = q(t(s))$, $\hat{v}(s) = v(t(s))$, and $s \in [0, 3l+2]$ $i = 1, \cdots, l$.
The constraints (4.5.6) - (4.5.10) reduce to

$$q_{min} \leq \hat{q}(s) \leq q_{max}, \tag{4.5.14}$$

$$0 < p = m(b_1\hat{v} + b_2\hat{v}^2 + a) < P \tag{4.5.15}$$

$$0 < \eta_i \qquad i = 1, \cdots, 3l+2, \tag{4.5.16}$$

$$\int_0^{3l+2} \hat{v}ds = S. \tag{4.5.17}$$

$$and \quad \int_0^{t_{3i-1}} v(t)dt = \tau_i \qquad i = 1, \cdots, l, \tag{4.5.18}$$

Let $\mathcal{V}^3$ be the set of all such function $\hat{v}(s)$, satisfy the conditions (4.5.14) - (4.5.18), and the Problem P_3 is now transformed into the following optimal control problem:

given the dynamical system (4.5.11) - (4.5.13), find an admissible element $(\boldsymbol{\eta}, \hat{v}) \in \Theta^* \times \mathcal{V}^3$ such that the cost function

$$\hat{g}_3(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{3l+2} \eta_i \tag{4.5.19}$$

is minimized subject to the constraints (4.5.14) - (4.5.18).

This problem is referred to as Problem (P_3^0.) It can be shown that the Problem (P_3^0) is equivalent to Problem (P_3). For the same reason as pointed out in Sections 4.3 and 4.4, we choose to solve Problem (P_3^0) rather than to solve Problem (P_3).

Theorem 4.5.1 *Problem (P_3^0) is equivalent to Problem (P_3) in the sense that $(\boldsymbol{\eta}^*, \hat{v}^*)$ is a solution of Problem (P_3^0) if and only if $(\boldsymbol{t}^*, v^*)$ is a solution of problem (P_3), and $\hat{g}_3(\boldsymbol{\eta}^*, \hat{v}^*) = g_3(\boldsymbol{t}^*, v^*)$, where*

$$\eta^*(s) = \begin{cases} t^*_{2i-2} - \zeta^*_{2i-2} & s \in [3i-3, 3i-2), \\ \zeta^*_{2i-2} - t^*_{2i-1} & s \in [3i-2, 3i-1) \\ t^*_{2i-1} - t^*_{2i} & s \in [3i-1, 3i) \\ t^*_{2l} - \zeta^*_{2l} & s \in [3l, 3l+1) \\ t^*_{3l+1} - t^*_{3l+2} & s \in [3l+1, 3l+2] \end{cases} \qquad i = 1, 2, \cdots, l$$

and

$$\hat{v}^*(s) = v(t^*(s)),\ s \in [0, 2l+1]$$

.

PROOF. The proof is similar to the proof for the Theorem 4.3.1. □

Remark : ζ_{2i-2} is related to the speed v and the acceleration a. If a is constant, ζ_{2i-2} is decided by v.

4.6 Undulating Road

In this section, the road connecting the two cities is assumed to be undulating. The distance between the two cities, the maximum allowable power to be applied

to the motor, and the number of battery cells are again assumed to be fixed. $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_l\}$, satisfying $0 < \tau_1 < \tau_2 < \cdots < \tau_l < S$, are recharge locations at which the vehicle can be recharged. These recharge locations include the turning points $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_m)$ of undulating road. The number l of recharge points is assumed to be fixed.

4.6.1 Problem Formulation

Since the road connecting the two cities is undulating, the power being applied to the motor (see (4.3.2)) becomes

$$\begin{aligned} p &= v \times r(v) \\ &= v \times m(b_1 + b_2 v + g\sin\theta), \end{aligned}$$

and the power flowing out from each of the n parallel battery cells (see (4.3.3)) becomes

$$p_1 = v \times \frac{m(b_1 + b_2 v + g\sin\theta)}{n}. \tag{4.6.1}$$

From (4.2.7) and (4.6.1), we have

$$\frac{dq}{dt} = \begin{cases} -c_1 \frac{(b_1 v + b_2 v^2 + g\sin\theta)\cdot m}{n}, & t \in [t_{2i-2}, t_{2i-1}) \\ \frac{c_2}{q+c_3} - \exp\{c_4 - c_5(t - t_{2i-1})\}, & t \in [t_{2i-1}, t_{2i}) \\ -c_1 \frac{(b_1 v + b_2 v^2 + g\sin\theta)\cdot m}{n}, & t \in [t_{2l}, t_{2l+1}] \end{cases}$$
$$\text{here} \qquad i = 1, \cdots, l \tag{4.6.2}$$

with initial condition

$$q(0) = q_0, \tag{4.6.3}$$

We may now specify the corresponding optimization problem formally as follows:

find a $(\boldsymbol{t}, v) \in \mathcal{T} \times \mathcal{V}$ such that the cost function

$$g_4(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v) \tag{4.6.4}$$

is minimized subject to the system (4.6.2) - (4.6.3) and the following constraints.

$$q_{\min} \leq q(t) \leq q_{\max}, \tag{4.6.5}$$

$$0 < p = m(av + bv^2 + g\sin\theta) < P, \tag{4.6.6}$$

$$0 < t_i - t_{i-1} \qquad i = 1, \cdots, 2l+1, \tag{4.6.7}$$

$$\int_0^{t_{final}} v dt = S, \tag{4.6.8}$$

$$\int_0^{t_{2i-1}} v(t)dt = \tau_i \qquad i = 1, \cdots, l, \tag{4.6.9}$$

where S is the total distance traveled, n and l are, respectively, the number of battery cells and number of recharge times, v is the speed, and $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_l\}$ is the vector of recharge points. Let this problem be referred to as Problem P_4.

4.6.2 Transformation

To find the optimal solution of Problem P_4, we need to determine the optimal time points $\boldsymbol{t} = (t_1, \cdots, t_{2l})$, and the optimal speed v.

We now apply the CPET [100, 50] to Problem P_4. Let $s \in [0, 2l+1]$ be a new time variable, and let $\eta(s)$ be defined by

$$\eta(s) = \sum_{i=1}^{2l+1} \eta_i \chi_{[i-1,i)}(s),$$

where $\chi_{[i-1,i)}(s)$ is the characteristic function on the interval [i-1,i), and the $\eta_i's$ are nonnegative constants. Clearly, $\eta(s)$, which is called the enhancing control, is a nonnegative piecewise constant function defined on [0,2l+1] with fixed switching points $\{1, 2, \cdots, 2l\}$. The CPET:

$$\begin{aligned} \frac{dt}{ds} &= \eta(s), \\ t(0) &= 0 \end{aligned}$$

maps $t \in [0, t_{final}]$ into $s \in [0, 2l+1]$, where

$$\eta(s) = t_i - t_{i-1}, \qquad s \in [i-1, i), \quad i = 1, 2, \cdots, 2l+1,$$

satisfying

$$\int_0^{2l+1} \eta(s) ds = t_{final}.$$

Let Θ denote the class of all such enhancing controls. Under the CPET, the system dynamics (4.6.2)–(4.6.3) becomes

$$\frac{d}{ds}\begin{pmatrix}\hat{q}(s)\\ t(s)\end{pmatrix} = \begin{pmatrix} \eta\begin{cases} -c_1\frac{(b_1\hat{v}+b_2\hat{v}^2+g\sin\theta)\cdot m}{n}, & s\in[2i-2,2i-1)\\ \frac{c_2}{\hat{q}+c_3}-\exp\{c_4-c_5(t-2i+1)\}, & s\in[2i-1,2i)\\ -c_1\frac{(b_1\hat{v}+b_2\hat{v}^2+g\sin\theta)\cdot m}{n}, & s\in[2l,2l+1]\end{cases}\\ \eta(s)\end{pmatrix} \tag{4.6.10}$$

with initial condition

$$\begin{pmatrix}\hat{q}(0)\\ t(0)\end{pmatrix} = \begin{pmatrix}q_0\\ 0\end{pmatrix}, \tag{4.6.11}$$

where $\hat{q}(s)=q(t(s))$, $\hat{v}(s)=v(t(s))$, $s\in[0,2l+1]$, and $i=1,\cdots,l$.
The constraints (4.6.5)–(4.6.9) reduce to

$$q_{min} \le \hat{q}(2i-1) \qquad i=1,\cdots,l+1, \tag{4.6.12}$$

$$\hat{q}(2i) \le q_{max} \qquad i=1,\cdots,l, \tag{4.6.13}$$

$$0<p=m(b_1\hat{v}+b_2\hat{v}^2+g\sin\theta)<P, \tag{4.6.14}$$

$$0<\eta_i \qquad i=1,\cdots,2l+1, \tag{4.6.15}$$

$$\int_0^{2l+1}\hat{v}ds=S, \tag{4.6.16}$$

$$\int_0^{2i-1}v(s)ds=\tau_i \qquad i=0,\cdots,l. \tag{4.6.17}$$

Let $\mathcal{V}_4$ be the set of all those $\hat{v}(s)$, such that the conditions (4.6.12)–(4.6.17) are satisfied. Problem P_4 is now transformed into the following optimal control problem:
given the dynamical system (4.6.10)–(4.6.11), find a $(\boldsymbol{\eta},\hat{v})\in\Theta\times\mathcal{V}_4$ such that the cost function

$$\hat{g}_4(\boldsymbol{\eta},\hat{v})=\sum_{i=1}^{k+1}\eta_i \tag{4.6.18}$$

is minimized subject to the constraints (4.6.12)–(4.6.17). This problem is referred to as Problem P_4^0.

Theorem 4.6.1 *Problem (P_4^0) is equivalent to Problem (P_4) in the sense that $(\boldsymbol{\eta}^*,\hat{v}^*)$ is a solution of Problem (P_4^0) if and only if $(\boldsymbol{t}^*,v^*)$ is a solution of*

problem (P_4), *and* $\hat{g}_4(\eta^*, \hat{v}^*) = g_4(\boldsymbol{t}^*, v^*)$, *where* $\eta^*(s) = t_i^* - t_{i-1}^*$, $s \in [i-1, i)$, $i = 1, 2, \cdots, 2l+1$, *and* $\hat{v}^*(s) = v(t^*(s))$, $s \in [0, 2l+1]$.

PROOF. The proof is similar to the proof for the Theorem 4.3.1. □

Note that the recharge points are fixed. However there may exist coalescence of the switching times. Thus, the exact optimal number of recharge points is equal to l minus the number of recharge points corresponding to the occurrence of redundancies.

4.7 Free Recharge Locations

With the notation defined in Section 4.5, we consider the situation in which the recharge location for a traveling vehicle are free to be optimized. The problem can be formulated as the following optimization problem:
given the dynamical system (4.6.2)–(4.6.3), find a $(\boldsymbol{t}, v, l) \in \mathcal{T} \times \mathcal{V} \times \mathbf{N}$ such that the cost function

$$g_5(\boldsymbol{t}, v, l) = t_{final}(\boldsymbol{t}, v, l) \tag{4.7.1}$$

is minimized subject to the constraints

$$q_{min} \leq q(t) \leq q_{max}, \tag{4.7.2}$$

$$0 < p = m(av + bv^2 + g\sin\theta) < P, \tag{4.7.3}$$

$$0 < t_i - t_{i-1} \qquad i = 1, \cdots, 2l+1, \tag{4.7.4}$$

$$\int_0^{t_{final}} v dt = S, \tag{4.7.5}$$

$$\int_0^{t_{i_j}} v(t) dt = \gamma_j \qquad j = 1, \cdots, m, \text{ and } t_{i_j} \in \boldsymbol{t}, \tag{4.7.6}$$

$$l \leq L, \tag{4.7.7}$$

where S is the total distance traveled, L is the the upper bound of the number of the recharge points, n and l are, respectively, the number of battery cells and the number of recharge points, v is the speed, and $\gamma = \{\gamma_1, \cdots, \gamma_m\}$ is the vector of the turning points of the undulated road. Let this problem be referred to as Problem P_5.

4.7.1 Transformation

To solve Problem P_5, we need to determine the optimal number l of recharge points, the optimal time points $\boldsymbol{t} = (t_1, \cdots, t_{2l})$, and the optimal speed v. Let us initially choose the optimal number of recharge points to be $k \geq m$. These recharge points include the turning points $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_m)$. Then, only the optimal switching time points with a fixed number of recharge points, and the corresponding optimal speed are to be determined optimally. This simplified problem is referred to as Problem P_5^k.

We now apply the CPET [100, 50] to Problem P_5^k. Let $s \in [0, 2k+1]$ be a new time variable, and define

$$\eta(s) = \sum_{i=1}^{2k+1} \eta_i \chi_{[i-1,i)}(s)$$

where $\chi_{[i-1,i)}(s)$ is the characteristic function on the interval [i-1,i), and the $\eta_i's$ are nonnegative constants. Clearly, $\eta(s)$, which is called the enhancing control, is a nonnegative piecewise constant function defined on [0,2k+1] with fixed switching times $\{1, 2, \cdots, 2\text{k}\}$. The CPET:

$$\begin{aligned} \frac{dt}{ds} &= \eta(s) \\ t(0) &= 0 \end{aligned}$$

maps $t \in [0, t_{final}]$ into $s \in [0, 2k+1]$, where

$$\eta(s) = t_i - t_{i-1}, \qquad s \in [i-1, i), \quad i = 1, 2, \cdots, 2k+1,$$

satisfying

$$\int_0^{2k+1} \eta(s) ds = t_{final}.$$

Let Θ denote the class of all such enhancing controls. Under the CPET, the system dynamics (4.6.2)-(4.6.3)(with $l = k$) becomes

$$\frac{d}{ds}\begin{pmatrix} \hat{q}(s) \\ t(s) \end{pmatrix} = \begin{pmatrix} \eta \begin{cases} -c_1 \frac{(b_1\hat{v} + b_2\hat{v}^2 + g\sin\theta)\cdot m}{n}, & s \in [2i-2, 2i-1) \\ \frac{c_2}{\hat{q}+c_3} - \exp\{c_4 - c_5(t - 2i + 1)\}, & s \in [2i-1, 2i) \\ -c_1 \frac{(b_1\hat{v} + b_2\hat{v}^2 + g\sin\theta)\cdot m}{n}, & s \in [2k, 2k+1] \end{cases} \\ \eta(s) \end{pmatrix} \tag{4.7.8}$$

with initial condition

$$\begin{pmatrix} \hat{q}(0) \\ t(0) \end{pmatrix} = \begin{pmatrix} q_0 \\ 0 \end{pmatrix} \tag{4.7.9}$$

where $\hat{q}(s) = q(t(s))$, $\hat{v}(s) = v(t(s))$, $s \in [0, 2k+1]$, and $i = 1, \cdots, k$.
The constraints (4.7.2)–(4.7.7) reduce to

$$q_{min} \leq \hat{q}(2i-1) \qquad i = 1, \cdots, k+1, \tag{4.7.10}$$

$$\hat{q}(2i) \leq q_{max} \qquad i = 1, \cdots, k. \tag{4.7.11}$$

$$0 < p = m(b_1\hat{v} + b_2\hat{v}^2 + g\sin\theta) < P, \tag{4.7.12}$$

$$0 < \eta_i \qquad i = 1, \cdots, 2k+1, \tag{4.7.13}$$

$$\int_0^{2k+1} \hat{v} ds = S, \tag{4.7.14}$$

$$\int_0^{i_j} v(t)dt = \gamma_j \qquad j = 1, \cdots, m, \text{ and } i_j \in \{1, \cdots, 2k-1\} \tag{4.7.15}$$

Let $\mathcal{V}_5$ be the set of all those $\hat{v}(s)$ such that the conditions (4.7.10)–(4.7.15) are satisfied. Problem P_5^k is now transformed into the following optimal control problem:
given the dynamical system (4.7.8)–(4.7.9), find a $(\boldsymbol{\eta}, \hat{v}) \in \Theta \times \mathcal{V}_4$ such that the cost function

$$\hat{g}_5(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{2k+1} \eta_i \tag{4.7.16}$$

is minimized subject to the constraints (4.7.11)–(4.7.15). This problem is referred to as Problem $P_5^{\tilde{k}}$.

Theorem 4.7.1 *Problem $(P_5^{\tilde{k}})$ is equivalent to Problem (P_5^k) in the sense that $(\boldsymbol{\eta}^*, \hat{v}^*)$ is a solution of Problem $(P_5^{\tilde{k}})$ if and only if $(\boldsymbol{t}^*, v^*)$ is a solution of Problem (P_5^k), and $\hat{g}_5(\boldsymbol{\eta}^*, \hat{v}^*) = g_5(\boldsymbol{t}^*, v^*)$, where $\eta^*(s) = t_i^* - t_{i-1}^*$, $s \in [i-1, i)$, $i = 1, 2, \cdots, 2k+1$, and $\hat{v}^*(s) = v(t^*(s))$, $s \in [0, 2k+1]$.*

PROOF. The proof is similar to the proof for the Theorem 4.3.1. □

The recharge points, besides the m turning points, can be located in any intervals of the undulated road. Let $\boldsymbol{t} = (t_1, \cdots, t_{2k})$ be the optimal solution of

Problem (P_5^k) for a fixed positive integer k, which denotes the number of the recharge points in the intervals of the undulated road. The optimal number of recharge points can be obtained by using the Algorithm A proposed in Section 4.3.

4.8 Numerical Experiments

We now consider some numerical examples of the problems formulated in Sections 4.3, 4.4, 4.5, 4.6 and 4.7. All the examples below were solved using MISER 3.2 ([41, 42]) in Fortran double precision on a Unix Workstation.

In Examples 4.1, 4.2, and 4.3, the road connecting the two cities is assumed to be even.

Example 4.1. Choose $S = 600$, $P = 50$, $m_1 = 1400$, $L = 5$, and $n = 80$ in Problem (P_1). The dynamical system is

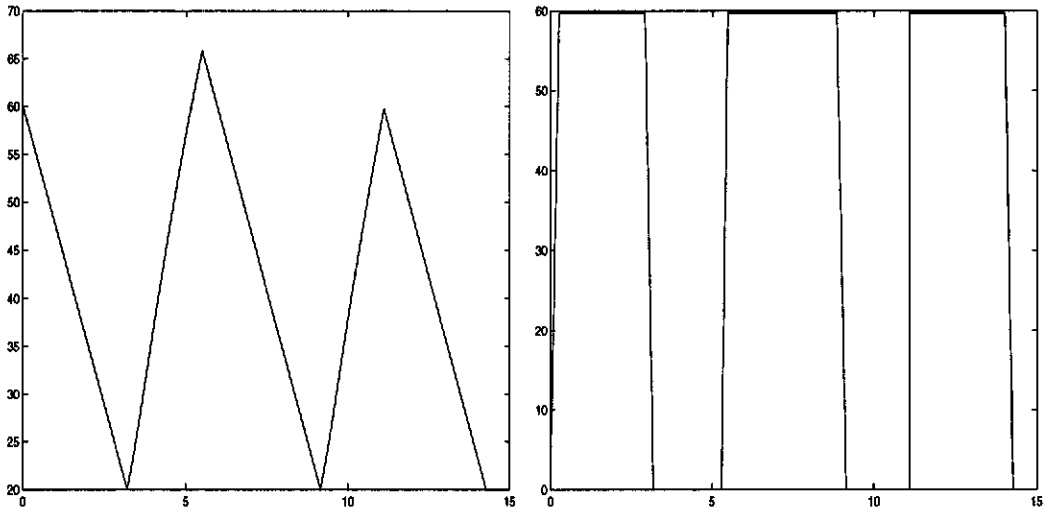
$$\frac{dq}{dt} = \begin{cases} -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200})\cdot 1400}{80}, & t \in [t_{2i-2}, t_{2i-1}) \\ \frac{2000}{q+50} - \exp\{2.5 - 3.0(t_{2i} - t_{2i-1})(t - t_{2i-1})\}, & t \in [t_{2i-1}, t_{2i}) \\ -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200})\cdot 1400}{80}, & t \in [t_{2l}, t_{2l+1}] \end{cases}$$
$$\text{here} \qquad i = 1, \cdots, l \tag{4.8.1}$$

with initial condition

$$q(0) = 60. \tag{4.8.2}$$

Then, Problem P_1 becomes:
given the dynamical system (4.8.1)–(4.8.2), find $\boldsymbol{t} = \{t_1, t_2, \cdots, t_{2l}\}$, v, l, such that

$$g_1(\boldsymbol{t}, v, l) = t_{final}(\boldsymbol{t}, v, l)$$



(a) Optimal recharge plan (b) Optimal driving speed

Figure 4.8.1: The optimal driving strategy for Example 4.1

is minimized subject to the constraints

$$
\begin{aligned}
& 20 \leq q(t) \leq 80, \\
& 0 < p < 50, \\
& 0 < t_i - t_{i-1} \qquad i = 1, \cdots, 2l+1, \\
& \int_0^{t_{final}} v dt = 600, \\
& l \leq 5.
\end{aligned}
$$

We choose $k_0 = 2$ in Algorithm A. Consider the case in which $k = 4$.

Under the CPET given in Section 4.3, we have:

given the dynamical system

$$
\frac{d\hat{q}}{ds} = \begin{cases} -\frac{1}{100}\eta\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200})\cdot 1400}{80}, & s \in [2i-2, 2i-1) \\ \eta\frac{2000}{q+50} - \exp\{2.5 - 3.0\eta_{2i}(s-2i+1)\}, & s \in [2i-1, 2i) \\ -\frac{1}{100}\eta\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200})\cdot 1400}{80}, & s \in [8, 9] \end{cases}
$$

here $\quad i = 1, \cdots, 4,$

with initial condition

$$\hat{q}(0) = 60,$$

find parameter vectors η and $\hat{v}$ such that

$$\hat{g}_1(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{9} \eta_i$$

i	1	2	3	4	5
t_i	3.2000	5.4852	9.1523	11.094	14.275
x_i	191. 08	191. 08	410.06	410. 06	600

Table 4.8.1: The optimal switching times and recharge points for Example 4.1

is minimized subject to

$$\begin{aligned} & 20 \le \hat{q}(2i-1) && i = 1, \cdots, 5, \\ & \hat{q}(2i) \le 80 && i = 1, \cdots, 4, \\ & 0 < p \le 50, \\ & 0 < \eta_i && i = 1, \cdots, 9, \\ & \int_0^9 \hat{v} ds = 600. \end{aligned}$$

The optimal switching times and recharge points are listed in Table 4.8.1, in which the optimal switching times are 3.200, 5.485, 9.153, and 11.09, while the optimal recharge points are 191.08 and 410.06. The corresponding optimal speed is 59.71 with minimum traveling time 14.275.

Figure 4.8.1 contains the optimal recharge plan, the optimal switching times and the optimal driving speed.

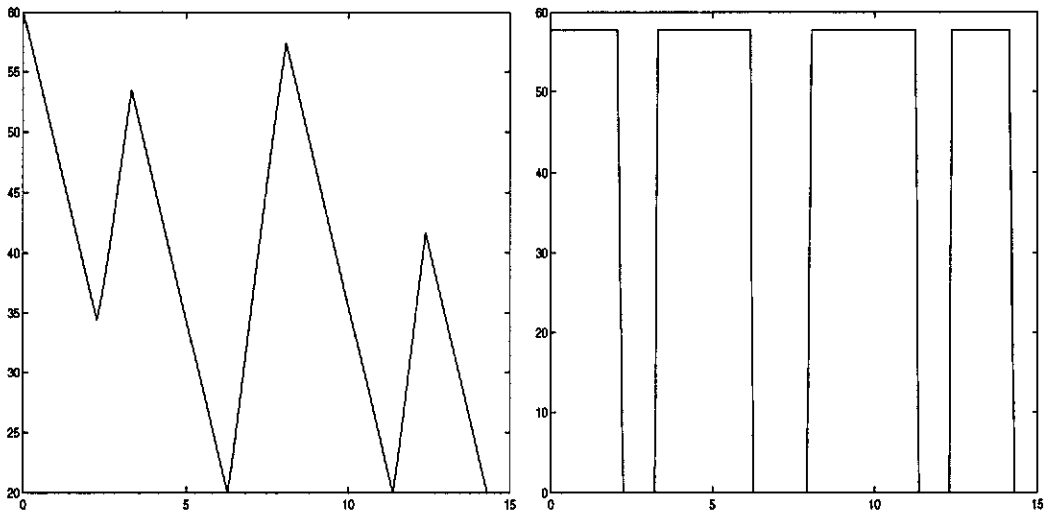
Example 4.2. In Problem P_2, choose $S = 600$, $P = 50$, $m = 1470$ and $n = 87$. The recharge locations are restricted at 130, 240, 300, 420, and 490. The dynamical system is:

$$\frac{dq}{dt} = \begin{cases} -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200})\cdot 1470}{87}, & t \in [t_{2i-2}, t_{2i-1}) \\ \frac{2000}{q+50} - \exp\{2.5 - 3.0(t_{2i} - t_{2i-1})(t - t_{2i-1}\}, & t \in [t_{2i-1}, t_{2i}) \\ -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200})\cdot 1470}{87}, & t \in [t_k, t_{k+1}] \end{cases}$$
$$\text{here} \qquad i = 1, \cdots, 5 \tag{4.8.3}$$

with initial condition

$$q(0) = 60, \tag{4.8.4}$$

Problem (P_2) becomes:



(a) Optimal recharge plan (b) Optimal driving speed

Figure 4.8.2: The optimal driving strategy for Example 4.2

given the dynamical system (4.8.3)–(4.8.4), find $\boldsymbol{t} = \{t_1, t_2, \cdots, t_{11}\}$, and v such that

$$g_2(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v)$$

is minimized subject to the constraints

$$\begin{aligned}
&20 \le q(t) \le 80 \\
&0 < p < 50, \\
&0 < t_i - t_{i-1} && i = 1, \cdots, 11, \\
&\int_0^{t_{11}} v dt = 600, \\
&\int_0^{t_{2i-1}} v dt = \tau_i && i = 1, 2, \cdots, 5,
\end{aligned}$$

where $(\tau_1, \cdots, \tau_5) = (130, 240, 300, 420, 490)$.

Under the CPET transform, we have the following problem:

given the dynamics

$$\frac{d\hat{q}}{ds} = \begin{cases} -\frac{1}{100}\eta^{\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200})\cdot 1470}{87}}, & s \in [2i-2, 2i-1) \\ \eta\frac{2000}{q+50} - \exp\{2.5 - 3.0\eta_{2i}(s-2i+1)\}, & s \in [2i-1, 2i) \\ -\frac{1}{100}\eta^{\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200})\cdot 1470}{87}}, & s \in [10, 11] \end{cases}$$

here $i = 1, \cdots, 5$.

i	1	2	3	4	5	6	7	8	9	10	11
t_i	2.24	3.33	5.23	5.23	6.26	8.09	10.16	10.16	11.37	12.39	14.29
x_i	130.	130.	240.	240.	300.	300.	418.7	418.7	488.4	488.4	600.

Table 4.8.2: The optimal switching times and recharge points for Example 4.2

with initial condition

$$\hat{q}(0) = 60,$$

find parameter vectors η, and $\hat{v}$ such that

$$\hat{g}_2(\eta, \hat{v}) = \sum_{i=1}^{11} \eta_i$$

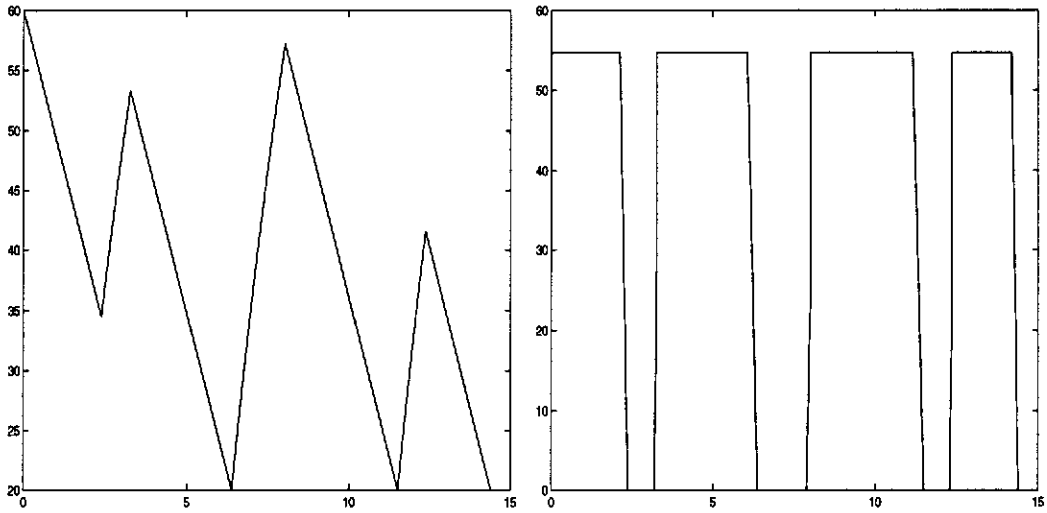
is minimized subject to

$$\begin{aligned}
&20 \leq \hat{q}(2i-1) && i = 1, \cdots, 6, \\
&\hat{q}(2i) \leq 80 && i = 1, \cdots, 5, \\
&0 < p =< 50 && \\
&0 < \eta_i && i = 1, \cdots, 11, \\
&\int_0^{11} \hat{v} ds = 600, && \\
&\int_0^{2i-1} \hat{v}(s) ds = \tau_i && i = 1, 2, \cdots, 5.
\end{aligned}$$

The optimal switching times and recharge points, listed in Table 4.8.2, show that there are four redundancies $t_3 = t_4$, $t_7 = t_8$. The lengths of the recharge intervals $[t_3, t_4]$, and $[t_7, t_8]$ are all equal to zero. Thus, the actual recharge intervals are $[t_1, t_2]$, $[t_5, t_6]$ and $[t_9, t_{10}]$. The optimal switching times are 2.242, 3.330, 6.262, 8.089, 11.36, and 12.39, while the optimal recharge points are 130, 300, and 488. The corresponding optimal speed is 57.76 with minimum traveling times 14.292.

Figure 4.8.2 contains the optimal recharge plan, the optimal switching times and the optimal driving speed.

Example 4.3. In Problem (P_3), we choose $S = 600$, $P = 50$, $m = 1410$, $n = 81$ recharge locations are (130, 300, 490), a=12000.



(a) Optimal recharge plan (b) Optimal driving speed

Figure 4.8.3: The optimal driving strategy for Example 4.3

The dynamic system is chosen to be

$$\frac{dq}{dt} = \begin{cases} -\frac{1}{600000}\frac{(240v+8v^2+12000v)\cdot 1410)}{81}, & t \in [t_{2i-2}, \zeta_{2i-2}) \\ -\frac{1}{600000}\frac{(240v+8v^2)\cdot 1410}{81}, & t \in [\zeta_{2i-2}, t_{2i-1}) \\ \frac{2000}{q+50} - \exp\{2.5 - 3.0(t_{2i} - t_{2i-1})(t - t_{2i-1}\}, & t \in [t_{2i-1}, t_{2i}) \\ -\frac{1}{600000}\frac{(240v+8v^2+12000v)\cdot 1410}{81}, & t \in [t_{2l}, \zeta_{2l}) \\ -\frac{1}{600000}\frac{(240v+8v^2)\cdot 1410}{81}, & t \in [\zeta_{2l}, t_{2l+1}] \end{cases}$$

here $i = 1, \cdots, 3.$ (4.8.5)

with initial condition

$$q(0) = q_0, \tag{4.8.6}$$

where ζ_{2i-2} are switching times at each of which the speed from zero monotonically increases from zero to a positive constant.

Problem (P_3) then becomes

given dynamics (4.8.5) - (4.8.6), find the switching times $\boldsymbol{t}$, and corresponding speed v, such that

$$g_3(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v) \tag{4.8.7}$$

is minimized subject to the constraints

$$20 \leq q(t) \leq 80, \tag{4.8.8}$$

$$0 < p < 50, \tag{4.8.9}$$

$$0 < t_i - t_{i-1} \qquad i = 1, \cdots, 3, \tag{4.8.10}$$

$$\int_0^{t_{final}} v dt = 600, \tag{4.8.11}$$

$$\int_0^{t_{2i-1}} = \tau_i \qquad i = 1, \cdots, 3, \tag{4.8.12}$$

where the $(\tau_1, \tau_2, \tau_3) = (130, 300, 490)$.

Under the CPET transformation, defined in Section 4.5, the above problem becomes:

given dynamics

$$\frac{d\hat{q}}{ds} = \begin{cases} -\frac{1}{600000}\eta\frac{(240\hat{v}+8\hat{v}^2+12000v)\cdot 1410}{81}, & s \in [3i-3, 3i-2) \\ -\frac{1}{600000}\eta\frac{(240\hat{v}+8\hat{v}^2)\cdot 1410}{81}, & s \in [3i-2, 3i-1) \\ \eta\frac{2000}{q+50} - \exp\{2.5 - 3.0\eta_{3i}(s-3i+1)\}, & s \in [3i-1, 3i) \\ -\frac{1}{600000}\eta\frac{(240\hat{v}+8\hat{v}^2+12000v)\cdot 1410}{81}, & s \in [9, 10) \\ -\frac{1}{600000}\eta\frac{(240\hat{v}+8\hat{v}^2)\cdot 1410}{81}, & s \in [10, 11] \\ \text{here} & i = 1, \cdots, 3. \end{cases}$$

with initial condition $\hat{q}(0) = 60$,

find the parameter vectors η, and $\hat{v}$, such that

$$\hat{g}_3(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{11} \eta_i$$

is minimized subject to

$$20 \leq \hat{q}(s) \leq 80,$$

$$0 < p =< 50$$

$$0 < \eta_i \qquad i = 1, \cdots, 11,$$

$$\int_0^{11} \hat{v} ds = 600,$$

$$\int_0^{3i-1} \hat{v}(s) ds = \tau_i \qquad i = 1, \cdots, 3.$$

The optimal switching times are 2.359, 3.292, 6.377, 8.055, 11.50, 12.39, the optimal speed is 55.15 with minimum traveling times 14.3836. The optimal

i	1	2	3	4	5	6	7
t_i	2.359	3.292	6.377	8.055	11.50	12.39	14.38
x_i	130	130	300	300	490	490	600

Table 4.8.3: The optimal switching times and recharge points for Example 4.3

switching times and recharge points are listed in Table 4.8.3. Figure 4.8.3 show the optimal switching times and optimal driving speed.

In Examples 4.4 and 4.5, the road connecting the two cities is assumed to be undulated with two tuning points located at 350 and 550, respectively. Therefore $\gamma = (350, 550)$. The angles of slope of the road on the intervals (0, 350), (350, 550), and (550, 600) are $4.4°, 174°$, and $5.86°$, respectively.

Example 4.4. In Problem (P_4), choose $P = 50$, $m = 1440$, and $n = 84$. The recharge locations are restricted at 150, 250, 350, and 470. All turning points of the undulated road are included in the recharge points. Thus, the recharge points are 150, 250, 350, 470, and 550. The dynamic system is

$$\frac{dq}{dt} = \begin{cases} -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200}+g\sin\theta)\cdot 1380}{78}, & t \in [t_{2i-2}, t_{2i-1}) \\ \frac{2000}{q+50} - \exp\{2.5 - 3.0(t_{2i} - t_{2i-1})(t - t_{2i-1}\}, & t \in [t_{2i-1}, t_{2i}) \\ -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200}+g\sin\theta)\cdot 1380}{78}, & t \in [t_8, t_{10}] \end{cases}$$

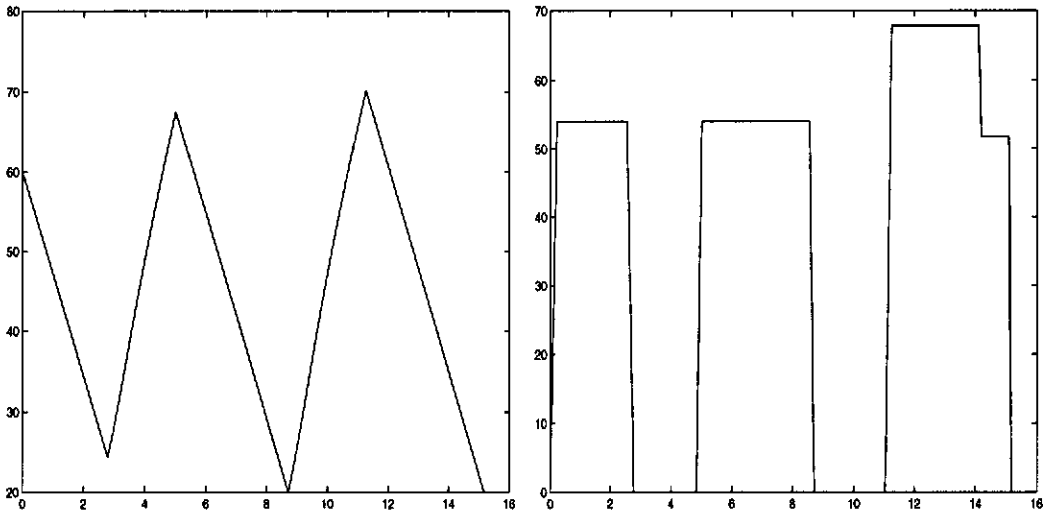
here $i = 1, \cdots, 4$ (4.8.13)

with initial condition

$$q(0) = 60, \tag{4.8.14}$$

Problem (P_4) then becomes:
given the dynamical system (4.8.13)–(4.8.14), find $\boldsymbol{t} = \{t_1, t_2, \cdots, t_9\}$, and v such that

$$g_4(\boldsymbol{t}, v) = t_{final}(\boldsymbol{t}, v)$$



(a) Optimal recharge plan (b) Optimal driving speed

Figure 4.8.4: The optimal driving strategy for Example 4.4

is minimized subject to the constraints

$$
\begin{aligned}
& 20 \leq q(t) \leq 80, \\
& 0 < p < 50, \\
& 0 < t_i - t_{i-1} && i = 1, \cdots, 10, \\
& \int_0^{t_{10}} v dt = 600, \\
& \int_0^{t_{2i-1}} v dt = \tau_i && i = 1, \cdots, 5,
\end{aligned}
$$

where $(\tau_1, \cdots, \tau_5) = (150, 250, 350, 470, 550)$.

Under the CPET given in Section 4.6, we have:

given the dynamics

$$
\frac{d\hat{q}}{ds} = \begin{cases} -\frac{1}{100}\eta\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200}+g\sin\theta)\cdot 1380}{78}, & s \in [2i-2, 2i-1) \\ \eta\frac{2000}{q+50} - \exp\{2.5 - 3.0\eta_{2i}(s-2i+1)\}, & s \in [2i-1, 2i) \\ -\frac{1}{100}\eta\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200}+g\sin\theta)\cdot 1380}{78}, & s \in [8, 10] \end{cases}
$$

here $i = 1, \cdots, 4$

with initial condition

$$\hat{q}(0) = 60,$$

i	1	2	3	4	5	6	7	8	9	10
t_i	2.78	5.01	6.86	6.86	8.71	11.26	13.02	13.02	14.20	15.17
x_i	150	150	250	250	350	350	470.1	470.1	550.1	600

Table 4.8.4: The optimal switching times and recharge points for Example 4.4

find parameter vectors η, and $\hat{v}$ such that

$$\hat{g}_4(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{10} \eta_i$$

is minimized subject to

$$\begin{aligned}
&20 \leq \hat{q}(i) && i = 1, 3, 5, 7, 9, 10, \\
&\hat{q}(2i) \leq 80 && i = 1, \cdots, 4, \\
&0 < p =< 50, \\
&0 < \eta_i && i = 1, \cdots, 10, \\
&\int_0^{10} \hat{v} ds = 600, \\
&\int_0^{2i-1} \hat{v}(s) ds = \tau_i && i = 1, \cdots, 5.
\end{aligned}$$

The optimal switching times and recharge points, listed in Table 4.8.4, show that there exist redundancies: $t_3 = t_4$, $t_7 = t_8$. The lengths of the recharge intervals $[t_3, t_4]$ and $[t_7, t_8]$ are all equal to zero. The actual recharge intervals are $[t_1, t_2]$ and $[t_5, t_6]$. Therefore, the optimal recharge locations are 150, and 350. The optimal switching times are 2.778, 5.010, 8.714, 11.256, and 14.204. The corresponding optimal speeds are 54.0, 67.8, and 51.8. The minimum traveling time is 15.170.

Figure 4.8.4 contains the optimal recharge plan, the optimal switching times, and the optimal driving speed.

Example 4.5. In Problem (P_5), choose $S = 600$, $P = 50$, $m_1 = 1440$, $L =$

5, and $n = 84$. The dynamical system is:

$$\frac{dq}{dt} = \begin{cases} -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200})\cdot 1440}{84}, & t \in [t_{2i-2}, t_{2i-1}) \\ \frac{2000}{q+50} - \exp\{2.5 - 3.0(t_{2i} - t_{2i-1})(t - t_{2i-1}\}, & t \in [t_{2i-1}, t_{2i}) \\ -\frac{1}{100}\frac{(\frac{v}{40}+\frac{v^2}{1200})\cdot 1440}{84}, & t \in [t_{2l}, t_{2l+1}] \end{cases}$$
$$\text{here} \qquad i = 1, \cdots, l \qquad (4.8.15)$$

with initial condition

$$q(0) = 60. \qquad (4.8.16)$$

Problem (P_5) becomes:
given the dynamical system (4.8.15)–(4.8.16), find $\boldsymbol{t} = \{t_1, t_2, \cdots, t_{2l}\}$, v, l, such that

$$g_5(\boldsymbol{t}, v, l) = t_{final}(\boldsymbol{t}, v, l)$$

is minimized subject to the constraints

$$\begin{aligned} & 20 \leq q(t) \leq 80, \\ & 0 < p < 50, \\ & 0 < t_i - t_{i-1} \qquad i = 1, \cdots, 2l+1, \\ & \int_0^{t_{final}} v dt = 600, \\ & \int_0^{t_{i_j}} v(t) dt = \gamma_j \qquad j = 1, 2, \text{ and } t_{i_j} \in \boldsymbol{t}, \\ & l \leq 10, \end{aligned}$$

where $\boldsymbol{\gamma} = (350, 550)$.
We choose $k_0 = 3$ in Algorithm B. Consider the case in which $k = 5$.

Under the CPET given in Section 4.7, the problem becomes:
given the dynamical system

$$\frac{d\hat{q}}{ds} = \begin{cases} -\frac{1}{100}\eta\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200})\cdot 1440}{84}, & s \in [2i-2, 2i-1) \\ \eta\frac{2000}{q+50} - \exp\{2.5 - 3.0\eta_{2i}(s - 2i + 1)\}, & s \in [2i-1, 2i) \\ -\frac{1}{100}\eta\frac{(\frac{\hat{v}}{40}+\frac{\hat{v}^2}{1200})\cdot 1440}{84}, & s \in [10, 11] \end{cases}$$
$$\text{here} \qquad i = 1, \cdots, 5$$

i	1	2	3	4	5	6
t_i	3.3600	5.2944	8.6227	11.018	14.028	15.026
x_i	175.83	175.83	350.00	350.00	550.00	600

Table 4.8.5: The optimal switching times and recharge points for Example 4.5

with initial condition

$$\hat{q}(0) = 60,$$

find parameter vectors η and $\hat{v}$, such that

$$\hat{g}_4(\boldsymbol{\eta}, \hat{v}) = \sum_{i=1}^{11} \eta_i$$

is minimized subject to

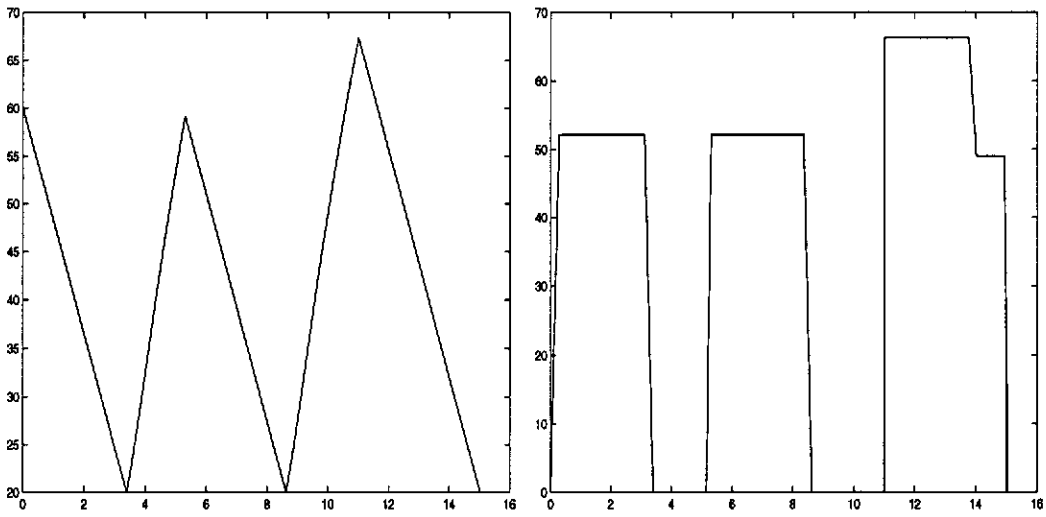
$$\begin{aligned}
&20 \leq \hat{q}(2i-1) && i = 1, \cdots, 6, \\
&\hat{q}(2i) \leq 80 && i = 1, \cdots, 5, \\
&0 < p \leq 50, && \\
&0 < \eta_i && i = 1, \cdots, 11, \\
&\int_0^{i_j} v(t)dt = \gamma_j && j = 1, 2, \text{ and } i_j \in \{1, 3, 5, 7, 9\}, , \\
&\int_0^{11} \hat{v} ds = 600. &&
\end{aligned}$$

The optimal switching times and recharge points are listed in Table 4.8.5. The optimal switching times are 3.360, 5.294, 8.623, 11.02, and 14.03, the optimal recharge locations are 175.8, and 350.0. The corresponding optimal speeds are 52.06, 66.22, and 49.79, respectively. The minimum traveling time is 15.026.

Figure 4.8.5 contains the optimal recharge plan, the optimal switching times and the optimal driving speed.

4.9 Conclusion

In this chapter, we construct a battery–powered electric vehicle model, in which driving strategy is to be obtained such that the total traveling time between



(a) optimal recharge plan (b) optimal driving speed

Figure 4.8.5: The optimal driving strategy for Example 4.5

two locations is minimized. We show that the recharge and driving plan for this battery–powered electrical vehicle model can be formulated as unconventional constrained optimal control problem, and the problem can be converted into conventional optimal control problem by using CPET. Numerical examples are solved using the proposed method.

From the above examples, it can be seen that the car traveling at a high speed will use more power per kilometer than traveling at a low speed, since the relationship between the rate of power used and the speed is a quadratic function. High speed will eventually increase the battery recharge time and result in the increase in the total traveling time. It is clear too that the low speed will increase the total running time and result in the increase in the total traveling time. The optimal speed will balance the battery recharge time and the driving time, and thus reduce the total traveling time. It can also be seen that the optimal speed on an uphill road will lower than on an even road, and the optimal speed on a downhill road is higher than on an even road, because of the effect of gravity. From the above examples, it can also be found that every recharge will cause some time delay. Therefore, the more the recharge time, the more the delay will be. On the other hand, the recharge rate will be lower, when the battery contains more power, and thus the recharge will

be slower when the battery close to the full capacity. The optimal number of recharge and optimal recharge times will balance the loss of time on the delay of recharges and the loss of time on the lower recharge rate when the battery is close to the full capacity and eventually reduce the total traveling time.

Chapter 5

Optimal Control Problem With Variable Time Points in the Objective Functions

5.1 Introduction

In this chapter, we consider the numerical solution of an optimal control problem involving variable time points. Its motivation comes from a situation in which a target is moving as a function of time in the space. A space-craft is launched into space, and its trajectory is maneuvered by certain control actions. The mission of the space-craft is to take measurements at various time points over a given mission period which is divided into a number of time subintervals. Each of the time points is to be selected from the respective time subinterval. Suppose, we wish to take the measurement in each time subinterval at the time point at which the distance between the moving target and the space-craft is a minimum. Let the sum of these distances be the cost function. Then, we have an optimal control problem, where the control actions of the space-craft and the variable time points are to be chosen optimally with respect to the given cost function. A different problem, also involving variable time points, has been discussed in [14] and [15] from the theoretical point of view. In that problem, each equation in the dynamical system is defined on an interval with variable initial and terminal time points which are decision variables. The dynamical system of the problem considered here has fixed initial and terminal time points,

but has some variable observation time points within the time interval. Also, the main focus of this chapter is to present some efficient numerical techniques for solving these optimal control problems with variable time points, while [14] and [15] are only concerned with the theory of necessary optimality conditions for their problems. The rest of this Chapter is organized as follows:

A general class of optimal control problems containing the situation just mentioned above as an example is formulated in Section 5.2, where the cost function includes multiple variable time points. The control parameterization enhancing technique is used to transform the problem into a form solvable by the control parameterization technique in Section 5.3. More specifically, the control parameterization enhancing transform [100] is first used to convert the optimal control problem with variable time points to an equivalent optimal control problem with fixed multiple characteristic time (MCT) (cf. [56]). The transformed problems are essentially optimal parameter selection problems with MCT. The gradient formulae for the objective function as well as the constraint functions with respect to relevant decision variables are derived in Section 5.4. With these gradient formulae, each of the transformed optimal control problems is solvable as an optimal parameter selection problem, and the software MISER 3.2 [42] can be modified for solving these optimal parameter selection problems. In Section 5.5, two examples are solved using the proposed method. Section 5.6 concludes this chapter.

5.2 Problem Formulation

Consider a process described by the following system of differential equations defined on [0,T].

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{z}(t)), \tag{5.2.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{5.2.2}$$

where T is a fixed terminal time, $\boldsymbol{x} = [x_1, \cdots, x_n]^T \in \mathbb{R}^n$, $\boldsymbol{u} = [u_1, \cdots, u_m]^T \in \mathbb{R}^m$ and $\boldsymbol{z} = [z_1, \cdots, z_p]^T \in \mathbb{R}^p$ are , respectively, state, control, and system pa-

rameter, while $\boldsymbol{f} = [f_1, \cdots, f_n]^T \in \mathbb{R}^n$ is a continuously differentiable function with respect to all its arguments, and $\boldsymbol{x}^0$ is a given vector.

Let $\underline{\tau_i}$ and $\overline{\tau_i}$, $i = 0, 1, \cdots, k$, be constants in the time interval [0,T] such that

$$\underline{\tau_0} = \overline{\tau_0} = 0 < \underline{\tau_1} < \overline{\tau_1} < \underline{\tau_2} < \overline{\tau_2} < \cdots < \underline{\tau_k} < \overline{\tau_k} < T \tag{5.2.3}$$

Furthermore, let a_i and b_i, $i = 1, \cdots, s$, c_i and d_i, $i = 1, \cdots, m$, be fixed constants. Define

$$\boldsymbol{V} = \{\boldsymbol{t} = [t_1, \cdots, t_k]^T \in \mathbb{R}^k : t_i \in [\underline{\tau_i}, \overline{\tau_i}],\ i = 1, \cdots, k\}$$

$$\boldsymbol{Z} = \{\boldsymbol{z} = [z_1, \cdots, z_r]^T \in \mathbb{R}^r : a_i \leq z_i \leq b_i,\ i = 1, \cdots, r\}$$

$$\boldsymbol{U} = \{\boldsymbol{u} = [u_1, \cdots, u_m]^T \in \mathbb{R}^m : c_i \leq u_i \leq d_i,\ i = 1, \cdots, m\}$$

Any Borel measurable function $\boldsymbol{u} : [0, T] \longrightarrow \boldsymbol{U}$ is called an admissible control. Let $\mathcal{U}$ be the class of all admissible controls. For each $(\boldsymbol{u}, \boldsymbol{z}) \in \mathcal{U} \times \boldsymbol{Z}$, let$\boldsymbol{x}(\cdot|\boldsymbol{u}, \boldsymbol{z})$ denote the corresponding solution of the system (5.2.1) –(5.2.2).

Our optimal control problem may now be formally stated as:
given the dynamical system (5.2.1) –(5.2.2), find a $(\boldsymbol{t}, \boldsymbol{u}, \boldsymbol{z}) \in \boldsymbol{V} \times \mathcal{U} \times \boldsymbol{Z}$ such that the cost function

$$g_0(\boldsymbol{t}, \boldsymbol{u}, \boldsymbol{z}) = \sum_{i=1}^{k} \Phi_i(t_i, \boldsymbol{x}(t_i|\boldsymbol{u}, \boldsymbol{z})) \tag{5.2.4}$$

is minimized subject to the constraints

$$g_j(t_i, \boldsymbol{x}(t_i|\boldsymbol{u}, \boldsymbol{z}), \boldsymbol{z}) \leq 0, \quad j = 1, \cdots, q; \quad i = 1, \cdots, k \tag{5.2.5}$$

$$a_i \leq z_i \leq b_i, \qquad i = 1, 2, \cdots, r, \tag{5.2.6}$$

$$c_i \leq u_i(t) \leq d_i \quad i = 1, 2, \cdots, m \qquad t \in [0, T] \tag{5.2.7}$$

$$\underline{\tau_i} \leq t_i \leq \overline{\tau_i} \quad i = 1, 2, \cdots, k, \tag{5.2.8}$$

where $\Phi_i(t_i, \boldsymbol{x})$, $i = 1, 2 \cdots, k$, and $g_j(t_i, \boldsymbol{x}, \boldsymbol{z})$ $j = 1, \cdots, q$, $i = 1, \cdots, k$, are continuously differentiable real valued functions on $[0, T] \times \mathbb{R}^n$ and $[0, T] \times \mathbb{R}^n \times \mathbb{R}^r$, respectively. Let this optimal control problem be referred to as Problem P.

5.3 Transformation

Let $s \in [0, k+1]$ be a new time variable, and let $\boldsymbol{v}(s)$ be defined by

$$\boldsymbol{v}(s) = \sum_{i=1}^{k+1} v_i \cdot \chi_{[i-1,i)}(s) \tag{5.3.1}$$

where $\chi_{[i-1,i)}$ is the indicator function of the interval [i-1,i), and the $v_i's$ are non-negative constants. Clearly, $v(s)$ is a nonnegative piecewise constant function, which is called the enhancing control, defined on [0,k+1] with fixed switching points located at $\{1, 2, \cdots, k\}$.

The control parameterization enhancing transform (CPET) maps $t \in [0, T]$ to $s \in [0, k+1]$ as follows:

$$\begin{aligned} &\frac{dt}{ds} = \boldsymbol{v}(s) \\ &t(0) = 0 \end{aligned}$$

where

$$\boldsymbol{v}(s) = \begin{cases} t_i - t_{i-1} & s \in [i-1, i) \\ T - t_k & s \in [k, k+1] \end{cases} \qquad i = 1, 2, \cdots, k$$

which satisfies

$$\underline{\tau_i} \leq \int_0^i \boldsymbol{v}(s)ds \leq \overline{\tau_i}, \qquad i = 1, \cdots, k, \tag{5.3.2}$$

and

$$\int_0^{k+1} \boldsymbol{v}(s)ds = T. \tag{5.3.3}$$

Such a function is called an enhancing control, let $\mathcal{V}^*$ denote the class of all such enhancing controls.

Under the CPET, the system dynamics changes to

$$\frac{d}{ds}\begin{pmatrix} \boldsymbol{y}(s) \\ t(s) \end{pmatrix} = \boldsymbol{v}(s)\begin{pmatrix} \boldsymbol{f}(t(s), \boldsymbol{y}(s), \boldsymbol{w}(s), \boldsymbol{z}) \\ 1 \end{pmatrix}, \; s \in [0, k+1] \tag{5.3.4}$$

with initial condition

$$\begin{pmatrix} \boldsymbol{y}(0) \\ t(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix} \tag{5.3.5}$$

where $\boldsymbol{y}(s) = \boldsymbol{x}(t(s))$ and $\boldsymbol{w}(s) = \boldsymbol{u}(t(s))$.
The constraints

$$c_i \leq u_i(t) \leq d_i, \quad i = 1, 2, \cdots, m, \quad t \in [0, T] \tag{5.3.6}$$

reduce to

$$c_i \leq w_i(s) \leq d_i \quad i = 1, 2, \cdots, m, \quad s \in [0, k+1]. \tag{5.3.7}$$

Define $\boldsymbol{w}(s) = (w_1(s), w_2(s), \cdots, w_m(s))$ where w_i, $i = 1, \cdots, m$, satisfy the constraints (5.3.7). Let $\mathcal{W}$ be the set of all such functions $\boldsymbol{w}(s)$, and the Problem P is now transformed into the following optimal control problem:
given the dynamical system (5.3.4)–(5.3.5), find an admissible element $(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z}) \in \mathcal{V}^* \times \mathcal{W} \times \boldsymbol{Z}$ such that the cost function:

$$\hat{g}_0(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z}) = \sum_{i=1}^{k} \hat{\Phi}_i(\sum_{j=1}^{i} v_j, \boldsymbol{y}(i|\boldsymbol{w}, \boldsymbol{z}), \boldsymbol{z}) \tag{5.3.8}$$

is minimized subject to the constraints:

$$\hat{g}_j(\sum_{j=1}^{i} v_j, \boldsymbol{y}(i), \boldsymbol{z}) \leq 0, \qquad j = 1, \cdots, l; \quad i = 1, \cdots, k, \tag{5.3.9}$$

$$a_i \leq z_i \leq b_i, \qquad i = 1, 2, \cdots, r, \tag{5.3.10}$$

$$c_i \leq w_i(s) \leq d_i, \qquad i = 1, 2, \cdots, m, \quad s \in [0, k+1] \tag{5.3.11}$$

$$\boldsymbol{v}(s) \in \mathcal{V}^* \tag{5.3.12}$$

This problem is referred to as Problem P^*.

An admissible element $(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z}) \in \mathcal{V}^* \times \mathcal{W} \times \boldsymbol{Z}$ (respectively, $(\boldsymbol{t}, \boldsymbol{u}, \boldsymbol{z})$) is called a feasible element of Problem P^* (respectively, Problem P) if the constraints (5.3.9)–(5.3.12) (respectively constraints (5.2.5) – (5.2.8)) are satisfied.

Theorem 5.3.1 *Problem P_* is equivalent to Problem P in the sense that $(\boldsymbol{v}^*, \boldsymbol{w}^*, \boldsymbol{z}^*)$ is a solution of problem (P_*) if and only if $(\boldsymbol{t}^*, \boldsymbol{u}^*, \boldsymbol{z}^*)$ is a solution of problem (P), and*

$$\hat{g}_0(\boldsymbol{v}^*, \boldsymbol{w}^*, \boldsymbol{z}^*) = g_0(\boldsymbol{t}^*, \boldsymbol{u}^*, \boldsymbol{z}^*)$$

PROOF. Let $(t_1, u_1, z_1) \in (\mathcal{V} \times \mathcal{U} \times \mathcal{Z})$ be a feasible element of Problem (P), and let $(v_1, w_1, z_1) \in \mathcal{V}^* \times \mathcal{W} \times \mathcal{Z}$ be the corresponding feasible element of problem (P_*). Then it is easy to check that $x(t)$ is the solution of (5.2.1)–(5.2.2) if and only if $y(s)$ is the solution of (5.3.4)–(5.3.5), and

$$\begin{aligned} g_0(t_1, u_1, z_1) &= \sum_{i=1}^{k} \Phi(t_i, x(t_i|u_1, z_1)) \\ &= \sum_{i=1}^{k} \hat{\Phi}_i(\sum_{j=1}^{i} v_j, y(i|w_1, z_1), z_1) = \hat{g}_0(v_1, w_1, z_1) \end{aligned}$$

Hence, the results follows readily. □

In Problem (P) the cost function (5.2.4) is to be minimized with respect to $(t, u, z) \in (\mathcal{V} \times \mathcal{U} \times \mathcal{Z})$ where $t = [t_1, \cdots, t_k]$ and t_i, $i = 1, \cdots, k$, are switching times. On the other hand, the cost function (5.3.8) in Problem P^* is to be minimized with respect to $(v, w, z) \in \mathcal{V}^* \times \mathcal{W} \times \mathcal{Z}$, where $v(t)$ is a nonnegative piecewise constant function. Since Problem P is equivalent to Problem (P^*), we choose to solve Problem P^* which is an optimal control problem with multiple characteristic times (see [56]). The main reason Problem P^* is numerically more tractable, is that it does not involve variable switching times.

In the classical control parameterization technique, each control function $w_i(s)$ is approximated by a zeroth order or first order spline function (that is, a piecewise constant function or a piecewise linear continuous function) defined on a set of knots $\{0 = s_0^i, s_1^i, \cdots, s_{p_i}^i = (k+1)\}$. Note that each component may have a different set of knots and the knots are not necessarily equally spaced. For the case of piecewise constant basis functions, we write the i-th control function as the sum of basis functions with coefficients or parameters $\{\sigma_{ij}, j = 1, 2, \cdots, p_i\}$:

$$w_i(s) = \sum_{j=1}^{p_i} \sigma_{ij} B_{ij}^{(0)}(s)$$

where $B_{ij}^{(0)}(s)$ is the indicator function for the j-th interval of the i-th set of knots defined by

$$B_{ij}^{(0)}(s) = \begin{cases} 1, & s_{j-1}^i \le s \le s_j^i, \\ 0, & \text{otherwise} \end{cases}$$

For piecewise linear continuous basis functions, we write the i-th control function as:

$$\boldsymbol{w}_i(s) = \sum_{j=0}^{p_i} \sigma_{ij} B_{ij}^{(1)}(s),$$

where $B_{ij}^{(1)}(s)$ are the witch's hat functions defined by

$$B_{ij}^{(1)}(s) = \begin{cases} (s - s_1^i)/(s_0^i - s_1^i), & s \in [s_0^i, s_1^i], \\ 0 & \text{otherwise}, \end{cases}$$

$$B_{ij}^{(1)}(s) = \begin{cases} (s - s_{j-1}^i)/(s_j^i - s_{j-1}^i), & s \in [s_{j-1}^i, s_j^i], \\ (s - s_{j+1}^i)/(s_j^i - s_{j+1}^i), & s \in [s_j^i, s_{j+1}^i], \\ 0 & \text{otherwise}, \end{cases}$$

$$B_{ip_i}^{(1)}(s) = \begin{cases} (s - s_{k_i-1}^i)/(s_{k_i}^i - s_{k_i-1}^i), & s \in [s_{p_i-1}^i, s_{p_i}^i], \\ 0 & \text{otherwise}, \end{cases}$$

Thus, $\boldsymbol{w}(s)$ can be uniquely identified with a control parameter vector $\boldsymbol{\sigma}$ and vice verses with:

$$\boldsymbol{\sigma} = [(\boldsymbol{\sigma}^1)^T, (\boldsymbol{\sigma}^2)^T, \cdots, (\boldsymbol{\sigma}^m)^T]^T$$

$$\boldsymbol{\sigma}^i = [\sigma_{i,1}, \sigma_{i,2}, \cdots, \sigma_{i,p_i}]^T$$

which satisfy conditions:

$$c_i \le \sigma_{ij} \le d_i, \quad i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, p_i \tag{5.3.13}$$

Let $\boldsymbol{\Sigma}$ denote the set of all such controls parameter vector $\boldsymbol{\sigma}$.

We now apply the Control Parameterization Enhancing Transform (CPET) to Problem P^*. Let q be the second new time scale which varies from 0 to k+1, the transformation from $s \in [0, k+1]$ to $q \in [0, k+1]$ can be defined by the differential equation:

$$\frac{ds(q)}{dq} = \eta(q)$$
$$s(0) = 0$$

where the scaling function $\eta(q)$ is called the enhancing control. It is a piecewise constant function with possible discontinuities at the pre-fixed knots $\xi_0, \cdots, \xi_M$, ie.

$$\eta(q) = \sum_{i=1}^{M} \eta_i \chi_i(q),$$

where $\chi_i(q)$ is the indicator function defined by

$$\chi_i(q) = \begin{cases} 1, & \text{if } q \in [\xi_{i-1}, \xi_i], \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$s(q) = \int_0^q \eta(\nu) d\nu = \sum_{j=1}^{i-1} \eta_j (\xi_j - \xi_{j-1}) + \eta_i (q - \xi_{i-1}).$$

Let Θ denote the class of all such enhancing controls $\eta(q)$ satisfying

$$\eta(i) = \int_0^i \eta(\nu) d\nu = i \qquad i = 1, \cdots, k+1 \tag{5.3.14}$$

Under the CPET, the system dynamics changes to :

$$\frac{d}{dq} \begin{pmatrix} \hat{\boldsymbol{y}}(q) \\ s(q) \end{pmatrix} = \eta(q) \begin{pmatrix} \boldsymbol{v}(s(q)) \boldsymbol{f}(t(s(q)), \hat{\boldsymbol{y}}(q), \boldsymbol{\sigma}, \boldsymbol{z}) \\ \boldsymbol{v}(s(q)) \end{pmatrix},$$
$$\text{here} \qquad q \in [0, k+1] \tag{5.3.15}$$

with initial condition

$$\begin{pmatrix} \hat{\boldsymbol{y}}(0) \\ s(0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_0 \\ 0 \end{pmatrix} \tag{5.3.16}$$

where $\hat{\boldsymbol{y}}(s) = \boldsymbol{x}(t(s)q))$.

The constraint (5.3.7) is reduced to (5.3.13). Problem (P^*) is now transformed into the following optimal control problem:

given the dynamical system (5.3.15) – (5.3.16) find an admissible element $(\boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{z}) \in \mathcal{V}^* \times \boldsymbol{\Theta} \times \boldsymbol{\Sigma} \times \boldsymbol{Z}$ such that the cost function

$$\hat{\hat{g}}_0(\boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{z}) = \sum_{i=1}^{k} \hat{\hat{\Phi}}_i(\sum_{j=1}^{i} v_j, \hat{\boldsymbol{y}}(i|\boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{z})\boldsymbol{z}) \tag{5.3.17}$$

is minimized subject to the constraints:

$$\hat{\hat{g}}_j(\sum_{j=1}^{i} v_j, \hat{\boldsymbol{y}}(i|\boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{z}), \boldsymbol{z}) \leq 0, \quad j = 1, \cdots, l; i = 1, \cdots, k, \tag{5.3.18}$$

$$a_i \leq z_i \leq b_i, \qquad i = 1, 2, \cdots, r, \tag{5.3.19}$$

$$c_i \leq \sigma_{ij} \leq d_i \qquad i = 1, \cdots, k, \quad j = 1, \cdots, p_i \tag{5.3.20}$$

$$\boldsymbol{v}(s) \in \mathcal{V}^* \tag{5.3.21}$$

$$\boldsymbol{\eta}(q) \in \Theta \tag{5.3.22}$$

This problem is referred to as Problem (P^{**}). Problem (P^{**}) is an optimal control problem with MCT, where the control functions are piecewise constant or piecewise linear continuous functions with pre-fixed partition points. Hence, it can be viewed as an optimal parameter selection problem with MCT.

5.4 Gradient Formulae

To solve Problem (P^{**}), we need to derive the gradient formulae of the cost function as well as the constraint functions $\hat{\hat{g}}_j(\sum_{j=1}^{i} v_j, \hat{\boldsymbol{y}}(i|\boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{z}), \boldsymbol{z})$, $j = 0, 1, \cdots, q$, $i = 1, \cdots, k$, (5.3.18). For simplicity, we simplify the notation, and hence the notation used in this section is applicable only to this section. For example, $\boldsymbol{\tau} = (\tau_1, \tau_2, \cdots, \tau_k)$ is used to denote the vector of multiple characteristic times. The parameter set $(\boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{z})$ is replaced by $\boldsymbol{z}$, and the cost

functional is reduced to

$$g_0(\boldsymbol{u}, \boldsymbol{z}) = \sum_{i=1}^{k} \Phi_{0,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u}, \boldsymbol{z}), \boldsymbol{z}), \tag{5.4.1}$$

where $0 = \tau_0 < \tau_1 < \ldots < \tau_k < \tau_{k+1} = T$. The state differential equation is

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}\big(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{z}\big) \tag{5.4.2}$$

with the initial conditions

$$\boldsymbol{x}(0) = \boldsymbol{x}^0(\boldsymbol{z}). \tag{5.4.3}$$

Let $\boldsymbol{U}$, and $\boldsymbol{Z}$ be as defined in Section 2. The optimal control problem is to find an admissible element $(\boldsymbol{u}, \boldsymbol{z}) \in \boldsymbol{U} \times \boldsymbol{Z}$ such that the cost functional (5.4.1) is minimized subject to the constraints:

$$g_j(\boldsymbol{u}, \boldsymbol{z}) = \sum_{i=1}^{k} \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u}, \boldsymbol{z}), \boldsymbol{z}), \tag{5.4.4}$$

where $j = 1, \cdots, q$.

Define a costate system given by following differential equations:

$$\dot{\boldsymbol{\lambda}}_j^T(t) = -\boldsymbol{\lambda}_j^T \frac{\partial \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{z})}{\partial \boldsymbol{x}(t)} \qquad j = 0, 1, \cdots, q, \tag{5.4.5}$$

where $t \in (\tau_{i-1}, \tau_i), \quad i = 1, 2, \cdots, k+1,$
subject to the internal jumps:

$$\boldsymbol{\lambda}_j^T(\tau_i^+) + \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i), \boldsymbol{z})}{\partial \boldsymbol{x}(\tau_i)} = \boldsymbol{\lambda}_j^T(\tau_i^-) \qquad i = 1, \cdots, k, \tag{5.4.6}$$

$$\boldsymbol{\lambda}_j^T(T) = 0 \tag{5.4.7}$$

Theorem 5.4.1 *The gradient of the cost functional (5.4.1) or constraints (5.4.4) with respect to control u is*

$$\nabla_{\boldsymbol{u}}\, g_j(\boldsymbol{u}, \boldsymbol{z}) = \int_0^T \boldsymbol{\lambda}_j^T \frac{\partial f}{\partial \boldsymbol{u}} dt \qquad j = 0, 1, \cdots, q.$$

PROOF. The main reference for this theorem is [56].

Let $\delta \boldsymbol{u}$ be a small variation in $\boldsymbol{u}$. Then δg_j is the corresponding change in g_j, and $\delta \boldsymbol{x}$ is the corresponding change in $\boldsymbol{x}$. Thus,

$$\begin{aligned}
\delta\, g_j &= \sum_{i=1}^{k} \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{x}} \delta \boldsymbol{x}(\tau_i) \\
&= \sum_{i=1}^{k} \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{x}} \delta \boldsymbol{x}(\tau_i) - \sum_{i=1}^{k+1} \left[\lambda_j^T \delta \boldsymbol{x}(t)\right]_{t=(\tau_{i-1})^+}^{t=\tau_i^-} \\
&\quad + \sum_{i=1}^{k+1} \int_{\tau_{i-1}}^{\tau_i} \left\{ (\dot{\lambda}_j^T + \lambda_j^T \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}) \delta \boldsymbol{x}(t) + \lambda_j^T \frac{\partial f}{\partial \boldsymbol{u}} \delta \boldsymbol{u} \right\} dt
\end{aligned}$$

Since $\boldsymbol{x}(t)$ is continuous in $t \in [0, T]$, we have

$$\begin{aligned}
\sum_{i=1}^{k+1} \left[\lambda_j^T \delta \boldsymbol{x}\right]_{t=\tau_{i-1}^+}^{t=\tau_i^-} &= \sum_{i=1}^{k+1} \{\lambda_j^T(\tau_i^-)\delta \boldsymbol{x}(\tau_i) - \lambda_j^T(\tau_{i-1}^+)\delta \boldsymbol{x}(\tau_{i-1})\} \\
&= \sum_{i=1}^{k} \{\lambda_j^T(\tau_i^-) - \lambda_j^T(\tau_i^+)\}\delta \boldsymbol{x}(\tau_i) \\
&\quad -\lambda_j^T(\tau_0^+)\delta \boldsymbol{x}(\tau_0) + \boldsymbol{\lambda}_j(\tau_{k+1}^-)\delta \boldsymbol{x}(\tau_{k+1}) \\
&= \sum_{i=1}^{k} \{\lambda_j^T(\tau_i^-) - \lambda_j^T(\tau_i^+)\}\delta \boldsymbol{x}(\tau_i) \\
&\quad -\lambda_j^T(0^+)\delta \boldsymbol{x}(0) + \boldsymbol{\lambda}_j(T^-)\delta \boldsymbol{x}(T) \\
&= \sum_{i=1}^{k} \{\lambda_j^T(\tau_i^-) - \lambda_j^T(\tau_i^+)\}\delta \boldsymbol{x}(\tau_i) \\
&\quad +\lambda_j^T(T^-)\delta \boldsymbol{x}(T) - \lambda_j^T(0^+)\delta \boldsymbol{x}(0)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\delta\, g_j &= \sum_{i=1}^{k} \left\{ \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}))}{\partial \boldsymbol{x}} - \lambda_j^T(\tau_i^-) + \lambda_j^T(\tau_i^+) \right\} \delta \boldsymbol{x}(\tau_i) \\
&\quad -\lambda_j^T(T)\delta \boldsymbol{x}(T) + \sum_{i=1}^{k+1} \int_{\tau_{i-1}}^{\tau_i} \left\{ (\dot{\lambda}_j^T + \lambda_j^T \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}) \delta \boldsymbol{x}(t) + \lambda_j^T \frac{\partial f}{\partial \boldsymbol{u}} \delta \boldsymbol{u} \right\} dt
\end{aligned}$$

From the definition of the costate system (5.4.5) – (5.4.7), all variations in $\delta \boldsymbol{x}$ and $\delta \boldsymbol{x}(\tau_i)$ vanish, yielding

$$\nabla_{\boldsymbol{u}}\, g_j(\boldsymbol{u},\boldsymbol{z}) = \sum_{i=1}^{k+1} \int_{\tau_{i-1}}^{\tau_i} \lambda_j^T \frac{\partial f}{\partial \boldsymbol{u}} dt = \int_0^T \lambda_j^T \frac{\partial f}{\partial \boldsymbol{u}} dt$$

The proof is complete. □

For the gradient of cost functional with respect to system parameters $\boldsymbol{z}$, we have the following theorem:

Theorem 5.4.2 *The gradient of the cost functional (5.4.1) or constraints (5.4.4) with respect to system parameters $\boldsymbol{z}$ is*

$$\nabla_{\boldsymbol{z}} g_j(\boldsymbol{u},\boldsymbol{z}) = \int_0^{k+1} \boldsymbol{\lambda}_j^T \frac{\partial f}{\partial \boldsymbol{z}} dt + \boldsymbol{\lambda}_j^T(0)\frac{\partial \boldsymbol{x}^0(\boldsymbol{z})}{\partial \boldsymbol{z}} + \sum_{i=1}^{k} \frac{\partial \Phi_{j,i}(\tau_i, \boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{z}}.$$

for $j = 0, \cdots, q.$

PROOF. Similar to the proof given for Theorem 5.4.1, let $\delta \boldsymbol{z}$ be a small variation in $\boldsymbol{z}$ and let the corresponding variations in $\boldsymbol{x}$ and g_j be as in Theorem 5.4.1. Then,

$$\begin{aligned}
\delta\, g_j &= \sum_{i=1}^{k} \left\{ \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{x}} \delta \boldsymbol{x}(\tau_i) + \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{z}} \delta \boldsymbol{z} \right\} \\
&= \sum_{i=1}^{k} \left\{ \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{x}} \delta \boldsymbol{x}(\tau_i) + \frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{z}} \delta \boldsymbol{z} \right\} \\
&\quad - \sum_{i=1}^{k+1} \left[\boldsymbol{\lambda}_j^T \delta \boldsymbol{x}(t) \right]_{t=(\tau_{i-1})^+}^{t=\tau_i^-} \\
&\quad + \sum_{i=1}^{k+1} \int_{\tau_{i-1}}^{\tau_i} \left\{ (\dot{\boldsymbol{\lambda}}_j^T + \boldsymbol{\lambda}_j^T \frac{\partial f}{\partial \boldsymbol{x}}) \delta \boldsymbol{x}(t) + \boldsymbol{\lambda}_j^T \frac{\partial f}{\partial \boldsymbol{z}} \delta \boldsymbol{z} \right\} dt
\end{aligned}$$

Since $\boldsymbol{x}(t)$ is continuous in $t \in [0,T]$, we have

$$\begin{aligned}
\sum_{i=1}^{k+1} \left[\boldsymbol{\lambda}_j^T \delta \boldsymbol{x} \right]_{t=\tau_{i-1}^+}^{t=\tau_i^-} &= \sum_{i=1}^{k+1} \{ \boldsymbol{\lambda}_j^T(\tau_i^-) \delta \boldsymbol{x}(\tau_i) - \boldsymbol{\lambda}_j^T(\tau_{i-1}^+) \delta \boldsymbol{x}(\tau_{i-1}) \} \\
&= \sum_{i=1}^{k} \{ \boldsymbol{\lambda}_j^T(\tau_i^-) - \boldsymbol{\lambda}_j^T(\tau_i^+) \} \delta \boldsymbol{x}(\tau_i) \\
&\quad - \boldsymbol{\lambda}_j^T(\tau_0^+) \delta \boldsymbol{x}(\tau_0) + \boldsymbol{\lambda}_j(\tau_{k+1}^-) \delta \boldsymbol{x}(\tau_{k+1}) \\
&= \sum_{i=1}^{k} \{ \boldsymbol{\lambda}_j^T(\tau_i^-) - \boldsymbol{\lambda}_j^T(\tau_i^+) \} \delta \boldsymbol{x}(\tau_i) \\
&\quad - \boldsymbol{\lambda}_j^T(0^+) \delta \boldsymbol{x}(0) + \boldsymbol{\lambda}_j(T^-) \delta \boldsymbol{x}(T) \\
&= \sum_{i=1}^{k} \{ \boldsymbol{\lambda}_j^T(\tau_i^-) - \boldsymbol{\lambda}_j^T(\tau_i^+) \} \delta \boldsymbol{x}(\tau_i) \\
&\quad + \boldsymbol{\lambda}_j^T(T^-) \delta \boldsymbol{x}(T) - \boldsymbol{\lambda}_j^T(0^+) \delta \boldsymbol{x}(0)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\delta\, g_j &= \sum_{i=1}^{k}\left\{\frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{x}} - \boldsymbol{\lambda}_j^T(\tau_i^-) + \boldsymbol{\lambda}_j^T(\tau_i^+)\right\}\delta\boldsymbol{x}(\tau_i) \\
&\quad -\boldsymbol{\lambda}_j^T(T)\delta\boldsymbol{x}(T) + \boldsymbol{\lambda}_j^T(0^+)\delta\boldsymbol{x}(0^+) + \sum_{i=1}^{k}\frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{z}}\delta\boldsymbol{z} \\
&\quad \sum_{i=1}^{k+1}\int_{\tau_{i-1}}^{\tau_i}\left\{(\dot{\boldsymbol{\lambda}}_j^T + \boldsymbol{\lambda}_j^T\frac{\partial f}{\partial \boldsymbol{x}})\delta\boldsymbol{x} + \boldsymbol{\lambda}_j^T\frac{\partial f}{\partial \boldsymbol{z}}\delta\boldsymbol{z}\right\}dt
\end{aligned}$$

From the definition of the costate system (5.4.5) – (5.4.7), all terms involving variation in $\boldsymbol{x}$ vanish, yielding

$$\begin{aligned}
\nabla_{\boldsymbol{z}}\, g_j(\boldsymbol{u},\boldsymbol{z}) &= \sum_{i=1}^{k+1}\int_{\tau_{i-1}}^{\tau_i}\boldsymbol{\lambda}_j^T\frac{\partial f}{\partial \boldsymbol{z}}dt + \boldsymbol{\lambda}_j^T(0^+)\frac{\partial \boldsymbol{x}(0^+)}{\partial \boldsymbol{z}} + \sum_{i=1}^{k}\frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{z}} \\
&= \int_0^T \boldsymbol{\lambda}_j^T\frac{\partial f}{\partial \boldsymbol{z}}dt + \boldsymbol{\lambda}^T(0)\frac{\partial \boldsymbol{x}^0(\boldsymbol{z})}{\partial \boldsymbol{z}} + \sum_{i=1}^{k}\frac{\partial \Phi_{j,i}(\boldsymbol{x}(\tau_i|\boldsymbol{u},\boldsymbol{z}),\boldsymbol{z})}{\partial \boldsymbol{z}}
\end{aligned}$$

The proof is complete. □

5.5 Numerical Experiments

Example 5.1. Let x_0 be a function of time given by

$$x_0(t) = \sin(4t) + 2t.$$

Consider a process described by the following differential equations

$$\dot{x_1} = x_2 \tag{5.5.1}$$

$$\dot{x_2} = -u_1(t)x_2 - x_1 + u_2(t) \tag{5.5.2}$$

with initial conditions

$$x_1(0) = 0.1 \tag{5.5.3}$$

$$x_2(0) = 0.2 \tag{5.5.4}$$

Suppose $\underline{\tau_k}$, $\overline{\tau_k}$, $i = 1, \cdots, 5$, are given by $\underline{\tau_i} = \frac{5i-2}{30}T$, $\overline{\tau_i} = \frac{5i+2}{30}T$, $i = 1, \cdots, 5$ and $T = 3$.

Our objective is to find observation times t_i, $i = 1, \cdots, k$, and the controls $u_1(t)$ and $u_2(t)$ such that the cost function

$$\sum_{i=1}^{5}\{(\sin 4t_i + 2t_i - x_0(t_i))^2\} \tag{5.5.5}$$

is minimized subject to the constraints:

$$\begin{aligned} &\underline{\tau_i} \leq t_i \leq \overline{\tau_i}, && i = 1, 2, \cdots, 5 \\ &-3.0 \leq u_1(t) \leq 4.0 \\ &-3.0 \leq u_2(t) \leq 11.5 \end{aligned}$$

Define the CPET transform which maps t to s:

$$\frac{dt}{ds} = \boldsymbol{v}(s)$$

Let

$$\boldsymbol{v}(s) = \sum_{i=1}^{6} v_i \chi_{[i-1,i)}(s)$$

where v_i, $i = 1, \cdots, 6$, are collectively referred to as the parameter vector $\boldsymbol{v}$. The equivalent transformed problem is:
given the dynamical system

$$\begin{aligned} \dot{y}_1 &= \boldsymbol{v} y_2 \\ \dot{y}_2 &= \boldsymbol{v}(-w_1(s)y_2 - y_1 + w_2(s)) \end{aligned}$$

with initial condition

$$\begin{aligned} y_1(0) &= 0.1 \\ y_2(0) &= 0.2 \end{aligned}$$

find a parameter vector $\boldsymbol{v}$ and control functions $w_1(s) = u_1(t(s))$, $w_2(s) = u_2(t(s))$ such that the cost function

$$\sum_{i=1}^{5}\{(\sin 4\sum_{j=1}^{i} v_j + 2\sum_{i=1}^{i} v_j - \hat{x}_0(i))^2\}$$

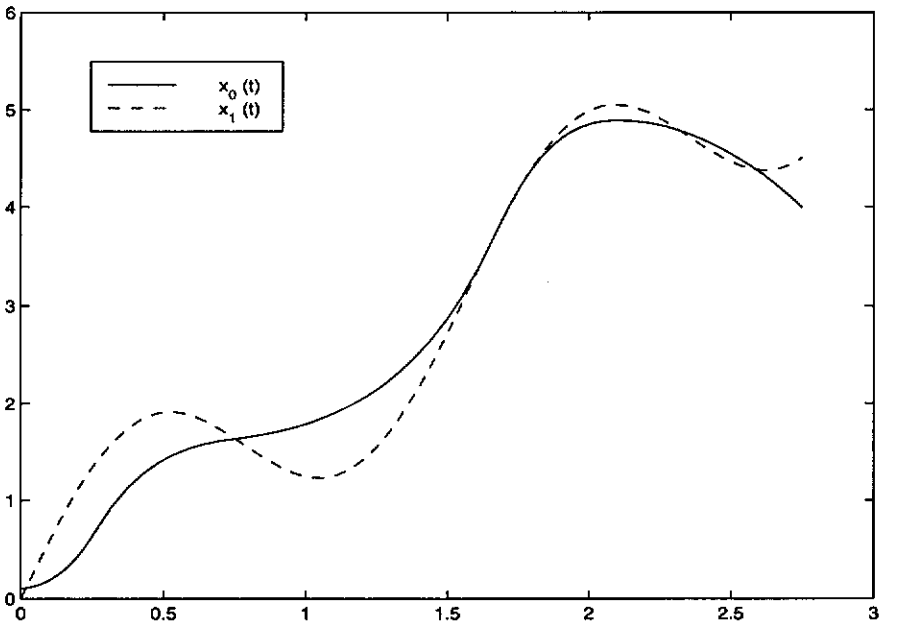


Figure 5.5.1: The target trajectory $x_0(t)$ and the optimal trajectory $x_1^*(t)$.

is minimized subject to

$$\underline{\tau_j} \leq \sum_{i=1}^{j} v_i \leq \overline{\tau_j} \qquad j = 1, 2, \cdots, 5,$$

$$\sum_{i=1}^{6} v_i = 3$$

$$-3.0 \leq w_1(s) \leq 4.0 \qquad \forall s \in [0, 6)$$

$$-3.0 \leq w_2(s) \leq 11.5 \qquad \forall s \in [0, 6)$$

Using the CPET transform again, which maps from s to q, with pre-fixed knots $\xi_0, \xi_1, \cdots, \xi_{10}$:

$$\frac{ds}{dq} = \eta(q)$$
$$s(0) = 0$$

where

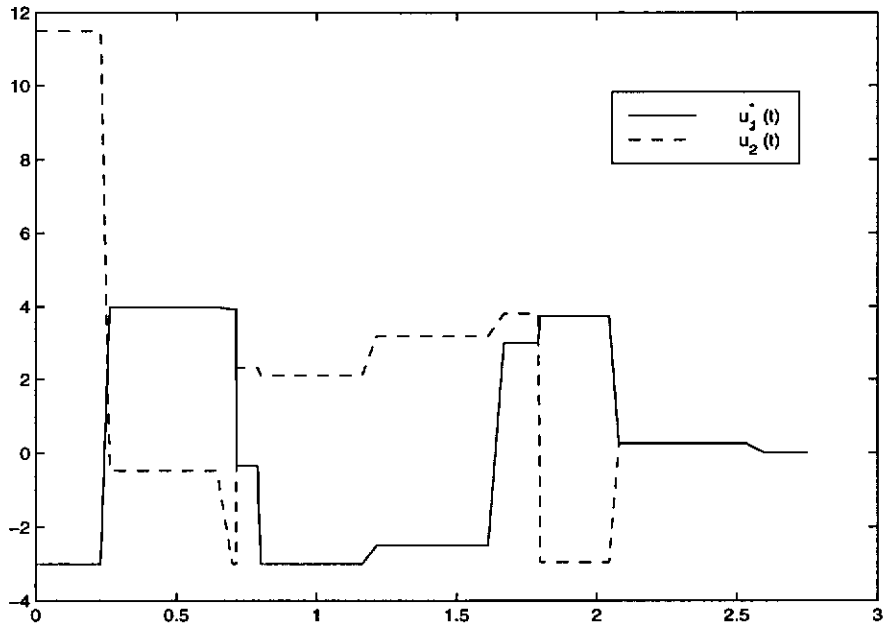


Figure 5.5.2: The optimal control $(u_1^*(t),\ u_2^*(t))$

$$\eta(q) = \sum_{i=1}^{10} \eta_i \chi_i(q)$$

$$\chi_i(q) = \begin{cases} 1 & q \in [\xi_{i-1}, \xi_i) \qquad i = 1, 2, \cdots, 10 \\ 0 & \text{otherwise} \end{cases}$$

Clearly

$$\int_0^i \eta(q)dq = i, \qquad i = 1, 2, \cdots, 6 \tag{5.5.6}$$

We thus obtain the optimal control problem with MCT : given the dynamical system

$$\dot{\hat{y}}_1 = \eta \upsilon y_2$$

$$\dot{\hat{y}}_2 = \eta \upsilon(-\sigma^1 \hat{y}_2 - \hat{y}_1 + \sigma^2)$$

with initial condition

$$\hat{y}_1(0) = 0.1$$

$$\hat{y}_2(0) = 0.2$$

find a $(\boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{\eta})$ such that the cost function

$$\sum_{i=1}^{5}\{(\sin 4\sum_{j=1}^{i} v_j + 2\sum_{i=1}^{i} v_j - \hat{x}_0(i))^2\}$$

is minimized subject to

$$\underline{\tau_j} \leq \sum_{i=1}^{j} v_i \leq \overline{\tau_j} \qquad j = 1, 2, \cdots, 5,$$

$$\sum_{i=1}^{6} v_i = 3$$

$$-3.0 \leq \sigma_{1,j} \leq 4.0 \qquad j = 1, \cdots, 11,$$

$$-3.0 \leq \sigma_{2,j} \leq 11.5 \qquad j = 1, \cdots, 11.$$

$$\int_0^i \eta(q)dq = i, \qquad i = 1, \cdots, 6.$$

This is an optimal control problem with MCT cost function. Using the gradient formulae obtained in Section 5.4, the optimal control software package MISER 3.2 can be adapted to solve this optimal control problem. Figure 5.5.1 shows the optimal trajectory of $x_1^*(t)$ and the trajectory for $x_0(t) = \sin(4t) + 2t$. The optimal observation times are $t_1 = 0.7, t_2 = 0.8, t_3 = 1.67, t_4 = 1.81, t_5 = 2.59, t_6 = 2.75$ with minimum cost function value of 0.027889, Figure 5.5.2 shows the optimal control functions $u_1^*(t)$ and $u_2^*(t)$.

Example 5.2. A three dimensional optimal control problem with variable characteristic times in the cost function.

Consider the dynamical system

$$\dot{x_1} = x_3 + z_1 \tag{5.5.7}$$

$$\dot{x_2} = z_2 x_3 + z_3 \tag{5.5.8}$$

$$\dot{x_3} = z_4 x_1 + x_2 + z_5 x_3 \tag{5.5.9}$$

with initial condition

$$x_1(0) = 0.1 \tag{5.5.10}$$

$$x_2(0) = 0.2 \tag{5.5.11}$$

$$x_3(0) = 0.1 \tag{5.5.12}$$

Let the target trajectory be specified as follows:

$$x_{01}(t) = \sin(4t) + 2t$$

$$x_{02}(t) = t^3 - 3t^2 + 3t$$

Let $\underline{\tau_i} = \frac{5i-2}{30}T$, $\overline{\tau_i} = \frac{5i+2}{30}T$, $i = 1, \cdots, 5$, and $T = 3$. Then, we formulate the following problem:

given the dynamic system (5.5.7)–(5.5.12), find observation times t_i , $i = 1, \cdots, 5$, and system parameters $z_i, i = 1, \cdots, 5$, such that the cost function

$$\sum_{i=1}^{5}\{(\sin 4t_i + 2t_i - x_{01}(t_i))^2 + ({t_i}^3 - 3{t_i}^2 + 3t_i - x_{02}(t_i))^2\} \tag{5.5.13}$$

is minimized subject to the constraints:

$$\underline{\tau_i} \le t_i \le \overline{\tau_i} \qquad i = 1, 2, \cdots, 5$$

$$-1.0 \le z_1 \le 2.0$$

$$-1.0 \le z_2 \le 10.0$$

$$-1.0 \le z_3 \le 1.0$$

$$-1.0 \le z_4 \le 1.0$$

$$-4.0 \le z_5 \le 1.0$$

Use the CPET transform to map t to s

$$\frac{dt}{ds} = v(s) :$$

Let

$$v(s) = \sum_{i=1}^{6} \sigma_i \chi_{[i-1,i)}(s)$$

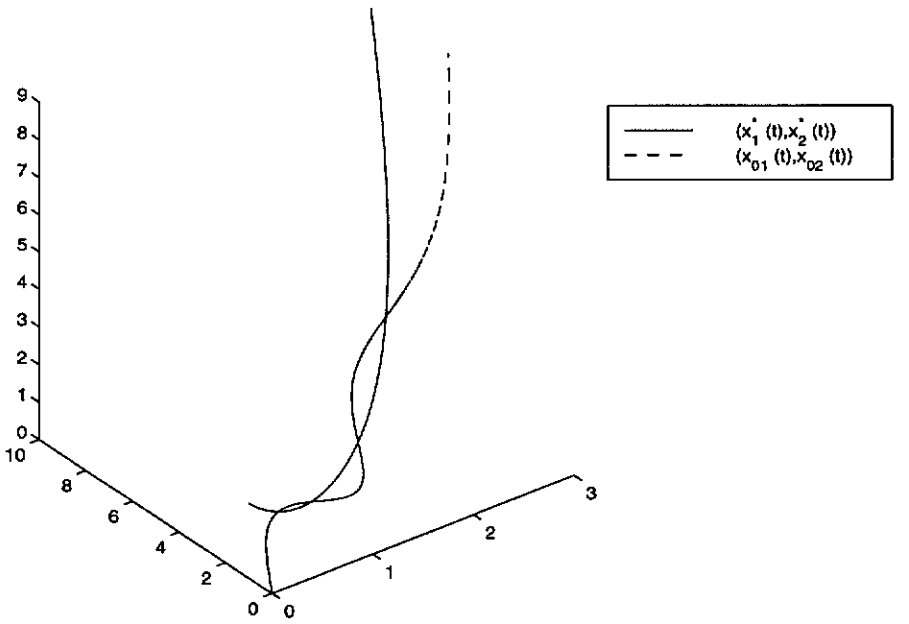
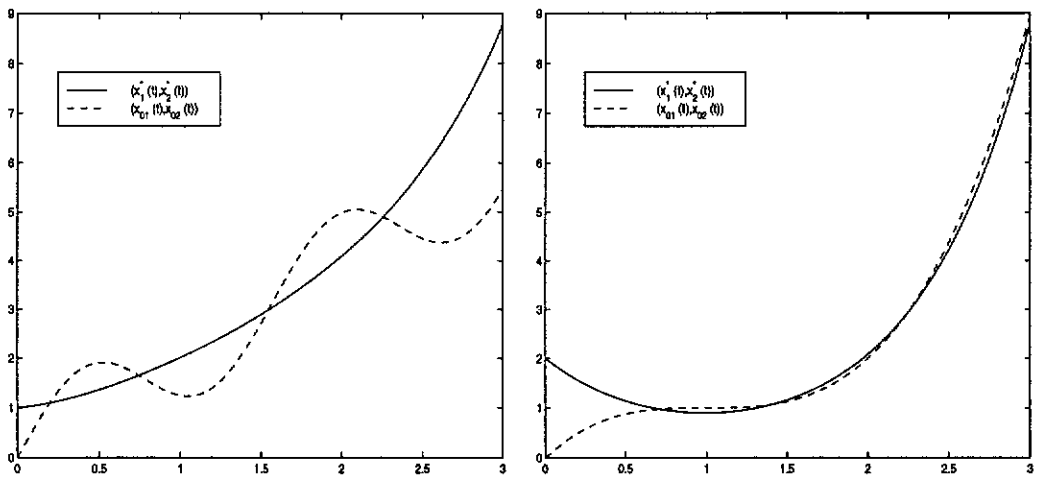


Figure 5.5.3: The optimal trajectory $(x_1^*(t), x_2^*(t))$ and target trajectory $(x_{01}(t), x_{02}(t))$.



(a) On the plane $\boldsymbol{X}_1 \times [0,T]$ (b) On the plane $\boldsymbol{X}_2 \times [0,T]$

Figure 5.5.4: The projections of trajectories depicted in Figure 5.5.3 onto the planes

where σ_i, $i = 1, \cdots, 6$, are collectively referred to as the parameter vector $\boldsymbol{\sigma}$. Furthermore, let z_i, $i = 1, \cdots, 5$, be collectively referred to as the system parameter $\boldsymbol{z}$. The equivalent transformed problem is:
given the dynamical system

$$\begin{aligned} \dot{y_1} &= \boldsymbol{v}(y_3 + z_1) & y_1(0) &= 0.1 \\ \dot{y_2} &= \boldsymbol{v}(z_2 y_3 + z_3) & y_2(0) &= 0.2 \\ \dot{y_3} &= \boldsymbol{v}(z_4 y_1 + y_2 + z_5 y_3) & y_3(0) &= 0.1 \end{aligned}$$

find a $(\boldsymbol{v}\ \boldsymbol{z})$ such that

$$\sum_{i=1}^{5} \Big\{ (\sin 4 \sum_{j=1}^{i} v_j + 2 \sum_{i=1}^{i} v_j - \hat{x}_{01}(i))^2 + ((\sum_{j=1}^{i}) v_j)^3 - 3(\sum_{j=1}^{i} v_j)^2 + 3 \sum_{j=1}^{i} v_j - \hat{x}_{02}(t_i))^2 \Big\}$$

is minimized, subject to

$$\begin{aligned} & \underline{\tau_j} \leq \sum_{i=1}^{j} v_i \leq \overline{\tau_j} && j = 1, 2, \cdots, 5 \\ & \sum_{i=1}^{6} v_i = 3 \\ & -1.0 \leq z_1 \leq 2.0 \\ & -1.0 \leq z_2 \leq 10.0 \\ & -1.0 \leq z_3 \leq 1.0 \\ & -1.0 \leq z_4 \leq 1.0 \\ & -4.0 \leq z_5 \leq 1.0 \end{aligned}$$

The gradient formulae obtained in Section 5.4 are applicable. Thus, MISER 3.2 can be adapted to solve this optimal control problem with MCT cost function. Figure 5.5.3 shows the optimal trajectory of $(x_1^*(t), x_2^*(t))$ and the trajectory for $x_{01}(t) = \sin(4t) + 2t$ and $x_{02}(t) = t^3 - 3t^2 + 3t$, and Figure 5.5.4 shows the projections of Figure 5.5.3 onto the planes $X_1 \times [0, T]$ and $X_2 \times [0, T]$. The optimal observation times are $t_1 = 0.7, t_2 = 0.8, t_3 = 1.544678, t_4 = 2.2, t_5 = 2.3$

with the minimum cost function value of 0.1915, and the optimal system parameters are $z_1 = 1.32478, z_2 = 2.13784, z_3 = -0.43432, z_4 = -0.00573633, z_5 = 0.073166$.

5.6 Conclusion

In this chapter, a computational method was obtained for solving the optimal control problem with multiple variables time points in objective function. The method is based on the combination of the control enhancing transform and the control parameterization technique. The method is efficient and supported by rigorous mathematical analysis. Two numerical examples are solved using the method.

Chapter 6

Solving a Class of Nonlinear Optimal Feedback Control Problems Using 3rd Order B-Splines with Optimal Partition Points

6.1 Introduction

In this chapter, we propose an approach to the approximation of an optimal feedback control law for a general nonlinear optimal control problem. In this approach, a control function is approximated by a linear combination of a 3rd order B-spline basis function constructed on a partition in state and time spaces. This approximation is then used to substitute the control to form an approximate feedback control problem. In this approximate optimal feedback control problem, the mesh points and the coefficients in the linear combination of the 3rd order B-spline basis functions are all decision variables. Once the initial condition is given, the resulting optimal control problem becomes an optimal parameter selection problem. The optimal control obtained is represented by a linear combination of the 3rd order B-splines with optimal coefficients and optimal partition points. This control is a function of the state and time. Thus, it is more robust in the presence of the system noise and parameter variation. The rest of the Chapter is organized as follows:

In Section 6.2, we state the optimal feedback control problem. The approximate optimal feedback control problem is discussed in Section 6.3. In Section 6.4, some examples are solved using the proposed approach.

6.2 The optimal feedback control problem

Consider a process described by the following system of differential equations defined on [0,T].

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{u}(\boldsymbol{x}, t), \boldsymbol{z}(t)), \tag{6.2.1}$$

$$\boldsymbol{x}(0) = \xi_0 \tag{6.2.2}$$

where T is a fixed terminal time, $\boldsymbol{x} = [x_1, \cdots, x_n]^T \in \mathbb{R}^n$, $\boldsymbol{u} = [u_1, \cdots, u_m]^T \in \mathbb{R}^m$ and $\boldsymbol{z} = [z_1, \cdots, z_p]^T \in \mathbb{R}^p$ are , respectively, state, control, and system parameter, while $\boldsymbol{f} = [f_1, \cdots, f_n]^T \in \mathbb{R}^n$ is a continuously differentiable function with respect to $\boldsymbol{x}$, $\boldsymbol{u}$ and $\boldsymbol{z}$ and piecewise continuous with respect to t, and ξ_0 is assumed to be randomly distributed according to a fixed probability density distribution $\rho(\xi_0)$ defined on some compact set $\Gamma \subset \mathbb{R}^n$. Let the compact sets $\boldsymbol{X} \subseteq \mathbb{R}^n$, $\boldsymbol{Z} \subseteq \mathbb{R}^p$ and $\boldsymbol{U} \subseteq \mathbb{R}^m$ be defined by

$$\begin{aligned}
\boldsymbol{X} &= \{\boldsymbol{x} = [x_1, \cdots, x_n]^T \in \mathbb{R}^n : x_{i,min} \leq x_i \leq x_{i,max}\} \\
\boldsymbol{Z} &= \{\boldsymbol{z} = [z_1, \cdots, z_r]^T \in \mathbb{R}^r : z_{i,min} \leq z_i \leq z_{i,max}\} \\
\boldsymbol{U} &= \{\boldsymbol{u} = [u_1, \cdots, u_m]^T \in \mathbb{R}^m : u_{i,min} \leq u_i \leq u_{i,max}\}
\end{aligned}$$

For a given $\boldsymbol{u}$ and $\boldsymbol{z}$, $\boldsymbol{x}(t)$ is determined by the (6.2.1) and (6.2.2). The control $\boldsymbol{u}(\boldsymbol{x}, t)$ in (6.2.1) is a feedback control obtained from the process of the system and is added to the process in order to improve the robustness in the presence of the noise and parameter variation. Any function $\boldsymbol{u} : \boldsymbol{X} \times [0, T] \longrightarrow \boldsymbol{U}$ is called an admissible control if it is continuously differentiable with respect to $\boldsymbol{x}$ and piecewise continuous with respect to t. Let $\mathcal{U}$ be the class of all such admissible controls.

Our optimal feedback control problem may now be formally stated as: given the dynamic system (6.2.1) –(6.2.2), find a $(\boldsymbol{u}, \boldsymbol{z}) \in \mathcal{U} \times \boldsymbol{Z}$ such that the cost function

$$G_0(\xi_o, \boldsymbol{u}, \boldsymbol{z}) = \phi_0(\boldsymbol{x}(T)) + \int_0^T g_0(t, \boldsymbol{x}(t), \boldsymbol{u}(x(t)), \boldsymbol{z})dt \tag{6.2.3}$$

is minimized subject to the constraints

$$g_j(\boldsymbol{u}(\cdot), \boldsymbol{z}) \leq 0, \quad j = 1, \cdots, q, \tag{6.2.4}$$

where $\phi_0(\boldsymbol{x})$ and g_j, $j = 0, 1, \cdots, q$, are continuously differentiable real valued functions on $\mathbb{R}^n$ and $[0, T] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$, respectively. Let this optimal control problem be referred to as Problem P_6.

6.3 Approximation of the optimal feedback control

We now consider the approximation of the control by a linear combination of the 3rd order spline basis functions. For simplicity, we restrict our consideration to the case in which the control function $\boldsymbol{u}$ in Problem P_6 is independent of time, i.e., $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$. Furthermore, for brevity, we also assume that n = 2 and m = 1, *i.e.*, the state is 2 dimensional and the control is a scalar. Now, the state region is chosen to be $\boldsymbol{X} = [c_1, d_1] \times [c_2, d_2]$ with constants c_1, c_2, d_1 and d_2 satisfying $c_1 \leq d_1$ and $c_2 \leq d_2$.

For $i = 1, 2$, let $\boldsymbol{v}_i := \{x_{i,j}\}_{j=0}^{N_i}$ be a set of partition points satisfying $c_i = x_{i,0} \leq x_{i,1} \leq \cdots \leq x_{i,N_i}\ x_{i,N_i+1} = d_i$. The 3rd order B-spline basis function (see. [47]), $B_{i,j}(x)$, associated with $x_{i,j}$, $j = 1, 2, \cdots, N_i$, $i = 1, 2$, is defined by

$$B_{i,j}(x) = \begin{cases} \frac{(x - x_{i,j-1})^2}{h_{i,j-1}(h_{i,j-1} + h_{i,j})} & x \in [x_{i,j-1}, x_{i,j}) \\ 1 - \frac{(x - x_{i,j})^2}{h_{i,j}(h_{i,j} + h_{i,j+1})} - \frac{(x_{i,j+1} - x)^2}{h_{i,j}(h_{i,j-1} + h_{i,j})} & x \in [x_{i,j}, x_{i,j+1}) \\ \frac{(x_{i,j+2} - x)^2}{h_{i,j+1}(h_{i,j} + h_{i,j+1})} & x \in [x_{i,j+1}, x_{i,j+2}] \\ 0 & \text{otherwise} \end{cases}$$

where $h_{i,j} = x_{i,j+1} - x_{i,j}$. These basis functions are globally smooth. Now, we approximate the control $u(x)$ by a linear combination of the tensor products of the basis functions, i.e.,

$$u \approx u_b(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{b}) = \sum_{i=-1}^{N_1+1} \sum_{j=-1}^{N_2+1} b_{i,j} B_{1,i}(x_1) B_{2,j}(x_2) \tag{6.3.1}$$

where $\boldsymbol{b}$ denotes the coefficient vector $\{b_{i,j}\}$. Note that in the above expression we have introduced a few auxiliary mesh points $x_{i,-2}$, $x_{i,-1}$, x_{i,N_i+2}, and x_{i,N_i+3}, $i = 1, 2$. (Mesh node $x_{i,-2}$ does not appear in the above explicitly, but it is needed for the construction of $\boldsymbol{B}_{i,-1}$.) With this set of functions, the above approximation preserves constants and first and second order polynomials within the range $[c_1, d_1] \times [c_2, d_2]$. Replacing u in Problem P_6 by u_b given above, we have the following approximate optimal feedback control problem: given N_1, and N_2, find $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{b}$ and the system parameter $\boldsymbol{z}$, such that

$$G_0(\xi_o, u_b, \boldsymbol{z}) = \phi_0(\boldsymbol{x}(T)) + \int_0^T g_0(t, \boldsymbol{x}(t), u_b(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{b}), \boldsymbol{z}) dt \tag{6.3.2}$$

is minimized subject to

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), u_b(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{b}), \boldsymbol{z}(t)), \tag{6.3.3}$$

$$\boldsymbol{x}(0) = \xi_0 \tag{6.3.4}$$

and

$$g_j(u_b, , \boldsymbol{z}) \leq 0, \quad j = 1, \cdots, q; \quad i = 1, \cdots, k \tag{6.3.5}$$

$$x_{i,0} = c_i, \qquad x_{i,N_i} = b_i, \tag{6.3.6}$$

$$x_{i,j} < x_{i,j+1}, \quad j = 1, 2, \cdots, N_{j-1}, \; i = 1, 2 \tag{6.3.7}$$

Let this optimal control problem be referred to as Problem P_6^*. This is an approximate optimal feedback control problem using 3rd order B-splines with variable partition points as controllers. The gradient of the function g_0 is given by

$$
\begin{aligned}
\frac{\partial g_0}{\partial b_{k,l}} &= \frac{\partial g_0}{\partial B}\frac{\partial B}{\partial b_{k,l}} = \frac{\partial g_0}{\partial B} B_{1,k}(x_1) B_{2,l}(x_2) \qquad \text{and,} \\
\frac{\partial g_0}{\partial x_{k,l}} &= \frac{\partial g_0}{\partial B}\frac{\partial B}{\partial x_{k,l}} \\
&= \frac{\partial g_0}{\partial B} \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} b_{i,j}\Big\{ \frac{\partial B_{1,i}(x_1)}{x_{k.l}} B_{2,j}(x_2) + B_{1,i}(x_1)\frac{\partial B_{2,j}(x_2)}{x_{k,l}} \Big\}
\end{aligned}
$$

where

$$
\frac{\partial B_{i,j}(x)}{\partial x_{i,j-1}} = \begin{cases}
\frac{-2(x-x_{i,j-1})}{h_{i,j-1}(h_{i,j-1}+h_{i,j})} + \frac{(x-x_{i,j-1})^2}{h_{i,j-1}^2(h_{i,j-1}+h_{i,j})} + \frac{(x-x_{i,j-1})^2}{h_{i,j-1}(h_{i,j-1}+h_{i,j})^2}, & x \in [x_{i,j-1}, x_{i,j}) \\
-\frac{(x_{i,j+1}-x)^2}{h_{i,j}(h_{i,j-1}+h_{i,j})^2}, & x \in [x_{i,j}, x_{i,j+1}) \\
0, & x \in [x_{i,j+1}, x_{i,j+2}] \\
0, \qquad \text{otherwise}
\end{cases}
$$

$$
\frac{\partial B_{i,j}(x)}{\partial x_{i,j}} = \begin{cases}
-1\frac{(x-x_{i,j-1})^2}{h_{i,j-1}^2(h_{i,j-1}+h_{i,j})}, & x \in [x_{i,j-1}, x_{i,j}) \\
2\frac{(x-x_{i,j})}{h_{i,j}(h_{i,j}+h_{i,j+1})} - \frac{(x-x_{i,j})^2}{h_{i,j}^2(h_{i,j}+h_{i,j+1})} - \frac{(x-x_{i,j})^2}{h_{i,j}(h_{i,j}+h_{i,j+1})^2} \\
-\frac{(x_{i,j+1}-x)^2}{h_{i,j}^2(h_{i,j-1}+h_{i,j})}, & x \in [x_{i,j}, x_{i,j+1}) \\
\frac{(x_{i,j+2}-x)^2}{h_{i,j+1}(h_{i,j}+h_{i,j+1})^2}, & x \in [x_{i,j+1}, x_{i,j+2}] \\
0, \qquad \text{otherwise}
\end{cases}
$$

$$
\frac{\partial B_{i,j}(x)}{\partial x_{i,j+1}} = \begin{cases}
-1\frac{(x-x_{i,j-1})^2}{h_{i,j-1}(h_{i,j-1}+h_{i,j})^2}, & x \in [x_{i,j-1}, x_{i,j}) \\
\frac{(x-x_{i,j})^2}{h_{i,j}^2(h_{i,j}+h_{i,j+1})} + 2\frac{(x_{i,j+1}-x)}{h_{i,j}(h_{i,j-1}+h_{i,j})} + \frac{(x_{i,j+1}-x)^2}{h_{i,j}^2(h_{i,j-1}+h_{i,j})} \\
+\frac{(x_{i,j+1}-x)^2}{h_{i,j}(h_{i,j-1}+h_{i,j})^2}, & x \in [x_{i,j}, x_{i,j+1}) \\
\frac{(x_{i,j+2}-x)^2}{h_{i,j+1}^2(h_{i,j}+h_{i,j+1})}, & x \in [x_{i,j+1}, x_{i,j+2}] \\
0, \qquad \text{otherwise}
\end{cases}
$$

$$
\frac{\partial B_{i,j}(x)}{\partial x_{i,j+2}} = \begin{cases}
0, & x \in [x_{i,j-1}, x_{i,j}) \\
\frac{(x-x_{i,j})^2}{h_{i,j}(h_{i,j}+h_{i,j+1})^2}, & x \in [x_{i,j}, x_{i,j+1}) \\
-2\frac{(x_{i,j+2}-x)}{h_{i,j+1}(h_{i,j}+h_{i,j+1})} - \frac{(x_{i,j+2}-x)^2}{h_{i,j+1}^2(h_{i,j}+h_{i,j+1})} - \frac{(x_{i,j+2}-x)^2}{h_{i,j+1}(h_{i,j}+h_{i,j+1})^2}, & x \in [x_{i,j+1}, x_{i,j+2}] \\
0, \qquad \text{otherwise}
\end{cases}
$$

with $h_{i,j} = x_{i,j} - x_{i,j-1}$ as defined above. Obviously, Problem P_6^* is an approximation of the problem (P_6). The accuracy of this approximation depends on the partition of the state region.

We comment that for a given initial value ξ_0^* of interest, the space region $\boldsymbol{X}$ is chosen properly so that the optimal trajectory $\boldsymbol{x}^*$ corresponding to ξ_0^* is contained in $\boldsymbol{X}$, *i.e.* $\boldsymbol{x}^*(t) \in \boldsymbol{X}$ for all $t \in (0, T]$. Then, Problem P_6^* becomes an

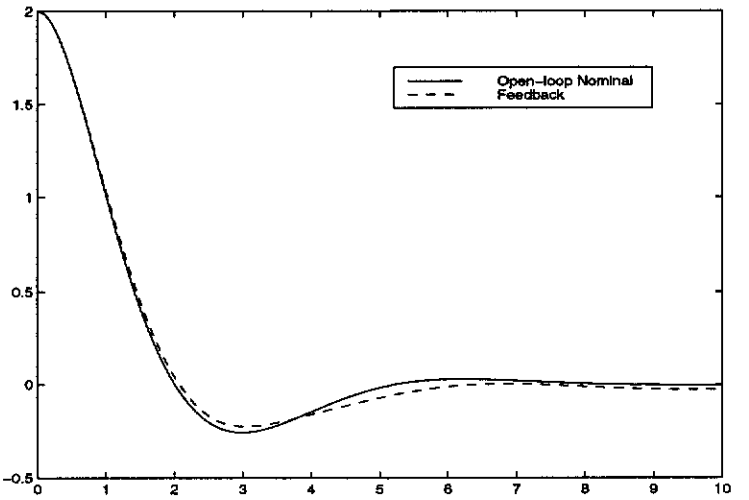


Figure 6.4.1: The trajectories of x_1(t)

optimal parameter selection problem, which can be viewed as a finite dimensional optimization problem. The solution to this optimal parameter selection problem gives rise to an optimal control defined at every point of $\boldsymbol{X}$. It is thus a feedback control defined on $\boldsymbol{X}$. The initial point ξ_0^* is sometimes referred to as an **operating point**.

For a large space region, we may choose a number of operating points distributed uniformly in the region, and then choose a sub-region associated with each of the operating points, so that the union of the sub-regions is equal to the original space region. Problem P_6^* can be solved for each operating point and the combination of all the controls obtained using these operating points will form a global feedback control.

6.4 Numerical Experiments

The numerical approach described in the previous section is applied to two and three dimensional state dependent examples. All computations were performed in Fortran double precision on a Unix workstation.

Example 6.1. The Duffing Equation

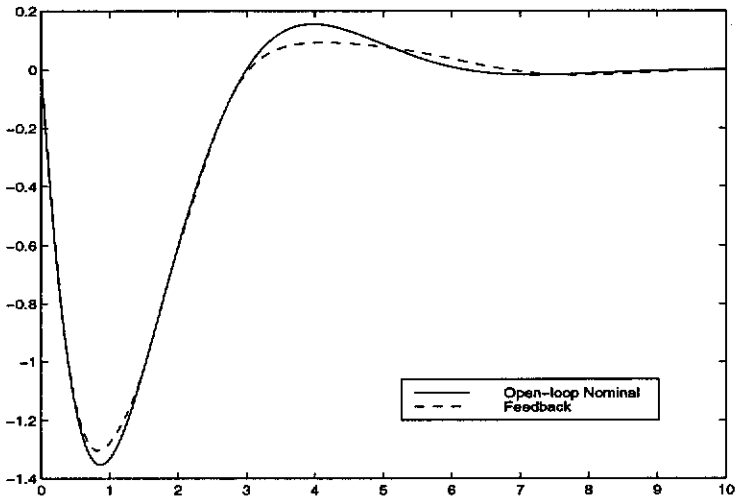


Figure 6.4.2: The trajectories of x_2(t)

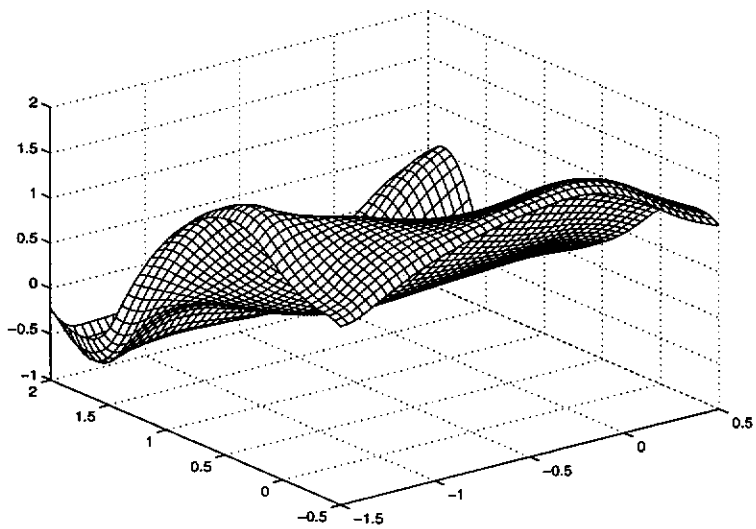


Figure 6.4.3: The optimal feedback control

Consider the following optimal control problem

$$\text{Minimize} \quad G_0(u) = x_1^2(10) + x_2^2(10) + \int_0^{10} (x_1^2 + x_2^2 + u^2)dt$$

subject to

$$\dot{x}_1 = x_2 \tag{6.4.1}$$

$$\dot{x}_2 = 0.2x_2 - x_1 - 0.1x_1^3 + u \tag{6.4.2}$$

with initial condition

$$x_1(0) = 2$$

$$x_2(0) = 0$$

The space region associated with this initial point is chosen to be $\boldsymbol{X} = [-0.5, 2] \times [-1.5, 0.5]$, and the control is approximated in this region $\boldsymbol{X}$ by (6.3.1) with 6 and 5 partition points along x_1 and x_2 axes, respectively. The corresponding optimal feedback control problem is essentially an optimal parameter selection problem. The solution of this problem using the previous method yields an approximate optimal feedback control on $\boldsymbol{X}$, as depicted in Figure 6.4.5.

For comparison, the minimum costs obtained from our method with both fixed and variable partition points are listed in Table 6.4.1. The optimal control problem (P_6) with the control $\boldsymbol{u}$ taken as only a function of time, t, is a conventional open - loop optimal control problem. Let this problem be referred to as problem $\tilde{P}_6$. The control obtained by solving $\tilde{P}_6$ with initial condition $(x_1(0), x_2(0)) = (2,0)$ is called the optimal control, and the corresponding trajectories and the cost are referred to the optimal trajectories and the optimal cost. The optimal cost is also listed in the Table 6.4.1. It is clear that the solution using variable partition points is much more accurate than that using fixed partition points. The trajectories corresponding to the feedback control with variable points and the optimal trajectories are plotted in Figures 6.4.1

	Approximate optimal feedback control with fixed partition points	Approximate optimal feedback control with variable partition points	Optimal control (obtained by solving of corresponding $\tilde{P}_6$)
Cost values	10.795	9.140	9.112
Relative error	0.1847	0.0030	

Table 6.4.1: Cost values and their relative errors

and 6.4.2. From these figures it is seen that the corresponding trajectories are very close to each other. However, the feedback control is defined on the whole space region $\boldsymbol{X}$, while the open-loop control is only a function of time.

To demonstrate the robustness of the computed approximate optimal feedback control depicted in Figure 6.4.3, we choose another initial condition $(x_1(0), x_2(0)) = (1, -0.1) \in \boldsymbol{X}$ and use this computed feedback control to solve the dynamic system (6.4.1) - (6.4.2) (without solving the original optimization problem). The computed trajectories for x_1 and x_2 are plotted in Figures 6.4.4 and 6.4.5, respectively. Furthermore, the problem $\tilde{P}_6$ with the new initial condition $(x_1(0), x_2(0))$ = (1,-0.1) is solved by using MISER3.2 to give the corresponding optimal open - loop control. The corresponding trajectories are used as nominal trajectories. These nominal trajectories for x_1 and x_2 are also depicted in Figures 6.4.3 and 6.4.4, respectively. From these figures, we see that the trajectories obtained using our method are very close to the reference trajectories. For further comparison, the costs evaluated using the feedback control solution and the true optimal cost, *i.e.* by solving the corresponding problem $\tilde{P}_6$ with the initial condition $(x_1(0), x_2(0))$ = (1,-0.1) are listed in Table 6.4.2. Again, the two results are very close each other with a relative error of around 3%.

Example 6.2. Maximum Rocket Height

In this example, we consider the vertical ascent of a rocket. The system dynamics are governed by:

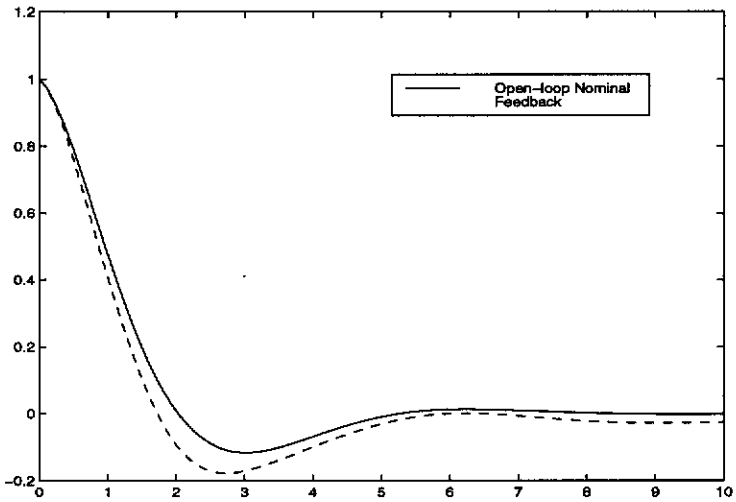


Figure 6.4.4: The trajectories of $x_1(t)$ after perturbation

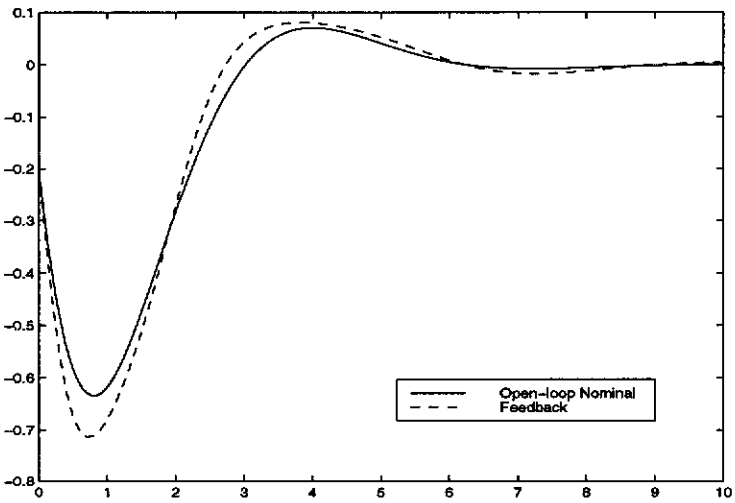


Figure 6.4.5: The trajectories of $x_2(t)$ after perturbation

	Approximate optimal feedback control	Optimal cost (obtained by solving the corresponding problem $\tilde{P}_6$ in that the change of initial value is taken into account)
Cost	2.140	2.073
Relative Error	0.0323	

Table 6.4.2: Robustness test of the approximate optimal feedback control

	Approximate optimal feedback control	Optimal cost (obtained by solving the corresponding problem $\tilde{P}_6$)
Cost	-28.25	-28.56
Relative Error	0.01080	

Table 6.4.3: The cost for applying the optimal feedback without perturbation

$$\dot{x}_1 = u \tag{6.4.3}$$

$$\dot{x}_2 = x_3 \tag{6.4.4}$$

$$\dot{x}_3 = -0.01 + \frac{2u - 0.05\exp(0.01x_2)x_3^2}{x_1} \tag{6.4.5}$$

with initial condition

$$\begin{aligned} x_1(0) &= 1.1 \\ x_2(0) &= 0 \\ x_3(0) &= 0 \end{aligned}$$

where $x_1(\cdot)$ is the rocket's mass , $x_2(\cdot)$ is the rocket's altitude in km above the earth's surface, $x_3(\cdot)$ is the rocket's vertical velocity and $u(\cdot)$ is the mass flow rate of the rocket's fuel. The terminal time T is 50, and the control is bounded by $0 \leq u(\cdot) \leq 0.04$. The objective of the problem is to minimize the cost function

$$G_0(u) = -x_2$$

subject to the constraint

$$0.2 - x_1 \leq 0.0.$$

The space region associated with this initial point is chosen to be $\boldsymbol{X} = [0.2, 1.1] \times [-5, 30] \times [0, 0.8]$, and the control is approximated in this region by (6.3.1) with 4, 10, and 4 partition points along x_1 ,x_2 and x_3 axes, respectively. The solution of this approximated optimal control problem yields an approximate optimal feedback control, defined on the region $\boldsymbol{X}$.

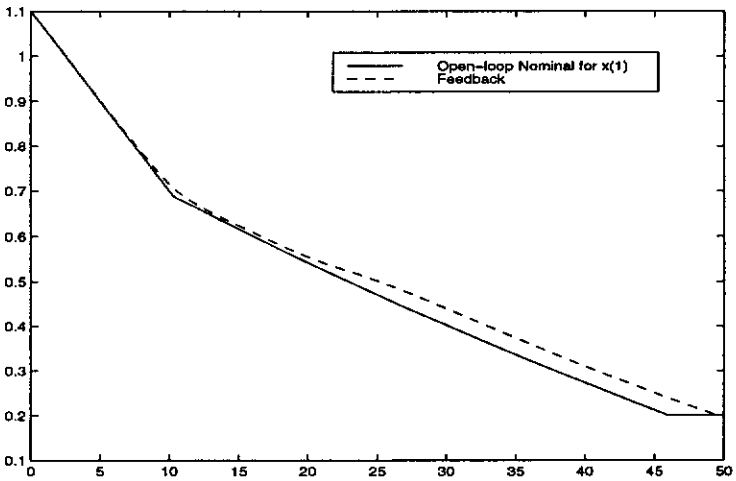


Figure 6.4.6: The trajectories of $x_1(t)$

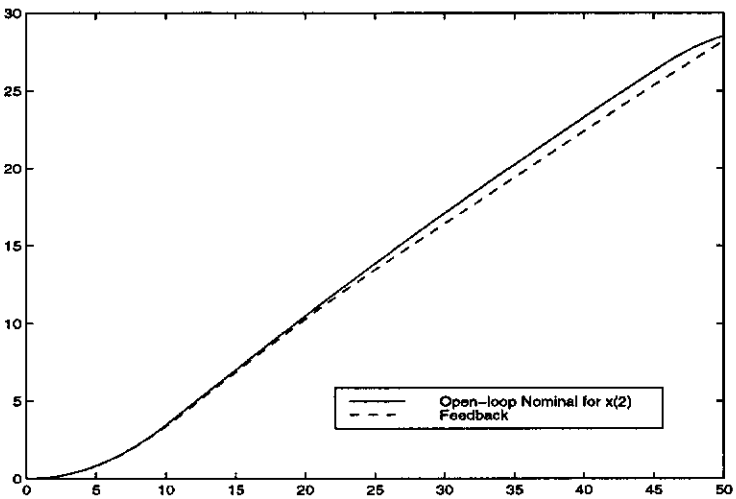


Figure 6.4.7: The trajectories of $x_2(t)$

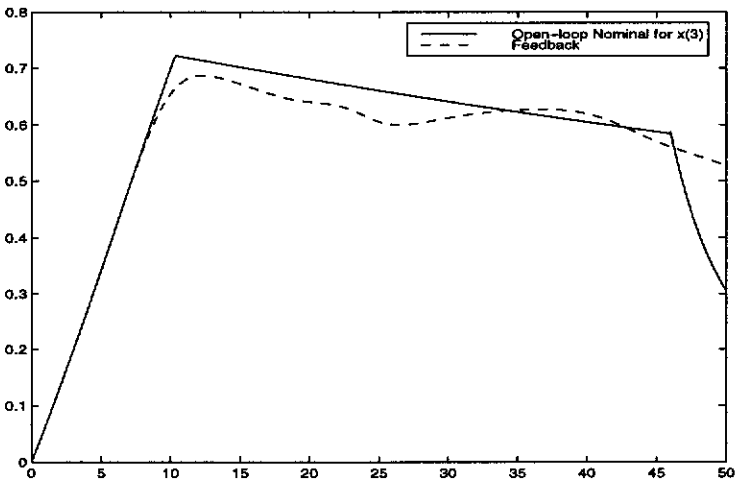


Figure 6.4.8: The trajectories of $x_3(t)$

	Approximate optimal feedback control	Optimal cost (obtained by solving the corresponding problem $\tilde{P}_6$ in that the change of the initial value is taken into account)
Cost	-27.89	-27.53
Relative Error	0.00228	

Table 6.4.4: The cost for applying the approximate optimal feedback after perturbation

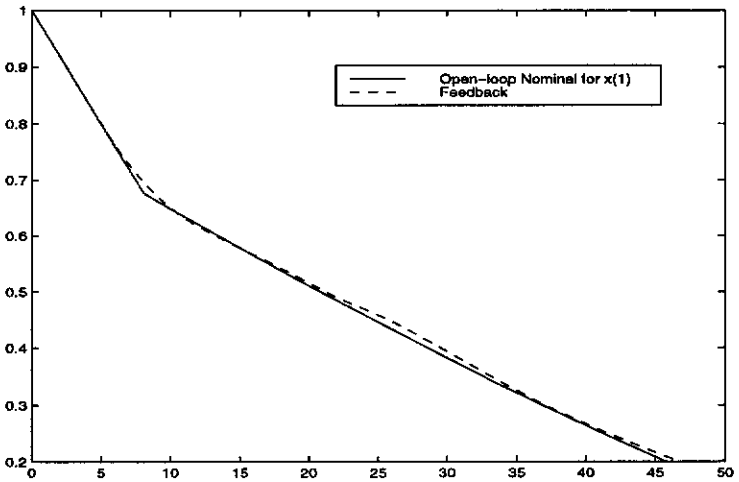


Figure 6.4.9: The trajectories of $x_1(t)$ after perturbation

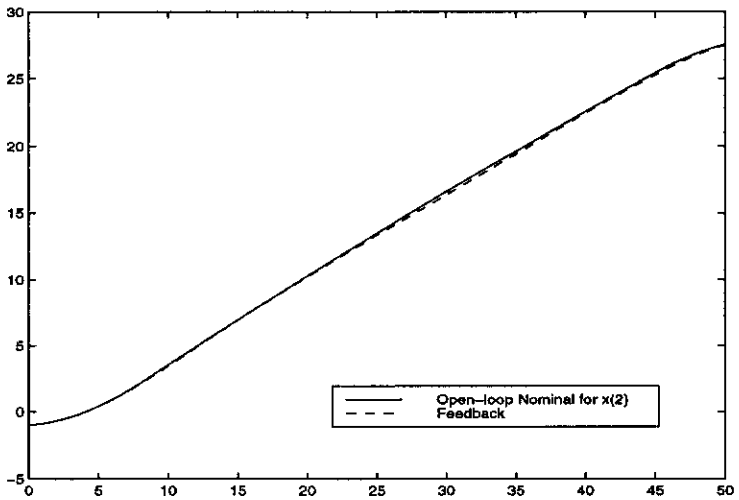


Figure 6.4.10: The trajectories of $x_2(t)$ after perturbation

The optimal control problem P_6 with the control u taken as only a function of time, t, is a conventional open - loop optimal control problem. Let this problem be referred to as problem $\tilde{P}_6$.

For comparison, problem $\tilde{P}_6$ with the initial condition $(x_1(0), x_2(0), x_3(0))$ = (1.1,0,0) is solved using MISER3.2. The trajectories for x_1, x_2 and x_3 obtained, and those obtained by using the approximate optimal feedback control are plotted in Figures 6.4.5 and 6.4.6. The cost obtained is to be referred to as the optimal cost. From these figures, it is seen that the corresponding trajectories are very close. The cost corresponding to the feedback control and the optimal cost are listed in Table 6.4.3. We note that the relative error is around 1%. However, the feedback control is defined on the whole region $\boldsymbol{X}$, while the optimal open - loop control is only a function of time, t.

To demonstrate the robustness of the computed approximate optimal feedback control, we choose a new initial condition $(x_1(0), x_2(0), x_3(0))$ = (1, -1, 0.1) $\in \boldsymbol{X}$ and use this computed feedback control to solve the dynamic system (6.4.3) - (6.4.5) (without solving the Problem 2 as an optimal parameter selection problem). The computed trajectories for x_1 and x_2 are plotted in Figures 6.4.9 and 6.4.10, respectively. For comparison, the corresponding problem $\tilde{P}_6$ with the new initial condition $(x_1(0), x_2(0), x_3(0))$ = (1, -1, 0.1) is solved as a conventional optimal control problem by using MISER3.2. The control obtained is an optimal control corresponding to this new initial condition, and hence is referred to as the optimal control. The corresponding optimal trajectories are plotted in Figures 6.4.9 and 6.4.10. From these figures we see that the trajectories corresponding to the optimal feedback control are still very close to the optimal open - loop trajectories. For further comparison, the cost corresponding to the optimal feedback control and the optimal open - loop cost are listed in Table 6.4.4. The relative error is around 0.23%.

Remark: When the initial values are changed, the constraint $2.0 - x_1 \leq 0$ may have some minor violation. In this case, we let the mass flow rate of fuel u turn to zero, when the state value of x_1 becomes 2.0. This is reasonable in practice,

since $x_1 = 2.0$ implies that the fuel runs out. In this case the mass flow rate of fuel will automatically turn to zero.

6.5 Conclusion

In this Chapter, the optimal feedback control for a general nonlinear optimal feedback control was approximated by a linear combination of the 3rd order B-spline basis functions in a neighborhood of an operating point. Both the coefficients and partition points of the B-splines are taken to be decision variables. Numerical experiments for two and three dimensional examples were performed, and the numerical results showed that the approximate optimal feedback control possesses a useful robustness property.

Chapter 7

Conclusion and Further Studies

7.1 Conclusion

In this thesis, we have developed efficient numerical methods for several real world problems arising from different areas. These real world problems are formulated as either static or dynamic optimization problems. For the problem of semiconductor device design, finding various parameters is time-consuming in the conventional design cycle, but in this thesis this problem is converted into a static optimization problem. Numerical examples illustrate that the methods used in this problem are efficient. For the problem of numerical integration, the variable partition points are introduced in numerical integration, and the problem is converted into a static optimization problem too, which produces much more accurate results than those from the uniform grid with the same number of partition points. As for the problem of planning recharge and driving strategies to minimize the total traveling time between two locations, the problem is formulated as an unconventional optimal control problem based on a battery-powered electric vehicle model, and furthermore the problem is transformed into a conventional optimal control problem. The solutions of the above problems demonstrate the usefulness of the approaches proposed in this thesis.

The optimization problem with variable time points in the objective function arises from many application areas. This is not a conventional optimal control problem. The problem is transformed into a conventional optimal control problem with multiple characteristic time by control parameterization en-

hancing transform (CPET). The gradients of the cost function and constraints are derived. The control parameterization with variable partition points combined with CPET are used again to convert the optimal control problem into an optimal parameter selection problem. Numerical experiments demonstrate the efficiency of this approach.

Feedback control is more suitable than open - loop control for many engineering applications, as a feedback control is dependent on both the state and time variables, while an open - loop control is a function of time only. Thus, feedback control is more robust in the presence of the system noise and parameter variation. However, to find an optimal feedback control law for a general non-linear problem is extremely difficult. In this thesis the optimal feedback control for a general nonlinear optimal control problem is approximated by a linear combination of the 3rd order B-spline basis functions constructed on a partition in state and time spaces. The partition points and coefficients of the B-spline are taken to be decision variables. Numerical experiments demonstrate the effectiveness of this method. Compared to the methods proposed in [33, 34, 49, 83], our method is simpler and much more efficient.

7.2 Further Studies

There are a number of difficult mathematical problems that remain. For the semiconductor device design, the problem is computable as seen in Chapter 2, but the solvability of the problem is theoretically difficult, since the dependence of the cost function on parameter θ and the applied bias v is complicated, and the solvability of the nonlinear PDE system (2.1.17)-(2.1.19) is still an open problem. The partial derivatives in the Jacobian for the Levenberg-Marquardt algorithm are approximated by forward finite differences in this thesis, but the exact formulas of these partial derivatives are still unknown and remain an open problem. Therefore, these problems can be further research topics.

The ideas of variable and optimal partition points are introduced in this thesis, and are applied to the problems in Chapters 3 - 6. As demonstrated

in these Chapters, the numerical results using the ideas improved on those without using the optimal partition points to a large extent. More specifically, to achieve a required accuracy our method needs much fewer partition points than those with fixed partition points. Thus, it has a much better chance of being applicable to high dimensional optimal control problems in which the optimal feedback control is taken as a function of both the time as well as the state variables. This is a challenging research direction. The results obtained will enhance the applicability of optimal control theory in solving practical engineering problems.

References

1. Alder R.B., Smith A.C., and Longini R.L., *Introduction to Semiconductor Physics*, Wiley, New York, (1964).

2. Allen D. N., and Southwell R. V., "Relaxation Methods Applied to determine the Motion, in Two Dimensions, of a Viscous Fluid past a Fixed Cylinder", *Quart. J. Mech. and Appl. Math.*, **VIII**, (1955) 129–145.

3. Anderson B.D.O., and Moore J.B., *Linear Optimal Control*, Prentice Hall, (1971).

4. Babuška I., and Rheinboldt W.C., "Analysis of optimal finite-element meshes in $\mathbb{R}^1$", *Math. Comp.*, **33** (1979) 435–463.

5. Bellman R., *Introduction to the Mathematical Theory of Control Processes*, **Vol. 1**, Academic Press, New York, (1971).

6. Bellman R. E., and Dryfus R. E., *Dynamic Programming and Modern Control Theory*, Academic Press, Orlando, Florida, (1977).

7. Biswas S.K., and Ahmed N.U., "Optimal Feedback Control of Power Systems Governed by Nonlinear Dynamics", *Optimal Control Applications & Methods*, **Vol. 7**, (1986) pp. 289–303.

8. Boggs P. T., Byrd R. H., and Schnabel R. B., "A Stable and Efficient Algorithm for Nonlinear Orthogonal Distance Regression", *SIAM J. Sci. Stat. Comput.* **Vol. 8** No. 6. (1987) 1052–1061.

9. Bonnans J. F., Panier E. R., Tits A. L., and Zhou J. L., " Avoiding the Maratos Effect by Means of a Nonmonotone Line search. II Inequality Constrained Problems –Feasible Iterates", *SIAM J. Numer. Anal.* **29** (1992) 1187–1202.

10. Bosarge W. E. Jr., and Johnson O. G., "Direct Method Approximation to the State Regulator Control Problem Using a Ritz–Trefftz Suboptimal Control", *IEEE Transactions on Automatic Control*, **Vol. AC–15** (1970) 627–631.

11. Bryson A. E. Jr., and Ho Y. C., *Applied Optimal Control*, Haisted Press, New York, (1969).

12. Bunch D.R. *et al* "Demand for clean fuelled vehicles in California: a discrete choice stated preference survey", *Conference on Transportation and Global Climate Change*, Pacific Grove California, (1991)

13. Carey G.F., and Dinh H.T., "Grading functions and mesh redistribution", *SIAM J. Num. Anal.*, **22** (1985) 1028–1040.

14. Clarke F.H., and Vinter R.B., "Optimal multiprocesses", *SIAM J. Control Optim.*, **Vol. 27** (1989) 1072–1091.

15. Clarke F.H., and Vinter R.B., "Applications of optimal multiprocesses", *SIAM J. Control Optim.*, **Vol. 27** (1989) 1048–1071.

16. Chen T.-F., Fix G.J., and Yang H.D., "Numerical studies of optimal grid construction", *Numer. Methods Partial Differential Equations*, **12** (1996) 191–206.

17. Ciarlel P.G., *The Finite element method for elliptic problems*, North - Holland, Amsterdam, (1978).

18. Craven B.D., *Control and Optimization*, Chapman & Hall, London. (1995).

19. Cullum J., "Finite–dimensional Approximations of State Constrained Continuous Optimal Problems ", *SIAM Journal on Control*, **Vol. 10**, (1972) 649–670.

20. Davis P.J., and Rabinowita P., *Methods of Numerical Integration*, Academic Press, New York, (1975)

21. de Doncker-Kapenga E., and Piessens R., "A bibliography on automatic integration" *J Comp. Appl. Math.*, **2** (1976) 273-280.

22. Delaunay B., "Sur la sphere vide", *Izv. Akad.Nauk.SSSR., Math.Nat.Sci.Div.*, **6** (1934) 793–800.

23. De M. A., " An Accurate Numerical Steady–State One–Dimensional Solution of the P–N junction", *Sol. St. Elect.*, **11** (1968) 33–58.

24. De M. A., " An Accurate Numerical Steady–State One–Dimensional Solution of the P–N junction under Arbitrary Transient Conditions", *Sol. St. Elect.*, **11** (1968) 1021–2053.

25. Dirichlet G.L., " Uber die Reduction der positive quadratischen Formen mit drei unbestimmten ganzen Zahlen", *J. Reine Angew. Math.*, **40** (1850) 209–227.

26. Dolezal Z., " On the Solution of Optimal Control Problems Involving Parameters and General Boundary Conditions", *Kybernetika.* **17** (1981) 71–81.

27. Dontchev A. L., "Error Estimates for a Discrete Approximation to Constrained Control Problems", *SIAM J. Num. Anal.* **Vol 18**, (1981) 500–514.

28. Engels H., *Numerical Quadrature and Cubature*, Academic Press, London, (1980).

29. Fitzsimons C.J., Miller J.J.H, Wang S. and Wu C.H., "Hexahedral Discretisations of the Stationary Semiconductor Device Equations", *Comp. Meth. Appl. Mech. Engrg.*, **84** (1990) 43–57.

30. Gilbarg D., and Trudinger N. S., *Elliptic Partial Differential Equations of Second Order*, **2nd ed.**, Berlin–Heidelberg–New York, Springer, (1984)

31. Gill P. E., Murray M. H., and Wright M. H., *Practical Optimization*, Academic Press, London, (1981).

32. Goh B. S., *Management and Analysis of Biological Populations*, Elsevier, Amsterdam, (1980).

33. Goh C. J., and Edwards N. J., "Feedback Control of Minimum-Time Optimal Control Problems Using Neural Networks", *Optimal Control Applications & Methods.* **Vol. 14** No. 1. (1993) Jan–Mar 1–16.

34. Goh C. J., and Edwards N. J., " Synthesis of Optimal Feedback Controller by Neural Networks", *INT. J. SYSTEMS SCI.*, **Vol. 25, No. 8,** (1994) 1235–1248.

35. Gonzalez S., and Miele A., "Sequential Gradient–Restoration Algorithm for Optimal Control Problems with General Boundary Conditions", *J. Optim. Theory Appl.*, **Vol 26**, (1978) 395–425.

36. Grippo L., Lampariello F., and Lucidi S., "A Nonmonotone Line Search Technique for Newton's Method", *SIAM J. Numer. Anal.* **23** (1986) 707–716.

37. Gummel H.K., "A self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculation", *IEEE Elec. Dev.*, **ED-11** (1964) 455-465.

38. Hager W. W., and Lanculescu G. D., "Dual Approximations in Optimal Control", *SIAM Journal on Control and Optimization*, **Vol 22**, (1984) 423–465.

39. Hasdorff L., *Gradient Optimization and Nonlinear Control*, John Wiley and Sons, (1976).

40. Jacobson D. H., and Lele M. M., "A Transformation Technique for Optimal Control Problems with a state Variable Inequality Constraint", *IEEE Transactions on Automatic Control*, **Vol AC –14**, (1969) 457–464.

41. Jennings L.S., Fisher M.E., Teo K.L., and Goh C.J., *MISER3 Optimal Control Software: Theory and User Manual,* EMCOSS Pty Ltd, (1990)

42. Jennings L.S., Fisher M.E., Teo K.L., and Goh C.J., *MISER3 - Version 2.0 (1997) Optimal Control Software: Theory and User Manual,* EMCOSS Pty Ltd, (1997)

43. Jennings L.S., Fisher M.E., Teo K.L., and Goh C.J., "MISER3: Solving Optimal Control Problems - An Update", *Advance Engineering Software,* **13:** (1991) 190-196.

44. Jenning L. S., and Teo K. L., "A Computational Algorithm for Functional Inequality Constrained Optimization Problems", *Automatica*, **26** (1990) 371–375.

45. Kurata M., *Numerical Analysis for Semiconductor Devices*, Lexington Books, D.C. Heath and Co., Lexington–Massachusetts–Toronto (1982).

46. Kwkanaak H., and Silvan R., *Linear Optimal Control System,* Wiley Interscience, (1972).

47. Lancaster P., and Šalkauskas K., *Curve and Surface Fitting: An Introduction*, Academic Press, London and Orlando, (1986).

48. Lawrence C. T., and Tits A. L., "Nonlinear Equality Constrains in Feasible Sequential Quadratic Programming", *Optimization Methods and Software* **6** (1996) 265–282.

49. Lee H. W. J., Teo K. L., and Rehbock V., "Sub-Optimal Local Feedback Control for a Class of Nonlinear Control Problems", *Dynamics of Continuous, Discrete and Impusive Systems* **1** (1995) 37–51.

50. Lee H.W.J., Teo K.L., Rehbock V., and Jennings L.S., "Control Parameterization Enhancing Technique for Optimal Discrete-Valued Control Problems", (*to appear*).

51. Levenberg K., "A method for the solution of certain nonlinear problems in least squares", *Quarterly of Applied Mathematics*, **2** (1944) 164–168.

52. Malanowski K., *Finite Difference Approximations to Constrained Optimal Control Problems*, Optimization and Optimal Control, Lecture Notes in Control and Information Sciences, Springer–Verlag, (1981) 243–254.

53. Marquardt D.W.,"An algorithm for least squares estimation of non-linear parameters", *SIAMJ. Appl. Math*, **11** (1963) 431–441.

54. Markowich P. A., Ringhofer Ch. A., Seberherr S., and Langer E., "A singularly pertubed boundary value problem modelling a semiconductor device", *Report 2388, MRC, Univ. of Wisconsin* (1982).

55. Markowich P.A., *The Stationary Semiconductor Device Equations*, Springer–Verlag, Vienna–New York (1986).

56. Martin R., and Teo K.L., *Optimal Control of Drug Administration in Cancer Chemotherapy* World Scientific, Singapore (1996).

57. Mehra R. K., and Davis R. E., "A Generalized Gradient Method for Optimal Control Problems with Inequality Constraints and Singular Ares ", *IEEE Transactions on Automatic Control*, **Vol. AC–17**, (1972) 69–78.

58. Miele A., "Recent Advances in Gradient Algorithms for Optimal Control Problems", *J. Optim. Theory Appl.*, **Vol. 17**, (1975) 361–430.

59. Miele A., Gradient Algorithms for Optimization of Dynamic Systems, in C.T. Leondes(ed), *Control and Dynamic Systems: Advances in Theory and Applications,* Academic Press, New York, **16,** (1980) 1-5, (1975).

60. Miele A., and Wang T., "An Elementary Proof of a Functional Analysis Result Having Interest for Minimax Optimal Control of Aeroassisted Orbital Transfer Vehicles", *Aero–Astronautics Report No. 182*, Rice University (1985).

61. Miele A., Wang T., and Basapur V. K., "Primal and Dual Formulations of Sequential Gradient–Restoration Algorithms for Trajectory Optimization Problems", *Acta Astronautica*, **Vol. 13**, (1986) 491–505.

62. Miele A., Wu A. K., and Liu C. T., "A Transformation Technique for Optimal Control with Partial Linear State Inequality Constraints", *J. Optim. Theory. Appl.*, **Vol 28**, (1979) 185–212.

63. Miller J. J. H., "Some Numerical Techniques for the Solution of Problems Related to Semiconductor Device", *Numerical Approximation of Partial Differential Equations*, (ed. Ortiz E.L.), North–Holland (1986) 69–82.

64. Miller J. J. H., O'Riordan E., and Shishkin G. I., *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, Singapore (1996).

65. Miller J.J.H., Wang S., "A Triangular Mixed Finite Element Method for the Stationary Semiconductor Device Equations", *RAIRO Modél. Math. Anal. Numér.*, **25** , No.4 (1991) 441–463.

66. Miller J.J.H., and Wang S., "An analysis of the Scharfetter-Gummel box method for the stationary semiconductor device equations", *RAIRO Modél. Math. Anal. Numér.*, **28** , No.2 (1994) 123–140.

67. Miller J.J.H., and Wang S., "An Exponentially Fitted Finite Volume Method for the Numerical Solution of 2d Incompressible Flow Problems", *J. Comp. Phys.*, **115**, No.1 (1994) 56–64.

68. Miller J.J.H., and Wang S., " A Tetrahedral Mixed Finite Element Method For The Stationary Semiconductor Continuity Equations", *SIAM J. Numer. Appl.*, **31**, No.1 (1994) 196–216.

69. Miller J.J.H., and Wang S., "A New Non–Conforming Petrov–Galerkin Finite Element Method with Triangular Elements fo a Singularly Pertubed Advection–Diffusion Problem", *IMA J. Numer. Anal.*, **14** (1994) 257–276.

70. Mock M. S., "On Equations Describing Steady–State carrier Distributions in Semiconductor Device", *Comm. Pure Appl. Math.*, **vol. XXV**, (1972) 781–792.

71. Mock M. S., *Analysis of Mathematical Models of Semiconductor Device*, Boole Press, Dublin (1983).

72. Murtagh B. A., and Saunders M. A., "A Projected Lagrangian Algorithm and its Implementation for Sparse Nonlinear Constraints", *Math. Prog. Study 16, Algorithms for Constrained Minimization of Smooth Nonlinear Functions*, (1982) 84–117.

73. Murtagh B. A., and Saunders M. A., "MINOS 5.4 user's guide " *Technical Report SOL 83-20R*, Systems Optimization Laboratory, Dept. of Operation Research, Stanford University (1995).

74. Narendra K. S., and Parthasarathy K., "Identification and Control of Dynamical Systems Using Neural Networks", *IEEE Trans. on Neural Networks,* **Vol. 1, No. 1,** (1990).

75. O'Connor L., "Energizing the batteries for electric cars, " *Mechanical Engineering,* (1993) 73-75.

76. Ortega J. M., and Rheinbolt W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York (1970).

77. Panier E. R., and Tits A. L., "On Combining Feasibility, Descent and Superlinear Convergence in Inequality Constrained Optimization", *Math. Programming* **59** (1993) 261–276.

78. Piessens R., de Doncker-Kapenga E., Überhuber C.W., and Kahaner D.K., *Quadpack*, Springer-Verlag, Berlin-Heidelberg-New York-Tokyo (1983).

79. Polak E., *Computatinal Methods in Optimization*, Academic Press, New York, (1971).

80. Polak E., and Mayne D. Q., "A Feasible Directions Algorithm for Optimal Control Problem with Control and Terminal Inequality Constraints", *IEEE Transactions on Automatic Control,* **AC-22,** (1971) 741–751.

81. Polak S. J., den Heijer C., Schilders W. H. A., and Markowich P., "Semiconductor Device Modelling From the Numerical Point of View", *Int. J. Numer. Meth. in Eng.*, **24** (1987) 763–838.

82. Press W.H., Teukolsky S.A., Vetterling W.T., and Flannery B.P., *Numerical recipes in FORTRAN. The art of scientific computing* **2nd ed** Cambridge Univ. Press. New York (1992).

83. Rehbock V., Teo K.L., and Jennings L. S., "Suboptimal Feedback Control for a Class of Nonlinear Systems Using Spline Interpolation", *Discrete and Continuous Dynamical Systems,* **Vol 1, No 2,** (1995).

84. Ritch P. S., "On the Numerical Solution of Optimal Control Problems with Multiple Inequality Constraints", Ph.D. Dissertation, University of New South Wales, Australia (1972).

85. Robinson S. M., " A Quadratically Convergent Algorithm for General Nonlinear Programming Problems", *Math. Prog. 3*, (1972) 145–156.

86. Sakawa Y., and Shindo Y., "On the Global Convergence of An Algorithm for Optimal Control,,, *IEEE Transactions on Automatic Control,* **AC-25,** (1980) 1149-1153.

87. Sakawa., "On Local Convergence of An Algorithm for Optimal Control", *Numerical Funct. Anal. Opt.*, **3** (1981) 301–319.

88. Scharfetter D. L., and Gummel H. K., "Large–Signal Analysis of Silicon Read Diode Oscillator", *IEEE Trans. Elec. Dev.*, **ED–16**, (1969) 64–77.

89. Seidman T. I., and Choo S. C., "Iterative Scheme For Computer Simulation of Semiconductor Device", *Sol. St. Elect.*, **15**, (1972) 1229–1235.

90. Selberherr S., *Analysis and Simulation of Semiconductor Devices*, Springer–Verlag, Wien–New York (1984).

91. Sirisena H. R., and Chou F. S., "Convergence of the Control Parametrization Ritz Method for Nonlinear Optimal Control Problems", *J. Optim. Theory. Appl.* , November, (1979).

92. Slotboom J. W., and de Graaf H. C., "Bandgap Narrowing in Silicon Bipolar Transistors", *IEEE Trans. Elect. Dev.*, **24** (1977) 1123–1125.

93. Styczen K., "Regularized Trigonometric Approximation of Optimal Control Problems", *International Journal of Control,* **Vol. 44**, (1986) 887–894.

94. Styczen K., "Characterization of the Smoothness of Optimal Periodic Control Problems", *International Journal of Control*, **Vol. 46**, (1987) 753–767.

95. Sze S.M., and Fix G. J., *The Physics of Semiconductor Devices*, **2nd ED.**, John Wiley & Sons, New York (1981)

96. Taylor, A., "Electric cars, special report." *Business Week*, (1994) 104-114.

97. Teo K. L., Fisher M. E., and Moore J. B., "A Suboptimal Feedback Stabilizing Controller for a Class of Nonlinear Regulator Problems", *Applied Mathematics and Computation*, (1993).

98. Teo K. L., and Goh C. J., "A Computational Method for Combined Optimal Parameter Selection and Optimal Control Problems with General Constraints", *J. Aust. Math. Soc.*, Series B, 30, (1989) 350-364.

99. Teo K.L., Goh C.J., and Wong K.H., *A Unified Computational Approach to Optimal Control Problems* Longman Scientific & Technical, New York, (1991).

100. Teo K.L., Jenning L.S., Lee H.W.J., and Rehbock V., "The Control parametrization Enhancing Transform for Constrained Optimal Control Problems", (to appear in Journal of the Australian Mathematical society, Series B.)

101. Teo K. L., and Jennings L. S., "Nonlinear Optimal Control Problems with Continuous State Inequality Constraints", *J. Optim. Theory. Appl.*, **63**, (1989) 1–22.

102. Teo K. L., and Jennings L. S., "Optimal Control with a Cost on Changing Control", *J. Optim. Theory. Appl.*, **68**, (1991) 335–357.

103. Teo K. L., and Wong K. H., "A Computational Method for Time–lag Control Problems with Control and Terminal Inequality Constraints", *Optimal Control: Applications and Methods*, **8**, (1987) 377–396.

104. Teo K. L., Wong K. H., and Clements D. J., "Optimal Control Computation for Linear Time–lag Systems with Linear Terminal Constraints", *J. of Opt. Theory and Applic.* **44** (1984) 509–526.

105. Teo K. L., Wong K. H., and Wu Z. S., *Computational Methods for Optimizing Distributed System*, Academic Press, New York. (1986).

106. Tyler, G. "Motor vehicle are poisoning us", *Management Services, Journal of Management Services, Section New Technol*, **38**(2) (1994) 26-28.

107. Van Roosbroeck W.V., "Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors", *Bell Syst. Tech. J.*, **29** (1950) 560–607.

108. Vasileva A. B., and Butuzov V. F., " Singularly Pertubed Equations in Critical Case", Translated Report 2039, MRC, University of Wisconsin (1978)

109. Wang S., "A Novel Exponentially Fitted Triangular Finite Element Method for an Advection–Diffusion Problem with Boundary Layers", *J. Comp. Phys.*, **134** (1997) 253–260.

110. Wang S., " Numerical Methods for the Solution of the Stationary Semiconductor Device Equations in Two and Three Dimensions", Ph.D Thesis, University of Dublin (1989).

111. Wilf H.S., "Exactness conditions in numerical quadrature", *Num. Math.*, **6** (1961) 315–319.

112. Wiwmer D.A., and Chattergy R., *Introduction to nonlinear Optimization* North–Holland, New York (1978).

113. Wong K. H., Kaji K., and Teo K. L., "Convergence Properties of the Sequential Gradient–Restoration Algorithm for a Class of Optimal Control Problems Involving Initial and Terminal Equality Constraints", *Applied Mathematics Research Report No. 3*, Department of Mathematics, the University of Western Australia (January 1990).

114. Zionts S., *Linear and integer programming*, Prentice–Hall, Inc., Englewood Cliffs, New Jersey (1974).

115. Zhou J. L., and Tits A. L., "Nonmonotone Line Search for Minimax Problems", *J. Optim. Theory Appl.* **76** (1993) 455–476.

116. Zhou J.L., Tits A.L., and Lawrence C.T., "User's guide for FFSQP version 3.7: A Fortran code for solving constrained nonlinear (minimax) optimization problems, generating iterates satisfying all inequality and linear constraints" *Technical Report TR-92-107r2*, Dept. of Elect. Eng., University of Maryland (1997).