Department of Mathematics and Statistics

A study of optimization and optimal control computation: exact penalty function approach

Changjun Yu

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University of Technology

June 2012

Declaration

I affirm that the material in this thesis is the result of my own original research and has not been submitted for any other degree, diploma, or award.

.....
Changjun Yu
June 2012

Abstract

In this thesis, We propose new computational algorithms and methods for solving four classes of constrained optimization and optimal control problems.

In Chapter 1, we present a brief review on optimization and optimal control.

In Chapter 2, we consider a class of continuous inequality constrained optimization problems. The continuous inequality constraints are first approximated by smooth function in integral form. Then, we construct a new exact penalty function, where the summation of all these approximate smooth functions in integral form, called the constraint violation, is appended to the objective function. In this way, we obtain a sequence of approximate unconstrained optimization problems. It is shown that if the value of the penalty parameter is sufficiently large, then any local minimizer of the corresponding unconstrained optimization problem is a local minimizer of the original problem. For illustration, three examples are solved using the proposed method. From the solutions obtained, we observe that the values of their objective functions are amongst the smallest when compared with those obtained by other existing methods available in the literature. More importantly, our method finds solutions which satisfy the continuous inequality constraints.

In Chapter 3, we consider a general class of nonlinear mixed discrete programming problems. By introducing continuous variables to replace the discrete variables, the problem is first transformed into an equivalent nonlinear continuous optimization problem subject to original constraints and additional linear and quadratic constraints. However, the existing gradient-based optimization techniques have difficulty to solve this equivalent nonlinear optimization problem effectively due to the new quadratic inequality constraint. Thus, an exact penalty function is employed to construct a sequence of unconstrained optimization problems, each of which can be solved effectively by unconstrained optimization techniques, such as conjugate gradient or quasi-Newton types of methods. It is shown that any local optimal solution of the unconstrained optimization problem is a local optimal solution of the transformed nonlinear constrained continuous optimization problem when the penalty parameter is sufficiently large. Numerical experiments are carried out to test the efficiency of the proposed method.

In Chapter 4, we investigate the optimal design of allpass variable fractional delay (VFD) filters with coefficients expressed as sums of signed powers-of-two terms, where the

weighted integral squared error is minimized. A new optimization procedure is proposed to generate a reduced discrete search region. Then, a new exact penalty function method is developed to solve the optimal design of allpass variable fractional delay filter with signed powers-of-two coefficients. Design examples show that the proposed method is highly effective. Compared with the conventional quantization method, the solutions obtained by our method are of much higher accuracy. Furthermore, the computational complexity is low.

In Chapter 5, we consider an optimal control problem in which the control takes values from a discrete set and the state and control are subject to continuous inequality constraints. By introducing auxiliary controls and applying a time-scaling transformation, we transform this optimal control problem into an equivalent optimal control problem subject to original constraints and additional linear and quadratic constraints, where the decision variables are taking values from a feasible region, which is the union of some continuous sets. However, due to the new quadratic constraints, standard optimization techniques do not perform well when they are applied to solve the transformed problem directly. We introduce a novel exact penalty function to penalize constraint violations, and then append this penalty function to the objective function, forming a penalized objective function. This leads to a sequence of approximate optimal control problems, each of which can be solved by using optimal control techniques, and consequently, many optimal control software packages, such as MISER 3.4, can be used. Convergence results show that when the penalty parameter is sufficiently large, any local solution of the approximate problem is also a local solution of the original problem. We conclude this chapter with some numerical results for two train control problems.

In Chapter 6, some concluding remarks and suggestions for future research directions are made.

List of publications

The following papers (which have been published or accepted for publication) were completed during PhD candidature:

- C. J. Yu, K. L. Teo, L. S. Zhang, and Y. Q. Bai, "A new exact penalty function method for continuous inequality constrained optimization problems," in *Journal of Industrial and Management Optimization*, vol. 6, no. 4, pp. 895-910, 2010.
- C. J. Yu, K. L. Teo, and Y. Q. Bai, "An exact penalty function method for nonlinear mixed discrete programming problems," *Optimization Letters*, DOI: 10.1007/s11590-011-0391-2.
- B. Li, C. J. Yu, K. L. Teo, and G. R. Duan, "An exact penalty function method for continuous inequality constrained optimal control problem," *Journal of Optimization Theory and Applications*, vol. 151, no. 2, pp. 260-291, 2011.
- C. J. Yu, B. Li, R. Loxton and K. L. Teo, "Optimal discrete-valued control computation," *Journal of Global Optimization*, DOI 10.1007/s10898-012-9858-7.
- C. H. Jiang, Q. Lin, C. J. Yu, K. L. Teo, and G. R. Duan, "An exact penalty method for free terminal time optimal control problem with continuous inequality constraints," *Journal of Optimization Theory and Applications*, DOI: 10.1007/s10957-012-0006-9.
- C. J. Yu, K. L. Teo, L. S. Zhang, and Y. Q. Bai, "On a refinement of the convergence analysis for the new exact penalty function method for continuous inequality constrained optimization problem," *Journal of Industrial and Management Optimization*, vol. 8, no. 2, pp. 485-491, 2012.

The following papers were completed during PhD candidature and are currently under review:

- C. J. Yu, K. L. Teo, and H. H. Dam, "Design of Allpass Variable Fractional Delay Filter with Signed Powers-of-Two Coefficients"
- Q. Lin, R. C. Loxton, K. L. Teo, Y. H. Wu, and C. J. Yu, "A new exact penalty method for semi-infinite programming problems"

Contents

Acknowledgements

The research reported in this thesis was carried out between March 2009 and April 2012, while I was a PhD student in the Department of Mathematics and Statistics, Curtin University and Department of Mathematics, Shanghai University. I really appreciate all kinds of help I have received from my supervisors, families, friends, and colleagues during this period of time.

I would like to express my heartfelt thanks to my supervisor, Prof. Kok Lay Teo and his wife Mrs. Lye-Hen Teo. Professor Teo has guided my research during the past three years with remarkable patience and enthusiasm. During the two years and seven months of my stay in Australia, he was not only a great supervisor of my research but also a gracious mentor of my life. I am also very grateful for the financial support that he provided from October 2011 to July 2012. Without it, this thesis would not have been possible.

I would like to thank Prof. Yanqin Bai, my co-supervisor in the Department of Mathematics, Shanghai University. I have known Prof. Bai since I was a post-graduate student in the Department of Mathematics, Shanghai University in 2005. It was her who led me to the road of research and helped me to apply for a scholarship from China Scholarship Council. With the support of this scholarship, I came to Australia, starting a new phase of my research in the Department of Mathematics and Statistics at Curtin University. I am also particularly grateful to Prof. Wancheng Sheng for his help in making the extension of my study in Australia possible.

I would like to give thanks to Prof. Liansheng Zhang in the Department of Mathematics, Shanghai University. It was he who first introduced me to Prof. Teo, and encouraged me to continue my research overseas. Prof. Zhang is a great mathematician. He has shared many of his novel and inspiring ideas in mathematics with me. These ideas have helped me greatly in my later research.

I wish to acknowledge the help that I received from Dr. Hai Huyen Dam. In particular, she has given me very valuable comments and suggestions on Chapter 4, where the problem considered is the optimal design of allpass variable fractional delay filter with coefficients expressed as sums of signed-powers-of-two terms.

I also wish to thank Prof. Yonghong Wu, the Postgraduate Coordinator in the Department of Mathematics and Statistics. He was also the Chair of my Thesis Committee. He has been very kind and most helpful.

I would like to express my special thanks to Dr. Ryan Loxton and Dr. Qun Lin. They are excellent young researchers, and always willing to help whenever I have difficulties in my research. I really appreciate all the unselfish helps I have received from them.

I would like to give thanks to my fellow PhD students and friends that I have worked with in Australia, especially, Prof. Honglei Xu and his wife Shaoli Wang, Prof. Fusheng Bai, Prof. Zhiyou Wu, A/Prof. Bin Li, Dr. Jingyang Zhou, Qinqin Chai, Xiangyu Gao, Yufei Sun, Ning Ruan, Xia Liu, and the department's academic visitors that I had the opportunity to meet. They are Dr. Canghua Jiang and his wife Chi Yuan, A/Prof. Chuanjiang Li, A/Prof. Tieqiao Tang, Prof. Xuegang Hu and A/Prof. Yonggang Li. In addition, I wish to thank my friends and teachers at Shanghai University, especially, A/Prof. Yirong Yao, A/Prof. Yongjian Yang, A/Prof. Boshun Han, Dr. Guoqiang Wang, Dr. Lipu Zhang, Yi Chen and Hui Dong. I really enjoyed the time that I spent with them.

I thank the staff in the Department of Mathematics and Statistics for making the work environment so friendly. The past and present administrative staff, in particular, Joyce Yang, Shuie Liu and Lisa Holling, deserve special thanks for providing such an effective and efficient administrative support to the department. They are always very kind and supporting whenever being approached.

Finally, on the personal note, I sincerely thank everyone in my family, especially my parents, my wife, Jing Xu, and daughter, Jiayue Yu, for their love, encouragement, understanding and support throughout the entire period of my PhD candidature in Australia.

CHAPTER 1

Introduction

1.1 Optimization

Optimization and optimal control have been studied intensively and many interesting and powerful results are now available in the literature. They have also been applied to a wide range of real world applications, which include portfolio optimization, minimization of energy consumption and maximization of system performance, structural engineering, robot arms control, DC/DC converters, resource allocation, and military defence. In both optimization and optimal control, a decision variable is to be chosen such that a cost function is minimized subject to a set of constraints. These constraints could be of equality and/or inequality forms. The main difference between optimization and optimal control is that there is a dynamic system involved in optimal control. Furthermore, the decision variable in optimal control is a measurable function. On the other hand, it is a vector, which is independent of time, in optimization.

A typical optimization problem can be stated as follows:

Problem P.

$$\begin{array}{lll} \text{Minimize} & f(\mathbf{x}) & \\ \text{subject to} & g_i(\mathbf{x}) \leq 0, & i \in \mathcal{I}, \\ & h_j(\mathbf{x}) = 0, & j \in \mathcal{E}, \\ & \mathbf{x} \in X, & \end{array} \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^r$ is the decision vector; $f(\mathbf{x})$, $g_i(\mathbf{x}), i \in \mathcal{I}$, and $h_j(\mathbf{x})$, $j \in \mathcal{E}$, are functions defined on $\mathbb{R}^r$; $\mathcal{I}$ and $\mathcal{E}$ are, respectively, the sets of indices for inequality and equality constraints. X is a subset of $\mathbb{R}^r$, which is often defined as boundedness constraints given by $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ are, respectively, the lower and upper bounds on the decision vector $\mathbf{x}$. The function $f(\mathbf{x})$ is called the *objective (or cost) function*. $g_i(\mathbf{x}) \leq 0$, $i \in \mathcal{I}$, are called *inequality constraints*, and $h_j(\mathbf{x}) = 0$, $j \in \mathcal{E}$, are called *equality constraints*.

1.1.1 Unconstrained optimization problems

Problem $\mathbf{P}$ with $\mathcal{I} = \mathcal{E} = \emptyset$ and $X = \mathbb{R}^r$ is called an *unconstrained optimization problem.* Let this problem be denoted as Problem $\mathbf{P}_U$.

For completeness, we shall present some basic concepts.

Definition 1.1. *A point $\mathbf{x}^\star$ is called a local minimizer of the unconstrained optimization Problem $\mathbf{P}_U$ if there exists an $\epsilon > 0$ such that,*

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}),$$

for all $\mathbf{x} \in \mathcal{N}_\epsilon(\mathbf{x}^\star)$, where $\mathcal{N}_\epsilon(\mathbf{x}^\star) = \{\mathbf{x} \in \mathbb{R}^r \mid |\mathbf{x} - \mathbf{x}^\star| \leq \epsilon\}$, and $|\cdot|$ denotes the usual Euclidean norm. A point $\mathbf{x}^\star$ is called a strict local minimizer if

$$f(\mathbf{x}^\star) < f(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{N}_\epsilon(\mathbf{x}^\star) \backslash \{\mathbf{x}^\star\}.$$

Definition 1.2. *A point $\mathbf{x}^\star$ is called a global minimizer of the unconstrained optimization Problem $\mathbf{P}_U$ if*

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^r.$$

A point $\mathbf{x}^\star$ is called a strict global minimizer if

$$f(\mathbf{x}^\star) < f(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^r \backslash \{\mathbf{x}^\star\}.$$

In the following theorem, we give the necessary conditions for optimality for the unconstrained optimization problem $\mathbf{P}_U$.

Theorem 1.1 (First-order necessary conditions)**.** *Suppose that $\mathbf{x}^\star$ is a local minimizer of the unconstrained optimization Problem $\mathbf{P}_U$, where the objective function f is continuously differentiable in an open neighborhood of $\mathbf{x}^\star$. Then, it holds that $\nabla f(\mathbf{x}^\star) = \mathbf{0}$, where*

$$\nabla f(\mathbf{x}) = [\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \dots \frac{\partial f(\mathbf{x})}{\partial x_r}]^\top$$

denotes the gradient of the objective function f at $\mathbf{x}$ and the superscript "$^\top$" denotes the transpose.

The *Hessian* of the objective function f and the *positive definite/semidefinite* matrices are defined in the next two definitions.

Definition 1.3. *Suppose that the function f is twice continuously differentiable. Then,*

the Hessian of the function f at $\mathbf{x}$ is defined by

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_r} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_r} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_r \partial x_1} & \dfrac{\partial^2 f}{\partial x_r \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_r^2} \end{bmatrix}.$$

Definition 1.4. *A matrix M is said to be positive semidefinite if*

$$\mathbf{x}^\top M \mathbf{x} \geq 0,$$

for all $\mathbf{x} \neq \mathbf{0}$. Furthermore, if the above inequality holds strictly, then the matrix M is said to be positive definite.

The *second-order necessary conditions* and the *second-order sufficient conditions* for unconstrained optimization problems are given below.

Theorem 1.2 (Second-order necessary conditions)**.** *Suppose that $\mathbf{x}^\star$ is a local minimizer of the unconstrained optimization Problem $\mathbf{P}_U$. If the objective function f is twice continuously differentiable in a neighborhood of $\mathbf{x}^\star$, then $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ and the Hessian $\nabla^2 f(\mathbf{x}^\star)$ is positive semidefinite.*

Theorem 1.3 (Second-order sufficient conditions)**.** *Let $\mathbf{x}^\star$ be a feasible solution of the unconstrained optimization Problem $\mathbf{P}_U$. Suppose that the objective function f is twice continuously differentiable in a neighborhood of $\mathbf{x}^\star$, that $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ and that $\nabla^2 f(\mathbf{x}^\star)$ is positive definite. Then, $\mathbf{x}^\star$ is a strict local minimizer.*

A point $\mathbf{x}^\star$ is called a *stationary point* if $\nabla f(\mathbf{x}^\star) = 0$. From Theorem 1.1, we see that any local minimizer is a stationary point.

We will present a briefly survey on some of the existing gradient-based algorithms for unconstrained optimization problems.

A typical optimization algorithm generates a sequence of points $\{\mathbf{x}^k\}$ such that the objective function value is reduced at each iteration. To obtain such a sequence, we need to find a *descent direction* at each iteration point.

Definition 1.5. *A direction $\mathbf{d}^k$ is called a descent direction of the objective function f at $\mathbf{x}^k$ if it satisfies $(\mathbf{d}^k)^\top \nabla f(\mathbf{x}^k) < 0$.*

For a descent direction $\mathbf{d}^k$, there exists an $\bar{\alpha} > 0$ such that $f(\mathbf{x}^k + \alpha \mathbf{d}^k) < f(\mathbf{x}^k)$ for each $\alpha \in (0, \bar{\alpha})$. Any chosen $\alpha^k \in (0, \bar{\alpha})$ is called a *step-length*. A typical descent

algorithm is given below.

Descent algorithm for Problem $\mathbf{P_u}$

Step 0:

Choose an initial guess $\mathbf{x}^0$ for Problem $\mathbf{P_U}$ and set $k = 0$ and the tolerance $\epsilon > 0$.

Step 1:

Check for convergence (*i.e.* if $|\nabla f(\mathbf{x}^k)| < \epsilon$). If it is satisfied, **Stop**, otherwise go to **Step 2**.

Step 2:

Determine a descent search direction $\mathbf{d}^k$, and then find a α^k such that $f(\mathbf{x}^k + \alpha^k \mathbf{d}^k) < f(\mathbf{x}^k)$.

Step 3:

Set $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$, and $k := k + 1$; go to **Step 1**.

Note that the finding of α^k in Step 2 is known as a line search, which is a one-dimensional optimization problem. However, finding the minimum of this one-dimensional optimization problem, which is referred to as the exact line search, is, in general, not implementable. In practice, it is chosen such that a sufficient decrease in the function value as well as an acceptable slope improvement are achieved. A popular scheme for finding an acceptable step length is known as the *Armijo Rule* [83].

Steepest descent method

From Definition 1.5, it is clear that the direction $-\boldsymbol{g}^{(k)}$, where $\boldsymbol{g}^{(k)} = \nabla f(\mathbf{x}^k)$, is a descent direction. In fact, it is the direction along which the objective function f decreases most rapidly. Thus, it is called the *Steepest Descent Direction* of the function f at $\mathbf{x}^k$

By choosing the search direction $\mathbf{d}^k$ as the steepest descent direction in the above descent algorithm, we have the *Steepest Descent Method*. Steepest descent method is the simplest one among all gradient-based unconstrained optimization methods. It only requires the gradient information of the function f at the current iteration point and the function value along a line segment. However, the convergence rate of the steepest descent method can be very slow [5, 83].

Newton's method

Newton's method is based on the quadratic approximation of the function f obtained by truncating the Taylor series expansion of $f(x)$ about $x^{(k)}$. That is, the objective function $f(\mathbf{x}^k + \boldsymbol{\delta})$ is approximated by the following quadratic function

$$q^k(\boldsymbol{\delta}) = f^k + \boldsymbol{\delta}^\top \boldsymbol{g}^k + \frac{1}{2}\boldsymbol{\delta}^\top G^k \boldsymbol{\delta},$$

where $G^{(k)} = \nabla^2 f(\mathbf{x}^k)$. The next iterate $\mathbf{x}^{k+1}$ is chosen such that $x^{k+1} = x^k + \delta^k$, where δ^k is the solution of

$$\nabla q^k(\boldsymbol{\delta}) = \mathbf{0}.$$

If $G^{(k)}$ is positive definite, then

$$\boldsymbol{\delta}^k = -(G^k)^{-1}\boldsymbol{g}^k.$$

Remark 1.1. (i). Newton's method requires the information on $f^{(k)}$, $g^{(k)}$ and $G^{(k)}$, i.e. function values, and first and second order partial derivatives.

(ii). The basic Newton's method does not involve a line search. The choice of $\boldsymbol{\delta}^{(k)}$ ensures that the minimum of the quadratic approximation is achieved.

(iii). If $G^\star$ is positive definite, it has a convergence rate of second order if the starting point is sufficiently close to $\mathbf{x}^\star$

(iv). Choosing δ^k as the solution of $\nabla q^k(\boldsymbol{\delta}) = \mathbf{0}$ is only appropriate and well-defined if the quadratic approximation has a minimum, i.e., G^k is positive definite. This may not be the case if $\mathbf{x}^k$ is remote from $\mathbf{x}^\star$ where $\mathbf{x}^\star$ is a local minimum.

Newton's method

Step 0:

Choose $\mathbf{x}^0$ and set $k = 0$.

Step 1:

If $\boldsymbol{g}^k = \mathbf{0}$, **Stop**.

Step 2:

Solve $G^k\boldsymbol{\delta} = -\boldsymbol{g}^k$ for $\boldsymbol{\delta} = \boldsymbol{\delta}^k$ where $\boldsymbol{g}^k = \nabla f(\mathbf{x}^k)$ and $G^k = \nabla^2 f(\mathbf{x}^k)$.

Step 3:

Set $\mathbf{x}^{k+1} = \mathbf{x}^k + \boldsymbol{\delta}^k$.

Step 4:

Set $k := k + 1$, go to **Step 1**.

Quasi-Newton methods

Quasi-Newton Methods might be the most popular unconstrained optimization methods among all the existing methods. They do not require the computation of Hessian at each iteration. Yet, they attain a super-linear convergence rate which is slightly inferior to that attained by Newton's method.

The search direction of a quasi-Newton method is of the form

$$\mathbf{d}^k = -(B^k)^{-1}\nabla f(\mathbf{x}^k). \tag{1.2}$$

Instead of choosing the matrix B^k as the Hessian of the objective function f as in Newton's method, B^k is a symmetric positive definite matrix which is updated at each iteration to approximate the Hessian of the objective function. Note that, the positive definiteness of the matrix B^k ensures that the search direction so generated is a descent direction.

In what follows, we shall present the updating formula for the approximation matrix B^k at each iteration, and the updating formula for its inverse at each iteration.

From Taylor's series expansion and the first mean value theorem for integration, we have

$$\begin{aligned}\nabla f(\mathbf{x}^k + \mathbf{d}^k) &= \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)\mathbf{d}^k + \int_0^1 [\nabla^2 f(\mathbf{x}^k + p\mathbf{d}^k) - \nabla^2 f(\mathbf{x}^k)]\mathbf{d}^k dp \\ &= \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)\mathbf{d}^k + o(|\mathbf{d}^k|), \end{aligned} \tag{1.3}$$

where

$$\lim_{|\mathbf{d}^k|\to 0} \frac{o(|\mathbf{d}^k|)}{|\mathbf{d}^k|} = 0.$$

Letting $\mathbf{d}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$ in (1.3), it gives

$$\nabla f(\mathbf{x}^{k+1}) = \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k) + o(|\mathbf{x}^{k+1} - \mathbf{x}^k|). \tag{1.4}$$

When $\mathbf{x}^k$ and $\mathbf{x}^{k+1}$ are sufficiently close to a local minimizer $\mathbf{x}^\star$, (1.4) can be approximately written as:

$$\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k) \approx \nabla^2 f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k) \tag{1.5}$$

The approximation matrix B^{k+1} is to be constructed such that the *quasi-Newton condition* is satisfied, i.e.,

$$B^{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k) = \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k).$$

The updating formula for the approximate matrix B^{k+1} is now available in any book on optimization (see, for example, [19, 20, 27]). It is given below.

$$B^{k+1} = B^k - \frac{B^k\mathbf{d}^k(\mathbf{d}^k)^\top B^k}{(\mathbf{d}^k)^\top B^k\mathbf{d}^k} + \frac{\boldsymbol{\gamma}^k(\boldsymbol{\gamma}^k)^\top}{(\boldsymbol{\gamma}^k)^\top\mathbf{d}^k} \tag{1.6}$$

where $\boldsymbol{\gamma}^k = \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)$. It can be shown (see, for example, [83]) that if the initial guess B^0 is a positive definite matrix, then the BFGS formula will generates a sequence of positive definite approximation matrices, where BFGS is the abbreviation of Broyden, Fletcher, Goldfarb and Shanno. Note that in the case of quadratic objective function, $B^k = \nabla^2 f(\mathbf{x}^k)$ if exact line search is adopted at each iteration.

From (1.2), we can see that it is $(B^k)^{-1}$, rather than B^k, is used to generate the search direction $\mathbf{d}^k$. The updating formula for the inverse matrix $H^k = (B^k)^{-1}$ is given below.

$$H^{k+1} = (I - \frac{\mathbf{d}^k(\boldsymbol{\gamma}^k)^\top}{(\boldsymbol{\gamma}^k)^\top \mathbf{d}^k}) H^k (I - \frac{\boldsymbol{\gamma}^k(\mathbf{d}^k)^\top}{(\boldsymbol{\gamma}^k)^\top \mathbf{d}^k}) + \frac{\mathbf{d}^k(\mathbf{d}^k)^\top}{(\boldsymbol{\gamma}^k)^\top \mathbf{d}^k}.$$

This is the well-known BFGS formula.

Conjugate gradient methods

Conjugate Gradient Methods are originally proposed in 1952 (see [21, 33, 44]) for solving systems of linear equations. This method was extended to solve general unconstrained optimization problems because the problem of minimizing a positive definite quadratic function is equivalent to solving a system of linear equations. Although these methods are normally less efficient when compared with Newton or quasi-Newton methods, they are much faster than the steepest descent method. Furthermore, conjugate gradient methods have very moderate storage requirements. Thus, they are often used for large-scale problems when quasi-Newton methods become problematic.

1.1.2 Constrained optimization problems

Problem **P** is called a *constrained optimization problem* when $\mathcal{I}$ or $\mathcal{E}$ or both of them are not empty.

For constrained optimization problem **P**, a vector $\mathbf{x}$ is called a *feasible solution* if it satisfies all the constraints of Problem **P**. The set of all feasible solutions is called the *feasible region*. If the objective function is linear and all the constraints are also linear, then we say that Problem **P** is a *linear programming problem*. Otherwise, Problem **P** is called a *nonlinear programming problem*.

Linear programming problems can be efficiently solved by many existing optimization algorithms. One of the most remarkable methods is the *simplex method* - developed by Dantzig in late 1940s [16]. Another type of efficient method for solving linear programming problems is the interior point method [27, 55]. Global solutions of linear programming problems can be obtained if the problems admit global solutions. For nonlinear programming problems, it is not the case anymore. A global optimal solution of a nonlinear programming problem is, in general, very difficult to obtained. Thus, for a nonlinear programming problem, it often aims to find a *local optimal solution* $\mathbf{x}^\star$ — a feasible solution that has less objective function value than all those feasible solutions in a neighborhood of $\mathbf{x}^\star$, rather than in the whole feasible region. An important mathematical result for the characterization of feasible solutions is the *Karush-Kuhn-*

Tucker (KKT) conditions [58], which is a set of necessary conditions for local optimal solutions. There are many methods available in the literature for solving nonlinear programming problems.

For general nonlinear programming problems, where both the objective function and constraint functions are nonlinear, the sequential quadratic approximation programming with active set strategy (see, for example, [31, 41, 42]) is a popular method. Methods developed based on Newton's method are also effective for solving nonlinear programming problems. For more details, see, for example, [7, 37].

Definition 1.6. *For any feasible solution* $\mathbf{x}$ *of Problem* $\mathbf{P}$*, let* $\mathcal{A}(\mathbf{x})$ *be the set of those indices defined by*

$$\mathcal{A}(\mathbf{x}) = \mathcal{E} \cup \{i \in \mathcal{I} \mid g_i(\mathbf{x}) = 0\}.$$

$\mathcal{A}(\mathbf{x})$ *is called the active set of* $\mathbf{x}$*. For a feasible solution* $\mathbf{x}$*, the inequality constraint* $g_i(\mathbf{x})$ *is said to be active if* $g_i(\mathbf{x}) = 0$*; otherwise, we say that* $g_i(\mathbf{x})$ *is inactive.*

Definition 1.7 (Feasible direction cone)**.** *For a feasible point* $\mathbf{x}$ *of Problem* $\mathbf{P}$*, a vector* $\mathbf{v}$ *is called a feasible direction of* $\mathbf{x}$ *if the following conditions are satisfied,*

$$\begin{aligned} \mathbf{v}^\top \nabla h_j(\mathbf{x}) &= 0, \qquad j \in \mathcal{E}, \\ \mathbf{v}^\top \nabla g_i(\mathbf{x}) &\leq 0, \qquad i \in \mathcal{I} \cap \mathcal{A}(\mathbf{x}). \end{aligned}$$

The set of all feasible directions of $\mathbf{x}$ *is called the feasible direction cone, denoted as* $\mathcal{F}(\mathbf{x})$*.*

To continue, we need to define *Linear Independent Constraint Qualification (LICQ)* and *Lagrangian function*. They are given below.

Definition 1.8 (LICQ)**.** *For a given feasible point* $\mathbf{x}$ *of Problem* $\mathbf{P}$*, suppose that the gradients of all the active constraints of the constraint functions at* $\mathbf{x}$ *are linearly independent. Then, it is said that the Linear Independent Constraint Qualification (LICQ) is satisfied at* $\mathbf{x}$*.*

Definition 1.9. *Consider Problem* $\mathbf{P}$*. The Lagrangian function is defined by*

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i \in \mathcal{I}} \alpha_i g_i(\mathbf{x}) + \sum_{j \in \mathcal{E}} \beta_j h_j(\mathbf{x}) \tag{1.7}$$

where α_i*,* $i \in \mathcal{I}$*, and* β_j*,* $j \in \mathcal{E}$*, are called the Lagrange multipliers for the constraints* $g_i(\mathbf{x})$ *and* $h_i(\mathbf{x})$*, respectively.*

Now, we are in the position to present the (KKT) conditions, also known as the first-order necessary conditions, in the following theorem.

Theorem 1.4 (Karush-Kuhn-Tucker conditions)**.** *Suppose that* $\mathbf{x}^\star$ *is a local optimal solution of Problem* $\mathbf{P}$*, and that the linear independent constraint qualification (LICQ) holds at* $\mathbf{x}^\star$*. Then, there exists vector* $\boldsymbol{\alpha}^\star$*, with components* $\alpha_i^\star$*,* $i \in \mathcal{I}$*, and vector* $\boldsymbol{\beta}^\star$*, with components* $\beta_j^\star$*,* $j \in \mathcal{E}$*, such that the following conditions are satisfied*

$$\nabla_{\mathbf{x}} L(\mathbf{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) \;\; = \;\; \mathbf{0}, \tag{1.8}$$

$$g_i(\mathbf{x}^\star) \leq 0, \quad i \in \mathcal{I}, \tag{1.9}$$
$$h_j(\mathbf{x}^\star) = 0, \quad j \in \mathcal{E}, \tag{1.10}$$
$$\alpha_i^\star \geq 0, \quad i \in \mathcal{I}, \tag{1.11}$$
$$\alpha_i^\star g_i(\mathbf{x}^\star) = 0, \quad i \in \mathcal{I}. \tag{1.12}$$

To state the second-order necessary conditions and the second-order sufficient conditions, we need the following definition.

Definition 1.10. *Let $\mathbf{x}^\star$ be a local optimal solution of Problem $\mathbf{P}$, and let $\boldsymbol{\alpha}^\star$ be the Lagrangian multipliers corresponding to the inequality constraints that satisfies the KKT conditions. Then, the critical cone $\mathcal{U}(\mathbf{x}^\star, \boldsymbol{\alpha}^\star)$ is defined by*

$$\mathcal{U}(\mathbf{x}^\star, \boldsymbol{\alpha}^\star) = \{\mathbf{u} \in \mathcal{F}(\mathbf{x}^\star) \mid \nabla g_i(\mathbf{x}^\star)^\top \mathbf{u} = 0, i \in \mathcal{I} \cap \mathcal{A}(\mathbf{x}^\star), \boldsymbol{\alpha}_i^\star > 0\}.$$

Theorem 1.5 (Second-order necessary conditions)**.** *Suppose that $\mathbf{x}^\star$ is a local optimal solution of Problem $\mathbf{P}$ and that the LICQ is satisfied. Let $\boldsymbol{\alpha}^\star$ be the Lagrangian multipliers corresponding to the inequality constraints such that the KKT conditions are satisfied. Then,*

$$\mathbf{u}^\top \nabla_{xx}^2 L(\mathbf{x}^\star, \boldsymbol{\alpha}^\star)\mathbf{u} \geq 0, \text{ for all } \mathbf{u} \in \mathcal{U}(\mathbf{x}^\star, \boldsymbol{\alpha}^\star).$$

Theorem 1.6 (Second-order sufficient conditions)**.** *Suppose that $\mathbf{x}^\star$ is a feasible solution of Problem $\mathbf{P}$. Let $\boldsymbol{\alpha}^\star$ be the Lagrangian multipliers corresponding to the inequality constraints such that the KKT conditions are satisfied. If*

$$\mathbf{u}^\top \nabla_{xx}^2 L(\mathbf{x}^\star, \boldsymbol{\alpha}^\star)\mathbf{u} > 0, \text{ for all } \mathbf{u} \in \mathcal{U}(\mathbf{x}^\star, \boldsymbol{\alpha}^\star)\backslash\{\mathbf{0}\},$$

then $\mathbf{x}^\star$ is a strict local optimal solution of Problem $\mathbf{P}$.

Due to a multitude of real world applications of nonlinear constrained optimization, it has attracted the interest of many mathematicians and engineers. Many interesting and important theoretical results as well as numerical algorithms are now available in the literature. Examples include *penalty and augmented Lagrangian methods* [6, 36, 43, 92, 93, 101] *sequential quadratic programming* methods [94–96] and nonlinear *interior-point* methods [1, 37]. Here, we shall give a brief review of the sequential quadratic programming approximation. For this, we shall first consider quadratic programming.

Quadratic programming

Quadratic programming problem is an optimization problem in which the objective function

is quadratic and the constraints are linear. It is typically stated as below.

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^\top Q\mathbf{x} + \mathbf{c}^\top\mathbf{x} \\ \text{subject to} & \mathbf{a}_i^\top\mathbf{x} \leq b_i, \quad i \in \mathcal{I}, \\ & \mathbf{a}_j^\top\mathbf{x} = b_j, \quad j \in \mathcal{E}, \end{array} \tag{1.13}$$

where $\mathbf{x} \in \mathbb{R}^r$ is a decision vector; Q is a positive definite symmetric $r \times r$ matrix; $\boldsymbol{c} \in \mathbb{R}^r$, $\mathbf{a}_i \in \mathbb{R}^r$, $i \in \mathcal{I}$, $\mathbf{a}_j \in \mathbb{R}^r$, $j \in \mathcal{E}$; and $\mathcal{I}$ and $\mathcal{E}$ are finite sets of indices. Let this problem be referred to as Problem $\mathbf{P_Q}$. If Problem $\mathbf{P_Q}$ contains only k equality constraints, $k < r$, *i.e.*, $\mathcal{E} = \{1, \ldots, k\}$ and $\mathcal{I} = \emptyset$, then Problem $\mathbf{P_Q}$ can be solved through solving the following system of KKT conditions for Problem $\mathbf{P_Q}$.

$$\begin{bmatrix} Q & A^\top \\ A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^\star \\ \boldsymbol{\beta}^\star \end{bmatrix} = \begin{bmatrix} -\mathbf{c} \\ \mathbf{b} \end{bmatrix}, \tag{1.14}$$

where $A = [\mathbf{a}_1, \ldots, \mathbf{a}_k]^\top$; $\mathbf{b} = [b_1, \ldots, b_k]^\top$; $\mathbf{x}^\star$ is the solution of Problem $\mathbf{P_Q}$; and $\boldsymbol{\beta}^\star$ is the vector of Lagrange multipliers. To ensure that the system of KKT conditions (1.14) has a solution, we have the following theorem [83].

Theorem 1.7. *Let A be a given $k \times r$ matrix which has full row rank, and let M be a $r \times k$ matrix such that $AM=0$. Suppose that the matrix $M^\top GM$ is positive definite. Then the system of KKT conditions* (1.14) *has a unique solution.*

For large scale problems, system (1.14) is often solved by using iterative methods. For more details, see, for example, [10, 34, 76, 102, 120].

Consider the general quadratic programming problem $\mathbf{P_Q}$, where $\mathcal{E} \neq \emptyset$ and $\mathcal{I} \neq \emptyset$. We introduce the Lagrangian function given below.

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} + \mathbf{c}^\top\mathbf{x} + \sum_{i\in\mathcal{I}} \alpha_i(\mathbf{a}_i^\top\mathbf{x}^\star - b_i) + \sum_{j\in\mathcal{E}} \beta_j(\mathbf{a}_j^\top\mathbf{x}^\star - b_j).$$

The active set $\mathcal{A}(\mathbf{x})$ for any feasible solution $\mathbf{x}$ is defined by

$$\mathcal{A}(\mathbf{x}) = \{i \in \mathcal{I} \mid \mathbf{a}_i^\top\mathbf{x}^\star - b_i = 0\} \cup \{j \in \mathcal{E} \mid \mathbf{a}_j^\top\mathbf{x}^\star - b_j = 0\}.$$

Let $\mathbf{x}^\star$ be an optimal solution. Then, it follows that there exist Lagrange multipliers $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\beta}^\star$ such that the following system of KKT conditions is satisfied.

$$\begin{array}{rcll} G\mathbf{x}^\star + \mathbf{c} + \displaystyle\sum_{i\in\mathcal{I}} \alpha_i^\star\mathbf{a}_i + \sum_{j\in\mathcal{E}} \beta_j^\star\mathbf{a}_j & = & \mathbf{0}, & \\ \mathbf{a}_j^\top\mathbf{x}^\star - b_j & = & 0, & \text{for all } j \in \mathcal{E} \\ \mathbf{a}_i^\top\mathbf{x}^\star - b_i & \leq & 0, & \text{for all } i \in \mathcal{I} \\ \alpha_i^\star(\mathbf{a}_i^\top\mathbf{x}^\star - b_i) & = & 0, & \text{for all } i \in \mathcal{I} \\ \alpha_i^\star & \geq & 0, & \text{for all } i \in \mathcal{I}. \end{array} \tag{1.15}$$

Due to the existence of inequality constraints, the KKT system (1.15) cannot be solved directly. General quadratic optimization problem with both equality and inequality constraints are solved via solving a sequence of linear equality constrained quadratic optimization problems based on *active set strategy*.

The main idea of this method is as follows: At each iteration, the corresponding active set is identified. This gives rise to an equality constrained quadratic programming problem. Then, the corresponding Lagrange multipliers are computed from the system (1.14). If all of the Lagrange multipliers associated with the active set are non-negative, then the KKT conditions are satisfied and an optimal solution is obtained. On the other hand, if some or all of the multipliers are strictly negative, then the constraint corresponding to the most negative multiplier is removed from the active set. The process is repeated until all of the Lagrange multipliers associated with the active set are non-negative. A typical algorithm is given below.

Active set strategy for quadratic programming problem $\mathbf{P_Q}$

Step 0:

Choose an initial feasible solution $\mathbf{x}^0$ of Problem $\mathbf{P_Q}$ and identify the corresponding active set $\mathcal{A}(\mathbf{x}^0)$. Set $k = 0$.

Step 1:

Compute the search direction $\mathbf{d}^k$ by solving the following problem:

$$\min_{\mathbf{d}} f(\mathbf{x}^k + \mathbf{d}) = \frac{1}{2}\mathbf{d}^\top Q\mathbf{d} + \mathbf{d}^\top(Q\mathbf{x}^k + \mathbf{c}) + f(\mathbf{x}^k) \tag{1.16a}$$

subject to

$$\mathbf{a}_i^\top(\mathbf{x}^k + \mathbf{d}) - b_i = 0, \quad i \in \mathcal{A}(\mathbf{x}^k). \tag{1.16b}$$

If $\mathbf{d} = 0$ solves problem (1.16), go to **Step 2**; otherwise, go to **Step 3**.

Step 2:

Use

$$Q\mathbf{x} + A^\top\boldsymbol{\lambda} = -\mathbf{c}$$

to compute the corresponding Lagrange multiplier vector $\boldsymbol{\lambda}^k = [\lambda_i^k, i \in \mathcal{A}(\mathbf{x}^k)]$. Let j be the index such that

$$\lambda_j^k = \min_{i \in \mathcal{A}(\mathbf{x}^k) \cap \mathcal{I}} \lambda_i^k.$$

If $\lambda_j^k \geq 0$, $\mathbf{x}^k$ is the optimal solution, **stop**; otherwise, set $\mathcal{A}(\mathbf{x}^k) = \mathcal{A}(\mathbf{x}^k)\backslash\{j\}$, go to **Step 3**.

Step 3:

Let $\mathbf{d}^k$ be the solution of problem (1.16). Compute the line search step length γ^k according to $\gamma^k = \min\{1, \bar{\gamma}^k\}$, where

$$\bar{\gamma}^k = \min_{i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^k)} \{ \frac{b_i - \mathbf{a}_i^\top \mathbf{x}^k}{\mathbf{a}_i^\top \mathbf{d}^k}, \mathbf{a}_i^\top \mathbf{d}^k < 0 \} \tag{1.17}$$

and set $\mathbf{x}^{k+1} = \mathbf{x}^k + \gamma^k \mathbf{d}^k$. If $\gamma^k < 1$, set $\mathcal{A}(\mathbf{x}^{k+1}) = \mathcal{A}(\mathbf{x}^k) + \{l\}$, where $l \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^k)$ is chosen such that the minimum of (1.17) is achieved. Otherwise, $\gamma^k = 1$, set $\mathcal{A}(\mathbf{x}^{k+1}) = \mathcal{A}(\mathbf{x}^k)$.

Step 4:

Set $k := k + 1$, go to **Step 1**.

Penalty Function Methods

For general constrained optimization problems, penalty methods use penalty functions to transform a constrained problem into a sequence of unconstrained problems or a single unconstrained problem. The constraints are appended to the objective function via a penalty parameter penalizing any violation of the constraints. By making this penalty parameter larger, the method penalizes constraint violations more severely, and hence forcing the minimizer of the penalty function to move closer to the feasible region of the constrained problem.

The most simple and intuitive penalty method is the *quadratic penalty method*. A typical quadratic penalty function for Problem **P** is given below:

$$\begin{aligned} \text{Minimize} \quad & f(\mathbf{x}) + \sigma \sum_{i \in \mathcal{I}} \max\{0, g_i(\mathbf{x})\}^2 + \sigma \sum_{j \in \mathcal{E}} h_j(\mathbf{x})^2 \\ \text{subject to} \quad & \mathbf{x} \in X. \end{aligned} \tag{1.18}$$

where $\sigma > 0$ is the *penalty parameter*. By increasing σ, the constraint violations will be penalized more and more severely. Thus, the satisfaction of the constraints will be achieved as $\sigma \to \infty$. Note that the quadratic penalty function is smooth, one can use any of the gradient-based unconstrained optimization techniques to find the local optimal solution for each σ.

The major disadvantage of the quadratic penalty method is that it requires the penalty parameter σ to approach to infinity for the satisfaction of the constraints. This might cause difficulties in actual numerical computation when σ is large. To overcome this drawback, a type of penalty function methods, called *exact penalty function method*, is developed. A typical exact penalty function is defined by

$$p(\mathbf{x}, \sigma) = f(\mathbf{x}) + \sigma \sum_{i \in \mathcal{I}} \max\{0, g_i(\mathbf{x})\} + \sigma \sum_{j \in \mathcal{E}} |h_j(\mathbf{x})|$$

Here, by the word "*exact*", it means that when the penalty parameter σ is sufficiently large, if the stationary point of the penalty problem is feasible for the original constrained optimization

problem, then it is a stationary point of the original constrained optimization problem. Since $p(\mathbf{x}, \sigma)$ consists of the term $\max\{0, g_i(\mathbf{x})\}$, it is non smooth. Thus, gradient-based optimization techniques are not applicable. In this thesis, we will introduce new exact penalty function which is differentiable.

Augmented Lagrangian Penalty Function Method is another important penalty function method. For simplicity, we consider Problem **P** where the index set of inequality constraints $\mathcal{I}$ is empty. Then, we employ the following quadratic penalty function

$$f(\mathbf{x}) + \sigma \sum_{j \in \mathcal{E}} h_j(\mathbf{x})^2.$$

Clearly, to obtain a stationary point for Problem **P**, it usually requires that $\sigma \to \infty$. However, if we perturb the constraint right-hand sides from $\mathbf{0}$ to $\boldsymbol{\delta} \in \mathbb{R}^{|\mathcal{E}|}$, where $|\mathcal{E}|$ denotes the size of $\mathcal{E}$, then the corresponding quadratic penalty function becomes

$$f(\mathbf{x}) + \sigma \sum_{j \in \mathcal{E}} (h_j(\mathbf{x}) - \delta_j)^2. \tag{1.19}$$

It is possible to obtain a stationary point of Problem **P**, without letting $\sigma \to \infty$. In fact, expanding (1.19) gives

$$f(\mathbf{x}) - \sum_{j \in \mathcal{E}} 2h_j(\mathbf{x})\sigma\delta_j + \sum_{j \in \mathcal{E}} \sigma h_j(\mathbf{x})^2 + \sum_{j \in \mathcal{E}} \sigma\delta_j^2. \tag{1.20}$$

For $j \in \mathcal{E}$, set $\beta_j = -2\sigma\delta_j$. Ignoring the last constant term, (1.20) can be written as:

$$L_{aLP}(\mathbf{x}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{j \in \mathcal{E}} \beta_j h_j(\mathbf{x}) + \sum_{j \in \mathcal{E}} \sigma h_j(\mathbf{x})^2. \tag{1.21}$$

Note that, if $(\mathbf{x}^\star, \boldsymbol{\beta}^\star)$ is a KKT solution of Problem **P**, then at $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$,

$$\nabla_{\mathbf{x}} L_{aLP}(\mathbf{x}^\star, \boldsymbol{\beta}^\star) = \nabla_{\mathbf{x}} f(\mathbf{x}^\star) + \sum_{j \in \mathcal{E}} \beta_j^\star \nabla_{\mathbf{x}} h_j(\mathbf{x}^\star) + 2\sigma \sum_{j \in \mathcal{E}} h_j(\mathbf{x}^\star) \nabla_{\mathbf{x}} h_j(\mathbf{x}^\star) = \mathbf{0},$$

for any σ, which means that $(\mathbf{x}^\star, \boldsymbol{\beta}^\star)$ is also a stationary point of the augmented Lagrangian penalty function.

However, this conclusion is valid only at $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$. It is not known a priori. For a given σ, it is known that the task of finding $\mathbf{x}^\star$ and $\boldsymbol{\beta}^\star$ through optimizing (1.21) with respective to both $\mathbf{x}$ and $\boldsymbol{\beta}$ simultaneously is not workable. In actual numerical computation, $\mathbf{x}^\star$ and $\boldsymbol{\beta}^\star$ are obtained iteratively as follows. Choose parameters σ and $\boldsymbol{\beta}$, and then optimize (1.21) with respect to $\mathbf{x}$. This gives rise to $\mathbf{x}_k$. With $\mathbf{x} = \mathbf{x}_k$, the parameter σ and $\boldsymbol{\beta}$ are updated according to the following updating rule:

$$\beta_j^{k+1} = \beta^k - \sigma_k h_j(\mathbf{x}_k), \quad j \in \mathcal{E},$$

where σ_k needs to be appropriately estimated. The whole process is to be repeated until a satisfactory result is obtained. This augmented Lagrangian penalty function method is quick

cumbersome to apply. Furthermore, since the desired solution depends critically on the accuracy of the estimate of $\boldsymbol{\beta}^\star$, the convergence may be slow.

Sequential quadratic programming methods

For small and medium sized general nonlinear constrained optimization problems, *sequential quadratic programming* (SQP) has been recognized as one of the most efficient methods.

Consider the general constrained optimization problem **P**. Its Lagrangian function is

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i \in \mathcal{I}} \alpha_i g_i(\mathbf{x}) + \sum_{j \in \mathcal{E}} \beta_j h_j(\mathbf{x}).$$

Let $\mathbf{x}^k$ be an estimate of the optimal solution $\mathbf{x}^\star$, and let $(\boldsymbol{\alpha}^k, \boldsymbol{\beta}^k)$ be an estimate of the optimal Lagrange multiplier vector $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$. The objective function at the current iteration point $\mathbf{x}^k$ can be approximated by the following quadratic function

$$f(\mathbf{x}^k + \mathbf{d}) \approx f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top B^k \mathbf{d} \tag{1.22}$$

where B^k is a positive definite matrix of the Hessian matrix of the Lagrangian function L evaluated at $(\mathbf{x}^k, \boldsymbol{\alpha}^k, \boldsymbol{\beta}^k)$. The matrix B^k is updated according to the BFGS formula (1.6). The constraints are linearized as follows:

$$g_i(\mathbf{x}^k + \mathbf{d}) \approx g_i(\mathbf{x}^k) + \nabla g_i(\mathbf{x}^k)^\top \mathbf{d} \;\leq\; 0, \quad i \in \mathcal{I}, \tag{1.23}$$

$$h_j(\mathbf{x}^k + \mathbf{d}) \approx h_j(\mathbf{x}^k) + \nabla h_j(\mathbf{x}^k)^\top \mathbf{d} \;=\; 0, \quad j \in \mathcal{E}, \tag{1.24}$$

Thus, Problem **P** is approximated as a quadratic programming problem given below.

$$\begin{array}{ll} \min\limits_{\mathbf{d}} & f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top B^k \mathbf{d} \\ \text{subject to} & g_i(\mathbf{x}^k) + \nabla g_i(\mathbf{x}^k)^\top \mathbf{d} \leq 0, \quad i \in \mathcal{I}, \\ & h_j(\mathbf{x}^k) + \nabla h_j(\mathbf{x}^k)^\top \mathbf{d} = 0, \quad j \in \mathcal{E}, \end{array} \tag{1.25}$$

The quadratic programming problem (1.25) is solvable by the active set strategy for quadratic programming. Let $\mathbf{d}^k$ be the solution of this quadratic programming problem and let $\bar{\boldsymbol{\lambda}}^k = [(\bar{\boldsymbol{\alpha}}^k)^\top, (\bar{\boldsymbol{\beta}}^k)^\top]^\top$ be the corresponding optimal multiplier vector. Then, the new estimates $\mathbf{x}^{k+1}$, $\boldsymbol{\lambda}^{k+1}$ and B^{k+1} can be determined by

$$\mathbf{x}^{k+1} \;=\; \mathbf{x}^k + \eta_k \mathbf{d}^k, \tag{1.26}$$

$$\boldsymbol{\lambda}^{k+1} \;=\; \boldsymbol{\lambda}^k + \eta_k(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k), \tag{1.27}$$

$$B^{k+1} \;=\; B^k + \frac{\mathbf{g}^k(\mathbf{g}^k)^\top}{(\mathbf{p}^k)^\top \mathbf{g}^k} - \frac{B^k \mathbf{p}^k (\mathbf{p}^k)^\top B^k}{(\mathbf{p}^k)^\top B^k \mathbf{p}^k} \tag{1.28}$$

where

$$\mathbf{p}^k = \mathbf{x}^{k+1} - \mathbf{x}^k, \tag{1.29}$$

$$\mathbf{g}^k = \nabla_{\mathbf{x}} L(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \nabla_{\mathbf{x}} L(\mathbf{x}^k, \boldsymbol{\lambda}^k) \tag{1.30}$$

For the step length η_k, it is chosen such that a sufficient decrease of the well-known Lagrangian multiplier penalty function is achieved:

$$P_{\sigma^k}(\mathbf{x}^k + \alpha \mathbf{d}^k; \lambda^k + \alpha(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k))$$

where

$$P_{\sigma^k}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{j \in \mathcal{E}} [\lambda_j h_j(\mathbf{x}) - \frac{1}{2}\sigma_j (h_j(\mathbf{x}))^2] - \sum_{i \in \mathcal{I}} \begin{cases} \lambda_i g_i(\mathbf{x}) - \frac{1}{2}\sigma_i (g_i((x)))^2, & \text{if } g_i(\mathbf{x}) \geq \lambda_i/\sigma_i, \\ \frac{1}{2}\lambda_i^2/\sigma_i, & \text{otherwise.} \end{cases}$$

Here, $\lambda_i = \theta_i \sigma_i, i \in \mathcal{I}$ and $\lambda_j = \theta_j \sigma_j, j \in \mathcal{E}$. $\sigma_i, i \in \mathcal{I}$ and $\sigma_j, i \in \mathcal{E}$ are, respectively, the penalty parameters of the inequality constraints g_i and equality constraints h_j. The parameters $\theta_i, i \in \mathcal{I}$ and $\theta_i, j \in \mathcal{E}$ correspond to the shift of the origin.

It is shown in [27] that if σ^k is appropriately updated, then the sequence $(\mathbf{x}^k, \boldsymbol{\lambda}^k, \sigma^k)$ converges to $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \sigma^\star)$, where $\mathbf{x}^\star$ is a local minimum of the function $P_{\sigma^\star}(\mathbf{x}, \boldsymbol{\lambda}^\star)$, which is also a local solution of Problem **P**.

For more details on the theory and computational algorithms, see, for example, [5, 83].

1.1.3 Optimal control

Optimal control problems are generally more complicated than static optimization problems. For an optimal control problem, it involves a dynamical system which does not appear in the formulation of a static optimization problem. Furthermore, the decision variables are functions of time in an optimal control problem, while they are constant vectors in a static optimization problem. In this section, we shall present a brief introduction to some of the fundamental results of optimal control theory.

Consider a dynamic system described by the following system of differential equations:

$$\dot{\boldsymbol{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad t \in [0, T] \tag{1.31}$$

with initial condition:

$$\mathbf{x}(0) = \mathbf{x}^0 \tag{1.32}$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_r(t)]^\top$ is called the *state vector* at time t; $\dot{\mathbf{x}}(t) = d\mathbf{x}(t)/dt$; $\mathbf{u}(t) = [u_1(t), \ldots, u_n(t)]^\top$ is called the *control vector* at time t; $\mathbf{f} = [f_1, \ldots, f_r]^\top$ is a given vector-valued function which is continuously differentiable; $\mathbf{x}^0 \in \mathbb{R}^r$ is a given vector which is referred to as the *initial state* or *initial condition* of the dynamic system. In this system, the process evolves

starting from the state $\mathbf{x}^0$ at $t = 0$ until the time $t = T$, where T is called the *terminal time*.

Let $\mathcal{S}$ be a bounded subset of $\mathbb{R}^n$. A measurable function $\mathbf{u} : [0, T] \to \mathcal{S}$ is called an admissible control function. Let $\mathcal{U}$ be the class of all such admissible controls.

A simple optimal control problem may now be stated formally as follows. Given the dynamic system (1.31)-(1.32), find an admissible control $\mathbf{u} \in \mathcal{U}$ such that the following *cost function*:

$$G_0(\mathbf{u}) = \Phi(\mathbf{x}(T)) + \int_0^\top \mathcal{L}(t, \mathbf{x}(t), \mathbf{u}(t))dt \tag{1.33}$$

is minimized, where Φ_0 and $\mathcal{L}$ are given continuously differentiable functions. Let this problem be referred to as Problem $\mathbf{P_C}$.

Pontryagin minimum principle

To state the Pontryagin minimum principle, we first introduce the *Hamiltonian function* given below:

$$H(t, \mathbf{x}, \boldsymbol{\lambda}, \mathbf{u}) = \mathcal{L}(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^\top \mathbf{f}(t, \mathbf{x}, \mathbf{u}), \tag{1.34}$$

where the time dependent Lagrange multiplier $\boldsymbol{\lambda}$ is called the *costate vector*.

The Pontryagin minimum principle is given in the following theorem, which is a first order necessary condition.

Theorem 1.8. *Consider Problem* $\mathbf{P_C}$. *Let* $\mathbf{u}^\star(t)$ *be an optimal control, and let* $\mathbf{x}^\star(t)$ *and* $\boldsymbol{\lambda}^\star(t)$ *be the corresponding state and costate. Then,*

- $\dot{\mathbf{x}}^\star(t) = \Big[\dfrac{\partial H(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t))}{\partial \lambda}\Big]^\top = \mathbf{f}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))$
- $\mathbf{x}^\star = \mathbf{x}^0$
- $\dot{\boldsymbol{\lambda}}^\star(t) = -\Big[\dfrac{\partial H(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t))}{\partial \mathbf{x}}\Big]^\top$
- $\boldsymbol{\lambda}^\star(T) = \Big[\dfrac{\partial \Phi(\mathbf{x}^\star(T))}{\partial \mathbf{x}}\Big]^\top$
- $\min\limits_{\mathbf{v} \in \mathcal{S}} H(t, \mathbf{x}^\star(t), \mathbf{v}, \boldsymbol{\lambda}^\star(t)) = H(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t))$ *for all* $t \in [0, T]$, *except possibly on a finite subset of* $[0, T]$.

For detailed information on the Pontryagin minimum principle, see, for example, [2,3,57,137].

Bellman's principle of optimality

By applying Bellman's principle of optimality to the optimal control problem $\mathbf{P_C}$, we obtain a sufficient condition for optimality for Problem $\mathbf{P_C}$. This sufficient condition for optimality is expressed as:

$$\frac{\partial V(t, \mathbf{x})}{\partial t} + \min_{v \in \mathcal{S}} \{\frac{\partial V(t, \mathbf{x})}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}, v) + \mathcal{L}(t, \mathbf{x}, \mathbf{v})\} = 0, \tag{1.35}$$

which is a nonlinear partial differential equation. It is to be solved with the boundary condition given by

$$V(T, \mathbf{x}) = \Phi(\mathbf{x}). \tag{1.36}$$

Equation (1.35) is the well-known *Hamilton-Jacobi-Bellman (HJB) equation*, and the function V is called the *value function*.

For most real world problems, they are, in general, much too complex to allow for analytical solutions by applying Pontryagin minimum principle or through solving the Hamilton-Jacobi-Bellman equation with given boundary condition. Furthermore, there exist various kinds of additional constraints in real world problems. Thus, numerical methods are inevitable for solving these real world problems. For this reason, the area of computational algorithms has attracted the interest of many engineers and mathematicians. As a result, many computational algorithms are now available in the literature. To solve the HJB equation, numerical methods based on finite-difference or finite-volume approximation are reported in [38, 123, 124]. However, these methods are applicable only to small dimensional problems. The multiple shooting methods are developed based on necessary conditions for optimality in [2, 57]. These multiple shooting methods tend to give good solutions. However, they are rather sensitive to the choice of initial guess of the optimal control.

The control parametrization technique [114] is a popular technique for developing computational methods for various optimal control problems. Its main idea is to approximate the control function by a finite number of basis functions, for example, piecewise constant functions. The coefficients of these basis functions are decision variables to be chosen optimally. By applying this approximation scheme, an approximate optimization problem is obtained. In the classical control parametrization technique, the times at which the approximate control changes its value—the *switching time*—are fixed. Intuitively, the switching times should also be regarded as decision variables. However, the computation of the gradient of the objective function with respect to the switching times is rather sensitive. Thus, any optimization technique using this gradient formula tends to perform poorly. Furthermore, it requires much more work to solve the dynamic system when the switching times are variable. To overcome these difficulties, a time-scaling transformation—it is originally called the control parametrization enhancing technique (CPET)—is developed in [61, 62]. By introducing a new time variable and a new control, this technique transform the time horizon of the optimal control problem into a new time horizon in such a way that the switching times can be chosen to be fixed in the new time horizon.

The constraint transcription method is originally developed in [113] to handle continuous inequality constraints on the state variables of the dynamical system. It is extended in [51] to handle optimal control problems subject to continuous inequality constraints on the state as well as on the control. There are many computational algorithms, which are derived based on the control parametrization technique, in conjunction with the time scaling transform and the constraint transcription method. See, for example, [24, 66, 72, 104, 109, 112, 115, 128, 130]. A general optimal control software package, MISER 3.4 [49], has been implemented based on some of these algorithms.

Recently, a new exact penalty function method [135] is used to handle continuous inequality state constraints in various optimal control problems (see, for example, [52,63,132]). It leads to effective computational algorithms for these optimal control problems with continuous inequality state constraints.

1.2 Overview of the thesis

In the previous sections, a brief introduction to optimization and optimal control is given. The purpose of this thesis is to develop new computational algorithms for four types of static and dynamic optimization problems. Some of their real world applications are also addressed.

In Chapter 2, we consider a class of continuous inequality constrained optimization problems (also known as semi-infinite programming problems) in the form given below:

$$\min \ f(x) \tag{1.37a}$$

$$\text{subject to} \qquad \phi_j(x,\omega) \leq 0, \ \forall \ \omega \in \Omega, \ j = 1, \ \ldots \ , \ m, \tag{1.37b}$$

where $x \in \mathbb{R}^n$ is the decision parameter vector, Ω is a compact interval in $\mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in x, and for each $j = 1, \ \ldots \ , \ m$, $\phi_j : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function in x and ω. Note that there are infinite many inequality constraints in (1.37b). Motivated by the idea reported in [116], a new exact penalty function approach, instead of the constraint transcription method, is introduced to handle the continuous inequality constraints. Furthermore, the summation of the integrals of the exact penalty functions, rather than the summation of the integrals of the smooth approximate functions as in the case of utilizing the constraint transcription method, is appended to the objective function forming a new objective function. This gives rise to a sequence of unconstrained optimization problems. It is shown that any local minimizer of the unconstrained optimization problem when the penalty parameter is sufficiently large is a local minimizer of the original problem. This result is not available for the constraint transcription approach reported in [116]. This is a major advancement in the study of the solution methods for semi-infinite optimization problems.

In Chapter 3, we consider a general class of nonlinear mixed discrete programming problems in the form given below:

$$\min f(\mathbf{x}, \ \mathbf{y}) \tag{1.38}$$

$$\begin{aligned} \text{subject to} \qquad & H_i(\mathbf{x},\mathbf{y}) = 0 \quad i = 1, 2, \ldots, M, \\ & G_i(\mathbf{x},\mathbf{y}) \leq 0 \quad j = 1, 2, \ldots, N. \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$ and $\mathbf{y} = [y_1, y_2, \ldots, y_m]^\top \in \mathbb{D}_1 \times \cdots \times \mathbb{D}_m$. Here, $\mathbb{R}^n$ is the n-dimensional Euclidean space, and for each $i = 1, 2, \ldots, m$, $\mathbb{D}_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,K_i}\}$, where $a_{i,j}, \ j = 1, \ldots, K_i$, are given discrete values. To solve this problem, we first define, for each

$i = 1, 2, \ldots, m$,

$$\bar{y}_i = \sum_{j=1}^{K_i} a_{i,j} w_{i,j}, \tag{1.39}$$

where, for each $i = 1, 2, \ldots, m$,

$$\sum_{j=1}^{K_i} w_{i,j} = 1, \tag{1.40a}$$

$$0 \leq w_{i,j} \leq 1, \quad j = 1, 2, \ldots, K_i, \tag{1.40b}$$

$$w_{i,j}(1 - w_{i,j}) \leq 0, \quad j = 1, 2, \ldots, K_i. \tag{1.40c}$$

Applying (1.39) and (1.40) to (1.38), we obtain an equivalent continuous nonlinear optimization problem subject to original constraints as well as the newly introduced linear and quadratic constraints. However, in view of the quadratic inequality constraints (1.40c), the equivalent nonlinear constrained optimization problem is very difficult to solve directly by using nonlinear optimization techniques, such as the sequential quadratic programming approximation scheme with active set strategy. This is because they fail to satisfy the linear independent constraint qualification. Thus, a new approach based on the exact penalty function method introduced in Chapter 2 is used to obtain a sequence of unconstrained optimization problems. Each of these unconstrained optimization problem is easier to solve.

In Chapter 4, we investigate the design of allpass variable fractional delay filters with sums of signed powers-of-two coefficients and the least square criterion. The design problem can be categorized as a constrained nonlinear integer programming problem, denoted by Problem **P**, where each coefficient $h_{n,m}$ of the filter can be expressed as

$$h_{n,m} = \sum_{i=1}^{b} d_{i,n,m} 2^{-i}, \tag{1.41}$$

where $d_{i,n,m} \in \{-1, 0, 1\}$, $i = 1, \ldots, b$, and b denotes the number of bits of the wordlength. Clearly, a larger b will give rise to a more accurate approximation. It can be shown that each coefficient has at most $2^{b+1} - 1$ options.

We solve this problem in the following three stages:

i. Consider Problem **P** with its decision variables assumed to take values from $\mathbb{R}$. Let this problem be referred to as Problem $\widehat{\mathbf{P}}$. Find the optimal solution, which is known as the infinite precision optimal solution, of Problem $\widehat{\mathbf{P}}$.

ii. Find a reduced search region around the minimizer of the infinite precision optimal solution obtained in Stage (i).

iii. Find a point that minimizes the objective function within the region obtained in Stage (ii).

For Stage (i), we use an approximation scheme reported in [17]. The objective function is approximated by a quadratic cost function which has a unique optimal solution. Based on this

optimal solution, a good search region containing the global solution is developed in Stage (ii) by using a two-step scheme. Then, in Stage (iii), a new exact penalty function method is proposed to solve the quadratic integer optimization problem containing the obtained search region as part of its constraints.

In Chapter 5, we consider a class of optimal discrete-valued control problems. It has many real world applications such as train control [46], switched amplifier design [110], submarine operation [99], sensor scheduling [126] and hybrid power system design [118, 127]. Our aim is to develop an effective solution method for solving this important class of discrete-valued control problems. To solve an optimal discrete-valued control problem, we need to determine the order in which the different control values operate, as well as the times at which the control switches from one value to the next. Since the ordering of control values is discrete in nature, classical optimal control methods are not applicable to this type of problem. In this chapter, we first apply the transformation reported in [125] so that the discrete-valued control is expressed as a linear combination of piecewise constant controls subject to a linear equality constraint and a set of quadratic inequality constraints. The original problem can then be written equivalently as an optimal control problem with piecewise constant controls subject to the original inequality constraints and the new additional constraints. Then, the time-scaling transformation [62] is applied to the transformed problem, yielding an optimal control problem with piecewise constant controls and fixed switching times. To solve this new problem, we introduce the exact penalty function method reported in Chapter 2 to construct a sequence of penalized optimal control problems. Convergence results show that when the penalty parameter is sufficiently large, a local optimal solution of the penalized problem is also a local optimal solution of the original optimal control problem. This penalized problem can be solved by using optimal control software packages, such as *MISER 3.4* where *fmincon*(MATLAB) (or *NLPQLP*(FORTRAN)) is used in its optimization process. Numerical results obtained from solving two train control problems indicate that this approach is effective.

In the last chapter, we summarize the main contributions of the thesis and discuss some possible future research directions.

CHAPTER 2

A new exact penalty function method for continuous inequality constrained optimization problems

2.1 Introduction

Many real world practical problems in engineering design such as the design of earthquake resistant structures; multi-input multi-output control systems; wide-band amplifiers; and robot trajectory planning, are considered in [45,89–91]. In [14,78], interesting applications in statistics, which include optimal experimental design in regression, constrained multinomial maximum-likelihood estimation, robustness in Bayesian statistics and actuarial risk theory, are investigated.

These problems can generally be formulated as continuous inequality constrained optimization problems in the form given below:

$$\min \ f(\mathbf{x}) \tag{2.1a}$$

$$\text{subject to} \qquad \phi_j(\mathbf{x}, \omega) \leq 0, \ \forall \ \omega \in \Omega, \ j = 1, \ \ldots \ , \ m, \tag{2.1b}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the decision parameter vector, Ω is a compact interval in $\mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in x, and for each $j = 1, \ \ldots \ , \ m$, $\phi_j : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function in $\mathbf{x}$ and ω. Let this problem be referred to as Problem **P**. This problem is also known as a semi-infinite optimization problem (SIP).

Since there are infinite many inequality constraints in (2.1b), it is, in general, impossible to solve Problem **P** analytically. In early 1970s, numerical methods for SIP are proposed in [39]. Since 1980, SIP has become an active research area in optimization both in theory and numerical algorithms. Many important publications have appeared in the literature. Examples include [4, 9, 22, 32, 97], and the relevant references cited therein. There are also several excellent review papers (see, for example, [45, 59, 88]) devoted to SIP.

A popular approach to solve semi-infinite optimization problem (2.1) is to replace the compact set Ω by a finite subset of Ω through certain systematic discretization scheme. This leads to

a problem that has only a finite number of constraints. Then, the resulting conventional problem can be solved by applying appropriate linear or nonlinear programming algorithms. There are basically four types of methods to generate finite subproblems for the original problem — *exchange methods*, *discretization methods*, *dual parametrization methods* and the *methods based on local reduction*. See, for example, [48, 69–71, 73, 88].

Note that the continuous inequality constraints (2.1b) can be written equivalently as

$$\int_{\Omega} \max\{0, \phi_j(\mathbf{x}, \omega)\} d\omega = 0, \quad j = 1, \ \ldots \ , \ m, \tag{2.2}$$

However, $\max\{0, \phi_j(\mathbf{x}, \omega)\}$, $j = 1, \ \ldots \ , \ m$, are non-smooth. Thus, Problem **P** with constraints (2.1b) replaced by their equivalent equality constraints (2.2) cannot be solved by using any smooth gradient-based optimization methods.

In [50], a constrained transcription method is introduced, where the continuous inequality constraints (2.1b) are first transformed into equivalent equality constraints in integral form (2.2). However, the integrands are nonsmooth. Thus, a local smoothing technique is used to approximate these nonsmooth integrands by smooth functions. In this way, Problem **P** is approximated by a sequence of optimization problems involving inequality constraints in integral form, where each of which can be solved by using conventional smooth gradient-based constrained optimization methods. There are two parameters, ϵ and τ, involved in these approximate constrained optimization problems, where $\epsilon > 0$ controls the accuracy of the approximation and $\tau > 0$ controls the feasibility. It is shown in [50] that, for any $\epsilon > 0$, if $\tau > 0$ is sufficiently small, then the solution obtained satisfies the continuous inequality constraints (2.1b). Furthermore, the global optimal solution of the approximate constrained optimization problem converges to the global optimal solution of the original problem as $\epsilon \to 0$. However, it is not known if a local optimal solution of the approximate constrained optimization problem will converge to a local optimal solution of the original problem. In [116], the smooth approximate functions in integral form are appended to the objective function by using the concept of the penalty function. This leads to a sequence of unconstrained optimization problems in integral form, where each of which is solvable by conventional smooth gradient-based unconstrained optimization techniques. Convergence results and the shortcomings similar to those reported in [50] are also valid. In [117, 131], discretization methods are used, and then the nonlinear Lagrangian functions are introduced. For all these algorithms, the feasibility condition is often missed in actual numerical calculation.

In [64, 82, 119, 129], numerical algorithms based on Newton method are developed to solve semi-infinite programming problems, where the *KKT* system is formulated as a system of nonsmooth equations. However, the number of Lagrange multipliers in *KKT* system is not known a priori. For this, a different formulation of *KKT* system is introduced in [26], where the equivalent nonsmooth function of the continuous inequality constraints are approximated by smooth functions. Then, a projected Newton-Type algorithm is used to solve the new *KKT* system.

For a semi-infinite optimization problem, where the objective function is quadratic and the continuous inequality constraints are linear, it is found that dual parametrization methods are effective (see, for example, [48, 69–71, 73]), where the dual problem of the linear-quadratic semi-

infinite optimization problem, called the primal problem, is transformed into an equivalent finite dimensional nonlinear programming problem. The global solution of the primal problem can be obtained from that of the dual problem. However, the dual problems are equally difficult to solve. Thus, discritization schemes of the primal problem are developed, and the corresponding dual formulations called parameterized dual problems are constructed on this basis, efficient computational methods, known as dual parametrization methods, are derived. It is shown in [48] that the suboptimal solutions generated by these dual parametrization methods converge to the optimal solution of the original semi-infinite programming problem.

For optimization problems with conventional smooth inequality constraints, the penalty function method is, in general, recognized as an efficient method. However, to ensure that the solution obtained is feasible, the penalty parameter σ is required to go to $+\infty$. This is clearly unsatisfactory. Thus, an exact penalty function, $f_\sigma(\mathbf{x})$, is introduced for these inequality constrained optimization problems (see, for example, [13] and [106]). A main advantage of the exact penalty function method is that a minimizer of the original problem can be obtained without requiring the penalty parameter σ to go to $+\infty$. In [47], by adding a new variable ϵ, a new exact penalty function, $f_\sigma(\mathbf{x}, \epsilon)$, is introduced to deal the optimization problem with inequality constraints as well as equality constraints, forming a new penalized cost function $f_\sigma(\mathbf{x}, \epsilon)$, where σ is the penalty parameter. Under some mild assumptions, it is shown in [47] that, if the value of the penalty parameter σ is sufficient large, then a local minimizer of the penalty problem such that $f_\sigma(\mathbf{x}^\star, \epsilon^\star)$ is finite is of the form $(\mathbf{x}^\star, 0)$, where $\mathbf{x}^\star$ is a local minimizer of the original problem.

In this chapter, a new exact penalty function approach is proposed for solving semi-infinite optimization problems, where an objective function is to be minimized subject to continuous inequality constraints. It is based on [135,136]. In this approach, the summation of the integrals of some smooth approximation functions is appended to the objective function forming an exact penalty objective function $f_\sigma(\mathbf{x}, \epsilon)$. This gives rise to a sequence of optimization problems subject to $\epsilon > 0$. We shall show that any local minimizer of these optimization problems is a local minimizer of the original problem when the penalty parameter is sufficiently large. This property is not shared by the approaches reported in [116], [117], [50] or [131]. Clearly, this is a major advancement in the study of solution methods for semi-infinite optimization problems.

The rest of the chapter is organized as follows. In Section 2.2, we give a new exact penalty function and analyze its convergent properties. In Section 2.3, we devise an algorithm for solving Problem **P** via solving a sequence of optimization problems subject to $\epsilon > 0$. Several examples are solved by using the algorithm proposed. Section 2.4 concludes the chapter.

2.2 New exact penalty function method

Consider Problem **P**. For each $\mathbf{x} \in \mathbb{R}^n$, $\max\{\phi_j(\mathbf{x}, \omega), 0\}$ is a continuous function of ω, since ϕ_j is continuously differentiable. Define

$$S_\epsilon = \{(\mathbf{x}, \epsilon) \in \mathbb{R}^n \times \mathbb{R}_+ : \phi_j(\mathbf{x}, \omega) \leq \epsilon^\gamma W_j, \ \forall\ \omega \in \Omega, \ j = 1, \ \ldots \ , \ m\}, \tag{2.3}$$

where $\mathbb{R}_+ = \{\alpha \in \mathbb{R} \ : \ \alpha \geq 0\}$, $W_j \in (0, 1)$, $j = 1, \ \ldots \ , \ m$, are fixed constants and γ is a positive real number. Clearly, Problem **P** is equivalent to the following problem, which is denoted as Problem $\hat{\mathbf{P}}$.

$$\min \ f(\mathbf{x}) \tag{2.4a}$$

subject to

$$(\mathbf{x}, \epsilon) \in S_0, \tag{2.4b}$$

where $S_0 = S_\epsilon$ with $\epsilon = 0$.

We assume that the following conditions are satisfied:

(A1). There exists a global minimizer of Problem **P**, implying that $f(\mathbf{x})$ is bounded from below on S_0.

(A2). The number of distinct local minimum values of the objective function of Problem **P** is finite.

Motivated by the exact penalty function introduced in [47] and the constraint transcription method for converting continuous inequality constraints into a sequence of inequality constraints in integral form (see [50] and [138]), we introduce a new exact penalty function $f_\sigma(\mathbf{x}, \epsilon)$ defined below:

$$f_\sigma(\mathbf{x}, \epsilon) = \begin{cases} f(\mathbf{x}), & \text{if } \epsilon = 0, \phi_j(\mathbf{x}, \omega) \leq 0 \ (\omega \in \Omega), \\ f(\mathbf{x}) + \epsilon^{-\alpha}\Delta(\mathbf{x}, \epsilon) + \sigma\epsilon^\beta, & \text{if } \epsilon > 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.5}$$

where $\Delta(\mathbf{x}, \epsilon)$, which is referred to as the constraint violation, is defined by

$$\Delta(\mathbf{x}, \epsilon) = \sum_{j=1}^{m} \int_\Omega \Big[\max\left\{0, \phi_j(\mathbf{x}, \omega) - \epsilon^\gamma W_j\right\}\Big]^2 d\omega, \tag{2.6}$$

α and γ are positive real numbers, $\beta > 2$, and $\sigma > 0$ is a penalty parameter. We now introduce a surrogate optimization problem, which is referred to as Problem $\mathbf{P}_\sigma$, as follows:

$$\min \ f_\sigma(\mathbf{x}, \epsilon) \tag{2.7a}$$

subject to

$$(\mathbf{x}, \epsilon) \in \mathbb{R}^n \times [0, +\infty). \tag{2.7b}$$

Intuitively, during the process of minimizing $f_\sigma(\mathbf{x},\epsilon)$, if σ is increased, ϵ^β should be reduced, meaning that ϵ should be reduced as β is fixed. Thus $\epsilon^{-\alpha}$ will be increased, and hence the constraint violation will also be reduced. This means that the value of

$$\Big[\max\big\{0,\phi_j(\mathbf{x},\omega)-\epsilon^\gamma W_j\big\}\Big]^2$$

must go down, leading to the satisfaction of the continuous inequality constraints, i.e.,

$$\phi_j(\mathbf{x},\omega)\le 0,\ \forall\ \omega\in\Omega,\ j=1,\ \ldots\ ,m.$$

In the next section, we shall show that, under some mild assumptions, if the parameter σ_k is sufficient large ($\sigma_k\to+\infty$ as $k\to+\infty$) and $(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ is a local minimizer of Problem $\mathbf{P}_{\sigma_k}$, then $\epsilon^{(k),\star}\to\epsilon^\star=0$, and $\mathbf{x}^{(k),\star}\to\mathbf{x}^\star$ with $\mathbf{x}^\star$ being a local minimizer of Problem $\mathbf{P}$. The importance of this result is quite obvious.

2.2.1 Convergence analysis

Taking the gradients of $f_\sigma(\mathbf{x},\epsilon)$ with respect to x and ϵ gives

$$\frac{\partial f_\sigma(\mathbf{x},\epsilon)}{\partial\mathbf{x}}=\frac{\partial f(\mathbf{x})}{\partial\mathbf{x}}+2\epsilon^{-\alpha}\sum_{j=1}^m\int_\Omega\max\big\{0,\phi_j(\mathbf{x},\omega)-\epsilon^\gamma W_j\big\}\frac{\partial\phi_j(\mathbf{x},\omega)}{\partial\mathbf{x}}d\omega, \tag{2.8}$$

$$\begin{aligned}\frac{\partial f_\sigma(\mathbf{x},\epsilon)}{\partial\epsilon} &= -\alpha\epsilon^{-\alpha-1}\sum_{j=1}^m\int_\Omega\Big[\max\big\{0,\phi_j(\mathbf{x},\omega)-\epsilon^\gamma W_j\big\}\Big]^2d\omega\\ &\quad -2\gamma\epsilon^{\gamma-\alpha-1}\sum_{j=1}^m\int_\Omega\max\big\{0,\phi_j(\mathbf{x},\omega)-\epsilon^\gamma W_j\big\}W_j d\omega+\sigma\beta\epsilon^{\beta-1}\\ &= \epsilon^{-\alpha-1}\Big\{-\alpha\sum_{j=1}^m\int_\Omega\Big[\max\big\{0,\phi_j(\mathbf{x},\omega)-\epsilon^\gamma W_j\big\}\Big]^2d\omega\\ &\quad +2\gamma\sum_{j=1}^m\int_\Omega\max\big\{0,\phi_j(\mathbf{x},\omega)-\epsilon^\gamma W_j\big\}(-\epsilon^\gamma W_j)d\omega\Big\}+\sigma\beta\epsilon^{\beta-1}.\end{aligned} \tag{2.9}$$

For every positive integer k, let $(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ be a local minimizer of Problem $\mathbf{P}_{\sigma_k}$.
To obtain our main result, we need

Lemma 2.1. *Let $(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ be a local minimizer of Problem $\mathbf{P}_{\sigma_k}$. Suppose that $f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ is finite and that $\epsilon^{(k),\star}>0$. Then*

$$(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})\notin S_\epsilon,$$

where S_ϵ is defined by (2.3).

Proof. Since $(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ is a local minimizer of Problem $\mathbf{P}_{\sigma_k}$ and $\epsilon^{(k),\star}>0$, we have

$$\frac{\partial f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})}{\partial\epsilon}=0. \tag{2.10}$$

On the contrary, we assume that the conclusion of the lemma is false. Then, we have

$$\phi_j(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \leq (\epsilon^{(k),\star})^{\gamma} W_j, \forall\ \omega \in \Omega,\ j = 1,\ \ldots\ , m.$$

Thus, by (2.9) and (2.10), we obtain

$$0 = \frac{\partial f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})}{\partial \epsilon} = \beta \sigma_k \epsilon^{\beta-1} > 0.$$

This is a contradiction, and hence completing the proof. □

To continue, we introduce

Definition 2.1. *It is said that the constraint qualification is satisfied for the continuous inequality constraints (2.1b) at* $\mathbf{x} = \bar{\mathbf{x}}$ *, if the following implication is valid. Suppose that*

$$\int_{\Omega} \sum_{j} \varphi_j(\omega) \frac{\partial \phi_j(\bar{x}, \omega)}{\partial x} d\omega = 0.$$

Then, $\varphi_j(\omega) = 0$, $\forall \omega \in \Omega$, $j = 1, \ldots, m$.

Let the conditions of Lemma 2.1 be satisfied. Then, we have

Theorem 2.1. *Suppose that* $(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})$ *is a local minimizer of Problem* $\mathbf{P}_{\sigma_k}$ *such that* $f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})$ *is finite. If* $(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \to (\mathbf{x}^{\star}, \epsilon^{\star})$ *as* $k \to +\infty$*, and the constraint qualification is satisfied for the continuous inequality constraints (2.1b) at* $\mathbf{x} = \mathbf{x}^{\star}$*, then* $\epsilon^{\star} = 0$ *and* $\mathbf{x}^{\star} \in S_0$.

Proof. From Lemma 2.1, it follows that $(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \notin S_{\epsilon^{(k),\star}}$. Furthermore,

$$\begin{aligned} & \frac{\partial f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})}{\partial \mathbf{x}} \\ = & \frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}} \\ & +2(\epsilon^{(k),\star})^{-\alpha} \sum_{j=1}^{m} \int_{\Omega} \max\left\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\right\} \frac{\partial \phi_j(\mathbf{x}^{(k),\star}, \omega)}{\partial \mathbf{x}} d\omega \\ = & \ 0, \end{aligned} \tag{2.11}$$

$$\begin{aligned} & \frac{\partial f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})}{\partial \epsilon} \\ = & -\alpha (\epsilon^{(k),\star})^{-\alpha-1} \sum_{j=1}^{m} \int_{\Omega} \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\big\} \Big]^2 d\omega \\ & - 2\gamma (\epsilon^{(k),\star})^{\gamma-\alpha-1} \sum_{j=1}^{m} \int_{\Omega} \max\big\{0, \phi_j(\mathbf{x}^{(k),*}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\big\} W_j d\omega \\ & + \sigma_k \beta (\epsilon^{(k),\star})^{\beta-1} \\ = & (\epsilon^{(k),\star})^{-\alpha-1} \Big\{ -\alpha \sum_{j=1}^{m} \int_{\Omega} \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\big\} \Big]^2 d\omega \end{aligned} \tag{2.12}$$

$$+ 2\gamma \sum_{j=1}^{m} \int_{\Omega} \max\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\}(-(\epsilon^{(k),\star})^{\gamma} W_j) d\omega \Big\}$$
$$+ \sigma_k \beta (\epsilon^{(k),\star})^{\beta-1}$$
$$= 0.$$

Suppose that $\epsilon^{(k),\star} \to \epsilon^{\star} \neq 0$. Then, by (2.12), we observe that its first term tends to a finite value, while the last term tends to infinity as $\sigma_k \to +\infty$, when $k \to +\infty$. This is impossible for the validity of (2.12). Thus, $\epsilon^{\star} = 0$.
Now, by (2.11), we obtain

$$(\epsilon^{(k),\star})^{\alpha} \frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}} + 2 \sum_{j=1}^{m} \int_{\Omega} \max\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\} \frac{\partial \phi_j(\mathbf{x}^{(k),\star}, \omega)}{\partial \mathbf{x}} d\omega$$
$$= 0. \tag{2.13}$$

Thus,

$$\lim_{k \to +\infty} \Big\{ (\epsilon^{(k),\star})^{\alpha} \frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}} + 2 \sum_{j=1}^{m} \int_{\Omega} \max\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma} W_j\} \frac{\partial \phi_j(\mathbf{x}^{(k),\star}, \omega)}{\partial \mathbf{x}} d\omega \Big\} \tag{2.14}$$
$$= 2 \sum_{j=1}^{m} \int_{\Omega} \max\{0, \phi_j(\mathbf{x}^{\star}, \omega)\} \frac{\partial \phi_j(\mathbf{x}^{\star}, \omega)}{\partial \mathbf{x}} d\omega = 0.$$

Since the constraint qualification is satisfied for the continuous inequality constraints (2.1b) at $\mathbf{x} = \mathbf{x}^{\star}$, it follows that, for each $j = 1, \ldots, m$,

$$\max\{0, \phi_j(\mathbf{x}^{\star}, \omega)\} = 0,$$

for each $\omega \in \Omega$. This, in turn, implies that, for each $j = 1, \ldots, m$, $\phi_j(\mathbf{x}^{\star}, \omega) \leq 0$, $\forall\, \omega \in \Omega$. The proof is completed. □

Corollary 2.1. *If* $\mathbf{x}^{(k),\star} \to \mathbf{x}^{\star} \in S_0$ *and* $\epsilon^{(k),\star} \to \epsilon^{\star} = 0$, *then* $\Delta(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \to \Delta(\mathbf{x}^{\star}, \epsilon^{\star}) = 0$.

Proof. The conclusion follows readily from the definition of $\Delta(\mathbf{x}, \epsilon)$ and the continuity of $\phi_j(\mathbf{x}, \omega)$. □

Remark 2.1. The existence of an accumulating point of the sequence $(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})$ is assured if the following condition is satisfied

$$f(\mathbf{x}) \to \infty, \text{ as } |\mathbf{x}| \to \infty.$$

In [47], the construction of the form of the exact penalty function $f_{\sigma}(\mathbf{x}, \omega)$ is such that it is continuously differentiable in S_{ϵ} when $\epsilon > 0$. Its limit is continuous on the part of the boundary when its values are finite. In particular, $f_{\sigma}(\mathbf{x}, 0)$ is finite when x is such that $\phi_j(\mathbf{x}, \omega) \leq 0$,

$\forall\, \omega \in \Omega, j = 1, \ldots, m$. In what follows, we shall turn our attention to the exact penalty function constructed in (2.5). We shall see that, under some mild conditions, $f_\sigma(\mathbf{x}, \omega)$ is continuously differentiable with continuous limits.

Theorem 2.2. *Assume that* $\phi_j(\mathbf{x}^{(k),\star}, \omega) = o((\epsilon^{(k),\star})^\delta)$, $\delta > 0$, $j = 1, \ldots, m$. *Suppose that* $\gamma > \alpha$, $\delta > \alpha$, $-\alpha - 1 + 2\delta > 0$, $2\gamma - \alpha - 1 > 0$. *Then*

$$f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \xrightarrow[\mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0]{\epsilon^{(k),\star} \to \epsilon^\star = 0} f_{\sigma_k}(\mathbf{x}^\star, 0) = f(\mathbf{x}^\star), \tag{2.15}$$

$$\nabla_{(\mathbf{x},\epsilon)} f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \xrightarrow[\mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0]{\epsilon^{(k),\star} \to \epsilon^\star = 0} \nabla_{(\mathbf{x},\epsilon)} f_{\sigma_k}(\mathbf{x}^\star, 0) = (\nabla f(\mathbf{x}^\star), 0). \tag{2.16}$$

Proof. By virtue of the conditions of the theorem, it follows that, for $\epsilon \neq 0$,

$$\begin{aligned}
& \lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) \\
& \lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} \Big\{ f(\mathbf{x}^{(k),\star}) \\
= \quad & + (\epsilon^{(k),\star})^{-\alpha} \sum_{j=1}^m \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^\gamma W_j\big\} \Big]^2 d\omega + \sigma_k (\epsilon^{(k),\star})^\beta \Big\} \\
= \quad & f(\mathbf{x}^\star) + \lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} \frac{\sum_{j=1}^m \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^\gamma W_j\big\} \Big]^2 d\omega}{(\epsilon^{(k),\star})^\alpha}.
\end{aligned} \tag{2.17}$$

For the second term of (2.17), it is clear from Lemma 2.1 that

$$\begin{aligned}
& \lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} \frac{\sum_{j=1}^m \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^\gamma W_j\big\} \Big]^2 d\omega}{(\epsilon^{(k),\star})^\alpha} \\
= \quad & \lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} \sum_{j \in J'} \int_\Omega \big[(\epsilon^{(k),\star})^{-\frac{\alpha}{2}} \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma - \frac{\alpha}{2}} W_j \big]^2 d\omega.
\end{aligned} \tag{2.18}$$

Here, J' denotes the index set such that for any $j \in J'$, $\max\big\{0, \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^\gamma W_j\big\} = \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^\gamma W_j$. Since $\gamma > \alpha$ and $\delta > \alpha$, we have

$$\lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} \sum_{j \in J'} \int_\Omega \big[(\epsilon^{(k),\star})^{-\frac{\alpha}{2}} \phi_j(\mathbf{x}^{(k),\star}, \omega) - (\epsilon^{(k),\star})^{\gamma - \frac{\alpha}{2}} W_j \big]^2 d\omega = 0. \tag{2.19}$$

Combining (2.17) and (2.19) gives

$$\lim_{\substack{\epsilon^{(k),\star} \to \epsilon^\star = 0 \\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0}} f_{\sigma_k}(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star}) = f_{\sigma_k}(\mathbf{x}^\star, 0) = f(\mathbf{x}^\star). \tag{2.20}$$

Similarly, we have

$$\begin{aligned} & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \nabla_{(\mathbf{x},\epsilon)} f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star}) \\ = & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \Big[\nabla_{\mathbf{x}} f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})\ \nabla_{\epsilon} f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})\Big]^T. \end{aligned} \tag{2.21}$$

where

$$\begin{aligned} & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \nabla_{\mathbf{x}} f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star}) \\ = & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \Big\{\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}} \\ & +2(\epsilon^{(k),\star})^{-\alpha}\sum_{j=1}^{m}\int_\Omega \max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^\gamma W_j\big\}\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}d\omega\Big\} \\ = & \nabla_{\mathbf{x}} f(\mathbf{x}^\star)+\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} 2\sum_{j\in J'}\int_\Omega\big[(\epsilon^{(k),\star})^{-\alpha}\phi_j(\mathbf{x}^{(k),\star},\omega) \\ & -(\epsilon^{(k),\star})^{\gamma-\alpha}W_j\big]\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}d\omega \\ = & \nabla_{\mathbf{x}} f(\mathbf{x}^\star). \end{aligned} \tag{2.22}$$

while

$$\begin{aligned} & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \nabla_{\epsilon} f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star}) \\ = & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \Big\{(\epsilon^{(k),\star})^{-\alpha-1}\Big\{-\alpha\sum_{j=1}^{m}\int_\Omega\Big[\max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^\gamma W_j\big\}\Big]^2 d\omega \\ & +2\gamma\sum_{j=1}^{m}\int_\Omega\max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^\gamma W_j\big\}(-(\epsilon^{(k),\star})^\gamma W_j)d\omega\Big\} \\ & +\sigma_k\beta(\epsilon^{(k),\star})^{\beta-1}\Big\} \\ = & \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{x}^{(k),\star}\to\mathbf{x}^\star\in S_0}} \Big\{-\alpha\sum_{j\in J'}\int_\Omega\big[\phi_j(\mathbf{x}^{(k),\star},\omega)(\epsilon^{(k),\star})^{-\frac{\alpha+1}{2}}-(\epsilon^{(k),\star})^{\gamma-\frac{\alpha+1}{2}}W_j\big]^2 d\omega \\ & +2\gamma\sum_{j\in J'}\int_\Omega\big[\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^\gamma W_j\big]\big(-(\epsilon^{(k),\star})^\gamma W_j\big)(\epsilon^{(k),\star})^{-\alpha-1}d\omega\Big\} \\ = & 0. \end{aligned} \tag{2.23}$$

Thus, the proof is completed. □

Theorem 2.3. *There exists a $k_0 > 0$, such that for any $k \geq k_0$, every local minimizer $(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ of the penalty problem with finite $f_{\sigma_k}(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ has the form $(\mathbf{x}^\star,0)$ where $\mathbf{x}^\star$ is a local minimizer of Problem* $\mathbf{P}$.

Proof. On the contrary, we assume that the conclusion is false. Then, there exists a subsequence of $\{(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})\}$, which is denoted by the original sequence, such that for any $k_0 > 0$, there

exists a $k' > k_0$ satisfying $\epsilon^{(k'),*} \neq 0$. By Theorem 2.1, we have

$$\epsilon^{(k),\star} \to \epsilon^\star = 0,\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0,\ \text{as } k \to +\infty.$$

Since $\epsilon^{(k),\star} \neq 0$ for all k, it follows from dividing (2.12) by $(\epsilon^{(k),\star})^{\beta-1}$ that

$$\begin{aligned}(\epsilon^{(k),\star})^{-\alpha-\beta}\Big\{ &-\alpha \sum_{j=1}^{m} \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\Big]^2 d\omega \\ &+2\gamma \sum_{j=1}^{m} \int_\Omega \max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\big(-(\epsilon^{(k),\star})^\gamma W_j\big) d\omega\Big\} + \sigma_k\beta = 0.\end{aligned} \tag{2.24}$$

This is equivalent to

$$\begin{aligned}&(\epsilon^{(k),\star})^{-\alpha-\beta}\Big\{ -\alpha \sum_{j=1}^{m} \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\Big]^2 d\omega \\ &+2\gamma \sum_{j=1}^{m} \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\big(-(\epsilon^{(k),\star})^\gamma W_j\big) \\ &+\max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\phi_j(\mathbf{x}^{(k),\star},\omega) \\ &-\max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\phi_j(\mathbf{x}^{(k),\star},\omega)\Big] d\omega\Big\} + \sigma_k\beta = 0.\end{aligned} \tag{2.25}$$

Rearranging (2.25) yields

$$\begin{aligned}&(\epsilon^{(k),\star})^{-\alpha-\beta}(2\gamma-\alpha)\Big\{ \sum_{j=1}^{m} \int_\Omega \Big[\max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) \\ &\qquad -(\epsilon^{(k),\star})^\gamma W_j\big\}\Big]^2 d\omega\Big\} + \sigma_k\beta \\ =\ & 2\gamma(\epsilon^{(k),\star})^{-\alpha-\beta} \sum_{j=1}^{m} \int_\Omega \max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\phi_j(\mathbf{x}^{(k),\star},\omega) d\omega.\end{aligned} \tag{2.26}$$

Letting $k \to +\infty$ in (2.26) gives

$$2\gamma(\epsilon^{(k),\star})^{-\alpha-\beta} \sum_{j=1}^{m} \int_\Omega \max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\}\phi_j(\mathbf{x}^{(k),\star},\omega) d\omega \to +\infty. \tag{2.27}$$

Define

$$y^k = (\epsilon^{(k),\star})^{-\alpha-\beta} \sum_{j=1}^{m} \int_\Omega \max\big\{0, \phi_j(\mathbf{x}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma W_j\big\} d\omega. \tag{2.28}$$

From (2.27) and (2.28), we have

$$y^k \to +\infty\ ,\ \text{as } k \to +\infty. \tag{2.29}$$

Define

$$z^k = y^k/|y^k|. \tag{2.30}$$

Clearly

$$\lim_{k\to+\infty}|z^k| = |z^*| = 1. \tag{2.31}$$

Dividing (2.13) by $|y^k|$ yields

$$\begin{aligned}\frac{\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}}}{|y^k|} + \frac{2(\epsilon^{(k),\star})^{-\alpha}}{|y^k|}\sum_{j=1}^{m}\int_{\Omega}\max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)\\ -(\epsilon^{(k),\star})^{\gamma}W_j\big\}\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}d\omega = 0.\end{aligned} \tag{2.32}$$

Note that $\mathbf{x}^{(k),\star}\to\mathbf{x}^{\star}$ as $k\to+\infty$ and that $\dfrac{\partial f(\mathbf{x})}{\partial\mathbf{x}}$ and, for each $j=1,\ \ldots,\ m$, ϕ_j and $\dfrac{\partial\phi_j(\cdot\ ,\omega)}{\partial\mathbf{x}}$ are continuous in $\mathbb{R}^n$ for each $\omega\in\Omega$, where Ω is a compact set. Then, it can be shown that there exist constants $\hat{K}$ and $\overline{K}$, independent of k, such that, for all $k=1,2,\cdots$,

$$|\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial\mathbf{x}}| \le \hat{K}, \tag{2.33}$$

$$|\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}| \le \overline{K},\ \text{for } j=1,\cdots,m. \tag{2.34}$$

By substituting (2.28) and (2.30) into (2.32), we obtain

$$\begin{aligned}\frac{\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}}}{|y^k|(\epsilon^{(k),\star})^{\beta}} + \frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|}\sum_{j=1}^{m}\int_{\Omega}\max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)\\ -(\epsilon^{(k),\star})^{\gamma}W_j\big\}\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}d\omega = 0.\end{aligned} \tag{2.35}$$

Note that

$$\begin{aligned}\frac{1}{|y^k|(\epsilon^{(k),\star})^{\beta}} &= \frac{1}{|(\epsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\big\}d\omega|(\epsilon^{(k),\star})^{\beta}}\\ &= \frac{1}{|\sum_{j=1}^{m}\int_{\Omega}\max\big\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\big\}d\omega|(\epsilon^{(k),\star})^{-\alpha}}.\end{aligned} \tag{2.36}$$

From Theorem 2.2, we have $\phi_j(\mathbf{x}^{(k),\star},\omega)=o((\epsilon^{(k),\star})^{\delta})$ and $\gamma>\alpha,\ \delta>\alpha$. Thus

$$\begin{aligned}
&\lim_{k\to+\infty}|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\right\}d\omega|(\epsilon^{(k),\star})^{-\alpha}\\
=&\lim_{k\to+\infty}|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,o((\epsilon^{(k),\star})^{\delta})-(\epsilon^{(k),\star})^{\gamma}W_j\right\}d\omega|(\epsilon^{(k),\star})^{-\alpha}\\
=&\lim_{k\to+\infty}|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,o((\epsilon^{(k),\star})^{\delta})(\epsilon^{(k),\star})^{-\alpha}-(\epsilon^{(k),\star})^{\gamma-\alpha}W_j\right\}d\omega|\\
=&\lim_{k\to+\infty}|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,\frac{o((\epsilon^{(k),\star})^{\delta})}{(\epsilon^{(k),\star})^{\delta}}(\epsilon^{\star})^{\delta-\alpha}-(\epsilon^{\star})^{\gamma-\alpha}W_j\right\}d\omega|\\
=&\ 0,
\end{aligned}\tag{2.37}$$

and hence,

$$\lim_{k\to\infty}\frac{1}{|y^k|(\epsilon^{(k),\star})^{\beta}}\to+\infty.\tag{2.38}$$

From (2.33) and (2.38), it is clear that

$$\frac{|\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial\mathbf{x}}|}{|y^k|(\epsilon^{(k),\star})^{\beta}}\to+\infty,\quad k\to+\infty.\tag{2.39}$$

On the other hand,

$$\begin{aligned}
&\Big|\frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\right\}\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}d\omega\Big|\\
\le&\frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|}\sum_{j=1}^{m}\int_{\Omega}\Big|\max\left\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\right\}\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}\Big|d\omega\\
=&\frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\right\}\Big|\frac{\partial\phi_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}\Big|d\omega\\
\le&\frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,\phi_j(\mathbf{x}^{(k),\star},\omega)-(\epsilon^{(k),\star})^{\gamma}W_j\right\}\overline{K}d\omega\\
=&\ 2\overline{K}z^k,
\end{aligned}\tag{2.40}$$

where z^k is defined by (2.30). Clearly, $|z^k|=1$. Thus, it follows from (2.40) that $2\overline{K}z^k$ is bounded uniformly with respect to k. This together with (2.39) is a contradiction to (2.35), and hence completing the first part of the proof .

For sufficiently large k, every local minimizer $(\mathbf{x}^{(k),\star},\epsilon^{(k),\star})$ has the form $(\mathbf{x}^{\star},0)$. It is obvious from Theorem 2.1 that $\mathbf{x}^{\star}$ is a feasible point of Problem **P**. This indicates that there is a neighborhood of $\mathbf{x}^{\star}$, such that for any feasible x of Problem **P**

$$f(\mathbf{x})=f_{\sigma_k}(\mathbf{x},0)\ge f_{\sigma_k}(\mathbf{x}^{\star},0)=f(\mathbf{x}^{\star}).$$

Therefore, $\mathbf{x}^{\star}$ is a local minimizer of Problem **P**. This completes the proof. □

We may now conclude that, under some mild assumptions and the constraint qualification condition, when the parameter σ is sufficiently large, a local minimizer of Problem $\mathbf{P}_\sigma$ is a local minimizer of Problem $\mathbf{P}$.

Results presented in Theorem 2.1, Corollary 2.1, Theorem 2.2 and Theorem 2.3. form the foundation for constructing a computational method to be presented in Section 2.3.

2.3 Algorithm and numerical results

Here, we use the optimization tool box *fmincon* within MATLAB environment to solve the optimization Problem $\mathbf{P}_\sigma$, where the integral appeared in $f_\sigma(\mathbf{x}, \epsilon)$ is calculated by using the *Simpson's Rule*. For *Simpson's Rule*, the global error is of order h^4, where h is the discretization step size. Thus, the required accuracy of the integrations can be easily achieved if the discretization step size is sufficient small.

In the following, we give definitions to the terms used.

σ − The penalty parameter which is to be increased in every iteration.

$\bar{\omega}$ −The point at which $\max\limits_{1\le j\le m} \phi_j(\mathbf{x}^{(k),\star}, \bar{\omega}) = \max\limits_{1\le j\le m} \max\limits_{\omega\in\Omega} \phi_j(\mathbf{x}^{(k),\star}, \omega)$.

g − The value of $\max\limits_{1\le j\le m} \max\limits_{\omega\in\Omega} \phi_j(\mathbf{x}^{(k),\star}, \omega)$.

f − The objective function value.

ϵ − A new variable which is introduced in the construction of the exact penalty function.

$\epsilon^\star$ − A lower bound of $\epsilon^{(k),\star}$, which is introduced for avoiding $\epsilon^{(k),\star} \to 0$.

With the new exact penalty function, we can construct an efficient algorithm, which is given below:

Algorithm 2.1

Step 1 set $\sigma^{(1)} = 10$, $\epsilon^{(1)} = 0.1$, $\epsilon^\star = 10^{-9}$, $\beta > 2$, choose an initial point $(\mathbf{x}_0, \epsilon_0)$, the iteration index $k = 0$. The values of γ and α are chosen depending on the specific structure of Problem $\mathbf{P}$ concerned.

Step 2 Solve Problem $\mathbf{P}_{\sigma_k}$, and let $(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})$ be the minimizer obtained.

Step 3 If $\epsilon^{(k),\star} > \epsilon^\star$, $\sigma^{(k)} < 10^8$,

set $\sigma^{(k+1)} = 10 \times \sigma^{(k)}$, $k := k + 1$. Go to **Step 2** with $(\mathbf{x}^{(k),\star}, \epsilon^{(k),\star})$ as the new initial point in the new optimization process

Else set $\epsilon^{(k),\star} := \epsilon^\star$, then go to **Step 4**

Step 4 Check the feasibility of $\mathbf{x}^{(k),\star}$ (i.e., whether or not $\max\limits_{1\le j\le m} \max\limits_{\omega\in\Omega} \phi_j(\mathbf{x}^{(k),\star}, \omega) \le 0$).

If $\mathbf{x}^{(k),\star}$ is feasible, then it is a local minimizer of Problem $\mathbf{P}$. Exit.

Else go to **Step 5**

Step 5: Adjust the parameters α, β and γ such that conditions of Lemma 2.1 are satisfied. Set $\sigma^{(k+1)} = 10\sigma^{(k)}$, $\epsilon^{(k+1)} = 0.1\epsilon^{(k)}$, $k := k + 1$. Go to **Step 2**.

Remark 2.2. In **Step 3**, if $\epsilon^{(k),\star} > \epsilon^\star$, we obtain from Theorem 2.1 and Theorem 2.3 that $\mathbf{x}^{(k),\star}$ is not a feasible point. This means that the penalty parameter σ may not be large enough.

Thus we need to increase σ. If $\sigma_k > 10^8$, but still $\epsilon^{(k),\star} > \epsilon^\star$, then we should adjust the value of α, β and γ, such that conditions assumed in Theorem 2.2 are satisfied. Go to **Step 2**.

Remark 2.3. Clearly, we cannot check the feasibility of $\phi_j(\mathbf{x}, \omega) \leq 0$, $j = 1, \ \ldots, \ m$, for every $\omega \in \Omega$. In practice, we choose a set $\hat{\Omega}$, which contains a dense enough of points in Ω. Check the feasibility of $\phi_j(\mathbf{x}, \omega) \leq 0$ over $\hat{\Omega}$ for each $j = 1, \ \ldots, \ m$.

Remark 2.4. Although we have proved that a local minimizer of the exact penalty function optimization problem $\mathbf{P}_{\sigma_k}$ will converge to a local minimizer of the original Problem $\mathbf{P}$, we need, in actual computation, set a lower bound $\epsilon^\star = 10^{-9}$ for $\epsilon^{(k),\star}$ so as to avoid the situation of being divided by $\epsilon^{(k),\star} = 0$, leading to infinity.

Example 2.1. The following example is taken from [35], and it is also used for testing the numerical algorithms in [116, 117, 131]. In this problem, the objective function:

$$f(\mathbf{x}) = \frac{x_2(122 + 17x_1 + 6x_3 - 5x_2 + x_1x_3) + 180x_3 - 36x_1 + 1224}{x_2(408 + 56x_1 - 50x_2 + 60x_3 + 10x_1x_3 - 2x_1^2)} \tag{2.41}$$

is to be minimized subject to

$$\phi(\mathbf{x}, \omega) \leq 0 \ , \ \forall \ \omega \in \Omega, \tag{2.42}$$

$$0 \leq x_1, x_3 \leq 100, \ 0.1 \leq x_2 \leq 100, \tag{2.43}$$

where $\Omega = [10^{-6}, 30]$ and $(i = \sqrt{-1})$, while

$$\phi(\mathbf{x}, \omega) = \Im T(\mathbf{x}, \omega) - 3.33[\Re T(\mathbf{x}, \omega)]^2 + 1.0,$$

$$T(\mathbf{x}, \omega) = 1 + H(x, i\omega)G(i\omega),$$

$$H(\mathbf{x}, s) = x_1 + x_2/s + x_3 s,$$

$$G(s) = \frac{1}{(s+3)(s^2 + 2s + 2)}.$$

Here, $\Im T(\mathbf{x}, \omega)$ and $\Re T(\mathbf{x}, \omega)$ are, respectively, the imaginary and real parts of $T(\mathbf{x}, \omega)$. The initial point is $[50, 50, 50]^\top$. Actually, we can start from any point within the boundedness constraints (2.43).

For the continuous inequality constraint (2.42), the corresponding exact penalty function $f_\sigma(\mathbf{x}, \epsilon)$ is defined by (2.5) with the constraint violation $\Delta(\mathbf{x}, \epsilon)$ given by

$$\Delta(\mathbf{x}, \epsilon) = \int_\Omega \Big[\max \left\{ 0, \Im T(\mathbf{x}, \omega) - 3.33[\Re T(\mathbf{x}, \omega)]^2 + 1.0 - \epsilon^\gamma W_j \right\} \Big]^2 d\omega.$$

Simpson's Rule with $\Omega = [10^{-6}, 30]$ being divided into 3000 equal subintervals is used to evaluate the integral. The value obtained is highly accurate. Also, these discretized points define a dense subset $\hat{\Omega}$ of Ω. We check the feasibility of the continuous inequality constraint by evaluating the values of the function ϕ over $\hat{\Omega}$. Results obtained are given in Table 2.1 and Table 2.2.

As we can see, as the penalty parameter, σ, is increased, the minimizer approaches to the boundary of the feasible region. When σ is sufficiently large, we obtain a feasible point. It has

σ	$\bar{\omega}$	g	f
10	5.35	1.7599e-005	0.178251096
10^2	5.64	8.2356e-006	0.174782133
10^3	5.63	-2.0612e-005	0.174778004

Table 2.1: Result for Example 2.1

σ	x_1	x_2	x_3	ϵ
10	21.796685	49.5750243	31.7018582	0.000264
10^2	17.3494249	48.9435269	34.5556544	0.0001
10^3	17.3937883	48.7713471	34.5227014	0.00001

Table 2.2: Result for Example 2.1

the same objective function value as that obtained in [117]. However, for the minimizer obtained in [117], there are some minor violations of the continuous inequality constraints (2.42).

Example 2.2. Consider the problem:

$$\begin{array}{ll} \min & x_1^2 + (x_2 - 3)^2 \\ \text{subject to} & x_2 - 2 + x_1 \sin(\frac{t}{x_2 - \omega}) \leq 0, \ \forall \ t \in [0, \pi] \\ & -1 \leq x_1 \leq 1, \ 0 \leq x_2 \leq 2. \end{array}$$

where ω is a parameter which controls the frequency of the constraint. As in [117], ω is chosen as 2.032.

In this case, the corresponding exact penalty function $f_\sigma(\mathbf{x}, \epsilon)$ is defined by (2.5) with the constraint violation given by

$$\Delta(\mathbf{x}, \epsilon) = \int_0^\pi \Big[\max \big\{0, x_2 - 2 + x_1 \sin(\frac{t}{x_2 - \omega}) - \epsilon^\gamma W_j \big\} \Big]^2 dt.$$

Simpson's Rule with interval $[0, \pi]$ being divided into 1000 equal subintervals is used to evaluate the integral. These discretized points also form a dense subset $\hat{\Omega}$ of the interval $[0, \pi]$. The feasibility check is carried over $\hat{\Omega}$. By using Algorithm 2.1 with the initial point taken as (x_1^0, x_2^0), the solution obtained is $(x_1^\star, x_2^\star) = [0, 2]^\top$ with the objective function value $f^\star = 1$. The results are presented in Table 2.3 and Table 2.4.

σ	$\bar{\omega}$	g	f
10	1.41	3.735773915e-008	1.000000669
10^2	1.41	3.735773916e-008	1.000006691
10^3	1.41	3.735773916e-008	1.00006691
10^4	1.41	3.735773916e-008	1.000669101
10^5	1.049	2.45667159e-007	1.000011501

Table 2.3: Result for Example 2.2

σ	x_1	x_2	ϵ
10	3.735773981e-008	2.0000	5.481e-004
10^2	3.735773981e-008	2.0000	5.481e-004
10^3	3.735773981e-008	2.0000	5.481e-004
10^4	3.735773981e-008	2.0000	5.481e-004
10^5	-5.504846644e-006	1.9999	10^{-7}

Table 2.4: Result for Example 2.2

It is observed that for sufficiently large σ, the minimizer obtained by the proposed method has the same minimum with the results obtained in [117]. Moreover, the continuous inequality constraints are satisfied for all $t \in [0, \pi]$.

Example 2.3. Consider the problem:

$$\begin{array}{ll} \min & (x_1 + x_2 - 2)^2 + (x_1 - x_2)^2 + 30[\min\{0, x_1 - x_2\}]^2 \\ \text{subject to} & x_1 \cos t + x_2 \sin t - 1 \leq 0, \ \forall\ t \in [0, \pi]. \end{array}$$

Again, *Simpson's Rule* with the interval $[0, \pi]$ being partitioned into 1000 equal subintervals is used to evaluate the corresponding constraint violation in the exact penalty function. These discretized points also define a dense subset $\hat{\Omega}$ of the interval $[0, \pi]$, which is also used for checking the feasibility of the continuous inequality constraint. Now, by using Algorithm 2.1 with the initial point taken as $[0.5, 0.5]^\top$, the result obtained are reported in Table 2.5 and Table 2.6.

σ	$\bar{\omega}$	g	f
10	0.786	0.02497208416	0.3292584852
10^2	0.786	0.00400356933	0.3409679661
10^3	0.78	-0.00029665527	0.3437506884
10^4	0.78	-0.00000024678	0.3432592109

Table 2.5: Result for Example 2.3

σ	x_1	x_2	ϵ
10	0.7247764975	0.7247530305	0.04447211922
10^2	0.7100525572	0.7098229283	0.006961707112
10^3	0.7113565666	0.7024091525	0.000000009999
10^4	0.7115629913	0.7026219620	0.00000000100

Table 2.6: Result for Example 2.3

By comparing our results with those obtained in [35, 50, 116, 117], it is observed that the objective values are almost the same. However, for our minimizer, it is a feasible point while those obtained in [35, 50, 116, 117] are not.

2.4 Conclusions

In this chapter, a new exact penalty method is proposed for solving an optimization problem with continuous inequality constraints. Compared with the existing schemes, our algorithm can be classified as an outer approximation method as the optimal solution is approached from outside the feasible region. Thus, there is no need to find an interior point to start with. Furthermore, our method is based on exact penalty function, so the penalty parameter σ doesn't need to go to ∞. Furthermore, any local minimizer of the penalized optimization problems is also a local minimizer of the original semi-infinite optimization problem when the penalty parameter is sufficiently large. This represents an important advancement in the solution method of semi-infinite optimization problems. From the numerical simulation, we observe that the minimizers obtained for all the test examples are feasible. This is an important feature of the method proposed, indicating that the proposed exact penalty method is effective when compared with other existing methods.

CHAPTER 3

An exact penalty function method for nonlinear mixed discrete programming problems

3.1 Introduction

For a vast number of applications in areas such as engineering design, computational chemistry, computational biology, communications and finance, some of the decision variables are continuous, while others are to be chosen from sets of discrete values. These problems can be formulated as mixed discrete nonlinear programming problem (MDNLP). In [53], an overview of applications of MDNLP is given, which include process design, process synthesis, process operations, facility location and allocation, facility planning and scheduling, topology of transportation networks, combinatorial optimization problems and other bilinear problems. For other applications, see, for example, [11, 28, 65, 107].

In a MDNLP, there involve discrete-valued variables. Thus, traditional gradient-based methods are not applicable. Theoretically, MDNLP is NP-hard, meaning that it is not possible to solve a MDNLP in polynomial time. Nevertheless, many efficient methods are now available in the literature for solving mixed discrete programming problems. In [105], Branch-and-Bound methods (BBM) are developed to solve mixed discrete linear programming problems and mixed discrete nonlinear programming problems.

In [100, 121], by regarding the discrete variables as continuous, the mixed discrete nonlinear programming problems are solved by continuous optimization techniques. Then, the discrete variables are obtained by rounding off those continuous variables to the closest discrete values. The idea is intuitive and has been widely used. However, the solution obtained may be far from optimal, and may even be infeasible. In [8, 74, 75], a method is proposed by combining linear programming technique with Branch-and-Bound method, where the Branch-and-Bound method is applied to linear subproblems. However, if the number of discrete variables is large, the number of nodes created in the branching process becomes very large, and subsequently the computational cost will be very high. A detailed literature survey on Branch-and-Bound

methods can be found in [68].

In [23, 30, 87], a special class of integer programming problems is considered where the objective function is quadratic and the constraints are linear. This class of integer programming problems is known as the linear quadratic integer programming problem. The convex relaxation and Lagrangian decomposition schemes are used in [111] and [139]. On the other hand, the canonical duality is used in [23] and [29].

In [125], a general class of mixed discrete nonlinear programming problems is considered. By introducing additional new variables, it is shown that the original mixed discrete nonlinear programming problem is transformed into an equivalent optimization problem involving only continuous and binary variables. For the binary variables, they are transformed into continuous variables subject to additional quadratic and linear constraints. Thus, an equivalent constrained nonlinear optimization problem with continuous variables is obtained, where the constraints consists of the original constraints plus the newly introduced quadratic and linear constraints. However, the resulting constrained optimization problem is very difficult to solve due to the additional quadratic constraints.

In [77] and [80], penalty function methods are employed for nonlinear optimization problems with binary variables, where a relaxation is made. In the relaxed problems, all variables are continuous. However, the obtained continuous constrained problem is also not easy to solve. For other continuous optimization approaches for solving discrete optimization problems, we refer to [84–86].

This chapter is based on [133]. We first use the idea in [125] to transform the mixed discrete nonlinear programming problem into a conventional nonlinear optimization problem. Then, a new approach based on the exact penalty function method introduced in [135] is used to obtain a sequence of unconstrained optimization problems. Each of these unconstrained optimization problem can be solved by gradient-based methods. We will show that, under some mild assumptions, any local minimizer of the unconstrained optimization problem is a local minimizer of the original problem when the penalty parameter is sufficiently large. Numerical experiments show that the method proposed is effective.

3.2 Mixed discrete nonlinear programming problems

Consider a mixed discrete nonlinear programming problem given below:

$$\min f(\mathbf{x}, \mathbf{y})$$

$$\begin{aligned} \text{subject to} \quad & H_i(\mathbf{x}, \mathbf{y}) = 0 \quad i = 1, 2, \ldots, M, \\ & G_i(\mathbf{x}, \mathbf{y}) \leq 0 \quad j = 1, 2, \ldots, N. \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$ and $\mathbf{y} = [y_1, y_2, \ldots, y_m]^\top \in \mathbb{D}_1 \times \cdots \times \mathbb{D}_m$. Here, $\mathbb{R}^n$ is the n-dimensional Euclidean space, and for each $i = 1, 2, \ldots, m$, $\mathbb{D}_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,K_i}\}$, where

$a_{i,j},\ j = 1, \ldots, K_i$, are given discrete values. Let this problem be denoted as Problem **P**.

To transform Problem **P** to a constrained optimization problem with continuous variables, we define, for each $i = 1, 2, \ldots, m$,

$$\bar{y}_i = \sum_{j=1}^{K_i} a_{i,j} w_{i,j}, \tag{3.2}$$

where, for each $i = 1, 2, \ldots, m$,

$$\sum_{j=1}^{K_i} w_{i,j} = 1, \tag{3.3a}$$

$$0 \le w_{i,j} \le 1, \quad j = 1, 2, \ldots, K_i, \tag{3.3b}$$

$$w_{i,j}(1 - w_{i,j}) \le 0, \quad j = 1, 2, \ldots, K_i. \tag{3.3c}$$

Now, consider the following problem, which is denoted as Problem $\bar{\mathbf{P}}$.

$$\min \bar{f}(\mathbf{x}, \boldsymbol{\omega}) \tag{3.4a}$$

$$\begin{aligned}
\text{subject to} \quad & \bar{h}_i(\mathbf{x}, \boldsymbol{\omega}) = 0 \quad i = 1, 2, \ldots, M, && (3.4b)\\
& \bar{g}_i(\mathbf{x}, \boldsymbol{\omega}) \le 0 \quad j = 1, 2, \ldots, N, && (3.4c)\\
& \sum_{j=1}^{K_i} w_{i,j} = 1,\ i = 1, 2, \ldots, m, && (3.4d)\\
& \sum_{j=1}^{K_i} w_{i,j}(1 - w_{i,j}) \le 0 \quad i = 1, 2, \ldots, m, && (3.4e)\\
& 0 \le w_{i,j} \le 1 \quad j = 1, 2, \ldots, K_i,\ i = 1, 2, \ldots, m. && (3.4f)
\end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$, $\boldsymbol{\omega}=[(\boldsymbol{\omega}_1)^\top, \ldots, (\boldsymbol{\omega}_m)^\top]^\top$ with $\boldsymbol{\omega}_i=[\omega_{i,1}, \ldots, \omega_{i,K_i}]^\top$, $i = 1, \ldots, m$, while

$$\bar{f}(\mathbf{x}, \boldsymbol{\omega}) = f(\mathbf{x}, \mathbf{y})$$

$$\bar{h}_i(\mathbf{x}, \boldsymbol{\omega}) = H_i(\mathbf{x}, \mathbf{y}), i = 1, \ldots, M$$

$$\bar{g}_j(\mathbf{x}, \boldsymbol{\omega}) = G_i(\mathbf{x}, \mathbf{y}), j = 1, \ldots, N$$

Here, $\mathbf{y} = [y_1, \ldots, y_k]^\top$ with $y_i = \sum_{j=1}^{K_i} a_{i,j}\omega_{i,j}, i = 1, \ldots, k$. Clearly, Problem $\bar{\mathbf{P}}$ is a nonlinear optimization problem with conventional equality and inequality constraints.

From Theorem 3.1 in [125], we note that, for each $i = 1, 2, \ldots, k$, the solution of (3.4) is that only one of the $w_{i,j}, j = 1, 2, \ldots, K_i$, can be taken as one, while others are all zeros. This indicates that for each $i = 1, \ldots, k$, $\bar{y}_i$ can only take a discrete value from the set $\mathbb{D}_i$, implying that Problem **P** is equivalent to Problem $\bar{\mathbf{P}}$.

In principle, the constrained optimization problem $\bar{\mathbf{P}}$ appears solvable by existing optimization techniques, such as those implemented in the optimization software packages. For example, *fmincon* within MATLAB or *NLPQLP* within FORTRAN environment. However, Problem

$\bar{\mathbf{P}}$ is not easy to be solved directly due to the quadratic constraints (3.4*e*). Numerous numerical experiments are carried out solving some test examples considered in Section 3.3. However, both of the optimization packages fail to find feasible solutions of the test problems due to the inequality constraints (3.4e).

Motivated by the idea presented in Chapter 2, we will introduce an exact penalty function to transform Problem $\bar{\mathbf{P}}$ into a sequence of unconstrained optimization problems, such that each of these unconstrained optimization problem becomes solvable by gradient-based optimization techniques. Furthermore, we will show that a local minimizer of the unconstrained optimization problem is a local minimizer of Problem $\bar{\mathbf{P}}$ if the penalty parameter is sufficiently large.

3.2.1 Exact penalty function method

Consider Problem $\bar{\mathbf{P}}$. It can be expressed as the following conventional constrained optimization problem, which is referred to as Problem $\hat{\mathbf{P}}$.

$$\begin{array}{lll} & \min F(\mathbf{z}) & \\ \text{subject to} & H_i(\mathbf{z}) = 0, & i = 1, 2, \ldots, \overline{M} \\ & G_i(\mathbf{z}) \leq 0, & i = 1, 2, \ldots, \overline{N} \end{array}$$

where $\mathbf{z} = [(\mathbf{x})^\top, (\boldsymbol{\omega})^\top]^\top \in \mathbb{R}^r$ with $r = n + \sum_{i=1}^{m} K_i$,

$$\begin{array}{lcl} F(\mathbf{z}) & = & \bar{f}(\mathbf{x}, \boldsymbol{\omega}) \\ H_i(\mathbf{z}) & = & \bar{h}_i(\mathbf{x}, \boldsymbol{\omega}),\ i = 1, \ldots, M \\ H_{i+M}(\mathbf{z}) & = & \sum_{j=1}^{K_i} \omega_{i,j} - 1,\ i = 1, \ldots, m \\ G_i(\mathbf{z}) & = & \bar{g}_i(\mathbf{x}, \boldsymbol{\omega}),\ i = 1, \ldots, N \\ G_{i+N}(\mathbf{z}) & = & \sum_{j=1}^{K_i} \omega_{i,j}(1 - \omega_{i,j}),\ i = 1, \ldots, m \\ G_{N_1+j+i}(\mathbf{z}) & = & \omega_{i,j} - 1,\ j = 1, \ldots, K_i;\ i = 1, \ldots, m \\ G_{N_2+j+i}(\mathbf{z}) & = & -\omega_{i,j},\ j = 1, \ldots, K_i;\ i = 1, \ldots, m \end{array}$$

Here, $\overline{N} = N + m + 2\sum_{i=1}^{m} K_i$, $\overline{M} = M + m$, $N_1 = N + \sum_{i=1}^{m} K_i$, and $N_2 = N_1 + \sum_{i=1}^{m} K_i$.

As in Chapter 2, we introduce an exact penalty function, which is denoted as $F_\sigma(\mathbf{z}, \epsilon)$, defined below:

$$F_\sigma(\mathbf{z}, \epsilon) = \begin{cases} F(\mathbf{z}) & \text{if } \epsilon = 0,\ \mathbf{z} \text{ is feasiable for Problem } (\hat{\mathbf{P}}) \\ F(\mathbf{z}) + \epsilon^{-\alpha}\Delta(\mathbf{z}, \epsilon) + \sigma\epsilon^{\beta} & \text{if } \epsilon > 0 \\ +\infty & \text{otherwise} \end{cases} \tag{3.6}$$

where ϵ is a newly introduced variable, and the constraint violation $\Delta(\mathbf{z}, \epsilon)$ is defined by

$$\Delta(\mathbf{z}, \epsilon) = \sum_{i=1}^{\overline{N}} \Big[\max\big\{0, G_i(\mathbf{z}) - \epsilon^{\gamma}\big\}\Big]^2 + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}) - \epsilon^{\gamma}\big)^2 \tag{3.7}$$

Here, α, β and γ are positive real numbers, and σ is a penalty parameter. Similarly, we define

$$S_{\epsilon} = \{[\mathbf{z}^{\top}, \epsilon]^{\top} : H_i(\mathbf{z}) = \epsilon^{\gamma}, i = 1, \ldots, \overline{M};\ G_i(\mathbf{z}) \leq \epsilon^{\gamma},\ i = 1, \ldots, \overline{N}\} \tag{3.8}$$

where $\mathbb{R}_+ = \{\alpha \in \mathbb{R} \ :\ \alpha \geq 0\}$. The definition below gives the linearly independent constraint qualification.

Definition 3.1. *For a given $\mathbf{z}^{\star} \in \mathbb{R}^r$, let $\mathcal{A}(\mathbf{z}^{\star})$ be the set of those indices $i \in \{1, \ldots, \overline{N}\}$ such that for $i \in \mathcal{A}(\mathbf{z}^{\star})$, $G_i(\mathbf{z}^{\star}) = 0$. Suppose that the gradients of the active constraints, i.e., $G_i(\mathbf{z}^{\star}) = 0$ for $i \in \mathcal{A}(\mathbf{z}^{\star})$, and the equality constraints $H_i(\mathbf{z}^{\star}) = 0$ for $i = 1, \ldots, \overline{M}$, which are evaluated at $\mathbf{z} = \mathbf{z}^{\star}$, are linearly independent. Then, it is said that the* linearly independent constraint qualification (LICQ) *is satisfied at $\mathbf{z} = \mathbf{z}^{\star}$.*

Now, consider the following optimization problem, which is denoted as Problem $\mathbf{P}_{\sigma}$.

$$\min\ F_{\sigma}(\mathbf{z}, \epsilon)$$

$$\text{subject to}\ \ (\mathbf{z}, \epsilon) \in \mathbb{R}^n \times [0, +\infty)$$

Clearly, Problem $\mathbf{P}_{\sigma}$ is a conventional unconstrained optimization problem. In fact, any local minimizer of Problem $\mathbf{P}_{\sigma}$ is a local minimizer of Problem $\hat{\mathbf{P}}$ if the penalty parameter is sufficiently large. This together with other relevant results are presented in the next section.

3.2.2 Convergence analysis

Let $\{\sigma_k\}_{k=1}^{\infty}$ be an increasing sequence of penalty parameters such that $\sigma_k \to \infty$. Furthermore, let $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ denote the solution of Problem $\mathbf{P}_{\sigma_k}$ corresponding to σ_k. We assume that the following hypotheses are satisfied:

(H_1) *F, G_i, $i = 1, \ \ldots\ , \overline{N}$, and H_i, $i = 1, \ \ldots\ , \overline{M}$, are continuously differentiable in $\mathbb{R}^r$. $F(\mathbf{z}) \to \infty$,* as $|\mathbf{z}| \to \infty$.
(H_2) *The linearly independent constraint qualification is satisfied at $\mathbf{z} = \mathbf{z}^{\star}$, where $\mathbf{z}^{\star}$ is a local minimizer of Problem $\hat{\mathbf{P}}$.*
(H_3) $\max\{0, G_i(\mathbf{z}^{(k),\star})\} = o\big((\epsilon^{(k),\star})^{\delta_1}\big)$, $i = 1, \ldots, \overline{N}$; $H_i(\mathbf{z}^{(k),\star}) = o\big((\epsilon^{(k),\star})^{\delta_2}\big)$, $i = 1, \ldots, \overline{M}$, *where δ_1 and δ_2 are positive constants, and*

$$\lim_{\eta \to 0} \frac{o(\eta^{\iota})}{\eta^{\iota}} = 0$$

with ι being δ_1 or δ_2.

The following two Lemmas show that the sequence $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ of local minimizers will converge to a feasible point of Problem $\hat{\mathbf{P}}$. They are needed in the proofs of Theorems 3.1-3.3 to be given below.

Lemma 3.1. *Let $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ be a local minimizer of Problem $\mathbf{P}_{\sigma_k}$. Suppose that $F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ is finite and that $\epsilon^{(k),\star} > 0$. Then*

$$(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) \notin S_{\epsilon^{(k),\star}}$$

where $S_{\epsilon^{(k),\star}}$ is defined by (3.8) with $\epsilon = \epsilon^{(k),\star}$.

Proof. Since $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ is a local minimizer of Problem $\mathbf{P}_{\sigma_k}$ and $\epsilon^{(k),\star} > 0$, it is clear that

$$\begin{aligned} & \frac{\partial F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})}{\partial \epsilon} \\ = \; & (\epsilon^{(k),\star})^{-\alpha-1}\Big\{ -\alpha\Delta(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) + 2\gamma\Big(\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big\}(-(\epsilon^{(k),\star})^{\gamma}) \\ & \qquad + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)(-(\epsilon^{(k),\star})^{\gamma})\Big)\Big\} + \sigma_k\beta(\epsilon^{(k),\star})^{\beta-1} \\ = \; & 0 \end{aligned} \tag{3.10}$$

If the conclusion of the lemma is false. Then, we have

$$\begin{aligned} H_i(\mathbf{z}) = \; & \epsilon^{\gamma}, \; i = 1, \ldots, \overline{M}, \\ G_i(\mathbf{z}) \le \; & \epsilon^{\gamma}, \; i = 1, \ldots, \overline{N}. \end{aligned} \tag{3.11}$$

Substituting (3.11) to (3.10) gives

$$0 = \frac{\partial F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})}{\partial \epsilon} = \sigma_k\beta(\epsilon^{(k),\star})^{\beta-1} > 0$$

This is a contradiction, and hence completing the proof. □

Lemma 3.2. *Let $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ be a local minimizer of Problem $\mathbf{P}_{\sigma_k}$ such that $F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ is finite and $\epsilon^{(k),\star} > 0$. Suppose that $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) \to (\mathbf{z}^{\star}, \epsilon^{\star})$ as $k \to +\infty$, and that the hypotheses (H_1)-(H_3) are satisfied. Then, $\epsilon^{\star} = 0$ and $\mathbf{z}^{\star} \in S_0$, where S_0 is defined by (3.8) with $\epsilon = 0$.*

Proof. It follows from Lemma 3.1 that $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) \notin S_{\epsilon^{(k),\star}}$. Moreover,

$$\begin{aligned} & \frac{\partial F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})}{\partial \mathbf{z}} \\ = \; & \frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + 2(\epsilon^{(k),\star})^{-\alpha}\Big[\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} \\ & \qquad + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big] \\ = \; & 0 \end{aligned} \tag{3.12}$$

Suppose that $\epsilon^{(k),\star} \to \epsilon^\star \neq 0$. Then, by (3.10), we observe that its first term tends to a finite value, while the last term tends to infinity as $\sigma_k \to +\infty$, when $k \to +\infty$. This is impossible for the validity of (3.10). Thus, $\epsilon^\star = 0$. Now, by (3.12), we obtain

$$\begin{aligned}(\epsilon^{(k),\star})^\alpha \frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + 2\Big[\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\} \frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} \\ + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big) \frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big] = 0\end{aligned} \tag{3.13}$$

Thus,

$$\begin{aligned}&\lim_{k\to+\infty} (\epsilon^{(k),\star})^\alpha \frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + 2\Big[\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\} \frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} \\ &\qquad + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big) \frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big] \\ = \;& 2\Big[\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^\star)\big\} \frac{\partial G_i(\mathbf{z}^\star)}{\partial \mathbf{z}} + \sum_{i=1}^{\overline{M}} H_i(\mathbf{z}^\star) \frac{\partial H_i(\mathbf{z}^\star)}{\partial \mathbf{z}}\Big] = 0.\end{aligned} \tag{3.14}$$

Since the LICQ is satisfied at $\mathbf{z} = \mathbf{z}^\star$, it follows that,

$$\begin{aligned}H_i(\mathbf{z}^\star) &= \;\; 0, \;\; i = 1, \ldots, \overline{M}, \\ G_i(\mathbf{z}^\star) &\leq \;\; 0, \;\; i = 1, \ldots, \overline{N}.\end{aligned} \tag{3.15}$$

The proof is completed. □

The main convergence results are presented in the following three theorems.

Theorem 3.1. *Suppose that the hypotheses* (H_1)*-*(H_3) *are satisfied, and that* $\gamma > \alpha$, $\delta = \min(\delta_1, \delta_2) > \alpha$, $-\alpha - 1 + 2\delta > 0$, $2\gamma - \alpha - 1 > 0$. *Then*

$$F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) \xrightarrow[\mathbf{z}^{(k),\star} \to \mathbf{z}^\star \in S_0]{\epsilon^{(k),\star} \to \epsilon^\star = 0} F_{\sigma_k}(\mathbf{z}^\star, 0) = F(\mathbf{z}^\star)$$

$$\nabla_{(\mathbf{z},\epsilon)} F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) \xrightarrow[\mathbf{z}^{(k),\star} \to \mathbf{z}^\star \in S_0]{\epsilon^{(k),\star} \to \epsilon^\star = 0} \nabla_{(\mathbf{z},\epsilon)} F_{\sigma_k}(\mathbf{z}^\star, 0) = (\nabla F(\mathbf{z}^\star), 0)$$

Proof. It follows from the conditions of the theorem that, for $\epsilon^{(k),\star} \neq 0$,

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star}) \\
&= \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \Big\{F(\mathbf{z}^{(k),\star}) + (\epsilon^{(k),\star})^{-\alpha}\Big[\sum_{i=1}^{\overline{N}}\Big(\max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}\Big)^2 \\
&\qquad\qquad + \sum_{i=1}^{\overline{M}}\Big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\Big)^2\Big] + \sigma_k(\epsilon^{(k),\star})^\beta\Big\} \\
&= F(\mathbf{z}^\star) + \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \frac{\sum_{i=1}^{\overline{N}}\Big(\max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}\Big)^2 + \sum_{i=1}^{\overline{M}}\Big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\Big)^2}{(\epsilon^{(k),\star})^\alpha}
\end{aligned} \tag{3.16}$$

For the second term of (3.16), it is clear from Lemma 3.1 that

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \frac{\sum_{i=1}^{\overline{N}}\Big(\max\big\{0, G_i(\mathbf{z}^{(k),\star},\omega) - (\epsilon^{(k),\star})^\gamma\big\}\Big)^2 + \sum_{i=1}^{\overline{M}}\Big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\Big)^2}{(\epsilon^{(k),\star})^\alpha} \\
&= \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \sum_{i\in I'}\big((\epsilon^{(k),\star})^{-\frac{\alpha}{2}} G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma-\frac{\alpha}{2}}\big)^2 \\
&\qquad + \sum_{i=1}^{\overline{M}}\big((\epsilon^{(k),\star})^{-\frac{\alpha}{2}} H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma-\frac{\alpha}{2}}\big)^2
\end{aligned} \tag{3.17}$$

Here, I' denotes the index set such that for any $i \in I'$, $\max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\} = G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma$. Since $\gamma > \alpha$ and $\delta > \alpha$, we have

$$\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \sum_{i\in I'}\big((\epsilon^{(k),\star})^{-\frac{\alpha}{2}} G_j(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma-\frac{\alpha}{2}}\big)^2 + \sum_{i=1}^{\overline{M}}\big((\epsilon^{(k),\star})^{-\frac{\alpha}{2}} H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma-\frac{\alpha}{2}}\big)^2 = 0 \tag{3.18}$$

Combining (3.16) and (3.18) gives

$$\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star}) = F_{\sigma_k}(\mathbf{z}^\star, 0) = F(\mathbf{z}^\star) \tag{3.19}$$

Similarly, we have

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \nabla_{(\mathbf{z},\epsilon)} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star}) \\
&= \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \Big[\nabla_{\mathbf{z}} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\ \nabla_\epsilon F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\Big]^\top
\end{aligned} \tag{3.20}$$

where

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \nabla_{\mathbf{z}} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\\
=\ &\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \Big\{\frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + 2(\epsilon^{(k),\star})^{-\alpha}\Big[\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\\
&+\sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big]\Big\}\\
=\ &\nabla_{\mathbf{z}} F(\mathbf{z}^\star) + \lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} 2\Big\{\sum_{i\in I'}\big[(\epsilon^{(k),\star})^{-\alpha} G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma-\alpha}\big]\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\\
&+\sum_{i=1}^{\overline{M}}\big((\epsilon^{(k),\star})^{-\alpha} H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma-\alpha}\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big]\Big\}\\
=\ &\nabla_{\mathbf{z}} f(\mathbf{z}^\star)
\end{aligned} \tag{3.21}$$

while

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \nabla_{\epsilon} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\\
=\ &\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} (\epsilon^{(k),\star})^{-\alpha-1}\Big\{-\alpha\Delta(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\\
&+2\gamma\Big(\sum_{i=1}^{\overline{N}}\max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}(-(\epsilon^{(k),\star})^\gamma)\\
&+\sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big)(-(\epsilon^{(k),\star})^\gamma)\Big)\Big\} + \sigma_k\beta(\epsilon^{(k),\star})^{\beta-1}\\
=\ &\lim_{\substack{\epsilon^{(k),\star}\to\epsilon^\star=0\\ \mathbf{z}^{(k),\star}\to\mathbf{z}^\star\in S_0}} \frac{-\alpha\Delta(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})}{(\epsilon^{(k),\star})^{\alpha+1}}\\
&+2\gamma\Big(\sum_{i\in I'}\max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}(-(\epsilon^{(k),\star})^{\gamma-\alpha-1})\\
&+\sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big)(-(\epsilon^{(k),\star})^{\gamma-\alpha-1})\Big)\\
&+\sigma_k\beta(\epsilon^{(k),\star})^{\beta-1}\\
=\ &0
\end{aligned} \tag{3.22}$$

Thus, the proof is completed. □

The above results indicate that the constructed exact penalty function is continuously differentiable with its gradients having finite limits.

From Lemmas 3.1, 3.2 and Theorem 3.1, we will show that the sequence $(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})$ of the local minimizers will converge to a feasible point of the original problem $\hat{\mathbf{P}}$ with finite objective function value. Furthermore, this feasible point is a local minimizer of Problem $\hat{\mathbf{P}}$. These results together with the exactness of the proposed penalty function (3.6) are presented in the following as a Theorem.

Theorem 3.2. *Let $(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})$ be a local minimizer of Problem $\mathbf{P}_{\sigma_k}$. Suppose that $(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\to(\mathbf{z}^\star,\epsilon^\star)$ as $k\to+\infty$, and that the parameters α, γ and δ satisfy the same conditions as in The-*

orem 3.1. Then, there exists a $k_0 > 0$, *such that* $\epsilon^{(k),\star} = 0$, *and* $\mathbf{z}^{(k),\star}$ *is a local minimizer of Problem* $\hat{\mathbf{P}}$, *for* $k \geq k_0$.

Proof. We follow the idea of the proof given for Theorem 2.3. To begin, we assume that the conclusion is false. Then, there exists a subsequence of $\{(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})\}$, which is denoted by the original sequence, such that for any $k_0 > 0$, there exists a $k' > k_0$ satisfying $\epsilon^{(k'),\star} \neq 0$. By Lemma 3.2, we have

$$\epsilon^{(k),\star} \to \epsilon^\star = 0, \ \mathbf{z}^{(k),\star} \to \mathbf{z}^\star \in S_0, \ \text{as } k \to +\infty$$

Since $\epsilon^{(k),\star} \neq 0$ for all k, we have

$$\begin{aligned} & \frac{\partial F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})}{\partial \mathbf{z}} \\ = \ & \frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + 2(\epsilon^{(k),\star})^{-\alpha}\Big[\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} \\ & \qquad + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big] \\ = \ & 0 \end{aligned} \tag{3.23}$$

$$\begin{aligned} & \frac{\partial F_{\sigma_k}(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})}{\partial \epsilon} \\ = \ & (\epsilon^{(k),\star})^{-\alpha-1}\Big\{ -\alpha\Delta(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) + 2\gamma\Big(\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}(-(\epsilon^{(k),\star})^\gamma) \\ & \qquad + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big)(-(\epsilon^{(k),\star})^\gamma)\Big)\Big\} + \sigma_k\beta(\epsilon^{(k),\star})^{\beta-1} \\ = \ & 0 \end{aligned} \tag{3.24}$$

Dividing (3.10) by $(\epsilon^{(k),\star})^{\beta-1}$, we obtain

$$\begin{aligned} (\epsilon^{(k),\star})^{-\alpha-\beta}\Big\{ & -\alpha\Delta(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star}) + 2\gamma\Big(\sum_{i=1}^{\overline{N}} \max\big\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big\}(-(\epsilon^{(k),\star})^\gamma) \\ & + \sum_{i=1}^{\overline{M}} \big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^\gamma\big)(-(\epsilon^{(k),\star})^\gamma)\Big)\Big\} + \sigma_k\beta = 0 \end{aligned} \tag{3.25}$$

This is equivalent to

$$
\begin{aligned}
&(\epsilon^{(k),\star})^{-\alpha-\beta}\Bigg\{-\alpha\Delta(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})+2\gamma\Bigg(\sum_{i=1}^{\overline{N}}\Big[\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}\big(-(\epsilon^{(k),\star})^{\gamma}\big)\\
&+\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}G_i(\mathbf{z}^{(k),\star})-\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}G_i(\mathbf{z}^{(k),\star})\Big]\\
&+\sum_{i=1}^{\overline{M}}\Big[\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)\big(-(\epsilon^{(k),\star})^{\gamma}\big)+\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)H_i(\mathbf{z}^{(k),\star})\\
&-\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)H_i(\mathbf{z}^{(k),\star})\Big]\Bigg)\Bigg\}+\sigma_k\beta=0
\end{aligned}
\tag{3.26}
$$

Rearranging (3.26) yields

$$
\begin{aligned}
&(\epsilon^{(k),\star})^{-\alpha-\beta}(2\gamma-\alpha)\Delta(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})+\sigma_k\beta\\
&=2\gamma(\epsilon^{(k),\star})^{-\alpha-\beta}\Bigg(\sum_{i=1}^{\overline{N}}\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}G_i(\mathbf{z}^{(k),\star})\\
&\qquad+\sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)H_i(\mathbf{z}^{(k),\star})\Bigg)
\end{aligned}
\tag{3.27}
$$

Letting $k\to+\infty$ in (3.27) gives

$$
\begin{aligned}
&2\gamma(\epsilon^{(k),\star})^{-\alpha-\beta}\Bigg(\sum_{i=1}^{\overline{N}}\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}G_i(\mathbf{z}^{(k),\star})\\
&\qquad+\sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)H_i(\mathbf{z}^{(k),\star})\Bigg)\to+\infty
\end{aligned}
\tag{3.28}
$$

Define

$$
y^{(k)}=(\epsilon^{(k),\star})^{-\alpha-\beta}\Big(\sum_{i=1}^{\overline{N}}\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}+\sum_{i=1}^{\overline{M}}\big|\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)\big|\Big) \tag{3.29}
$$

From (3.28) and (3.29), we have

$$
y^{(k)}\to+\infty\ ,\ \text{as } k\to+\infty \tag{3.30}
$$

Dividing (3.12) by $|y^{(k)}|(\epsilon^{(k),\star})^{\beta}$ yields

$$
\begin{aligned}
&\frac{\frac{\partial F(\mathbf{z}^{(k),\star})}{\partial\mathbf{z}}}{|y^{(k)}|(\epsilon^{(k),\star})^{\beta}}+\frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^{(k)}|}\Big[\sum_{i=1}^{\overline{N}}\max\big\{0,G_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial\mathbf{z}}\\
&\qquad+\sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star})-(\epsilon^{(k),\star})^{\gamma}\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial\mathbf{z}}\Big]=0
\end{aligned}
\tag{3.31}
$$

This is equivalent to

$$\Big|\frac{\frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}}{|y^{(k)}|(\epsilon^{(k),\star})^{\beta}}\Big| = \frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^{(k)}|}\Big|\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + \sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big|. \tag{3.32}$$

Note that, $\mathbf{z}^{(k),\star} \to \mathbf{z}^{\star}$ as $k \to +\infty$. Thus, for a $\xi > 0$, there exists a sufficiently large $\bar{K}$, such that for all $k > \bar{K}$, $\mathbf{z}^{(k),\star} \in \mathcal{N}_{\xi}(\mathbf{z}^{\star})$, where $\mathcal{N}_{\xi}(\mathbf{z}^{\star})$ is a ξ - neighborhood of $\mathbf{z}^{\star}$. It is clear from hypothesis (H_1) that there exists a constant C, independent of $k > \bar{K}$, such that, for all $k > \bar{K}$,

$$\Big|\frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big| \leq C \tag{3.33}$$

$$\big|\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\big| \leq C, i = 1, \ldots, \overline{N}, \text{ and } \big|\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\big| \leq C, i = 1, \ldots, \overline{M} \tag{3.34}$$

For the RHS of (3.32), when $k > \bar{K}$, we have

$$\frac{2(\epsilon^{(k),\star})^{-\alpha-\beta}}{|y^{(k)}|}\Big|\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + \sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big|$$

$=$

$$\frac{2\Big|\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\}\frac{\partial G_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}} + \sum_{i=1}^{\overline{M}}\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\frac{\partial H_i(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}\Big|}{\Big|\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\} + \sum_{i=1}^{\overline{M}}\big|\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\big|\Big|}$$

$$\leq \frac{2\Big[\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\}C + \sum_{i=1}^{\overline{M}}\Big|\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\Big|C\Big]}{\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\} + \sum_{i=1}^{\overline{M}}\Big|\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\Big|}$$

$$\leq 2C \tag{3.35}$$

On the other hand, from (3.29), we note that

$$= \frac{\frac{1}{|y^{(k)}|(\epsilon^{(k),\star})^{\beta}}}{\Big|\sum_{i=1}^{\overline{N}}\max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\} + \sum_{i=1}^{\overline{M}}\big|\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\big|\Big|(\epsilon^{(k),\star})^{-\alpha}}$$

From the hypothesis (H_3), we have $\max\{0, G_i(\mathbf{z}^{(k),\star})\} = o((\epsilon^{(k),\star})^{\delta_1})$, $H_i(\mathbf{z}^{(k),\star}) = o((\epsilon^{(k),\star})^{\delta_2})$ and $\gamma > \alpha,\ \delta = \min(\delta_1, \delta_2) > \alpha$. Thus

$$\begin{aligned} & \lim_{k\to+\infty} \Big| \sum_{i=1}^{\overline{N}} \max\left\{0, G_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\right\} + \sum_{i=1}^{\overline{M}} \Big|\big(H_i(\mathbf{z}^{(k),\star}) - (\epsilon^{(k),\star})^{\gamma}\big)\Big| \Big| (\epsilon^{(k),\star})^{-\alpha} \\ = & \lim_{k\to+\infty} \Big| \sum_{i=1}^{\overline{N}} \max\left\{0, G_i(\mathbf{z}^{(k),\star})(\epsilon^{(k),\star})^{-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\right\} + \sum_{i=1}^{\overline{M}} \Big|\big((\epsilon^{(k),\star})^{\delta-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\big)\Big| \Big| \end{aligned}$$

For any $i \in 1, \ldots, \overline{N}$, if $G_i(\mathbf{z}^{(k),\star}) \leq 0$, then it is clear that

$$\lim_{k\to\infty} \max\left\{0, G_i(\mathbf{z}^{(k),*})(\epsilon^{(k),\star})^{-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\right\} = 0.$$

On the other hand, we have

$$\max\{0, G_i(\mathbf{z}^{(k),\star})\} = G_i(\mathbf{z}^{(k),\star}) = o((\epsilon^{(k),\star})^{\delta_1}),$$

and

$$\lim_{k\to\infty} \max\left\{0, G_i(\mathbf{z}^{(k),\star})(\epsilon^{(k),\star})^{-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\right\} = \lim_{k\to\infty} \max\left\{0, (\epsilon^{(k),\star})^{\delta-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\right\} = 0.$$

Thus,

$$\begin{aligned} & \lim_{k\to+\infty} \Big| \sum_{i=1}^{\overline{N}} \max\left\{0, G_i(\mathbf{z}^{(k),\star})(\epsilon^{(k),\star})^{-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\right\} + \sum_{i=1}^{\overline{M}} \Big|\big((\epsilon^{(k),\star})^{\delta-\alpha} - (\epsilon^{(k),\star})^{\gamma-\alpha}\big)\Big| \Big| \\ = & \ 0 \end{aligned}$$

which means that

$$\frac{1}{|y^{(k)}|(\epsilon^{(k),\star})^{\beta}} \to +\infty, \quad k \to +\infty \tag{3.36}$$

From (3.33) and (3.36), it is clear that

$$\frac{|\frac{\partial F(\mathbf{z}^{(k),\star})}{\partial \mathbf{z}}|}{|y^{(k)}|(\epsilon^{(k),\star})^{\beta}} \to +\infty, \quad k \to +\infty \tag{3.37}$$

Thus, (3.35) together with (3.37) contradicts the validity of (3.32), and hence completing the first part of the proof .

For sufficiently large k, every local minimizer $(\mathbf{z}^{(k),\star}, \epsilon^{(k),\star})$ has the form $(\mathbf{z}^{\star}, 0)$. It is obvious from Lemma 3.2 that $\mathbf{z}^{\star}$ is a feasible point of Problem $\hat{\mathbf{P}}$. This indicates that there is a neighborhood of $\mathbf{z}^{\star}$, such that for any feasible $\mathbf{z}$ of Problem $\hat{\mathbf{P}}$

$$F(\mathbf{z}) = F_{\sigma_k}(\mathbf{z}, 0) \geq F_{\sigma_k}(\mathbf{z}^{\star}, 0) = F(\mathbf{z}^{\star}).$$

Therefore, $\mathbf{z}^{\star}$ is a local minimizer of Problem $\hat{\mathbf{P}}$. This completes the proof. □

Theorem 3.2 indicates that, under some mild assumptions, a local minimizer of the penalty Problem $\mathbf{P}_\sigma$ is a local minimizer of Problem $\hat{\mathbf{P}}$, when the parameter σ is sufficiently large.

3.3 Numerical results

To test the method proposed, we consider some examples in this section. The equivalent constrained continuous optimization problems are solved by using the optimization tool box *fmincon* within MATLAB environment. For the newly introduced variables $w_{i,j}$ in (3.2) and (3.3), a natural way is to set their initial values as:

$$w_{i,j} = 1/K_i,\ \ j = 1, 2, \ldots, K_i \tag{3.38}$$

Example 3.1 (Pressure vessel design problem)

$$\min\ \ f(\mathbf{x}, \mathbf{y}) = 0.6224x_1x_2y_1 + 1.7781x_2^2y_2 + 3.1611x_2y_1^2 + 19.84x_1y_1$$

subject to

$$\begin{aligned}
g_1(\mathbf{x}, \mathbf{y}) &= 0.0193x_1 - y_1 \le 0 \\
g_2(\mathbf{x}, \mathbf{y}) &= 0.00954x_1 - y_2 \le 0 \\
g_3(\mathbf{x}, \mathbf{y}) &= 750 \times 1728 - \pi x_1^2 x_2 - \tfrac{4}{3}\pi x_1^3 \le 0 \\
g_4(\mathbf{x}, \mathbf{y}) &= x_2 - 240 \le 0 \\
& x_1 \in [0, \infty),\ \ x_2 \in [0, \infty) \\
& y_1 \in \{1.125 + 0.0625(j-1) : j = 1, 2, \ldots, 7\} \\
& y_2 \in \{0.625 + 0.0625(j-1) : j = 1, 2, \ldots, 7\}.
\end{aligned}$$

Using transformation (3.2), the discrete variables y_1 and y_2 are replaced by the newly introduced variables $w_{1,j}, j = 1, \ldots, 7$, and $w_{2,j}, j = 1, \ldots, 7$. That is,

$$y_1 = \sum_{j=1}^{7} [1.125 + 0.0625(j-1)]w_{1,j}, \tag{3.39a}$$

$$y_2 = \sum_{j=1}^{7} [0.625 + 0.0625(j-1)]w_{2,j}. \tag{3.39b}$$

Substituting (3.39) into the original problem, we obtain an equivalent nonlinear constrained optimization problem with continuous variables. Then, by introducing the corresponding penalty function defined by (3.6), we obtain a sequence of unconstrained optimization problems. Each of these unconstrained optimization problems is solved by the optimization toolbox $fmincon$ within MATLAB environment. We set the initial values for $x_i, i = 1, 2$, $w_{i,j}, j = 1, \ldots, 7; i = 1, 2$, as $x_1 = 50,\ x_2 = 100,\ w_{i,j} = 1/7,\ i = 1, 2; j = 1, 2, \ldots, 7$. The penalty parameter is chosen as 10^8.

Applying our method, only one minimizer is found, which is $\mathbf{x}^\star = [67.6351, 1.51 \times 10^{-7}]^\top$, $\mathbf{y}^\star = [1.375, 0.875]^\top$ with $f(\mathbf{x}^\star, \mathbf{y}^\star) = 1845.1$. From Table 3.1, we see that a substantial improvement

is achieved when compared with the results obtained in [105] and [125]. In fact, we have obtained the global minimizer.

Table 3.1: Result for Example 3.1

	$\mathbf{x}^\star$	$\mathbf{y}^\star$	$f(\mathbf{x}^\star,\mathbf{y}^\star)$
Our result	$[67.6351,\ 1.51\times 10^{-7}]^\top$	$[1.375,\ 0.875]^\top$	1845.1
Result in [125]	$[58.2902,\ 43.6972]^\top$	$[1.125,\ 0.625]^\top$	7198.0
Result in [105]	$[48.3515,\ 111.9893]^\top$	$[1.125,\ 0.625]^\top$	7790.6

Example 3.2 (Three-bar truss problem)

$$\min\ f(\mathbf{x}) = 2x_1 + x_2 + \sqrt{3}x_3$$

subject to

$$\begin{aligned}
g_1(\mathbf{x}) &= -1 + \frac{\sqrt{3}x_2 + 1.932x_3}{1.5x_1x_2 + \sqrt{2}x_2x_3 + 1.319x_1x_3} \le 0\\
g_2(\mathbf{x}) &= -1 + \frac{0.634x_1 + 2.828x_3}{1.5x_1x_2 + \sqrt{2}x_2x_3 + 1.319x_1x_3} \le 0\\
g_3(\mathbf{x}) &= -1 + \frac{0.5x_1 - 2x_2}{1.5x_1x_2 + \sqrt{2}x_2x_3 + 1.319x_1x_3} \le 0\\
g_4(\mathbf{x}) &= -1 - \frac{0.5x_1 - 2x_2}{1.5x_1x_2 + \sqrt{2}x_2x_3 + 1.319x_1x_3} \le 0\\
&x_i \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1.0, 1.2\}, i = 1, 2, 3.
\end{aligned}$$

As it is pointed out in [125], the global minimizer is $x^\star = [1.2, 0.5, 0.1]^\top$ with $f(\mathbf{x}^\star) = 3.0732$. Using our method with the initial points chosen as:

$$w_{i,1} = w_{i,2} = \cdots = w_{i,7} = 1/7, \text{for } i = 1, 2, \ldots, 3.$$

and the penalty parameter chosen as 10^8, a local minimizer, which is $\hat{\mathbf{x}} = [1.2, 0.5, 0.2]^\top$ with $f(\hat{\mathbf{x}}) = 3.2464$, is obtained. This local minimizer obtained by our method is slightly inferior to the global minimizer obtained in [125].

In the following, three large scale nonlinear integer programming problems with 100 discrete variables are considered to test the performance of our method. These three problems are modified from those considered in [81], where the discrete sets for x are integers uniformly distributed from -5 to 5. In this situation, the discrete variables could be regarded as continuous ones. Then, the optimal values of the discrete variables are obtained by searching around the optimal values of the continuous variables. In the modified examples considered in this section, the discrete variables are chosen from $\mathbb{D}_1 \times \cdots \times \mathbb{D}_{100}$, where $\mathbb{D}_i = \{-4, -1, 0, 1, 4\}, i = 1, \ldots, 100$. These integer values are not uniformly distributed.

Example 3.3

$$\min\ f(\mathbf{x}) = (x_1 - 1)^2 + (x_{100} - 1)^2 + n\sum_{i=1}^{100}(100 - i)(x_i^2 - x_{i+1})^2$$

subject to

$$x_i \in \mathbb{D}_i = \{-4, -1, 0, 1, 4\},\ i = 1, \ldots, 100.$$

The global minimizer is $\mathbf{x}^\star = [1, 1, \ldots, 1]^\top$ with global minimum $f(\mathbf{x}^\star) = 0$

Example 3.4

$$\min\ f(\mathbf{x}) = \sum_{i=1}^{99}\left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right]$$

subject to

$$x_i \in \mathbb{D}_i = \{-4, -1, 0, 1, 4\},\ i = 1, \ldots, 100.$$

The global minimizer is $\mathbf{x}^\star = [1, 1, \ldots, 1]^\top$ with global minimum $f(\mathbf{x}^\star) = 0$

Example 3.5

$$\min\ f(\mathbf{x}) = \sum_{i=1}^{100} x_i^4 + \left(\sum_{i=1}^{100} x_i\right)^2$$

subject to

$$x_i \in \mathbb{D}_i = \{-4, -1, 0, 1, 4\},\ i = 1, \ldots, 100.$$

The global minimizer is $\mathbf{x}^\star = [0, 0, \ldots, 0]^\top$ with global minimum $f(\mathbf{x}^\star) = 0$.
For Examples $3.3 - 3.5$, our method is used with the initial values chosen as:

$$w_{i,1} = w_{i,2} = \cdots = w_{i,5} = 1/5,\ i = 1, \ldots, 100.$$

and the penalty parameter chosen as 10^9 for each example. The results obtained for these examples are shown in Table 3.2. From which, we see that our method finds global minimizers

Table 3.2: Results for Example 3.3, 3.4 and 3.5

Example	$\mathbf{x}^\star$	$f(\mathbf{x}^\star)$
3.3	$[1, 1, \ldots, 1]^\top$	0
3.4	$[1, 1, \ldots, 1]^\top$	0
3.5	$[0, 0, \ldots, 0]^\top$	0

for all these three examples. This indicates that the proposed method is an effective approach for large scale nonlinear integer programming problems.

3.4 Conclusion

This chapter considered a class of nonlinear mixed discrete programming problems. It is first transformed into an equivalent constrained optimization problem involving only continuous variables. However, this transformed problem is difficult to solve by using standard optimization techniques. A new exact penalty function method is proposed to construct a sequence of unconstrained optimization problems, each of which can be solved effectively by standard unconstrained optimization techniques, such as conjugate gradient or quasi-Newton methods. From the numerical simulation studies, we see that the proposed method is effective.

CHAPTER 4

Design of allpass variable fractional delay filter with signed powers-of-two coefficients

4.1 Introduction

Digital filters with tunable fractional phase-delay or fractional group delay, referred to as variable fractional delay (VFD) filters, are useful in various signal processing applications, including timing offset recovery in digital receivers, comb filter design, sampling rate conversion, speech coding, time delay estimation, one-dimensional digital signal interpolation and image interpolation. For details, see [15, 17], where a range of applications have been considered. For finite impulse response (FIR) based VFD filters, an appropriate optimization problem can be formulated. It is relatively easy to solve this approximate problem, meeting the desired characteristics [15, 25]. The design of allpass VFD filters is more involved. It has been investigated in [54, 79]. The key advantage of allpass VFD filters is that they can achieve higher design accuracy than FIR filters, yielding smaller frequency response errors for applications that require unity gain. However, since an allpass VFD filter has infinite impulse response, adjusting its coefficients will cause transients. In general, the transients depend on the magnitude of the input signals, how often and how large the coefficients are changed and how fast the impulse responses decay. Efforts to minimize the transient can be found in [98].

In [12], the design of allpass VFD filters with least squares and minimax group delay errors is investigated. The design of minimax phase error allpass VFD filters is discussed in [108]. In [17] and [122], the design of an allpass VFD filter with minimum integral squared error is developed. The obtained filters might have large deviation from the desired response, especially at the cutoff frequencies. In addition, several restrictions are required for the VFD filter specification. In [18], the minimax optimization problem is solved by fixing the coefficient of the denominator and iteratively updating the numerator coefficients. The designed allpass filters might have large integral squared error. These papers are mainly concentrate on the design of allpass VFD filters with infinite precision coefficients.

For ease in practical implementation, we investigate, in this chapter, the design of allpass VFD filters with signed powers-of-two coefficients and the least square criterion. It is based on [134]. By using the approximation scheme obtained in [17], the objective function is approxi-

mated by a quadratic cost function which has a unique optimal solution. Based on this optimal solution, a good search region containing the global solution is then developed by using a two-step scheme. Then, a new exact penalty function method is proposed to solve the quadratic integer optimization problem with the constraints being the obtained search region. Design examples demonstrate the effectiveness of the proposed method over the traditional quantization method.

The outline of this chapter is given as follows. The problem formulation is given in Section 4.2. The proposed solution method is given in Section 4.3. Simulation results are discussed in Section 4.4 and finally some concluding remarks are made in Section 4.5.

4.2 Problem formulation

Consider the design of an allpass filter with coefficients $a_n(p)$, $1 \leq n \leq N$, which depend on a tuning parameter p. More specifically, each coefficient $a_n(p)$ is expressed as a polynomial of p given below:

$$a_n(p) = \sum_{m=1}^{M} h_{n,m} p^m \tag{4.1}$$

where the parameter p is varied in the range $\mathcal{P} = [p_1, p_1 + 1]$, and p_1 denotes the lower bound. For ease in practical implementation, the coefficients $h_{n,m}$ are expressed in the form of sum of signed powers-of-two terms as given below:

$$h_{n,m} = \sum_{i=1}^{b} d_{i,n,m} 2^{-i} \tag{4.2}$$

where $d_{i,n,m} \in \{-1, 0, 1\}$, $i = 1, \ldots, b$; b denotes the number of bits of the wordlength; $n = 1, \ldots, N$; and $m = 1, \ldots, M$. Let N_1 denote the total allowable number of signed-powers-of-two terms used. Then, we have the constraint

$$\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{i=1}^{b} |d_{i,n,m}| \leq N_1 \tag{4.3}$$

The frequency response of the allpass filter is given by

$$\begin{aligned} H(\omega, p) &= \frac{a_N(p) + \cdots + a_1(p) e^{-j(N-1)\omega} + e^{-jN\omega}}{1 + a_1(p) e^{-j\omega} + \cdots + a_N(p) e^{-jN\omega}} \\ &= e^{-jN\omega} \frac{1 + \sum_{n=1}^{N} a_n(p) e^{jn\omega}}{1 + \sum_{n=1}^{N} a_n(p) e^{-jn\omega}} \\ &= e^{-jN\omega} \frac{1 + \sum_{n=1}^{N} \sum_{m=1}^{M} h_{n,m} p^m e^{jn\omega}}{1 + \sum_{n=1}^{N} \sum_{m=1}^{M} h_{n,m} p^m e^{-jn\omega}}. \end{aligned} \tag{4.4}$$

Let

$$\begin{aligned} R(\omega,p) &= 1+\sum_{n=1}^{N}\sum_{m=1}^{M} h_{n,m}p^m e^{jn\omega} \\ &= 1+\sum_{n=1}^{N}\sum_{m=1}^{M} h_{n,m}\cos(n\omega)p^m + j\sum_{n=1}^{N}\sum_{m=1}^{M} h_{n,m}\sin(n\omega)p^m \\ &= 1+\mathbf{c}^\top\mathbf{H}\mathbf{p}+j\boldsymbol{s}^\top\mathbf{H}\mathbf{p} \end{aligned} \tag{4.5}$$

with

$$\begin{aligned} \mathbf{p}^\top &= [p\ p^2\ \cdots\ p^M] \\ \mathbf{c}^\top &= [\cos(\omega)\ \cos(2\omega)\ \cdots\ \cos(N\omega)\ ] \\ \mathbf{s}^\top &= [\sin(\omega)\ \sin(2\omega)\ \cdots\ \sin(N\omega)\ ] \\ \mathbf{H} &= \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1M} \\ h_{21} & h_{22} & \cdots & h_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{NM} \end{bmatrix}. \end{aligned}$$

The equation (4.4) can be stated as:

$$H(\omega,p) = e^{-jN\omega}\cdot\frac{R(\omega,p)}{R^*(\omega,p)} \tag{4.6}$$

where * is the complex conjugate operator.

Let us specify the desired frequency response $H_d(\omega,p)$ which is given by

$$H_d(\omega,p) = e^{-j(N+p)\omega} \tag{4.7}$$

for all $\omega\in\Omega=[0,\alpha\pi]$, where $\alpha>0$ is a real number. The design objective is to choose the coefficients $h_{n,m}$ in the form of (4.2) such that

$$\int_{p_1}^{p_1+1}\int_{0}^{\alpha\pi} W(\omega,p)\big(H(\omega,p)-H_d(\omega,p)\big)^2 d\omega dp \tag{4.8}$$

is minimized, subject to constraint (4.3), where $W(\omega,p)$ is a positive weighting function. It is assumed that $W(\omega,p)$ is separable, i.e.,

$$W(\omega,p) = W_1(\omega)W_2(p)$$

where $W_1(\omega)$ and $W_2(p)$ are piecewise constant functions. Let this problem be referred to as Problem **P**.

Noting that, for each $n=1,\ldots,N$ and $m=1,\ldots,M$, $h_{n,m}$ has at most $2^{b+1}-1$ options. Thus, Problem **P** is a constrained nonlinear integer programming problem, which is very difficult to solve. A natural way to reduce the complexity is to reduce the number of options for each $h_{n,m}$. Since the objective function of Problem **P** is quadratic, the discrete points in the neighborhoods

of the infinite precision solution of Problem **P** are good choices for each $h_{m,n}$. We shall solve Problem **P** in three stages:

i. Find the optimal infinite precision solution for Problem **P**.

ii. Find a reduced search region around the minimizer obtained in stage (i).

iii. Find a point that minimize the objective function (4.8) within the region obtained in stage (ii).

For stage (i), many existing methods (see, for example [17] and [12]) can be used, producing good approximations to the infinite precision solution of Problem **P**. The best approximate solution obtained is by the noniterative method reported in [17]. The main idea of the noniterative method is summarized as follows. Using $H(\omega, p)$ to approximated $H_d(\omega, p)$ can be equivalently stated as using

$$\frac{R(\omega, p)}{R^*(\omega, p)}$$

to approximate

$$\frac{e^{-j(\omega p/2)}}{e^{j(\omega p/2)}}.$$

If

$$\frac{R(\omega, p)}{R^*(\omega, p)} \approx \frac{e^{-j(\omega p/2)}}{e^{j(\omega p/2)}},$$

then,

$$R(\omega, p)e^{j(\omega p/2)} \approx R^*(\omega, p)e^{-j(\omega p/2)}.$$

Since

$$R(\omega, p)e^{j(\omega p/2)} = \big[R^*(\omega, p)e^{-j(\omega p/2)}\big]^*,$$

it follows that

$$\mathcal{I}[R^*(\omega, p)e^{-j(\omega p/2)}] \approx 0,$$

where $\mathcal{I}[R^*(\omega, p)e^{-j(\omega p/2)}]$ denotes the imaginary part of $R^*(\omega, p)e^{-j(\omega p/2)}$.

Thus, the minimization of the expression (4.8) can be achieved approximately through the minimization of the error given below:

$$\begin{aligned} G(\mathbf{H}) &= \int_{p_1}^{p_1+1}\int_{0}^{\alpha\pi} W(\omega, p)\mathcal{I}[R^*(\omega, p)e^{-j(\omega p/2)}]^2 d\omega dp \\ &= \int_{p_1}^{p_1+1}\int_{0}^{\alpha\pi} W(\omega, p)\big[-\sin\Big(\frac{\omega p}{2}\Big)(1+\mathbf{c}^\top\mathbf{H}\mathbf{p}) - \cos\Big(\frac{\omega p}{2}\Big)\mathbf{s}^\top\mathbf{H}\mathbf{p}\big]^2 d\omega dp. \end{aligned} \tag{4.9}$$

Expanding (4.9) gives

$$G(\mathbf{H}) = \int_{p_1}^{p_1+1} \int_{0}^{\alpha\pi} \sum_{i=1}^{6} W_1(\omega) W_2(p) g_i(\omega, p) d\omega dp \tag{4.10}$$

where

$$\begin{aligned}
g_1(\omega, p) &= \sin^2\left(\frac{\omega p}{2}\right) \\
g_2(\omega, p) &= 2\sin^2\left(\frac{\omega p}{2}\right)\mathbf{c}^\top \mathbf{H}\mathbf{p} \\
g_3(\omega, p) &= \sin^2\left(\frac{\omega p}{2}\right)\mathbf{c}^\top \mathbf{H}\mathbf{p}\mathbf{p}^\top \mathbf{H}^\top \mathbf{c} \\
g_4(\omega, p) &= \cos^2\left(\frac{\omega p}{2}\right)\mathbf{s}^\top \mathbf{H}\mathbf{p}\mathbf{p}^\top \mathbf{H}^\top \mathbf{s} \\
g_5(\omega, p) &= \sin(\omega p)\mathbf{s}^\top \mathbf{H}\mathbf{p} \\
g_6(\omega, p) &= \sin(\omega p)\mathbf{c}^\top \mathbf{H}\mathbf{p}\mathbf{p}^\top \mathbf{H}^\top \mathbf{s}.
\end{aligned} \tag{4.11}$$

Applying Taylor series expansion to the sine and cosine terms within $g_2(\omega, p)$ to $g_6(\omega, p)$ gives

$$\begin{aligned}
G(\mathbf{H}) = \int_{p_1}^{p_1+1} \int_{0}^{\alpha\pi} W_1(\omega) W_2(p) \sin^2\left(\frac{\omega p}{2}\right) d\omega dp + \sum_{i=1}^{+\infty} \Big(& \mathrm{tr}[\mathbf{H}A_{1i}] + \mathrm{tr}[\mathbf{H}A_{2i}\mathbf{H}^\top A_{3i}] \\
& + \mathrm{tr}[\mathbf{H}A_{4i}\mathbf{H}^\top A_{5i}] + \mathrm{tr}[\mathbf{H}A_{6i}] + \mathrm{tr}[\mathbf{H}A_{7i}\mathbf{H}^\top A_{8i}] \Big)
\end{aligned}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The definitions of A_{1i} to A_{9i} are given below:

$$\begin{aligned}
A_{1i} &= \frac{(-1)^{i-1}}{(2i)!} \int_{p_1}^{p_1+1} W_2(p) p^{2i} \mathbf{p} dp \int_{0}^{\alpha\pi} W_1(\omega) \omega^{2i} \mathbf{c}^\top d\omega \\
A_{2i} &= \int_{p_1}^{p_1+1} W_2(p) p^{2i} \mathbf{p}\mathbf{p}^\top dp \\
A_{3i} &= \frac{(-1)^{i-1}}{2(2i)!} \int_{0}^{\alpha\pi} W_1(\omega) \omega^{2i} \mathbf{c}\mathbf{c}^\top d\omega \\
A_{4i} &= \begin{cases} \displaystyle\int_{p_1}^{p_1+1} W_2(p)\mathbf{p}\mathbf{p}^\top dp, & \text{if } i = 1, \\ \\ A_{2(i-1)}, & \text{if } i = 2, 3, \cdots . \end{cases}
\end{aligned}$$

$$A_{5i} = \begin{cases} \int_0^{\alpha\pi} W_1(\omega)\mathbf{s}\mathbf{s}^\top d\omega, & \text{if } i = 1, \\ \dfrac{(-1)^{i-1}}{2(2i-2)!}\int_0^{\alpha\pi} W_1(\omega)\omega^{2(i-1)}\mathbf{s}\mathbf{s}^\top d\omega, & \text{if } i = 2, 3, \cdots . \end{cases}$$

$$A_{6i} = \frac{(-1)^{i-1}}{(2i)!}\int_{p_1}^{p_1+1} W_2(p)p^{2i-1}\mathbf{p}dp \int_0^{\alpha\pi} W_1(\omega)\omega^{2i-1}\mathbf{s}^\top d\omega$$

$$A_{7i} = \int_{p_1}^{p_1+1} W_2(p)p^{2i-1}\mathbf{p}\mathbf{p}^\top dp$$

$$A_{8i} = \frac{(-1)^{i-1}}{(2i-1)!}\int_0^{\alpha\pi} W_1(\omega)\omega^{2i-1}\mathbf{s}\mathbf{c}^\top d\omega$$

$$A_{9i} = \frac{A_{8i}^\top + A_{8i}}{2}$$

It is reported in [17] that for a moderate large L, for example, $L = 9$, a sufficiently accurate approximation of the optimal solution $\mathbf{H}$ can be obtained by the following equation

$$\mathrm{cs}(\mathbf{H}) = \Big\{ \sum_{i=1}^{L} \Big[A_{2i} \otimes A_{3i} + A_{4i} \otimes A_{5i} + A_{7i} \otimes A_{9i} \Big] \Big\}^{-1} \Big\{ -\mathrm{cs}\Big[\sum_{i=1}^{L} \Big(\frac{A_{1i}^\top + A_{6i}^\top}{2} \Big) \Big] \Big\}, \tag{4.12}$$

where $\otimes$ and $\mathrm{cs}(\mathbf{H})$ denotes, respectively, the Kronecker product and the column string of the matrix $\mathbf{H}$.

We now consider the following problem, which is referred to as Problem $\tilde{\mathbf{P}}$.

$$\min \tilde{G}(\mathbf{H})$$

subject to

$$h_{n,m} = \sum_{i=1}^{b} d_{i,n,m} 2^{-i}, \tag{4.13}$$

$$\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{i=1}^{b} |d_{i,n,m}| \le N_1 \tag{4.14}$$

where

$$\tilde{G}(\mathbf{H}) = \int_{p_1}^{p_1+1}\int_0^{\alpha\pi} W_1(\omega)W_2(p)\sin^2(\frac{\omega p}{2})d\omega dp + \sum_{i=1}^{9}\Big(\mathrm{tr}[\mathbf{H}A_{1i}] + \mathrm{tr}[\mathbf{H}A_{2i}\mathbf{H}^\top A_{3i}] + \mathrm{tr}[\mathbf{H}A_{4i}\mathbf{H}^\top A_{5i}] + \mathrm{tr}[\mathbf{H}A_{6i}] + \mathrm{tr}[\mathbf{H}A_{7i}\mathbf{H}^\top A_{8i}] \Big),$$

b denotes the number of bits in the wordlength, N_1 is the maximum allowable number of nonzero

coefficients.

In the next section, we shall introduce a novel computational approach to obtain a reduced discrete search region which contains the optimal solution of Problem $\tilde{\mathbf{P}}$. Then, an exact penalty function method is developed to search for the optimal solution of Problem $\tilde{\mathbf{P}}$ from the obtained reduced discrete search region.

4.3 Solution method for problem $\tilde{\mathbf{P}}$

Since equation (4.12) gives a very good approximation to the solution to the minimization of (4.8), our method for solving Problem $\tilde{\mathbf{P}}$ is now divided into two steps: (i) Searching for a desirable reduced discrete search region around the solution of (4.12); and (ii) Finding the optimal solution from the reduced discrete search region.

4.3.1 Construct reduced search region

To construct the reduced discrete search region, it is carried out by two algorithms. They are based on the fundamental results to be presented in Theorem 4.1 and Theorem 4.2 below. For the proof of Theorem 4.1, we need the following lemma.

Lemma 4.1. *Let i and j be any integers such that $i, j > 0$. If $i < j$, then*

$$2^{-i} - 2^{-j} = 2^{-(i+1)} + \cdots + 2^{-j}. \tag{4.15}$$

Proof. Since $i < j$, we have

$$\begin{aligned} 2^{-i} &= 2^{-(i+1)} + 2^{-(i+1)} \\ &= 2^{-(i+1)} + 2^{-(i+2)} + 2^{-(i+2)} \\ &= 2^{-(i+1)} + 2^{-(i+2)} + 2^{-(i+3)} + 2^{-(i+3)} \\ &= 2^{-(i+1)} + 2^{-(i+2)} + 2^{-(i+3)} + \cdots + 2^{-j} + 2^{-j} \end{aligned} \tag{4.16}$$

Thus,

$$2^{-i} - 2^{-j} = 2^{-(i+1)} + \cdots + 2^{-j}.$$

□

To proceed, we need the following definition:

$$\mathcal{A}_C = \Big\{ \mathbf{x} \in \mathbb{R}^b \;\Big|\; \sum_{i=1}^{b} x_i 2^{-i} = C, x_i \in \{-1, 0, 1\} \Big\},$$

where $C \in \mathcal{S} = \Big\{ \sum\limits_{i=1}^{b} x_i 2^{-i} \;\Big|\; x_i \in \{-1, 0, 1\} \Big\}$.

Theorem 4.1. *Let* n_+ *and* n_- *denote, respectively, the numbers of "* 1 *" and "* -1 *" in* $\mathbf{x}$*. Then, for any* $C \in \mathcal{S}$*, there exists a unique* $\mathbf{x}^\star \in \mathcal{A}_C$ *such that* $n_+^\star n_-^\star = 0$*.*

Proof. If $C = 0$, we just let $\mathbf{x}^\star = [0, \cdots, 0]^\top$. Thus,

$$n_+^\star = n_-^\star = 0,$$

and hence the existence is established. For the uniqueness of $\mathbf{x}^\star$, assume that there exists another $\mathbf{x}' = [x_1', \cdots, x_b']^\top \neq \mathbf{x}^\star$ such that

$$\sum_{i=1}^{b} x_i' 2^{-i} = 0, \quad x_i' \in \{-1, 0, 1\}, \quad i = 1, \cdots, b. \tag{4.17}$$

Clearly, $n_+' > 0$ and $n_-' > 0$. Let r denote the smallest index such that $x_r' \neq 0$. By applying Lemma 4.1, we can always obtain a vector $\bar{\mathbf{x}} = [\bar{x}_1, \cdots, \bar{x}_b]^\top$, satisfying

$$\sum_{i=1}^{b} \bar{x}_i 2^{-i} = 0,$$

and for each $i = 1, \cdots, b$,

$$\begin{cases} \bar{x}_i \in \{0, 1\}, \text{ and } \bar{n}_+ > 0 & \text{if } x_r' = 1, \\ \bar{x}_i \in \{0, -1\}, \text{ and } \bar{n}_- > 0 & \text{if } x_r' = -1. \end{cases} \tag{4.18}$$

This is impossible. Thus, there exists no such $\bar{\mathbf{x}}$ and hence $\mathbf{x}'$. This shows the uniqueness for the case of $C = 0$.

Now, suppose that

$$C = \sum_{i=1}^{b} x_i 2^{-i} > 0, \quad x_i \in \{-1, 0, 1\}.$$

Clearly, $n_+ > 0$. To prove our result, we assume, on the contrary, that $n_- > 0$ for any $\mathbf{x} \in \mathcal{A}_C$. Let l denote the smallest index such that $x_l = -1$. Note that $C > 0$, and that

$$2^{-l} = 2^{-(l+1)} + 2^{-(l+2)} + \cdots + 2^{-(l+n)} + \cdots. \tag{4.19}$$

There must exist an index $k < l$ such that

$$x_k = 1, x_{k+1} = \cdots = x_{l-1} = 0.$$

From Lemma 4.1, we have

$$x_k 2^{-k} + \cdots + x_l 2^{-l} = 2^{-k} - 2^{-l} = 2^{-(k+1)} + \cdots + 2^{-l}. \tag{4.20}$$

It is clear that all the coefficients in the RHS of (4.20) are 1. One can always apply this procedure until all the resulting coefficients, denoted as $x_i^\star, i = 1, \cdots, b$, are greater or equal to zero. This

is a contradiction to $n^{\star}_{-} > 0$. The existence of $\mathbf{x}^{\star}$ is proved.

To prove the uniqueness of $\mathbf{x}^{\star}$, we assume that there exist another $\mathbf{x}' \in \mathcal{A}_C$ such that

$$\sum_{i=1}^{b} x_i^{\star} 2^{-i} = \sum_{i=1}^{b} x_i' 2^{-i} = C, \quad x_i^{\star}, x_i' \in \{0,1\}, \quad i = 1, \cdots, b. \tag{4.21}$$

It follows from (4.21) that

$$\sum_{i=1}^{b} (x_i^{\star} - x_i') 2^{-i} = 0.$$

From previous result, we have $\mathbf{x}^{\star} = \mathbf{x}'$. The uniqueness of $\mathbf{x}^{\star}$ is proved.

For the case of $C < 0$, the proof is similar. □

From Theorem 4.1, we see that for any $C \in \mathcal{S}$, there exists a unique $\mathbf{y}$ such that

$$\sum_{i=1}^{b} y_i 2^{-i} = C, \quad \begin{cases} y_i \geq 0, & \text{if } C > 0, \\ y_i = 0, & \text{if } C = 0, \\ y_i \leq 0, & \text{if } C < 0. \end{cases} \tag{4.22}$$

It follows from Lemma 4.1 and Theorem 4.1 that for $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_b\}$, and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_b\}$, any equivalent transform

$$\sum_{i=1}^{b} \alpha_i 2^{-i} = \sum_{i=1}^{b} \beta_i 2^{-i} = C, \quad \alpha_i, \beta_i \in \{-1, 0, 1\},$$

can be achieved by applying equation (4.15) in Lemma 4.1. Furthermore, the non-zero elements in the LHS of equation (4.15) is less than or equal to that of the RHS of (4.15) when $j - i \geq 2$. Based on this nice property, we shall first devise an algorithm for finding a $\bar{\mathbf{y}}$ such that the number of non-zero elements is minimized. This least number is denoted as χ_C, where

$$\chi_C = \sum_{i=1}^{b} |\bar{y}_i| = \min \sum_{i=1}^{b} |x_i|,$$

and

$$\sum_{i=1}^{b} \bar{y}_i 2^{-i} = \sum_{i=1}^{b} x_i 2^{-i} = C, \quad x_i \in \{-1, 0, 1\}.$$

Algorithm 4.1

Step 1:

For any C such that $C > 0$, find the $\boldsymbol{y}$ according to (5.27).

Step 2:

Find all the terms "0, $\underbrace{1, \cdots, 1}_{m}$" in $(y_1, \cdots, y_b)$, where $m \geq 2$. Replace each of them by "1, $\underbrace{0, \cdots, 0,}_{m-1}$ -1". Let the resulting coefficients be denoted as $(\tilde{y}_1, \cdots, \tilde{y}_b)$.

Step 3:

Find all the terms "-1, 1" in $(\tilde{y}_1, \cdots, \tilde{y}_b)$. Replace each of them by "0, -1". Let the resulting coefficients be denoted as $(\bar{y}_1, \cdots, \bar{y}_b)$.

Stop.

The following theorem shows that the $\bar{\boldsymbol{y}} = [\bar{y}_1, \cdots, \bar{y}_b]^\top$ obtained by Algorithm 4.1 has the least number of non-zero elements.

Theorem 4.2. *For any $C > 0$, where $C \in \mathcal{S}$, let $\bar{\boldsymbol{y}} = [\bar{y}_1, \cdots, \bar{y}_b]^\top$ be the coefficient vector obtained by Algorithm 4.1 such that*

$$\sum_{i=1}^{b} \bar{y}_i 2^{-i} = C, \bar{y}_i \in \{-1, 0, 1\}.$$

Then, $\bar{\boldsymbol{y}}$ has the least number of non-zero elements .

Proof. Clearly, after Step 2 of Algorithm 4.1, the replaced coefficients $(\tilde{y}_1, \cdots, \tilde{y}_b)$ has following features:

(a) Suppose that $(\tilde{y}_1, \cdots, \tilde{y}_b)$ contains the term "1, 1". Then, the "1, 1" must be contained in the structure of "0, 1, 1, $\underbrace{0, \cdots, 0,}_{k}$ -1", where $k \geq 1$. In this case, there is no need to replace the term "0, 1, 1" within "0, 1, 1, $\underbrace{0, \cdots, 0,}_{k}$ -1" by "1, 0, -1". This is because if we do so, we get "1, 0, -1, $\underbrace{0, \cdots, 0,}_{k}$ -1", which has the same number of non-zero elements.

(b) Suppose that $(\tilde{y}_1, \cdots, \tilde{y}_b)$ contains the term "-1, 1". Then, this term must be contained in the structure of "0, -1, 1, $\underbrace{0, \cdots, 0,}_{k}$ -1", where $k \geq 1$.

For (b), we can apply Step 3 to convert the term "0, -1, 1, $\underbrace{0, \cdots, 0,}_{k}$ -1" into "0, 0, -1, $\underbrace{0, \cdots, 0,}_{k}$ -1", where $k \geq 1$. Let the resulting coefficients be denoted as $(\bar{y}_1, \cdots, \bar{y}_b)$.

Now, it is easy to see that Lemma 4.1 cannot be applied to $(\bar{y}_1, \cdots, \bar{y}_b)$ to reduce the non-zero elements any further, this means that $(\bar{y}_1, \cdots, \bar{y}_b)$ has the least non-zero elements such that $\sum_{i=1}^{b} |\bar{y}_i| = C$. □

Remark 4.1. For the case when $E < 0$, a procedure similar to that reported in Algorithm 4.1 can be used to obtain a $\bar{\boldsymbol{y}}$ which contains the least number of non-zero elements. Also note that, the results presented in Theorem 1 and Theorem 2 are important properties of binary representations, Algorithm 1 is the Canonical-Signed Digit (CSD) representation.

Now, we are in the position to present an algorithm to find the desired reduced discrete search region for Problem **P**.

Algorithm 4.2:

Step 1:

Find the infinite precision solution $H^\star$ of Problem **P**.

Step 2:

For each $h_{n,m}$, $n = 1, \cdots, N$ and $m = 1, \cdots, M$, consider the following two cases:

I. $h^\star_{n,m} \in \mathcal{S}$. Define $\mathcal{M}_{n,m} = \left\{ \begin{pmatrix} h^\star_{n,m} \\ \chi_{h^\star_{n,m}} \end{pmatrix} \right\}$ where $\chi_{h^\star_{n,m}}$ is obtained from $h^\star_{n,m}$ by applying Algorithm 4.1.

II. $h^\star_{n,m} \notin \mathcal{S}$. There are three cases to be considered.

1. There exist two constants $C_1^{n,m}$, $C_2^{n,m} \in S$ such that $h^\star_{n,m} \in (C_1^{n,m}, C_2^{n,m})$. Let $\mathcal{M}_{n,m} = \left\{ \begin{pmatrix} C_l^{n,m} \\ \chi_{C_l^{n,m}} \end{pmatrix}, \begin{pmatrix} C_u^{n,m} \\ \chi_{C_u^{n,m}} \end{pmatrix} \right\}$, where $C_l^{n,m}$ is the largest feasible lower bound of $h^\star_{n,m}$ in $\mathcal{S}$ and $C_u^{n,m}$ is the least feasible upper bound of $h^\star_{n,m}$ in $\mathcal{S}$.
2. $h^\star_{n,m} > \max(\mathcal{S})$. Let $\mathcal{M}_{n,m} = \left\{ \begin{pmatrix} \max(\mathcal{S}) \\ \chi_{\max(\mathcal{S})} \end{pmatrix} \right\}$.
3. $h^\star_{n,m} < \min(\mathcal{S})$. Let $\mathcal{M}_{n,m} = \left\{ \begin{pmatrix} \min(\mathcal{S}) \\ \chi_{\min(\mathcal{S})} \end{pmatrix} \right\}$.

Step 3:

For each $n = 1, \cdots, N$ and $m = 1, \cdots, M$, let $\mathcal{M}^1_{n,m}$ be the set that contains the first element of each of all 2-dimensional vectors in the set $\mathcal{M}_{n,m}$. Furthermore, let $\mathcal{M}^2_{n,m}$ be the set which contains the second element of each of all 2-dimensional vectors in the set $\mathcal{M}_{n,m}$, and let

$$\chi_{\mathcal{M}} = \sum_{n=1}^{N} \sum_{m=1}^{M} \max(\mathcal{M}^2_{n,m}).$$

If $\chi_{\mathcal{M}} \leq N_1$, then, for each $n = 1, \cdots, N$ and $m = 1, \cdots, M$, $h_{n,m} \in \mathcal{M}^1_{n,m}$. Stop. Otherwise, go to Step 4.

Step 4:

Increase the size of $\mathcal{M}_{n,m}$ as follows. For each $n = 1, \cdots, N$ and $m = 1, \cdots, M$, consider the following two cases:

(I) $h^{\star}_{n,m} \in \mathcal{S}$. If $h^{\star}_{n,m} > 0$, define the set

$$\mathcal{M}_{n,m} = \Big\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \cdots, \begin{pmatrix} C^{n,m}_{s-1} \\ \chi_{C^{n,m}_{s-1}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{s} \\ \chi_{C^{n,m}_{s}} \end{pmatrix}, \begin{pmatrix} h^{\star}_{n,m} \\ \chi_{h^{\star}_{n,m}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t} \\ \chi_{C^{n,m}_{t}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t+1} \\ \chi_{C^{n,m}_{t+1}} \end{pmatrix}, \cdots \Big\}. \tag{4.23}$$

On the other hand, if $h^{\star}_{n,m} < 0$, define the set

$$\mathcal{M}_{n,m} = \Big\{ \cdots, \begin{pmatrix} C^{n,m}_{s-1} \\ \chi_{C^{n,m}_{s-1}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{s} \\ \chi_{C^{n,m}_{s}} \end{pmatrix}, \begin{pmatrix} h^{\star}_{n,m} \\ \chi_{h^{\star}_{n,m}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t} \\ \chi_{C^{n,m}_{t}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t+1} \\ \chi_{C^{n,m}_{t+1}} \end{pmatrix}, \cdots, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Big\}. \tag{4.24}$$

Here,

$$C^{n,m}_{s} = \max\{C \in \mathcal{S} | C < h^{\star}_{n,m}, \chi_C < \chi_{h^{\star}_{n,m}}\},$$

$$C^{n,m}_{s-1} = \max\{C \in \mathcal{S} | C < C^{n,m}_{s}, \chi_C < \chi_{C^{n,m}_{s}}\},$$

and

$$C^{n,m}_{t} = \min\{C \in \mathcal{S} | C > h^{\star}_{n,m}, \chi_C < \chi_{h^{\star}_{n,m}}\},$$

$$C^{n,m}_{t+1} = \min\{C \in \mathcal{S} | C > C^{n,m}_{t}, \chi_C < \chi_{C^{n,m}_{t}}\}.$$

(II) $h^{\star}_{n,m} \notin \mathcal{S}$. There are three cases to be considered.

1. There exist two constants $C^{n,m}_1$, $C^{n,m}_2 \in S$ such that $h^{\star}_{n,m} \in (C^{n,m}_1, C^{n,m}_2)$. If $h^{\star}_{n,m} > 0$, let

$$\mathcal{M}_{n,m} = \Big\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \cdots, \begin{pmatrix} C^{n,m}_{s-1} \\ \chi_{C^{n,m}_{s-1}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{s} \\ \chi_{C^{n,m}_{s}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{l} \\ \chi_{C^{n,m}_{l}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{u} \\ \chi_{C^{n,m}_{u}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t} \\ \chi_{C^{n,m}_{t}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t+1} \\ \chi_{C^{n,m}_{t+1}} \end{pmatrix}, \cdots \Big\}.$$

   On the other hand, if $h^{\star}_{n,m} < 0$, let

$$\mathcal{M}_{n,m} = \Big\{ \cdots, \begin{pmatrix} C^{n,m}_{s-1} \\ \chi_{C^{n,m}_{s-1}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{s} \\ \chi_{C^{n,m}_{s}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{l} \\ \chi_{C^{n,m}_{l}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{u} \\ \chi_{C^{n,m}_{u}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t} \\ \chi_{C^{n,m}_{t}} \end{pmatrix}, \begin{pmatrix} C^{n,m}_{t+1} \\ \chi_{C^{n,m}_{t+1}} \end{pmatrix}, \cdots \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Big\}.$$

   Here, $C^{n,m}_{l}$ is the largest feasible lower bound of $h^{\star}_{n,m}$ in $\mathcal{S}$ and $C^{n,m}_{u}$ is the least feasible upper bound of $h^{\star}_{n,m}$ in $\mathcal{S}$, where

$$C^{n,m}_{s} = \max\{C \in \mathcal{S} | C < C^{n,m}_{l}, \chi_C < \chi_{C^{n,m}_{l}}\},$$

$$C_{s-1}^{n,m} = \max\{C \in \mathcal{S} | C < C_s^{n,m}, \chi_C < \chi_{C_s^{n,m}}\},$$

and

$$C_t^{n,m} = \min\{C \in \mathcal{S} | C > C_u^{n,m}, \chi_C < \chi_{C_u^{n,m}}\},$$

$$C_{t+1}^{n,m} = \min\{C \in \mathcal{S} | C > C_t^{n,m}, \chi_C < \chi_{C_t^{n,m}}\}.$$

2. $h_{n,m}^\star > \max(\mathcal{S})$. Let

$$\mathcal{M}_{n,m} = \left\{ \binom{0}{0}, \cdots, \binom{C_{s-1}^{n,m}}{\chi_{C_{s-1}^{n,m}}}, \binom{C_s^{n,m}}{\chi_{C_s^{n,m}}}, \binom{\max(\mathcal{S})}{\chi_{\max(\mathcal{S})}} \right\}.$$

3. $h_{n,m}^\star < \min(\mathcal{S})$. Let

$$\mathcal{M}_{n,m} = \left\{ \binom{\min(\mathcal{S})}{\chi_{\min(\mathcal{S})}}, \binom{C_t^{n,m}}{\chi_{C_t^{n,m}}}, \binom{C_{t+1}^{n,m}}{\chi_{C_{t+1}^{n,m}}}, \cdots, \binom{0}{0} \right\}.$$

Here,

$$C_s^{n,m} = \max\{C \in \mathcal{S} | C < \max(\mathcal{S}), \chi_C < \chi_{\max(\mathcal{S})}\},$$

$$C_{s-1}^{n,m} = \max\{C \in \mathcal{S} | C < C_s^{n,m}, \chi_C < \chi_{C_s^{n,m}}\},$$

and

$$C_t^{n,m} = \min\{C \in \mathcal{S} | C > \min(\mathcal{S}), \chi_C < \chi_{\min(\mathcal{S})}\}$$

$$C_{t+1}^{n,m} = \min\{C \in \mathcal{S} | C > C_t^{n,m}, \chi_C < \chi_{C_t^{n,m}}\}.$$

Stop.

Remark 4.2. The idea of Algorithm 4.2 is somewhat similar to that of quantization method. In the traditional quantization method, a discrete feasible solution is directly assigned to the coefficient $c_{n,m}$ which has the largest absolute value. Instead, our algorithm find a search region for each of the coefficient, where the search region is obtained by adopting the idea of quantization—that is to expand the search region of the coefficient which has the least deviation from its infinite precision solution. Thus, our algorithm produces more options for each of the coefficients. It is easy to see that the quantization solution is contained in the obtained search region. After applying Algorithm 4.2, the search region of Problem (P) is approximated by a greatly condensed set. However, it is still a difficult integer programming.

In the next subsection, we will introduce a newly developed exact penalty function method to solve the approximate problem.

4.3.2 A new exact penalty function method

For each $h_{n,m}$, where $n = 1, \cdots, N$ and $m = 1, \cdots, M$, let the set $\mathcal{M}^1_{n,m}$ be obtained by Algorithm 4.2. Suppose that

$$\mathbb{M} = \{\mathcal{M}^1_{1,1}, \cdots, \mathcal{M}^1_{N,M}\}.$$

Then, Problem $\tilde{\mathbf{P}}$ can be equivalently stated as follows:

$$\min \ \tilde{G}(\mathbf{H}) \tag{4.25}$$

where $h_{n,m} \in \mathcal{M}^1_{n,m}$, $n = 1, \cdots, N$ and $m = 1, \cdots, M$. Let this problem be referred to as Problem $\mathbf{P_d}$.

Clearly, Problem $\mathbf{P_d}$ is a standard integer programming problem. We adopt the idea introduced in Chapter 3 to solve this problem.

First, we assume that for each $n = 1, \cdots, N$ and $m = 1, \cdots, M$, $\mathcal{M}^1_{n,m}$ has $l_{n,m}$ distinct elements, i.e.

$$\mathcal{M}^1_{n,m} = \{u^1_{n,m}, \cdots, u^{l_{n,m}}_{n,m}\}, \quad n = 1, \cdots, N \text{ and } m = 1, \cdots, M.$$

Then, we introduce new variables $\alpha_{n,m,j}$ satisfying

$$\sum_{j=1}^{l_{n,m}} \alpha_{n,m,j} = 1, \quad n = 1, \cdots, N, \quad m = 1, \cdots, M, \tag{4.26}$$

$$\alpha_{n,m,j}(1 - \alpha_{n,m,j}) \leq 0, \quad n = 1, \cdots, N, \quad m = 1, \cdots, M, \quad j = 1, \cdots, l_{n,m}, \tag{4.27}$$

$$0 \leq \alpha_{n,m,j} \leq 1, \quad n = 1, \cdots, N, \quad m = 1, \cdots, M, \quad j = 1, \cdots, l_{n,m}. \tag{4.28}$$

Now, we consider the following problem:

$$\min \ \tilde{G}(\mathbf{H}) = \bar{G}(\boldsymbol{\alpha})$$

where

$$\boldsymbol{\alpha} = [\alpha_{1,1,1}, \cdots, \alpha_{1,1,l_{1,1}}, \cdots\cdots, \alpha_{N,M,1}, \cdots, \alpha_{N,M,l_{N,M}}]^\top, \tag{4.29}$$

$$h_{n,m} = \sum_{j=1}^{l_{n,m}} \alpha_{n,m,j} u^j_{n,m}, \ n = 1, \cdots, N; \quad m = 1, \cdots, M. \tag{4.30}$$

subject to

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{j=1}^{l_{n,m}} \alpha_{n,m,j} \chi_{u^j_{n,m}} \leq N_1 \tag{4.31}$$

and constraints (4.26), (4.27) and (4.28). Let this problem be referred to as Problem $\bar{\mathbf{P}}$.

Noting that, for each $n = 1, \cdots, N$ and $m = 1, \cdots, M$, the solution of (4.26)-(4.28) is that there exists only one $k \in \{1, 2, \cdots, l_{n,m}\}$ such that $\alpha_{n,m,k} = 1$, while $\alpha_{n,m,j} = 0$ for all $j \neq k$.

This indicates that for each $n = 1, \cdots, N$ and $m = 1, \cdots, M$, $h_{n,m}$ can only take a discrete value from the set $\mathcal{M}^1_{n,m}$, implying that Problem $\mathbf{P_d}$ is equivalent to Problem $\bar{\mathbf{P}}$.

As it is noted in Chapter 3, the inequality constraints (4.27) are very difficult to be satisfied by using existing optimization techniques. Thus, as in Chapter 3, we shall introduce a new exact penalty function given below:

$$F_\kappa(\boldsymbol{\alpha}, \epsilon) = \begin{cases} \bar{G}(\boldsymbol{\alpha}), & \textit{if } \epsilon = 0,\ \boldsymbol{\alpha} \textit{ is feasible for Problem } \bar{\mathbf{P}}, \\ \bar{G}(\boldsymbol{\alpha}) + \epsilon^{-\eta}\Delta(\boldsymbol{\alpha}, \epsilon) + \kappa\epsilon^{\beta}, & \textit{if } \epsilon > 0, \\ +\infty, & \textit{otherwise}, \end{cases}$$

where $\epsilon > 0$ is a new decision variable, and the constraint violation $\Delta(\boldsymbol{\alpha}, \epsilon)$ is defined by

$$\begin{aligned} \Delta(\boldsymbol{\alpha}, \epsilon) = & \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{j=1}^{l_{n,m}} \max\left\{0,\ \alpha_{n,m,j}(1-\alpha_{n,m,j}) - \epsilon^{\gamma}\right\}^2 \\ & + \sum_{n=1}^{N}\sum_{m=1}^{M}\Big(\sum_{j=1}^{l_{n,m}} \alpha_{n,m,j} - 1 - \epsilon^{\gamma}\Big)^2 \\ & + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{j=1}^{l_{n,m}} \max\left\{0,\ \alpha_{n,m,j} - 1 - \epsilon^{\gamma}\right\}^2 \\ & + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{j=1}^{l_{n,m}} \max\left\{0,\ -\alpha_{n,m,j} - \epsilon^{\gamma}\right\}^2 \\ & + \max\left\{0,\ \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{j=1}^{l_{n,m}} \alpha_{n,m,j}\chi_{u^1_{n,m}} - N_1 - \epsilon^{\gamma}\right\}^2. \end{aligned}$$

Here, β, γ and η are positive real numbers, and κ is a penalty parameter.

Now, consider the following problem:

$$\begin{array}{ll} \min & F_\kappa(\boldsymbol{\alpha}, \epsilon) \\ \text{subject to} & \epsilon > 0 \end{array} \tag{4.32}$$

Let this problem be called Problem $\mathbf{P}_\kappa$.

In what follows, we shall give a brief introduction on the convergence result of the proposed method.

A Convergence Analysis

Let $\{\kappa_k\}_{k=1}^{\infty}$ be an increasing sequence of penalty parameters such that $\kappa_k \to \infty$. Furthermore, let $(\boldsymbol{\alpha}^{(k),\star}, \epsilon^{(k),\star})$ denote the solution of Problem $\mathbf{P}_{\kappa_k}$ corresponding to κ_k. We assume that the following hypotheses are satisfied:

(H_1) *The objective function as well as all constraint functions are continuously differentiable with respect to their respective augments.*
(H_2) *The linearly independent constraint qualification (LICQ) given in Definition 3.1 is satisfied*

at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$, *where* $\boldsymbol{\alpha}^\star$ *is a local minimizer of Problem* $\bar{\mathbf{P}}$.

(H_3) *Let* G_i, $i = 1, \cdots, 3\sum_{n=1}^{N}\sum_{m=1}^{M} l_{n,m}$ *and* H_i, $i = 1, \cdots, N \times M$, *denote, respectively, the inequality and equality constraints in Problem* $\bar{\mathbf{P}}$. *Then, it holds that*

$$\max\{0, G_i(\boldsymbol{\alpha}^{(k),\star})\} = o\big((\epsilon^{(k),\star})^{\delta_1}\big), i = 1, \cdots, 3\sum_{n=1}^{N}\sum_{m=1}^{M} l_{n,m};$$

and

$$H_i(\boldsymbol{\alpha}^{(k),\star}) = o\big((\epsilon^{(k),\star})^{\delta_2}\big), i = 1, \cdots, N \times M,$$

where δ_1 *and* δ_2 *are positive constants, and*

$$\lim_{\varsigma \to 0} \frac{o(\varsigma^\iota)}{\varsigma^\iota} = 0,$$

with ι *being* δ_1 *or* δ_2.

The main convergence results are presented in the following three Theorems. Their proofs are similar to those given for relevent theorems in Chapter 3, and hence are omitted.

Theorem 4.3. *Suppose that the hypotheses* (H_1)-(H_3) *are satisfied, and that* $\gamma > \eta$, $\delta = \min(\delta_1, \delta_2) > \eta$, $-\eta - 1 + 2\delta > 0$, *and* $2\gamma - \eta - 1 > 0$. *Then, as* $\boldsymbol{\alpha}^{(k),\star} \to \boldsymbol{\alpha}^\star \in S_0$ *and* $\epsilon^{(k),\star} \to \epsilon^\star = 0$, *it holds that*

$$F_{\kappa_k}(\boldsymbol{\alpha}^{(k),\star}, \epsilon^{(k),\star}) \longrightarrow F_{\kappa_k}(\boldsymbol{\alpha}^\star, 0) = F(\boldsymbol{\alpha}^\star),$$

$$\nabla_{(\boldsymbol{\alpha},\epsilon)} F_{\kappa_k}(\boldsymbol{\alpha}^{(k),\star}, \epsilon^{(k),\star}) \longrightarrow \nabla_{(\boldsymbol{\alpha},\epsilon)} F_{\kappa_k}(\boldsymbol{\alpha}^\star, 0) = (\nabla F(\boldsymbol{\alpha}^\star), 0).$$

Proof. The proof is similar to that given for Theorem 3.1 and hence is omitted. □

The above results indicate that the constructed exact penalty function is continuously differentiable with its gradients having finite limits.

In the next theorem, it is shown that the sequence $(\boldsymbol{\alpha}^{(k),\star}, \epsilon^{(k),\star})$ of the local minimizers will converge to a feasible point of the original problem $\bar{\mathbf{P}}$ with finite objective function value. Furthermore, this feasible point is a local minimizer of Problem $\bar{\mathbf{P}}$.

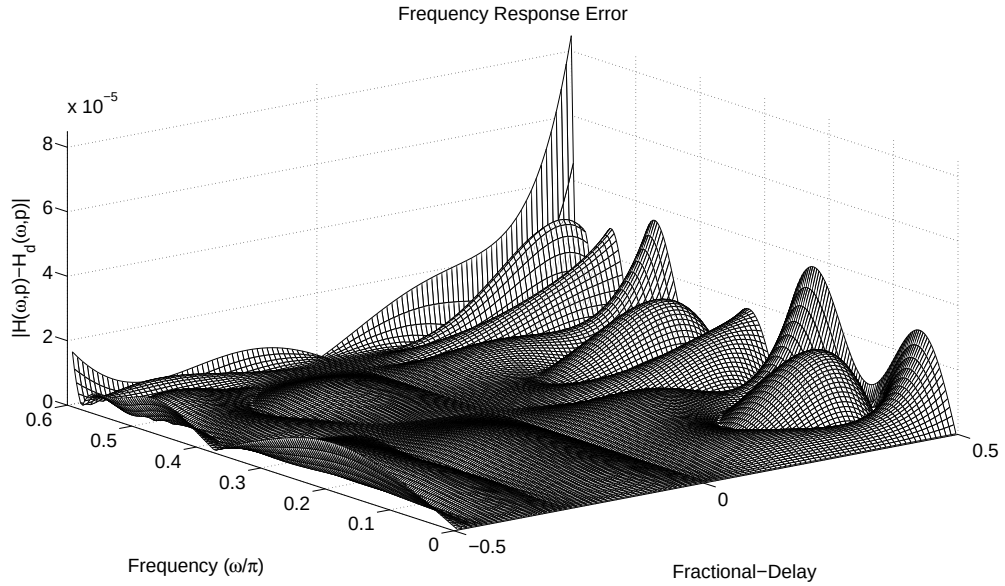
Theorem 4.4. *Let* $\epsilon^{(k),\star} \to \epsilon^\star = 0$, $\boldsymbol{\alpha}^{(k),\star} \to \boldsymbol{\alpha}^\star \in S_0$ *be such that* $F_{\kappa_k}(\boldsymbol{\alpha}^\star, \epsilon^\star)$ *is finite. Then,* $\boldsymbol{\alpha}^\star$ *is a local minimizer of the original Problem* $\bar{\mathbf{P}}$.

Proof. The proof is similar to that given for Lemma 3.2 and hence is omitted. □

The exactness of the proposed penalty function is given in the following theorem.

Theorem 4.5. *Let* $(\boldsymbol{\alpha}^{(k),\star}, \epsilon^{(k),\star})$ *be a local minimizer of Problem* $\mathbf{P}_{\kappa_k}$. *Suppose that* $(\boldsymbol{\alpha}^{(k),\star}, \epsilon^{(k),\star}) \to (\boldsymbol{\alpha}^\star, \epsilon^\star)$ *as* $k \to +\infty$, *and that the parameters* η, γ *and* δ *satisfy the same conditions as stated in Theorem 4.3. Then, there exists a* $k_0 > 0$, *such that for* $k \geq k_0$, $\epsilon^{(k),\star} = 0$, *and* $\boldsymbol{\alpha}^{(k),\star}$ *is a local minimizer of Problem* $\bar{\mathbf{P}}$.

Figure 4.1: Absolute error of variable frequency response (Infinite precision solution)



Proof. The proof is similar to that given for Theorem 3.2 and hence is omitted. □

Theorem 4.5 indicates that, under some mild assumptions, a local minimizer of the penalty Problem (P_κ) is a local minimizer of Problem $\bar{\mathbf{P}}$, when the penalty parameter κ is sufficiently large.

4.4 Simulation result

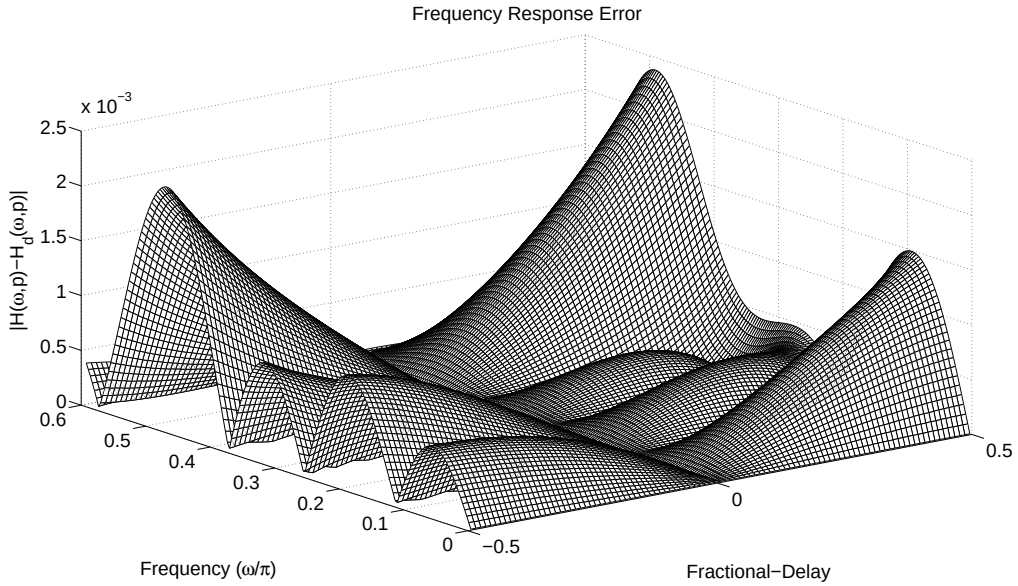
Consider the design of an allpass variable fractional delay filter, where the bandwidth under consideration for the filter is from 0 to 0.6π. The length of each FIR filters used in the Farrow structure is $L = 4$ with $N = 8$. The number of bits is $b = 10$ and the range for p is chosen as $\Delta = [-0.5, 0.5]$. The maximum allowable number of nonzero SPT term is $N_1 = 67$. The weighting functions are set as:

$$W_1(\omega) = 1, \qquad \text{for}\, \omega \in [0, 0.6\pi],$$
$$W_2(p) = 1, \qquad \text{for}\, p \in [-0.5, 0.5].$$

From (4.12), it follows that the infinite precision solution of (4.25) is -223.2442 dB. Figure 4.1 shows the corresponding absolute error of the variable frequency response.

For comparison, we apply our method and quantization method to Problem $\tilde{\mathbf{P}}$. The basic idea of the quantization procedure (see [67]) is briefly stated below. First, obtain the infinite-precision solution of Problem $\tilde{\mathbf{P}}$. Then, the algorithm assigns one SPT term at a time to the coefficient which has the largest absolute value of the solution, such that the difference between the SPT term and the coefficient is minimized. After a coefficient has received a SPT term, the corresponding value of the coefficient is decreased by the allocated SPT term. The process is

Figure 4.2: Absolute error of variable frequency response (Proposed method)



repeated until the maximum allowable number of the SPT term is reached.

The results obtained by proposed method and the quantization procedure are given in Table 4.1. Figure 4.2 shows the absolute error of variable frequency response obtained by our method.

Table 4.1: Objective function value [dB]

Proposed Method	Quantization
-134.9939	-114.0715

Figure 4.3 shows the absolute error of variable frequency response obtained by quantization method. Figure 4.4 shows the maximum radius of the poles of the filter obtained by our method as the value of the fractional-delay varies. Obviously, all the poles are inside the unit circle, meaning that the filter obtained is stable.

It is clearly seen from Figure 4.2 and Figure 4.3 that our method can achieve a much higher accuracy when compared with that obtained by the quantization method. To make a more comprehensive comparison between the quantization method and the proposed method, a range of values of N_1 is chosen. The result is shown in Table 4.2 and Figure 4.5

It is clear from Table 4.2 that the curve generated by the proposed method is monotonically decreasing. Except the case of having the same objective function value for both methods when the maximum allowable number of nonzero SPT terms is $N_1 = 1$, the proposed method can always achieve a much better objective function value when compared with those obtained by the traditional quantization method.

Figure 4.3: Absolute error of variable frequency response (Quantization method)

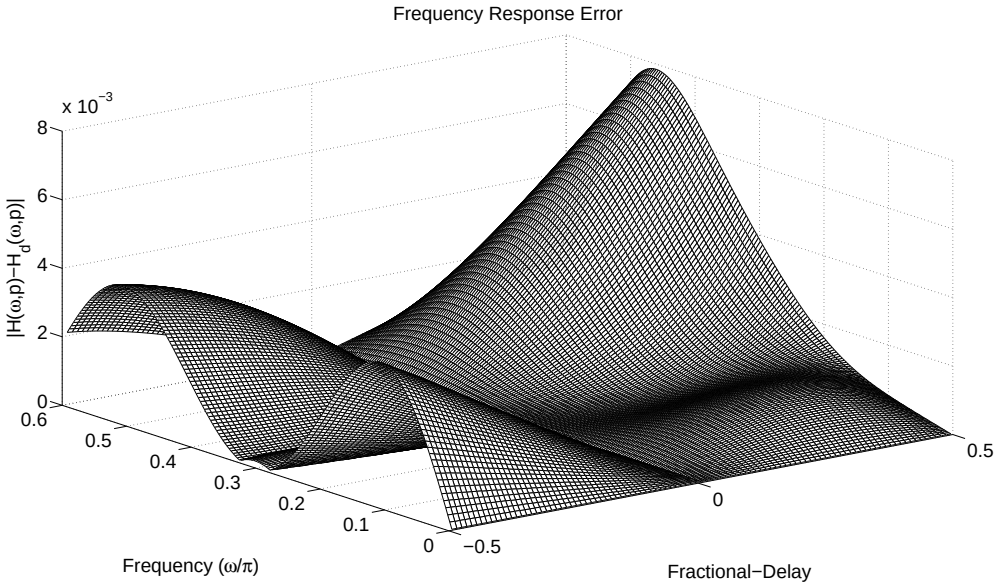


Figure 4.4: Maximum pole radius (Proposed method)

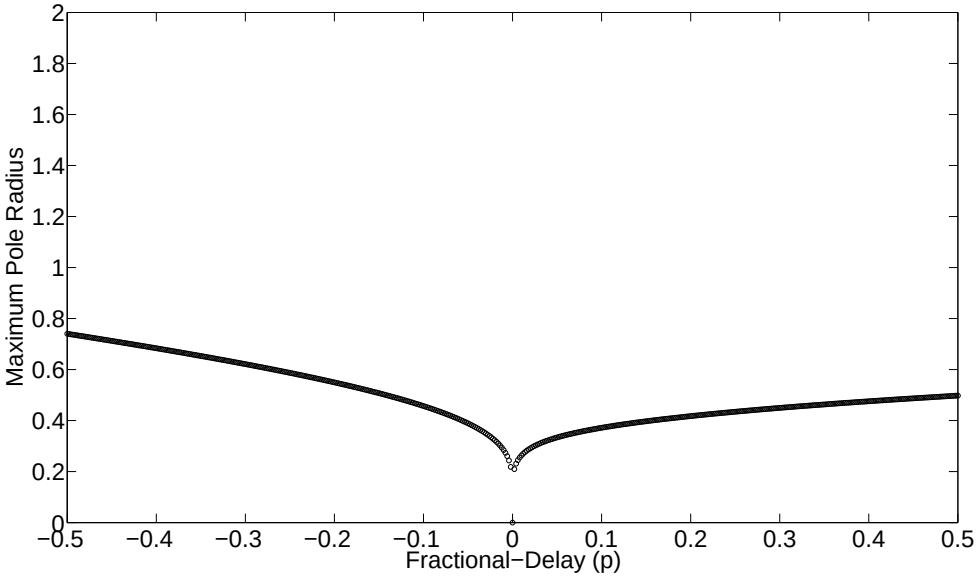


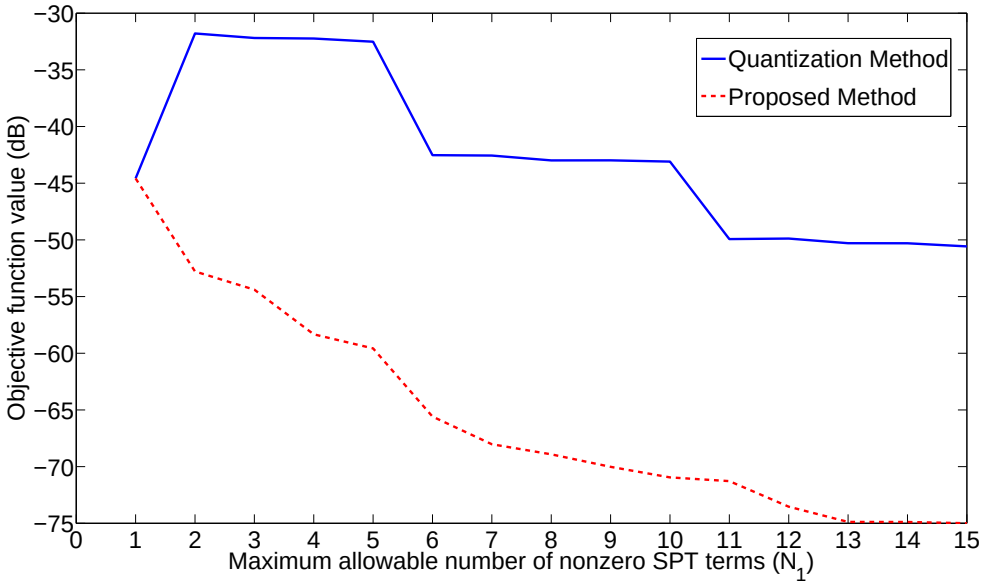
Table 4.2: Objective function value [dB]

N_1	Proposed Method	Quantization	Stability
1	-44.57932465	-44.57932465	stable
2	-52.78539948	-31.78914912	stable
3	-54.39156437	-32.19192863	stable
4	-58.33215624	-32.23734926	stable
5	-59.57329168	-32.51561199	stable
6	-65.58700394	-42.52760972	stable
7	-68.02584876	-42.55855436	stable
8	-68.9149	-42.99028226	stable
9	-70.0226	-42.98284857	stable
10	-70.94982679	-43.09320262	stable
11	-71.27931426	-49.93136406	stable
12	-73.53893253	-49.88524948	stable
13	-74.8817593	-50.29020691	stable
14	-74.88505584	-50.29843054	stable
15	-74.99477304	-50.57351345	stable

4.5 Conclusion

In this chapter, the design of allpass variable fractional delay filter with signed powers-of-two coefficients is approximated by a quadratic integer programming problem. We developed a two-step scheme for constructing a desired reduced discrete search region containing the global minimizer of the problem. Then, an exact penalty function method is introduced to solve the quadratic integer programming problem from within the obtained reduced discrete search region. Simulation result shows that the proposed method is effective.

Figure 4.5: Comparison of Objective function value



CHAPTER 5

Optimal discrete-valued control computation

5.1 Introduction

In many practical optimal control problems, the control is only allowed to assume values from a finite number of values. Such problems are called *optimal discrete-valued control problems*. Optimal discrete-valued control problems arise in many applications, including train control [46], switched amplifier design [110], submarine operation [99], sensor scheduling [126] and hybrid power system design [118, 127]. To solve an optimal discrete-valued control problem, we need to determine the order in which the different control values are operated, as well as the times at which the control switches from one value to another. Since the ordering of control values is discrete in nature, classical optimal control methods are not applicable to this type of problem.

In [46], the driving strategy for a diesel train traveling on a level track is considered. The train only has three modes of operation— accelerate, coast and brake — and thus the problem of controlling the train so that fuel consumption is minimized is an optimal discrete-valued control problem. An optimality condition is derived in [46] for solving this problem. However, this condition is only applicable to the train problem, and is not applicable to general optimal discrete-valued control problems.

In [62], a time-scaling transformation technique is developed for solving optimal discrete-valued control problems. Under this transformation, the original problem with variable control switching points is transformed into an ordinary optimal control problem with known and fixed switching points. Thus, the transformed problem can be solved by many existing optimal control methods. However, the time-scaling transformation introduces many additional switches, and therefore the transformed problem is not equivalent to the original problem.

In [125], a new approach is proposed for solving nonlinear mixed discrete programming problems. The idea is to introduce a set of new continuous variables and transform the mixed discrete programming problem into a conventional optimization problem involving only continuous variables. In principle, this new problem can be solved by using existing nonlinear programming techniques. However, the transformation introduces additional equality and inequality constraints, for which the quadratic inequality constraints are extremely difficult to satisfy in practice.

In Chapter 2, an exact penalty method is proposed for solving semi-infinite programming

problems. This method is adapted in [63] to develop an effective algorithm for solving optimal control problems with continuous inequality constraints via solving a sequence of penalized optimal control problems. It is shown that, under some mild assumptions, if the penalty parameter is sufficiently large, the solution obtained for the corresponding penalized optimal control problem will satisfy the continuous inequality constraints of the original optimal control problem. Furthermore, a local optimal solution of the penalized optimal control problem is also a local optimal solution of the original optimal control problem.

This chapter is based on [63, 133, 135, 136]. We consider a class of optimal discrete-valued control problems, where there is an upper bound on the maximum number of control switches. We first apply the transformation reported in [125], under which the discrete-valued control is expressed as a linear combination of piecewise constant controls subject to a linear equality constraint and a set of quadratic inequality constraints. The original problem can then be written equivalently as an optimal control problem with piecewise constant controls subject to the original inequality constraints and the new constraints. Then, the time-scaling transformation [62] is applied to the transformed problem, yielding an optimal control problem with piecewise constant controls and fixed switching times. To solve this new problem, we introduce an exact penalty function to construct a sequence of penalized optimal control problem. Convergence results show that when the penalty parameter is sufficiently large, the penalized optimal control problem is equivalent to the original problem. This penalized optimal control problem can be solved by existing optimal control software packages. Numerical results obtained from solving two train control problems show that the approach proposed is effective.

5.2 Problem formulation

5.2.1 A discrete-valued control problem

Consider the following dynamic system on the time horizon $[0, T]$:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \tag{5.1}$$

with the initial and terminal conditions

$$\mathbf{x}(0) = \mathbf{x}^0, \quad \mathbf{x}(T) = \mathbf{x}^f, \tag{5.2}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the state vector, T is a given terminal time, and $\mathbf{x}^0$ and $\mathbf{x}^f$ are given vectors. We assume that the function $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n$ is continuously differentiable with respect to its arguments.

Let

$$\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m\},$$

where each $\mathbf{u}_j \in \mathbb{R}^r$ is a given vector. We assume that the control $\mathbf{u}$ is a *discrete-valued control* taking values in $\mathbf{U}$. Thus, $\mathbf{u}$ is completely determined by specifying:

- The order in which it assumes the different values in $\mathbf{U}$ (the so-called *switching sequence*); and
- The times at which it switches from one value in $\mathbf{U}$ to another (the so-called *switching times*).

In this chapter, we assume that there is an upper bound N on the maximum number of control switches. A function $\mathbf{u} : [0, T] \to \mathbf{U}$ with at most N switches/discontinuities is called an *admissible control*. Let $\mathscr{U}$ denote the class of all such admissible controls.

Our optimal discrete-valued control problem is stated as follows: Given the dynamic system (5.1)-(5.2), find an admissible control $\mathbf{u} \in \mathscr{U}$ such that the cost function

$$J(\mathbf{u}) = \int_0^T L_0(\mathbf{x}(t), \mathbf{u}(t))dt \tag{5.3}$$

is minimized subject to the constraints

$$g_i(\mathbf{x}(t), \mathbf{u}(t)) \leq 0, \quad t \in [0, T], \quad i = 1, 2, \cdots, p. \tag{5.4}$$

Let this problem be referred to as Problem $\mathbf{P}$. Here, we assume that the functions L_0 and $g_i,\ i = 1, \cdots, p$, are continuously differentiable with respect to each of their arguments.

Most numerical techniques for solving nonlinear optimal control problems— for example, the control parametrization (see [114]) and the state discretization (see [40, 56]) — are applicable only when the control range is a continuous set. Thus, such methods are not applicable to Problem $\mathbf{P}$, in which the control range consists of a finite number of discrete points.

The time-scaling transform introduced in [62], which is also called the control parametrization enhancing technique (CPET), is an effective method for solving optimal discrete-valued control problems. This transformation involves expanding the number of control switches to allow for every possible switching sequence, and then mapping the switching times to fixed points in a new time horizon. This yields a new optimal control problem that can be solved using standard optimal control techniques, see, for example, [114]. However, this transformation introduces many "artificial" switches, and thus the optimal control obtained is always having many more switches than the maximum allowable number of switches. Consequently, the transformed optimal control problem obtained by using the time-scaling transformation introduced in [62] is not equivalent to the original problem. We will introduce an equivalent transformation in the next section.

5.2.2 Problem transformation

Let $\mathscr{V}$ denote the class of all piecewise constant functions mapping $[0, T]$ into $\mathbb{R}^m$ with no more than N switches/discontinuities. Let $\mathbf{v} \in \mathscr{V}$, where $\mathbf{v}(t) = \big[v_1(t), v_2(t), \cdots, v_m(t)\big]^\top$, be an auxiliary control function.

We impose the following constraints:

$$\sum_{j=1}^{m} v_j(t) = 1, \quad t \in [0,T], \tag{5.5a}$$

$$v_j(t)(1 - v_j(t)) \leq 0, \quad t \in [0,T], \quad j = 1,2,\cdots,m, \tag{5.5b}$$

$$0 \leq v_j(t) \leq 1, \quad t \in [0,T], \quad j = 1,2,\cdots,m. \tag{5.5c}$$

The constraints (5.5) ensure that at each time $t \in [0,T]$, there exists exactly one $j \in \{1,\cdots,m\}$ such that $v_j(t) = 1$ and $v_k(t) = 0$ for all $k \neq j$.

To continue, we let

$$\bar{\mathbf{u}}(t) = \sum_{j=1}^{m} v_j(t)\mathbf{u}_j. \tag{5.6}$$

Since $\mathbf{v} \in \mathscr{V}$ and constraints (5.5) hold, $\bar{\mathbf{u}}(t) \in \mathbf{U}$ for all $t \in [0,T]$. Moreover, since $\mathbf{v}$ contains at most N switches, so does $\bar{\mathbf{u}}$. It follows that $\bar{\mathbf{u}}$ is an admissible control for Problem $\mathbf{P}$. In fact, it is easy to see that *any* admissible control for Problem $\mathbf{P}$ can be written in the form of (5.6). Thus, by substituting $\mathbf{u}(t) = \bar{\mathbf{u}}(t)$ into the dynamical system (5.1), we obtain

$$\dot{\mathbf{x}}(t) = \sum_{j=1}^{m} v_j(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}_j). \tag{5.7}$$

Similarly, the constraints (5.4) become

$$\sum_{j=1}^{m} v_j(t) g_i(\mathbf{x}(t), \mathbf{u}_j) \leq 0, \quad t \in [0,T], \quad i = 1,2,\cdots,p. \tag{5.8}$$

Our new optimal control problem is stated as follows: Given the dynamic system (5.7) with the initial and terminal conditions (5.2), find a control $\mathbf{v} \in \mathscr{V}$ such that the cost function

$$\bar{J}(\mathbf{v}) = J(\bar{\mathbf{u}}) = \sum_{j=1}^{m} \int_0^T v_j(t) L_0(\mathbf{x}(t), \mathbf{u}_j) dt$$

is minimized subject to constraints (5.5) and (5.8). Let this problem be referred to as Problem $\bar{\mathbf{P}}$.

It is clear that Problems $\bar{\mathbf{P}}$ and $\mathbf{P}$ are equivalent. Thus, we have the following result.

Theorem 5.1. *Let* $\mathbf{v}^\star = \left[v_1^\star, v_2^\star, \cdots, v_m^\star\right]^\top \in \mathscr{V}$ *and*

$$\bar{\mathbf{u}}^\star(t) = \sum_{j=1}^{m} v_j^\star(t)\mathbf{u}_j.$$

Then $\mathbf{v}^\star$ *is an optimal control for Problem* $\bar{\mathbf{P}}$ *if and only if* $\bar{\mathbf{u}}^\star$ *is an optimal control for Problem* $\mathbf{P}$.

Problem $\bar{\mathbf{P}}$ is a standard optimal control problem subject to the continuous inequality con-

straints (5.8) and the newly introduced constraints (5.5). In principle, many optimal control software packages — for example, MISER [49]— can be used to solve this problem. However, in reality, there are three major difficulties that prevent us from solving Problem $\bar{\mathbf{P}}$ directly:

- The switching times for the new controls v_j are decision variables.
- The feasible region defined by the constraints (5.5) is a disconnected set.
- The newly introduced quadratic constraints (5.5b) are very difficult to deal with by standard gradient-based optimization techniques.

We can overcome the first difficulty by applying the time-scaling transformation (see [62]), in which the variable switching times are mapped into fixed switching times. For the second and third difficulties, we will introduce an exact penalty function method as in [63] and Chapter 2-4. The details are given in the next section.

5.3 Solution procedure

5.3.1 Time-scaling transformation

Recall that the control $\mathbf{v} \in \mathscr{V}$ in Problem $\bar{\mathbf{P}}$ has at most N switches. Let τ_k denote the kth switching time. Then

$$0 = \tau_0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_{N+1} = T.$$

We map these switching times to fixed time points as follows. Let $s \in [0, N+1]$ be a new time variable, and let t be related to s through the following differential equation:

$$\begin{aligned} \dot{t}(s) &= \mu(s), \\ t(0) &= 0, \end{aligned} \tag{5.9}$$

where $\mu(s) = \theta_k = \tau_k - \tau_{k-1}$ for $s \in [k-1, k)$, $k = 1, \cdots, N+1$. We can express the piecewise constant function μ as follows:

$$\mu(s) = \sum_{k=1}^{N+1} \theta_k \chi_{[k-1,k)}(s),$$

where χ_I is the indicator function of I defined by

$$\chi_I(s) = \begin{cases} 1, & \text{if } s \in I, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_{N+1}]^\top \in \mathbb{R}^{N+1}$, and note that $\theta_k = \tau_k - \tau_{k-1}$ is the duration of the kth control value. For each $k = 1, \cdots, N+1$, we have

$$\begin{aligned} t(k) &= \int_0^k \mu(s)ds \\ &= \int_0^k \big[\theta_1 \chi_{[0,1)}(s) + \cdots + \theta_{N+1}\chi_{[N,N+1]}(s)\big]ds \\ &= \theta_1 + \cdots + \theta_k = \tau_k. \end{aligned}$$

This shows that the transformation (5.9) maps each integer k to the kth switching time. Furthermore,

$$t(N+1) = \int_0^{N+1} \mu(s)ds = \sum_{l=1}^{N+1} \theta_l = T. \tag{5.10}$$

Clearly,

$$0 \le \theta_k = \tau_k - \tau_{k-1} \le T, \quad k = 1, \cdots, N+1. \tag{5.11}$$

Thus,

$$0 \le \mu(s) \le T, \quad s \in [0, N+1].$$

Under the time-scaling transform, the control v_j in Problem $\bar{\mathbf{P}}$ becomes

$$\tilde{v}_j(s) = v_j(t(s)) = \sum_{k=1}^{N+1} \xi_{jk}\chi_{[k-1,k)}(s),$$

where ξ_{jk} is the value of v_j on $[\tau_{k-1}, \tau_k)$. Constraints (5.5) become:

$$\sum_{j=1}^{m} \xi_{jk} = 1, \quad k = 1, \cdots, N+1, \tag{5.12a}$$

$$\xi_{jk}(1 - \xi_{jk}) \le 0, \quad j = 1, \cdots, m, \quad k = 1, \cdots, N+1, \tag{5.12b}$$

$$0 \le \xi_{jk} \le 1, \quad j = 1, \cdots, m, \quad k = 1, \cdots, N+1. \tag{5.12c}$$

Define

$$\boldsymbol{\xi}_j = [\xi_{j1}, \cdots, \xi_{j(N+1)}]^\top \in \mathbb{R}^{N+1}$$

and

$$\boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \cdots, \boldsymbol{\xi}_m^\top]^\top \in \mathbb{R}^{m\times(N+1)}.$$

Now, by applying the time-scaling transform to Problem $\bar{\mathbf{P}}$, the dynamical system (5.7) becomes

$$\frac{d\tilde{\mathbf{x}}(s)}{ds} = \mu(s)\sum_{j=1}^{m} \tilde{v}_j(s)\mathbf{f}(\tilde{\mathbf{x}}(s), \mathbf{u}_j) = \sum_{k=1}^{N+1}\sum_{j=1}^{m} \theta_k \xi_{jk}\mathbf{f}(\tilde{\mathbf{x}}(s), \mathbf{u}_j)\chi_{[k-1,k)}(s), \tag{5.13}$$

where

$$\tilde{\mathbf{x}}(s) = \mathbf{x}(t(s)).$$

The initial and terminal conditions (5.2) become

$$\tilde{\mathbf{x}}(0) = \mathbf{x}^0, \quad \tilde{\mathbf{x}}(N+1) = \mathbf{x}^f. \tag{5.14}$$

Problem $\bar{\mathbf{P}}$ may now be written equivalently as the following problem, which we call Problem $\tilde{\mathbf{P}}$: Given the dynamic system (5.13)-(5.14), find $\boldsymbol{\theta} \in \mathbb{R}^{N+1}$ and $\boldsymbol{\xi} \in \mathbb{R}^{m\times(N+1)}$ such that the cost function

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{\xi}) = \int_0^{N+1} \tilde{L}_0(s, \tilde{\mathbf{x}}(s), \boldsymbol{\theta}, \boldsymbol{\xi}) ds, \tag{5.15}$$

where

$$\tilde{L}_0(s, \tilde{\mathbf{x}}(s), \boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{k=1}^{N+1} \sum_{j=1}^{m} \theta_k \xi_{jk} L_0(\tilde{\mathbf{x}}(s), \mathbf{u}_j) \chi_{[k-1,k)}(s),$$

is minimized subject to the constraints

$$\begin{aligned} \tilde{g}_i(s, \tilde{\mathbf{x}}(s), \boldsymbol{\xi}) = \sum_{k=1}^{N+1} \sum_{j=1}^{m} \xi_{jk} g_i(\tilde{\mathbf{x}}(s), \mathbf{u}_j) \chi_{[k-1,k)}(s) \leq 0, \\ s \in [0, N+1], \quad i = 1, \cdots, p, \end{aligned} \tag{5.16}$$

and constraints (5.10), (5.11) as well as (5.12).

In the next section, we will introduce an exact penalty function for Problem $\tilde{\mathbf{P}}$.

5.3.2 An exact penalty function

Problem $\tilde{\mathbf{P}}$ is an optimal control problem subject to the linear constraints (5.10), (5.12a) and (5.12c), the quadratic constraints (5.12b), and the nonlinear continuous inequality constraints (5.16). The continuous inequality constraints (5.16) are continuously differentiable with respect to each of their arguments. By adopting the idea introduced in Chapter 2 and [63], we construct the following exact penalty function:

$$F_\kappa(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon) = \begin{cases} \tilde{J}(\boldsymbol{\theta}, \boldsymbol{\xi}), & \text{if } \epsilon = 0, \text{ and } (\boldsymbol{\theta}, \boldsymbol{\xi}) \text{ is feasible} \\ & \quad \text{for Problem } \tilde{\mathbf{P}}, \\ \tilde{J}(\boldsymbol{\theta}, \boldsymbol{\xi}) + \epsilon^{-\alpha} \Delta(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon) + \kappa \epsilon^{\beta}, & \text{if } \epsilon > 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\epsilon > 0$ is a new decision variable, and the constraint violation $\Delta(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon)$ is defined by

$$\begin{aligned}\Delta(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon) = &\sum_{k=1}^{N+1}\sum_{j=1}^{m} \max\left\{0, \xi_{jk}(1-\xi_{jk}) - \epsilon^{\gamma}\right\}^2 + \sum_{k=1}^{N+1}\sum_{j=1}^{m} \max\left\{0, \xi_{jk} - 1 - \epsilon^{\gamma}\right\}^2 \\ &+ \sum_{k=1}^{N+1}\sum_{j=1}^{m} \max\left\{0, -\xi_{jk} - \epsilon^{\gamma}\right\}^2 + \sum_{k=1}^{N+1}\left\{\sum_{j=1}^{m}\xi_{jk} - 1 - \epsilon^{\gamma}\right\}^2 \\ &+ \sum_{i=1}^{p}\int_0^{N+1} \max\left\{0, \tilde{g}_i(s, \tilde{\mathbf{x}}(s), \boldsymbol{\xi}) - \epsilon^{\gamma}\right\}^2 ds + \left(t(N+1) - T - \epsilon^{\gamma}\right)^2 \\ &+ \sum_{k=1}^{N+1} \max\{0, -\theta_k - \epsilon^{\gamma}\}^2 + (\tilde{\mathbf{x}}(N+1) - \mathbf{x}^f - \epsilon^{\gamma})^2.\end{aligned}$$

Here, α, β and γ are positive real numbers, and κ is a penalty parameter. Next, we define

$$\begin{aligned}S_\epsilon = \Big\{&(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon) \in \mathbb{R}^{N+1} \times \mathbb{R}^{m\times(N+1)} \times [0, \infty) : \\ &t(N+1) - T = \epsilon^{\gamma} \\ &\tilde{\mathbf{x}}(N+1) - \mathbf{x}^f = \epsilon^{\gamma} \\ &-\theta_k \le \epsilon^{\gamma}, \quad k = 1, \cdots, N+1, \\ &\sum_{j=1}^{m}\xi_{jk} - 1 = \epsilon^{\gamma}, \quad k = 1, \cdots, N+1, \\ &\xi_{jk}(1-\xi_{jk}) \le \epsilon^{\gamma}, \quad j = 1, \cdots, m, \quad k = 1, \cdots, N+1, \\ &\xi_{jk} - 1 \le \epsilon^{\gamma}, \quad j = 1, \cdots, m, \quad k = 1, \cdots, N+1, \\ &-\xi_{jk} \le \epsilon^{\gamma}, \quad j = 1, \cdots, m, \quad k = 1, \cdots, N+1, \\ &\sum_{k=1}^{N+1}\sum_{j=1}^{m}\xi_{jk} g_i(\tilde{\mathbf{x}}(s), \mathbf{u}_j)\chi_{[k-1,k)}(s) \le \epsilon^{\gamma}, \quad i = 1, 2, \cdots, p, \quad s \in [0, N+1].\Big\}.\end{aligned} \tag{5.17}$$

Now, consider the following problem: Given the dynamical system (5.13)-(5.14), find a triple $(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon) \in \mathbb{R}^{N+1} \times \mathbb{R}^{m\times(N+1)} \times [0, \infty)$ such that the penalty function $F_\kappa(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon)$ is minimized. This problem is referred to as Problem $\tilde{\mathbf{P}}_\kappa$.

In the next section, we will see that, under some mild assumptions, when the penalty parameter κ is sufficiently large, the satisfaction of the constraints (5.10), (5.11), (5.12) and (5.16) will be achieved, i.e. $\Delta(\boldsymbol{\theta}, \boldsymbol{\xi}, \epsilon) = 0$ for $\epsilon = 0$. Furthermore, an optimal solution of Problem $\tilde{\mathbf{P}}_\kappa$ is an optimal solution of Problem $\tilde{\mathbf{P}}$.

5.3.3 Convergence results

To obtain our main result, we need the following definition.

Definition 5.1. *Suppose that the following implication holds:*

$$\sum_{\iota=1}^{\overline{M}}\int_0^{N+1}\varphi_\iota(s)\frac{\partial G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)}{\partial\boldsymbol{\xi}}ds+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}\varphi_\eta(s)\frac{\partial H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)}{\partial\boldsymbol{\xi}}ds=0$$

$$\Longrightarrow\ \varphi_\iota(s)=0\ \textit{and}\ \varphi_\eta(s)=0$$

for all $s\in[0,N+1]$. *Then, we say that the* constraint qualification *is satisfied for the constraints* G_ι *and* H_η *at* $(\boldsymbol{\theta},\boldsymbol{\xi})=(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)$, *where* G_ι, $\iota=1,\cdots,\overline{M}=p+(3m+1)(N+1)$, *and* H_η, $\eta=1,\cdots,\overline{N}=N+3$, *are, respectively, the inequality constraints and the equality constraints of Problem* $\tilde{\mathbf{P}}$.

Let $\{\kappa_l\}_{l=1}^{\infty}$ be an increasing sequence of penalty parameters such that $\kappa_l\to\infty$. Furthermore, let $(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})$ denote a local optimal solution of Problem $\tilde{\mathbf{P}}_{\kappa_l}$. We assume that the following hypotheses are satisfied.

(H_1) *The constraint qualification defined in Definition* 5.1 *is satisfied at* $(\boldsymbol{\theta},\boldsymbol{\xi})=(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)$, *where* $(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)$ *is a local optimal solution of Problem* $\tilde{\mathbf{P}}$.

(H_2) *There exists real numbers* $\delta^1>0$ *and* $\delta^2>0$ *such that*

$$\lim_{l\to\infty}\frac{\max\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\}}{(\epsilon^{(l),\star})^{\delta^1}}=0,\ \iota=1,\ldots,\overline{M},$$

and

$$\lim_{l\to\infty}\frac{H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{(\epsilon^{(l),\star})^{\delta^2}}=0,\ \eta=1,\ldots,\overline{N}.$$

Theorem 5.2. *Suppose that* $(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star,\epsilon^\star)$ *as* $l\to+\infty$, *and that the hypotheses* (H_1)-(H_2) *are satisfied. Then,* $\epsilon^\star=0$ *and* $(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0$, *where* S_0 *is defined by* (5.17) *with* $\epsilon=0$.

Proof. From Lemma 2.1 and Theorem 2.1, we can follow similar arguments to show that $(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\notin S_{\epsilon^{(l),\star}}$. Thus, we have

$$\begin{aligned}
&\frac{\partial F_\kappa(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})}{\partial\boldsymbol{\xi}}\\
&=\int_0^{N+1}\frac{\partial\tilde{\mathcal{H}}\big(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)}{\partial\boldsymbol{\xi}}ds\\
&=\int_0^{N+1}\frac{\partial\tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}ds\\
&\quad+2(\epsilon^{(l),\star})^{-\alpha}\int_0^{N+1}\sum_{\iota=1}^{\overline{M}}\max\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&\quad-(\epsilon^{(l),\star})^{\gamma}\}\frac{\partial G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}+\sum_{\eta=1}^{\overline{N}}\big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})
\end{aligned}\tag{5.18}$$

$$
\begin{aligned}
&- (\epsilon^{(l),\star})^{\gamma}\big)\frac{\partial H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\\
&+\int_{0}^{N+1}\big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top}\frac{\partial \tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\\
=&0,
\end{aligned}
$$

where $\tilde{\mathcal{H}}\big(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)$ is the Hamiltonian function for the exact penalty function ($\epsilon>0$) given by

$$
\begin{aligned}
&\tilde{\mathcal{H}}\big(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)\\
=\;&\tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta},\boldsymbol{\xi})+(\epsilon^{(l),\star})^{-\alpha}\Delta(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\\
&+\Big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\Big)^{\top}\tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}),
\end{aligned}
\tag{5.19}
$$

$\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})$ is the costate vector determined by the following system of co-state differential equations:

$$
\Big(\frac{d\tilde{\boldsymbol{\lambda}}(s)}{ds}\Big)^{\top}=-\frac{\partial\tilde{\mathcal{H}}}{\partial\tilde{\mathbf{x}}},
$$

with the boundary condition

$$
\big(\tilde{\boldsymbol{\lambda}}(N+1)\big)^{\top}=0,
$$

where $d\tilde{\mathbf{x}}(s)/ds{=}\tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})$, and

$$
\begin{aligned}
&\frac{\partial F_{\kappa}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})}{\partial\epsilon}\\
=\;&(\epsilon^{(l),\star})^{-\alpha-1}\\
&\Big(-\alpha\int_{0}^{N+1}\sum_{\iota=1}^{\overline{M}}[\max\{0,G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma}\}]^2\\
&+\sum_{\eta=1}^{\overline{N}}[H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma}]^2ds\\
&+2\gamma\int_{0}^{N+1}\sum_{\iota=1}^{\overline{M}}\max\{0,G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma}\}(-(\epsilon^{(l),\star})^{\gamma})\\
&+\sum_{\eta=1}^{\overline{N}}(H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma})(-(\epsilon^{(l),\star})^{\gamma})ds\Big)+\kappa_l\beta(\epsilon^{(l),\star})^{\beta-1}\\
=\;&0,
\end{aligned}
\tag{5.20}
$$

Suppose that $\epsilon^{(k),\star}\to\epsilon^{\star}\neq 0$. Then, by (5.20), it can be shown by invoking hypotheses (H_2) and Lebesgue dominated convergence theorem [103] that its first term tends to a finite value, while the last term tends to infinity as $\kappa_l\to+\infty$, when $l\to+\infty$. This is impossible for the validity of (5.20). Thus, $\epsilon^{\star}=0$.

From (5.18), we have

$$\begin{aligned}
&\int_0^{N+1} \frac{\partial \tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \\
&+2(\epsilon^{(l),\star})^{-\alpha} \int_0^{N+1} \sum_{\iota=1}^{\overline{M}} \max\{0, G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}) \\
&-(\epsilon^{(l),\star})^{\gamma}\} \frac{\partial G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} + \sum_{\eta=1}^{\overline{N}} \big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}) \\
&-(\epsilon^{(l),\star})^{\gamma}\big) \frac{\partial H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \\
&+\int_0^{N+1} \big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top} \frac{\partial \tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \\
=\ & 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\lim_{l\to\infty} \Bigg\{ \int_0^{N+1} \frac{\partial \tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \\
&+ 2(\epsilon^{(l),\star})^{-\alpha} \int_0^{N+1} \sum_{\iota=1}^{\overline{M}} \max\{0, G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}) \\
&- (\epsilon^{(l),\star})^{\gamma}\} \frac{\partial G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} + \sum_{\eta=1}^{\overline{N}} \big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}) \\
&- (\epsilon^{(l),\star})^{\gamma}\big) \frac{\partial H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \\
&+ \int_0^{N+1} \big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top} \frac{\partial \tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \Bigg\} \\
&= 0.
\end{aligned}$$

Again, by invoking Lebesgue dominated convergence theorem, it follows that the first and third terms converge to some finite values. On the other hand, the second term tends to infinite, which is impossible, and hence

$$\begin{aligned}
&\int_0^{N+1} \lim_{l\to\infty} \Big\{ \sum_{\iota=1}^{\overline{M}} \max\{0, G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}) \\
&- (\epsilon^{(l),\star})^{\gamma}\} \frac{\partial G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} + \sum_{\eta=1}^{\overline{N}} \big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}) \\
&- (\epsilon^{(l),\star})^{\gamma}\big) \frac{\partial H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}} ds \Big\} \\
=\ & \sum_{\iota=1}^{\overline{M}} \int_0^{N+1} \max\{0, G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})
\end{aligned}$$

$$
\begin{aligned}
&\}\frac{\partial G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)}{\partial\boldsymbol{\xi}}ds+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\\
&\frac{\partial H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)}{\partial\boldsymbol{\xi}}ds\\
&=0.
\end{aligned}
$$

Since the constraint qualification is satisfied for the constraints G_ι and H_η at $(\boldsymbol{\theta},\boldsymbol{\xi})=(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)$, it follows that, for each $\iota=1,\cdots,p+(3m+1)(N+1)$ and $\eta=1,\cdots,N+3$,

$$
\max\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\}=0,\qquad H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)=0,
$$

for each $s\in[0,N+1]$. This, in turn, implies that, for each $\iota=1,\cdots,p+(3m+1)(N+1)$ and $\eta=1,\cdots,N+3$,

$$
G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\le 0,\qquad H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)=0,
$$

for each $s\in[0,N+1]$. The proof is completed. □

Theorem 5.3. *Suppose that* $\gamma>\alpha,\ \delta=\min(\delta^1,\delta^2)>\alpha,\ 2\delta>\alpha+1,\ 2\gamma>\alpha+1$. *Then*

$$
F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\xrightarrow[(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0]{\epsilon^{(l),\star}\to\epsilon^\star=0}F_{\kappa_l}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star,0)=\tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star),
$$

$$
\nabla F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\xrightarrow[(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0]{\epsilon^{(l),\star}\to\epsilon^\star=0}\nabla F_{\kappa_l}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star,0)=(\nabla\tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star),0).
$$

Proof. From the conditions of the theorem and the definition of $F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})$, it follows that, for $\epsilon^{(l),\star}\neq 0$,

$$
\begin{aligned}
&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}}F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\\
=&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}}\Big\{\tilde{J}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})+(\epsilon^{(l),\star})^{-\alpha}\Delta(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})+\kappa(\epsilon^{(l),\star})^\beta\Big\}\\
=&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}}\Big\{\tilde{J}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})+(\epsilon^{(l),\star})^{-\alpha}\Big[\sum_{\iota=1}^{\overline{M}}\int_0^{N+1}\big(\max\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&-(\epsilon^{(l),\star})^\gamma\}\big)^2ds+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}\big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^\gamma\big)^2ds\Big]+\kappa(\epsilon^{(l),\star})^\beta\Big\}.
\end{aligned}
\tag{5.21}
$$

By arguments similar to those given for the proofs of Lemma 6.4.3 and Lemma 6.4.4 in [114], we can show that

$$
\lim_{(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}\tilde{J}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})=\tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star).
\tag{5.22}
$$

Substituting (5.22) into (5.21) gives

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} F_{\sigma_k}(\mathbf{z}^{(k),\star},\epsilon^{(k),\star})\\
=\;& \tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)+\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \Big\{(\epsilon^{(l),\star})^{-\alpha}\Big[\sum_{\iota=1}^{\overline{M}}\int_0^{N+1}\big(\max\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&\qquad -(\epsilon^{(l),\star})^\gamma\}\big)^2ds+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}\big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^\gamma\big)^2ds\Big]\Big\}\\
=\;& \tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)+\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \Big\{\sum_{\iota=1}^{\overline{M}}\int_0^{N+1}\big(\max\{0,(\epsilon^{(l),\star})^{-\frac{\alpha}{2}}G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&\qquad -(\epsilon^{(l),\star})^{\gamma-\frac{\alpha}{2}}\}\big)^2ds+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}\big((\epsilon^{(l),\star})^{-\frac{\alpha}{2}}H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma-\frac{\alpha}{2}}\big)^2ds\Big\}.
\end{aligned}\tag{5.23}$$

Since $\gamma>\alpha,\delta>\alpha$, applying Lebesgue dominated convergence theorem to (5.23) gives

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \Big\{\sum_{\iota=1}^{\overline{M}}\int_0^{N+1}\big(\max\{0,(\epsilon^{(l),\star})^{-\frac{\alpha}{2}}G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&\qquad -(\epsilon^{(l),\star})^{\gamma-\frac{\alpha}{2}}\}\big)^2ds+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}\big((\epsilon^{(l),\star})^{-\frac{\alpha}{2}}H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma-\frac{\alpha}{2}}\big)^2ds\Big\}\\
=\;& \sum_{\iota=1}^{\overline{M}}\int_0^{N+1}\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}}\big(\max\{0,(\epsilon^{(l),\star})^{-\frac{\alpha}{2}}G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma-\frac{\alpha}{2}}\}\big)^2ds\\
&+\sum_{\eta=1}^{\overline{N}}\int_0^{N+1}\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}}\big((\epsilon^{(l),\star})^{-\frac{\alpha}{2}}H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^{\gamma-\frac{\alpha}{2}}\big)^2ds\\
=\;& 0.
\end{aligned}\tag{5.24}$$

Combining (5.23) and (5.24) gives

$$\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})=F_{\kappa_l}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star,0)=\tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star).$$

For the second part of the theorem, we need the gradient formulas of $\tilde{J}(\boldsymbol{\theta},\boldsymbol{\xi})$. They are:

$$\frac{\partial\tilde{J}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}=\int_0^{N+1}\frac{\partial\bar{\mathcal{H}}\big(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\big)}{\partial\boldsymbol{\xi}}ds,\tag{5.25}$$

$$\frac{\partial\tilde{J}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\theta}}=\int_0^{N+1}\frac{\partial\bar{\mathcal{H}}\big(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\big)}{\partial\boldsymbol{\theta}}\mathrm{d}s,\tag{5.26}$$

where $\bar{\mathcal{H}}$ is the Hamiltonian function defined by

$$\begin{aligned}&\bar{\mathcal{H}}\big(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\big)\\ =\;&\tilde{L}_0(s,\tilde{\mathbf{x}}(x),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})+\Big(\bar{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\Big)^{\top}\tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star}),\end{aligned}\tag{5.27}$$

$\bar{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})$ is the costate vector determined by the following system of co-state differential equations:

$$\Big(\frac{d\bar{\boldsymbol{\lambda}}(s)}{ds}\Big)^{\top}=-\frac{\partial\bar{\mathcal{H}}}{\partial\tilde{\mathbf{x}}},$$

with the boundary condition

$$\big(\bar{\boldsymbol{\lambda}}(N+1)\big)^{\top}=0.$$

By an augment similar to that given for the proof of Theorem 5.2 in [63], we can show that, for each $s\in[0,N+1]$,

$$\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}|\bar{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})|=0.\tag{5.28}$$

By (5.18) and (5.19), we have

$$\begin{aligned}&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\nabla_{\boldsymbol{\xi}}F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\\ =&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\Big\{\int_0^{N+1}\frac{\partial\tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}ds\\ &+2(\epsilon^{(l),\star})^{-\alpha}\int_0^{N+1}\sum_{\iota=1}^{\overline{M}}\max\{0,G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\ &-(\epsilon^{(l),\star})^{\gamma}\}\frac{\partial G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}+\sum_{\eta=1}^{\overline{N}}\big(H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\ &-(\epsilon^{(l),\star})^{\gamma}\big)\frac{\partial H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}ds\\ &+\int_0^{N+1}\big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top}\frac{\partial\tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}ds\Big\}\\ =&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\Big\{\int_0^{N+1}\frac{\partial\tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}ds\\ &+\int_0^{N+1}\big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top}\frac{\partial\tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}ds\Big\}\\ &+\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\Big\{2(\epsilon^{(l),\star})^{-\alpha}\int_0^{N+1}\sum_{\iota=1}^{\overline{M}}\max\{0,G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\ &-(\epsilon^{(l),\star})^{\gamma}\}\frac{\partial G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial\boldsymbol{\xi}}+\sum_{\eta=1}^{\overline{N}}\big(H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\end{aligned}\tag{5.29}$$

$$- (\epsilon^{(l),\star})^{\gamma}\big)\frac{\partial H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\Big\}.$$

Then, by Lebesgue dominated convergence theorem, it follows from (5.28) that

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}} \Big\{\int_0^{N+1}\frac{\partial \tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\\
&+\int_0^{N+1}\big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top}\frac{\partial \tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\Big\}\\
=&\int_0^{N+1}\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\frac{\partial \tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\\
&+\int_0^{N+1}\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\big(\tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\big)^{\top}\frac{\partial \tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\\
=&\int_0^{N+1}\frac{\partial \tilde{L}_0(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})}{\partial \boldsymbol{\xi}}ds+\int_0^{N+1}\big(\bar{\boldsymbol{\lambda}}(\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\big)^{\top}\frac{\partial \tilde{\mathbf{f}}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})}{\partial \boldsymbol{\xi}}ds\\
=&\nabla_{\boldsymbol{\xi}}\tilde{J}(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star}).
\end{aligned}\tag{5.30}$$

Similarly, since $\delta>\alpha$, $\gamma>\alpha$, it follows from the Lebesgue dominated convergence theorem that

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}} 2(\epsilon^{(l),\star})^{-\alpha}\int_0^{N+1}\sum_{\iota=1}^{\overline{M}}\max\{0,G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&-(\epsilon^{(l),\star})^{\gamma}\}\frac{\partial G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}+\sum_{\eta=1}^{\overline{N}}\big(H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&-(\epsilon^{(l),\star})^{\gamma}\big)\frac{\partial H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}ds\Big\}\\
=&2\int_0^{N+1}\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\Big\{\sum_{\iota=1}^{\overline{M}}\max\{0,G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&-(\epsilon^{(l),\star})^{\gamma}\}(\epsilon^{(l),\star})^{-\alpha}\frac{\partial G_{\iota}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}+\sum_{\eta=1}^{\overline{N}}\big(H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\\
&-(\epsilon^{(l),\star})^{\gamma}\big)(\epsilon^{(l),\star})^{-\alpha}\frac{\partial H_{\eta}(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})}{\partial \boldsymbol{\xi}}\Big\}ds.\\
=&0.
\end{aligned}\tag{5.31}$$

Substituting (5.30) and (5.31) into (5.29) gives

$$\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^{\star}=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star})\in S_0}}\nabla_{\boldsymbol{\xi}}F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})=\nabla_{\boldsymbol{\xi}}\tilde{J}(\boldsymbol{\theta}^{\star},\boldsymbol{\xi}^{\star}).\tag{5.32}$$

Similarly, we can show that

$$\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \nabla_{\boldsymbol{\theta}} F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star}) = \nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star). \tag{5.33}$$

On the other hand, we note that

$$\begin{aligned}
&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \nabla_{\boldsymbol{\theta}} F_{\kappa_l}(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\\
=&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \Big\{(\epsilon^{(l),\star})^{-\alpha-1}\Big\{-\alpha\Delta(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\\
&+2\gamma\Big(\sum_{\iota}\max\big\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^\gamma\big\}(-(\epsilon^{(l),\star})^\gamma)\\
&+\sum_{\eta=1}^{\overline{N}}\big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^\gamma\big)(-(\epsilon^{(l),\star})^\gamma)\Big)\Big\}+\kappa_l\beta(\epsilon^{(l),\star})^{\beta-1}\Big\}\\
=&\lim_{\substack{\epsilon^{(l),\star}\to\epsilon^\star=0\\(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)\in S_0}} \Big\{\frac{-\alpha\Delta(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})}{(\epsilon^{(l),\star})^{\alpha+1}}\\
&+2\gamma\Big(\sum_{\iota}\max\big\{0,G_\iota(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^\gamma\big\}(-(\epsilon^{(l),\star})^{\gamma-\alpha-1})\\
&+\sum_{\eta=1}^{\overline{N}}\big(H_\eta(s,\tilde{\mathbf{x}}(s),\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})-(\epsilon^{(l),\star})^\gamma\big)(-(\epsilon^{(l),\star})^{\gamma-\alpha-1})\Big)+\kappa_l\beta(\epsilon^{(l),\star})^{\beta-1}\Big\}\\
=&0.
\end{aligned}$$

Thus, the proof is completed. □

Theorem 5.4. *Suppose that* $(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star},\epsilon^{(l),\star})\to(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star,\epsilon^\star)$ *as* $l\to+\infty$*, and that the parameters* α *and* γ *satisfy the same conditions as in Theorem 5.3. Then, there exists a* $l_0>0$ *such that* $\epsilon^{(l),\star}=0$ *and* $(\boldsymbol{\theta}^{(l),\star},\boldsymbol{\xi}^{(l),\star})=(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)$*, for all* $l\geq l_0$*. Furthermore* $(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^\star)$ *is a local optimal solution of Problem* $\tilde{\mathbf{P}}$*.*

Proof. The proof is similar to that given for Theorem 3.2 and hence it is omitted. □

From the results above, we can conclude that under some mild assumptions, for a sufficiently large κ, a local optimal solution of Problem $\tilde{\mathbf{P}}_\kappa$ is a local optimal solution of Problem $\tilde{\mathbf{P}}$. This solution can then be used to construct a corresponding local solution of Problem $\mathbf{P}$.

Problem $\tilde{\mathbf{P}}_\kappa$ is a standard optimal control problem with fixed switching points and can be readily solved by various existing optimal control techniques. Here, the optimal control software package MISER 3.3 [49] is used. In the next section, two practical problems concerning optimal driving strategies for trains are solved by the method proposed.

ζ_1	ζ_2	ζ_3	ζ_4	ζ_5	ζ_6	ζ_7
1.5	1	1.4	0.1	-0.015	-0.00003	-0.000006

Table 5.1: Values of ζ_i, $i = 1, \cdots, 7$.

5.4 Numerical results

5.4.1 Optimal train control on a level track

The following model for the motion of a train is given in references [46, 62]:

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \varphi(x_2)u_1 + \zeta_2 u_2 + \rho(x_2), \end{aligned}$$

where x_1 is the train's distance along the track, x_2 is the train's speed, u_1 is the fuel setting and u_2 models the deceleration applied to the train by the brakes. The function φ, which models the tractive effort, is defined by

$$\varphi(x_2) = \begin{cases} \zeta_1/x_2, & \text{if } x_2 \geq \zeta_3 + \zeta_4, \\ \zeta_1/\zeta_3 + \eta_1(x_2 - \zeta_3 + \zeta_4)^2 & \\ \qquad + \eta_2(x_2 - \zeta_3 + \zeta_4)^3, & \text{if } \zeta_3 - \zeta_4 \leq x_2 < \zeta_3 + \zeta_4, \\ \zeta_1/\zeta_3, & \text{if } x_2 < \zeta_3 - \zeta_4, \end{cases}$$

where $\zeta_1, \zeta_2, \zeta_3$ and ζ_4 are constants, and

$$\eta_1 = \zeta_1\Big[\Big(\frac{1}{\zeta_3 + \zeta_4} - \frac{1}{\zeta_3}\Big)\frac{3}{4\zeta_4^2} + \frac{1}{2\zeta_4(\zeta_3 + \zeta_4)^2}\Big],$$

and

$$\eta_2 = \zeta_1\Big[-\Big(\frac{1}{\zeta_3 + \zeta_4} - \frac{1}{\zeta_3}\Big)\frac{3}{4\zeta_4^3} - \frac{1}{4\zeta_4^2(\zeta_3 + \zeta_4)^2}\Big].$$

The function ρ, which models the resistive deceleration due to friction, is given by

$$\rho(x_2) = \zeta_5 + \zeta_6 x_2 + \zeta_7 x_2^2.$$

The constants $\zeta_i, i = 1, \cdots, 7$, are defined in Table 5.1. The initial and terminal states are

$$\mathbf{x}(0) = [0, 0]^\top, \quad \mathbf{x}(1500) = [18000, 0]^\top.$$

This means that the train starts from the origin at rest and comes to rest again 18,000 meters away at $t = 1500$. Since the train is not allowed to go backwards, a non-negativity constraint is imposed on the speed,

$$x_2(t) \geq 0, \quad t \in [0, 1500].$$

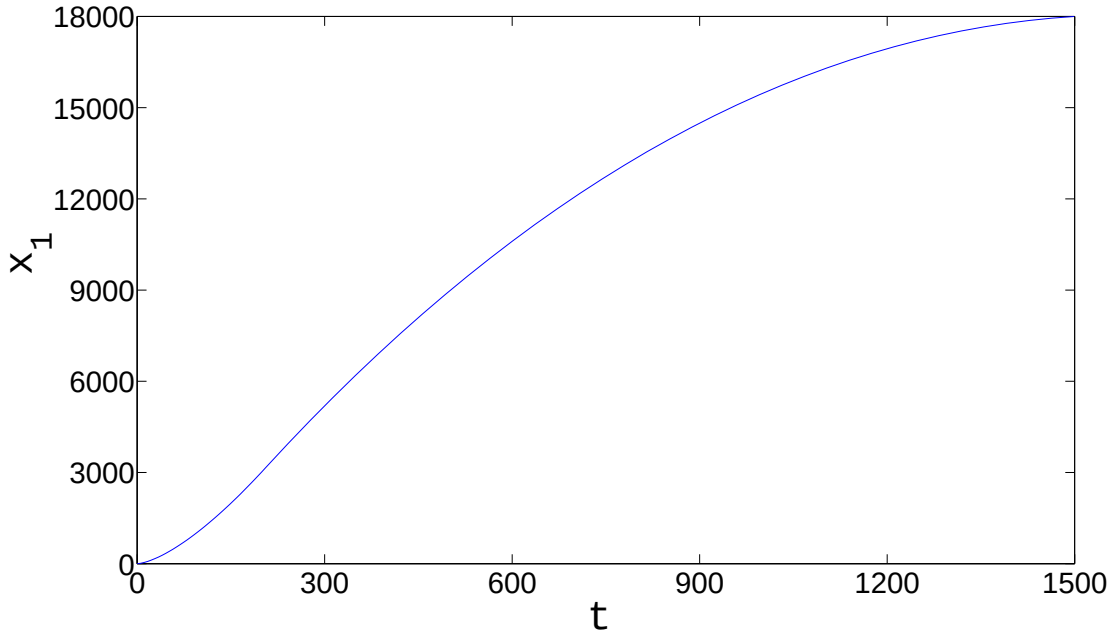


Figure 5.1: The trajectory $x_1(t)$ against t

The train driver can choose from three operation modes for the train: accelerate (powered by the engine), coast (no power), and brake (decelerate by the brakes). These three modes correspond to the following values for $\mathbf{u} = [u_1, u_2]^\top$, i.e.,

$$\mathbf{U} = \left\{[1, 0]^\top, [0, 0]^\top, [0, -1]^\top\right\}.$$

The objective is to minimize the fuel consumption, i.e.,

$$\min: \ J(\mathbf{u}) = \int_0^{1500} u_1(t)dt.$$

Here, we assume that the maximum number of switches is $N = 2$. We apply our method in conjunction with MISER 3.3 to solve the problem.

Figure 5.1 and Figure 5.2 show the optimal trajectory of x_1 and x_2, respectively. From the figures, we see that the train accelerates for the first quarter of the journey, then coasts almost until the end. Figure 5.3 shows that the brakes are applied briefly at the end before the train stops.

Figure 5.4 and Figure 5.5 show the optimal controls u_1 and u_2, respectively. We see that the control u_2 stays zero for almost the entire time horizon, and assumes the value -1 less than two seconds before the end.

The minimum fuel consumption is 205.06. This is slightly higher than the result of 202.67 reported in [62], which was obtained using the time scaling transform (also called the control parametrization enhancing transform) directly with 6 switching points. It is worth noting that our method obtains the same result as in [62] when we increase the maximum number of switches

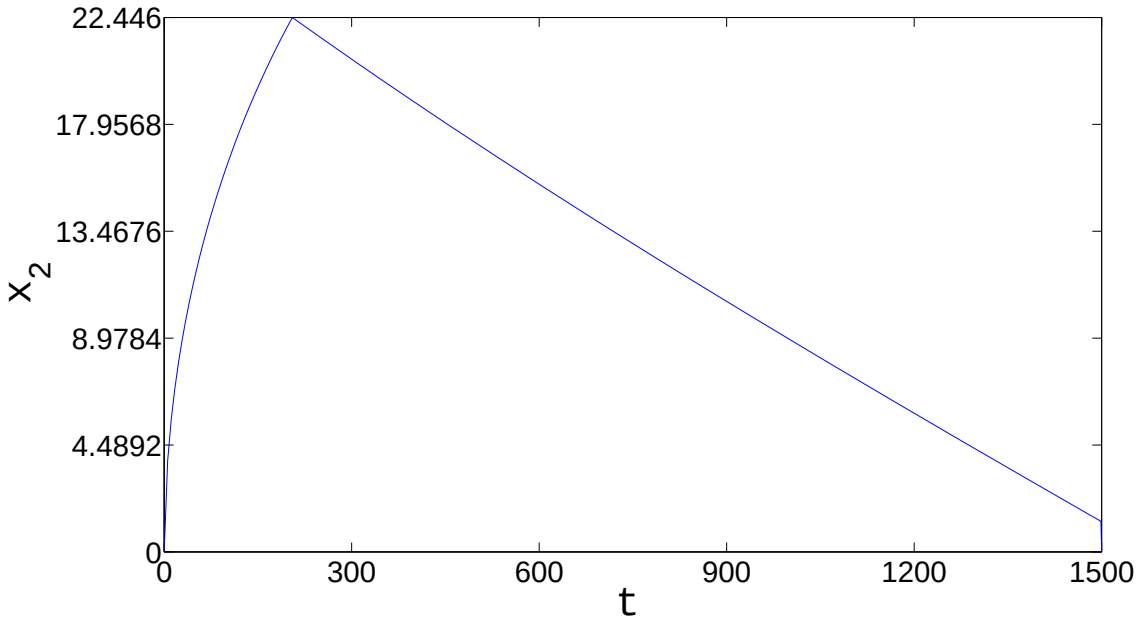


Figure 5.2: The speed $x_2(t)$ against t

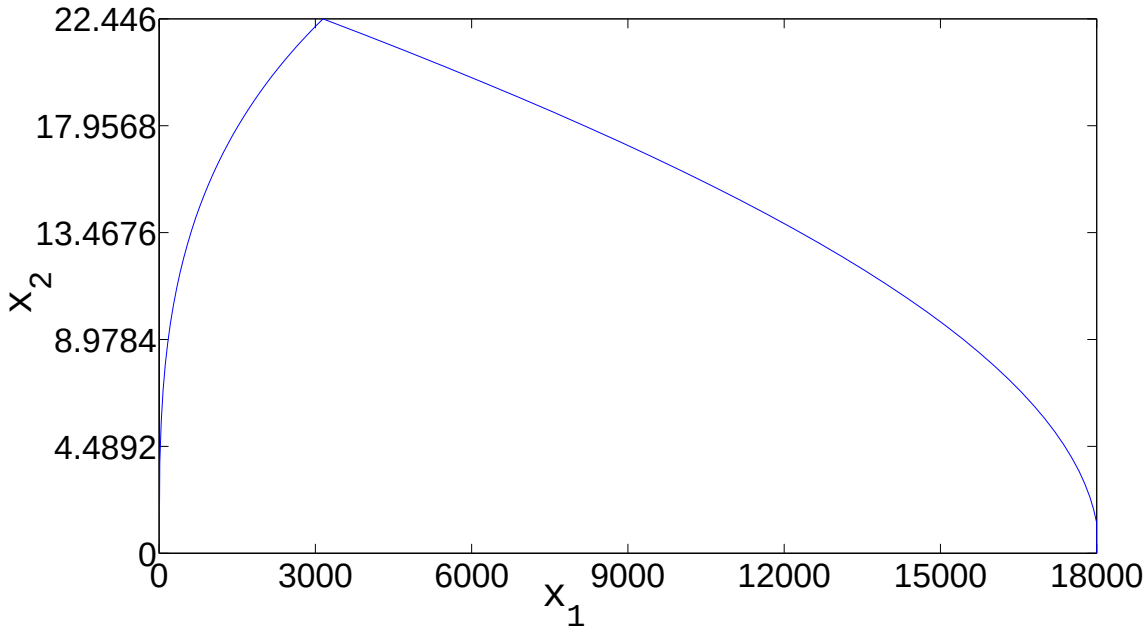


Figure 5.3: The state space plot of x_2 against x_1

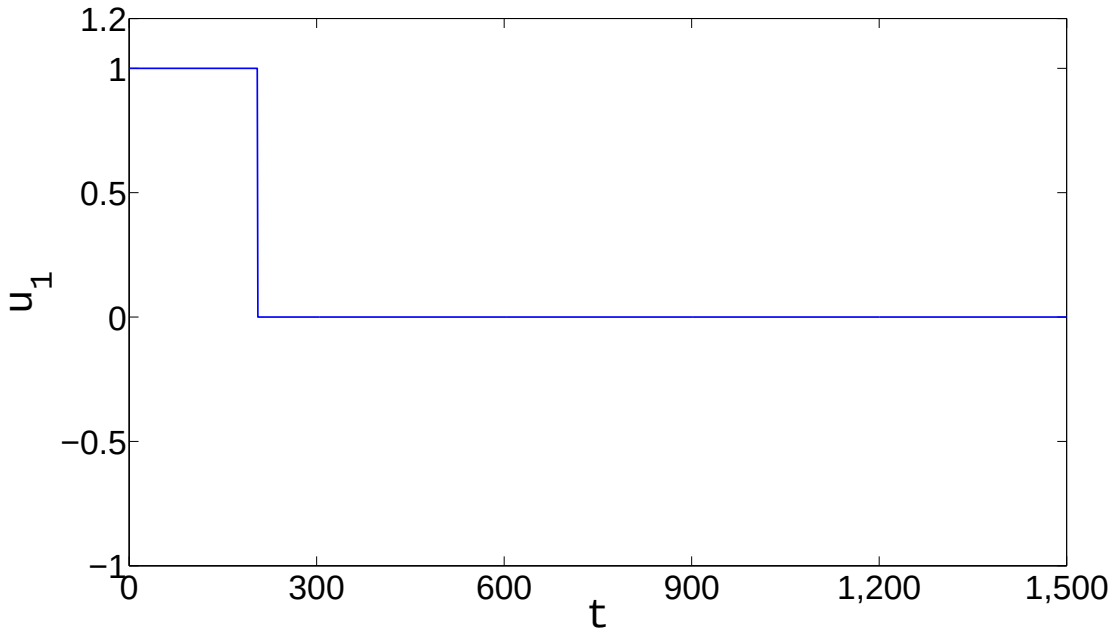


Figure 5.4: The optimal control u_1 against t

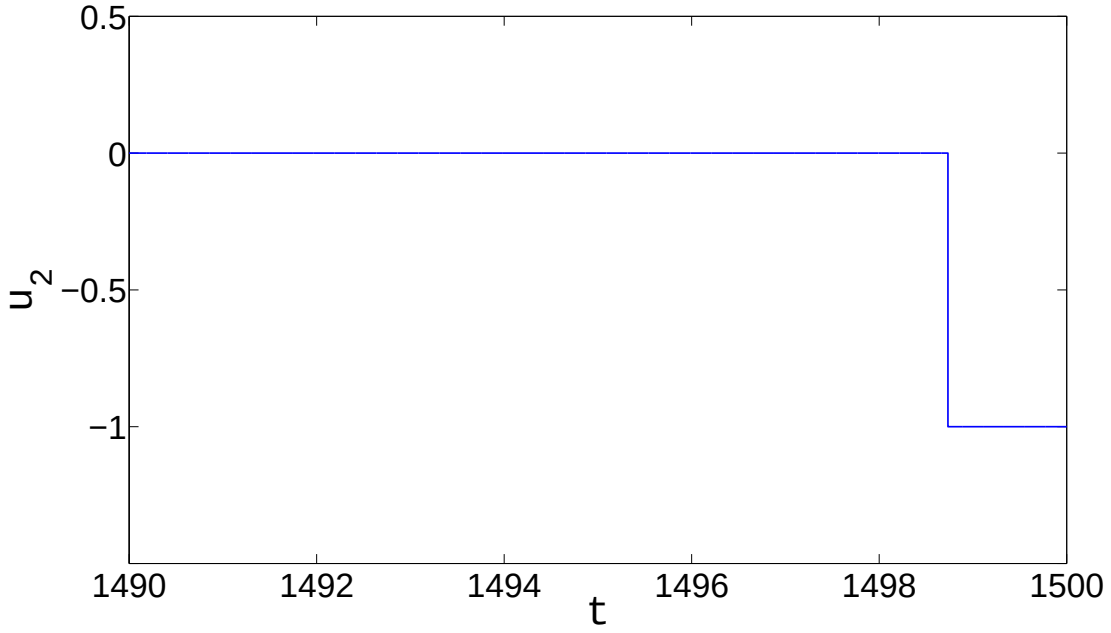


Figure 5.5: The optimal control u_2 near the terminal time

to $N = 6$. More importantly, unlike the direct application of the time scaling transform, our method ensures that the constraint on the maximum number of switches is always satisfied.

5.4.2 Optimal train control on an uneven track

We now consider a more complicated train control problem [46, 60]. The dynamics for this problem are

$$\begin{aligned}\dot{x}_1 &= x_2,\\ \dot{x}_2 &= \varphi(x_2)u_1 + \zeta_2 u_2 + \rho(x_2) + \vartheta(x_1),\end{aligned}$$

where $x_1, x_2, u_1, u_2, \varphi(\cdot)$ and $\rho(\cdot)$ are as defined in Section 5.4.1, and $\xi_i, i = 1, \cdots, 7$, are as defined in Table 5.1. The function $\vartheta(\cdot)$ is the gravitational acceleration due to the non-constant gradient of the track given by

$$\vartheta(x_1) = \begin{cases} 0, & \text{if } x_1 \le 20000 - \zeta_8,\\ -0.05\{\frac{(x_1-20000)^2}{\zeta_8^2} + \frac{(x_1-20000)}{\zeta_8} + 1\}, & \text{if } 20000 - \zeta_8 < x_1 \le 20000,\\ -0.05\{-\frac{(x_1-20000)^2}{\zeta_8^2} + \frac{(x_1-20000)}{\zeta_8} + 1\}, & \text{if } 20000 < x_1 \le 20000 + \zeta_8,\\ -0.1, & \text{if } 20000 + \zeta_8 < x_1 \le 25000 - \zeta_8,\\ -0.05\{-\frac{(x_1-25000)^2}{\zeta_8^2} - \frac{(x_1-25000)}{\zeta_8} + 1\}, & \text{if } 25000 - \zeta_8 < x_1 \le 25000,\\ -0.05\{\frac{(x_1-25000)^2}{\zeta_8^2} - \frac{(x_1-25000)}{\zeta_8} + 1\}, & \text{if } 25000 < x_1 \le 25000 + \zeta_8,\\ 0, & \text{if } x_1 > 25000 + \zeta_8,\end{cases}$$

where $\zeta_8 = 300$.

The initial and terminal states are

$$\mathbf{x}(0) = [0, 0]^\top, \quad \mathbf{x}(2800) = [50000, 0]^\top.$$

Again, we have a non-negativity constraint on x_2 to prevent the train from going backwards:

$$x_2(t) \ge 0, \quad t \in [0, 2800].$$

We also impose a speed limit on the train that decreases as the train moves further along the track:

$$0.0002x_1(t) + x_2(t) \le 28, \quad t \in [0, 2800].$$

The control $\mathbf{u} = [u_1, u_2]^\top$ is now restricted to the discrete set

$$\mathbf{U} = \left\{[1, 0]^\top, [0, 0]^\top, [0, -1]^\top, [2, 0]^\top\right\}.$$

The objective is

$$\min \; J(\mathbf{u}) = \int_0^{2800} u_1(t)dt$$

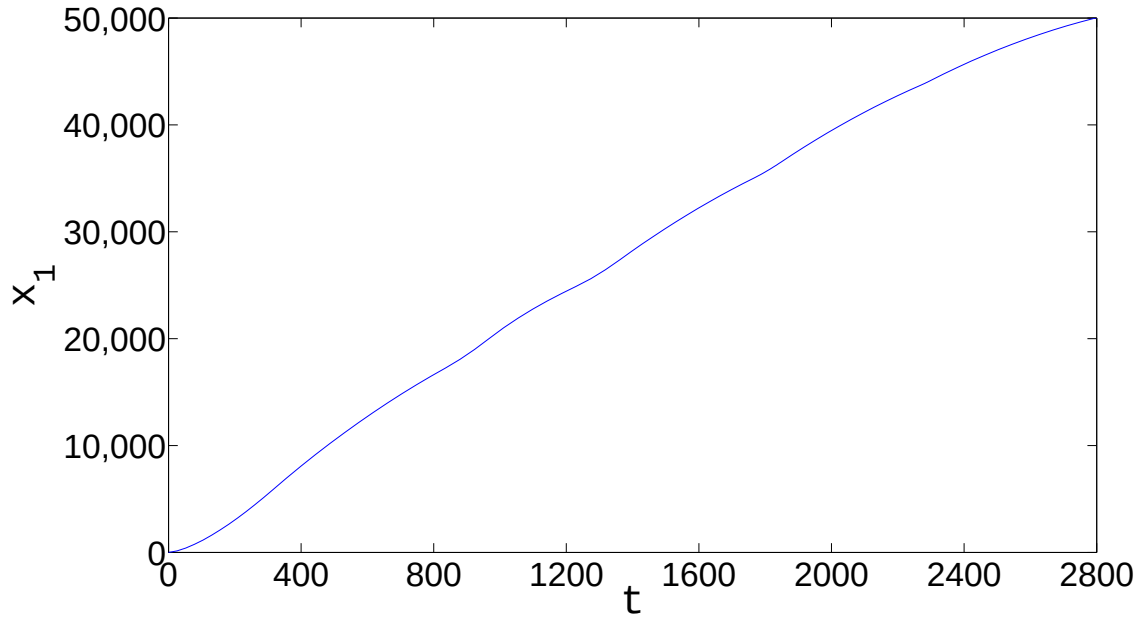


Figure 5.6: The trajectory $x_1(t)$ against t

Here, we assume that the maximum allowable number of switches is $N = 8$. Using our method, the problem is again solved by MISER 3.3. Figure 5.6 and Figure 5.7 show the optimal trajectory of x_1 and x_2, respectively. Figure 5.8 and Figure 5.9 show the optimal controls u_1 and u_2, respectively. Note that the optimal control does not assume the value $[2, 0]^\top$. From Figure 5.10, we can see that the continuous inequality constraint is satisfied throughout the entire period of the time horizon.

To solve this highly complex problem, we first used our method to determine the optimal switching sequence. After identifying the optimal switching sequence, we then applied the time scaling transform directly with the control sequence fixed to refine the switching times. The minimum fuel consumption is 937.42. This is better than the result obtained in [60], which uses the time scaling transform directly. There are 18 switching points, giving rise to a larger fuel consumption of 938.63.

5.5 Conclusion

In this chapter, a new computational method was proposed for solving optimal discrete-valued control problems. By introducing new controls and applying an equivalent transformation, the original problem becomes a standard optimal control problem subject to equality and inequality constraints. Then, an exact penalty method is employed to solve the transformed problem. Our numerical results for the train control problems in Section 5.4 show that this approach is superior to the direct application of the time scaling transform, which leads to many artificial switches. Our optimal solutions require less switchings and always satisfy the constraint on the maximum allowable number of switchings.

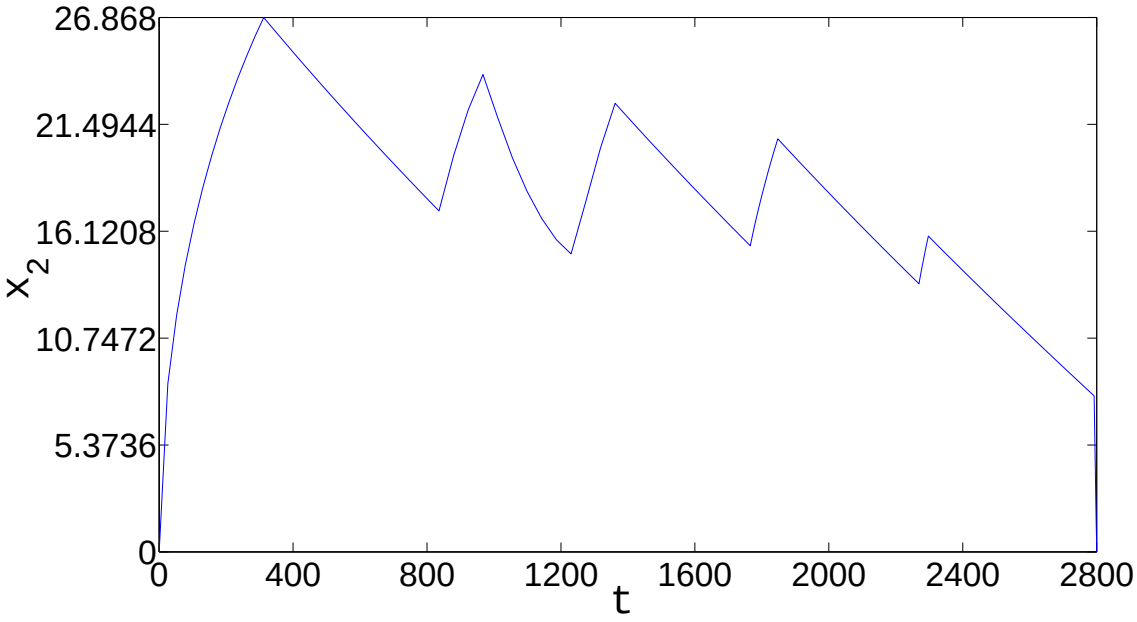


Figure 5.7: The speed $x_2(t)$ against t

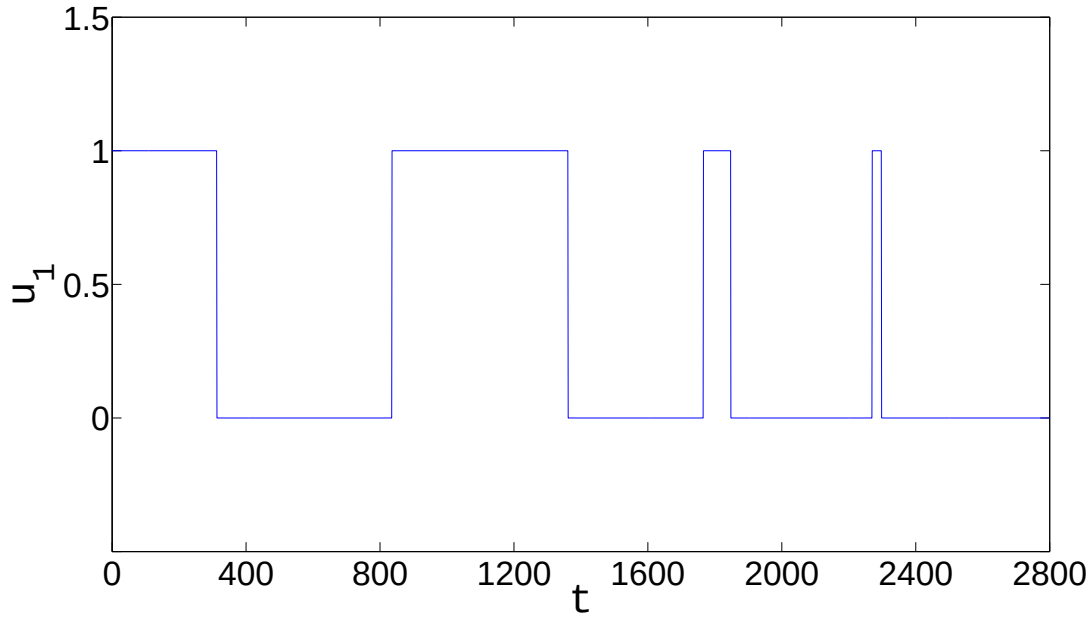


Figure 5.8: The optimal control u_1 against t

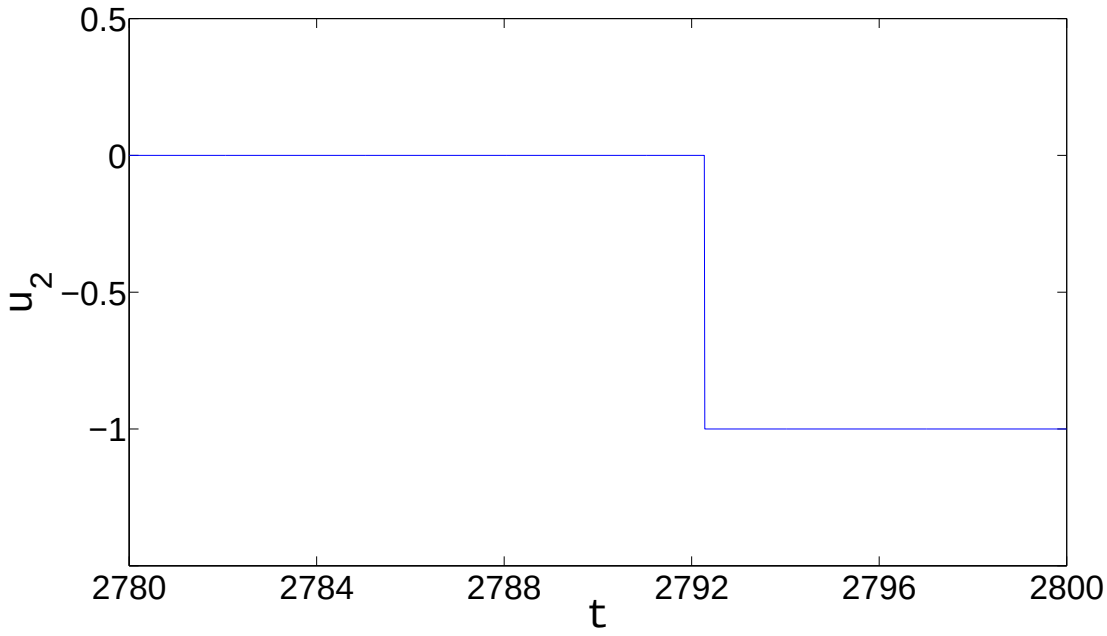


Figure 5.9: The optimal control u_2 near the terminal time

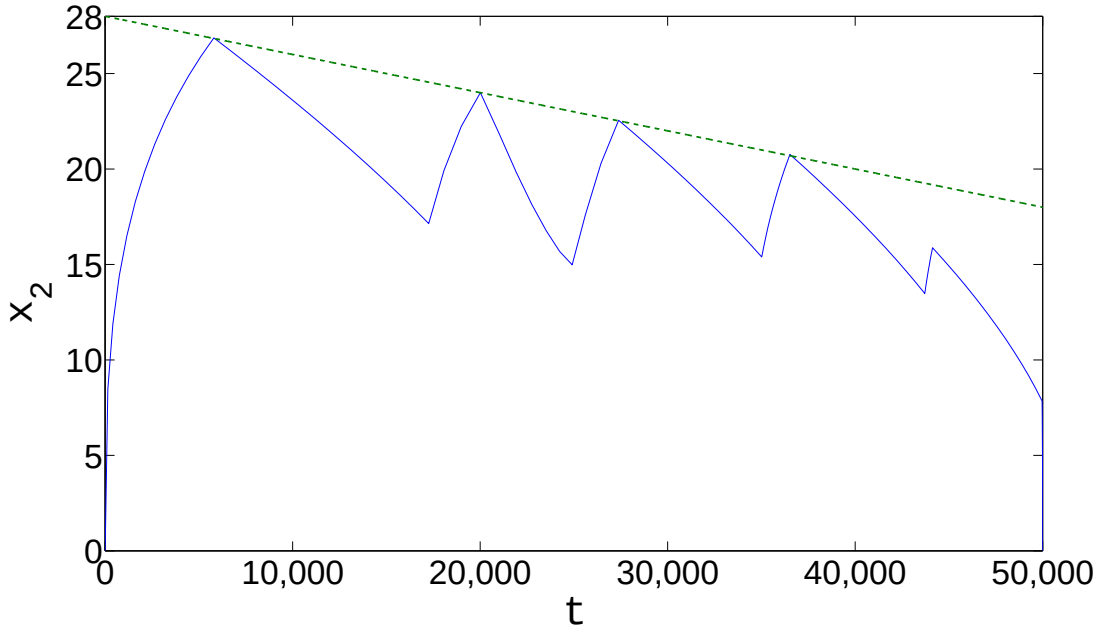


Figure 5.10: The plot of speed limit constraint

CHAPTER 6

Summary and suggestions for future research directions

6.1 Summary of the main contributions

In this thesis, we considered three optimization problems and a discrete-valued optimal control problem. We developed new algorithms and methods to solve these problems numerically. This involved a variety of novel techniques, including a new exact penalty function and the way of generating reduced search region for a particular application in signal processing. We summarize our main contributions below.

In Chapter 1, we provided a brief survey on optimization and optimal control.

In Chapter 2, we considered a class of continuous inequality constrained optimization problems. The major challenge for this type of problems is that they contain infinite many inequality constraints. Instead of using the well-known constrained transcription method, we developed a computational scheme based on a new exact penalty function for solving this class of problems. To handle the continuous inequality constraints, we introduced a new variable and append the constraint violation to the objective function, forming a new objective function subject to the nonnegativity constraint on the new variable. We have shown that under some mild assumptions, a local minimizer of the new optimization problem is a local minimizer of the original problem when the penalty parameter is sufficiently large. This property is not shared by the approaches reported in [116], [117], [50] or [131]. Clearly, this is a major advancement in the study of solution methods for semi-infinite optimization problems.

In Chapter 3, we considered a class of nonlinear mixed integer programming problems. Since discrete-valued variables are involved, traditional gradient-based optimization methods are not applicable. To overcome this difficulty, we first introduce new variables to transform the mixed discrete nonlinear programming problem into an equivalent conventional nonlinear optimization problem. Then, we applied a new exact penalty function method to obtain a sequence of unconstrained optimization problems. Each of these unconstrained optimization problems can be solved by gradient-based optimization methods such as quasi-Newton methods. We also showed that under some mild assumptions, a local minimizer of the unconstrained optimization problem is a local minimizer of the transformed nonlinear constrained optimization problem

which is equivalent to the original problem when the penalty parameter is sufficiently large. Several numerical experiments were carried out, the results show that the method proposed is effective.

In Chapter 4, we considered the design of allpass variable fractional delay filters with sums of signed powers-of-two coefficients and the least square criterion. This problem is a typical integer programming problem. However, this particular problem is not easy to solve due to the following two reasons:

i) Each element of the decision variable is to be chosen from a corresponding set which contains a tremendous number of options. These options are not uniformly distributed.

ii) Due to the specific structure of these coefficients together with the constraints on the total allowable number of signed-powers-of-two terms, the problem is extremely difficult to solve by conventional integer programming techniques.

To reduce the computational complexity, we investigated the problem and develop a two-step computational scheme to find reduced search region. The size of the obtained search region for each element is much smaller, and hence the computation complexity is greatly reduced. Furthermore, we have shown that under some mild assumptions, the new reduced search region still contains the global minimizer of the design problem. Then, we applied the techniques introduced in Chapter 3 to transform this problem into an equivalent conventional continuous optimization problem. Finally, an exact penalty function method is introduced to solve the new problem. Simulation was carried out to test the efficiency of the proposed method. Comparing our results with those obtained by the traditional quantization method, it is clearly seen that our results are much superior to those obtained by the quantization method.

In Chapter 5, we considered a class of discrete-valued optimal control problems, where there is an upper bound on the maximum number of control switches. The time-scaling transform introduced in [62], which is also called the control parametrization enhancing technique (CPET), is an effective method for solving optimal discrete-valued control problems. However, it introduces many more "artificial" switches, and hence the optimal control obtained is always having many more switches than the maximum number of allowable switches. Thus, the transformed optimal control problem obtained by using the time-scaling transformation is not equivalent to the original problem.

To obtain an equivalent transformation, we first introduce new control functions taking values from a compact set. Then, the original controls are replaced by the newly introduced controls to form a conventional optimal control problem. Furthermore, additional constraints are imposed such that the problem with new control functions is equivalent to the original discrete-valued optimal control problem. Finally, we applied the exact penalty function method to solve this problem. Numerical results obtained from solving two real practical train control problems show that our approach is effective.

6.2 Future research directions

In this thesis, our main work is in the development of computational algorithms for solving several types of optimization and optimal control problems based on a new exact penalty function method. It is observed that these algorithms are computationally very effective for solving all the problems under consideration. To make significant advancement, it requires further understanding of the properties of this penalty function. On this basis, new and more efficient computational algorithms could be derived for solving existing optimization and optimal control problems and new unconventional optimization and optimal control problems arising in the study of real world practical problems.

In Chapter 2, the optimization problem under consideration is an optimization problem subject to continuous inequality constraints. The exact penalty function is introduced to these continuous inequality constraints. In the construction of the penalty function, some approximate functions of the continuous inequality constraint functions are constructed. Then, the sum of their integrations is appended to the cost function, forming a penalized cost function with a new decision variable. It gives rise to a sequence of penalized optimization problems, each of which can be solved by gradient-based optimization techniques. It is known that many optimal filter design problems in signal processing can be formulated as optimization problems subject to continuous inequality constraints, and hence the method proposed in Chapter 2 is applicable. However, in many of these signal processing problems, the argument ω involved in the continuous inequality constraints is two, rather than one, dimensional as given below.

$$g_i(\omega, \mathbf{x}) \leq 0, \text{ for all } \omega \in \Omega \subset \mathbb{R}^2, \ i = 1, \ldots, m.$$

Furthermore, these continuous inequality constraints are very sensitive with respect to ω near the cut-off frequency. The integrations of the approximate functions constructed from the continuous inequality functions cannot be carried out analytically. Thus, it is inevitable to use the numerical integration scheme. Due to the specific structures of these signal processing problems (see, for example, the one considered in Chapter 4), it is found that the approach based on the numerical integration is not satisfactory. Thus, it is important to devise a systematic approach to deal with the continuous inequality constraints based on the idea proposed in [17]. Furthermore, can this idea be applied to deal with continuous inequality constraints with structures different from those of the signal processing problems? These questions are both mathematically challenging and practically significant.

A second future research direction is to develop new exact penalty functions for optimization problems and optimal control problems subject to inequality constraints, which are more effective and contain better mathematical properties. In particular, the constraint qualifications introduced in Definition 2.1 and Definition 5.1 are rather strong. Could the constraint qualification given in Definition 2.1 and Definition 5.1 be relaxed such as the one given below?

Let $\bar{\mathbf{x}}$ *be such that* $\frac{\partial \phi_j(\bar{\mathbf{x}},\omega)}{\partial \mathbf{x}}$, $j = 1, \ldots, m$. *are linearly independent for each* $\omega \in \Omega$. *Then it*

is said that the constraint qualification is satisfied for the continuous inequality constraints ϕ_j, $j = 1, \ldots, m$, *at* $\mathbf{x} = \bar{\mathbf{x}}$.

Similarly, could the constraint qualification given in Definition 3.1 be relaxed?

In Chapter 3 and Chapter 5, we develop a novel transformation to convert the discrete optimization and discrete optimal control problems into ones with continuous decision variables. The transformed continuous optimization and continuous optimal control problems are then solved by conventional gradient-based optimization techniques. A natural question to ask is whether or not methods based on highly efficient interior point type of method can be developed? Furthermore, what are the limitations of this approach?

All the optimization methods developed in this thesis are for finding local optimal solutions at the very best. In practice, a local optimal solution, if found, may be very far away from the global optimal solution, yielding an unsatisfactory cost value. Thus, it is of practical important to incorporate global optimization methods in the study of the solution methods for these optimization and optimal control problems. This is an interesting future research direction with great practical significance.

Bibliography

[1] P. E. G. A. Forsgren and M. H. Wright, "Interior methods for nonlinear optimization," *SIAM Review*, vol. 44, pp. 525–597, 2003.

[2] N. U. Ahmed, *Elements of finite-dimensional systems and control theory.* Essex: Longman Scientific and Technical, 1988.

[3] ——, *Dynamic systems and control with applications.* Singapore: World Scientific, 2006.

[4] K. K. E. A.V. Fiacco, *Semi-Infinite Programming and Applications.* Lecture Notes in Economics and Mathematical Systems, vol. 215, 1983.

[5] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms.* John Wiley & Sons, New Jersey, 2006.

[6] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods.* New York: Academic Press, 1982.

[7] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta Numerica*, vol. 4, pp. 1–51, 1996.

[8] M. Bremicker, P. Y. Papalambros, and H. T. Loh, "Solution of mixed-discrete structural optimization problems with a new sequential linearization algorithms," *Computer and Strucrures*, vol. 37, no. 4, pp. 451–461, 1990.

[9] B. Brosowski, *Parametric Semi-infinite Optimization.* Frankfurt, France: Verlag Peter Lang, 1982.

[10] N. I. M. G. C. Keller and A. J. Wathen, "Constraint preconditioning for indefinite linear systems," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, pp. 1300–1317, 2000.

[11] J. Cai and G. Thierauf, "Discrete optimization of structures using an improved penalty function method," *Engineering Optimization*, vol. 21, no. 4, pp. 293–306, 1993.

[12] C. H. Chan, S. C. Pei, and J. J. Shyu, "A new method for least-squares and minimax group design error design of allpass variable fractional-delay digital filters," *EURASIP Journal on Advances in Signal Processing*, Article ID: 976913 2010.

[13] A. R. Conn, N. I. M. Gould, and P. L. Toint, "Methods for nonlinear constraints in optimization calculations." RAL, Chilton, Tech. Rep. RAL-TR-96-042, Jun 1996.

[14] M. Dall'Aglio, "On some applications of lsip to probability and statistics," *in: M. A. Goberna, M. A. López (Eds), Semi-infinite Programming, Recent Advances, Nonconvex Optimization and Its Applications, 57, Kluwer, Dordrecht.*

[15] H. H. Dam, A. Cantoni, K. L. Teo, and S. Nordholm, "Variable digital filter with group delay flatness specification or phase constraints," *IEEE Transactions on Circuits and Systems II*, vol. 55, no. 5, pp. 442–446, May 2008.

[16] G. B. Dantzig, *Linear Programming and Extensions.* Princeton University Press, Princeton, NJ., 1963.

[17] T. B. Deng, "Noniterative wls design of allpass variable fractional-delay digital filters," *IEEE Trans. Circuits Systems*, vol. 53, no. 2, pp. 358–371, February 2006.

[18] ——, "Minimax design of low-complexity allpass variable fractional-delay digital filters," *IEEE Trans. Circuits Systems I*, vol. 57, no. 8, pp. 2075–2086, August 2010.

[19] J. E. Dennis and J. J. More, "Quasi-newton methods, motivation and theory," *SIAM Review*, vol. 19, pp. 46–89, 1977.

[20] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[21] D.Luenberger, *IntroductiontoLinearandNonlinearProgramming*, Addison Wesley, second ed., 1984.

[22] R. H. (Ed.), "Semi-infinite Programming," *Lecture Notes in Control and Information Sciences, Springer, Berlin*, 1979.

[23] S. C. Fang, D. Y. Gao, R. L. Sheu, and S. Y. Wu, "Canonical dual approach to solving 0-1 quadratic programming problems," *Journal of Industry and Management Optimization*, vol. 4, no. 1, February 2008.

[24] B. Farhadinia, K. L. Teo, and R. C. Loxton, "A computational method for a class of non-standard time optimal control problems involving multiple time horizons," *Mathematical and Computer Modelling*, vol. 49, no. 7-8, pp. 1682–1691, 2009.

[25] C. W. Farrow, "A continuously variable digital delay element," *Proc. IEEE Int. Symp. Circuits Syst*, vol. 3, pp. 2641–2645, June 1988.

[26] Z. G. Feng, K. L. Teo, and V. Rehbock, "A smoothing approach for semi-infinite programming with projected Newton-type algorithm," *Journal of Industrial and Management Optimization*, vol. 5, no. 1, pp. 141–151, 2009.

[27] R. Fletcher, *Practical Methods of Optimization.* John Wiley & Sons, New York, second ed., 1987.

[28] J. F. Fu, R. G. Fenton, and W. L. Cleghorn, "A mixed integer discrete continuous programming method and its application to engineering design optimization," *Engineering Optimization*, vol. 17, no. 4, pp. 263–280, 1991.

[29] D. Y. Gao, *Duality Principles in Nonconvex Systems - Theory, Methods and Applications*. Dordrecht: Kluwer Academic Publishers, 2000.

[30] D. Gao and N. Ruan, "Solutions to quadratic minimization problems with box and integer constraints," *Journal of Global Optimization*, vol. 47, no. 3, pp. 463–484, 2010.

[31] U. M. Garcia-Palomares and O. L. Mangasarian, "Super linearly convergent quasi-newton methods for nonlinearly constrained optimization problems," *Mathematical Programming*, vol. 11, pp. 1–13, 1976.

[32] K. Glashoff and S. A. Gustafson, "Linear optimization and approximation," 1983.

[33] G. Golub and D. O'leary, "Some history of the conjugate gradient methods and the lanczos algorithms: 1948c1976," *SIAM Review*, vol. 31, pp. 50–100, 1989.

[34] G. H. Golub and C. F. V. Loan, *Matrix Computations*. The Johns Hopkins University Press, Baltimore, third ed., 1996.

[35] C. Gonzaga, E. Polak, and R. Trahan, "An improved algorithm for optimization problems with functional inequality constraints," *Automatic Control, IEEE Transactions on*, vol. 25, no. 1, pp. 49–54, February 1980.

[36] N. I. M. Gould, "On the accurate determination of search directions for simple differentiable penalty functions," *I.M.A. Journal on Numerical Analysis*, vol. 6, pp. 357–372, 1986.

[37] N. I. M. Gould, D. Orban, and P. L. Toint, "Numerical methods for large-scale nonlinear optimization," *Acta Numerica*, vol. 14, pp. 299–361, 2005.

[38] B. Z. Guo and T. T. Wu, "Approximation of optimal feedback control: a dynamic programming approach," *Journal of Global Optimization*, vol. 46, no. 3, pp. 395–422, 2010.

[39] S. A. Gustafson and K. O. Kortanek, "Numerical treatment of a class of semi-infinite programming problems," *Naval Research Logistics Quarterly*, vol. 20, pp. 477–504, 1973.

[40] W. Hager, "Runge-kutta methods in optimal control and the transformed adjoint system," *Numerische Mathematik*, vol. 87, no. 2, pp. 247–282, 2000.

[41] S. P. Han, "Superlinearly convergent variable metric algorithms for general nonlinear programming problems," *Mathematical Programming*, vol. 11, pp. 263–282, 1976.

[42] ——, "A globally convergent method for nonlinear programming," *Journal of Optimization Theory and Applications*, vol. 22, pp. 297–309, 1977.

[43] M. R. Hestenes, "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, pp. 303–320, 1969.

[44] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, pp. 409–436, 1952.

[45] R. Hettich and K. Kortanek, "Semi-infinite programming: theory, methods, and applications," *SIAM Review*, vol. 35, no. 3, pp. 380–429, 1993.

[46] P. Howlett, "Optimal strategies for the control of a train," *Automatica*, vol. 32, no. 4, pp. 519–532, 1996.

[47] W. Huyer and A. Neumaier, "A new exact penalty function," *SIAM Journal on Optimization*, vol. 13, no. 4, pp. 1141–1158, 2003.

[48] S. Ito, Y. Liu, and K. L. Teo, "A dual parametrization method for convex semi-infinite programming," *Annals of Operations Research*, vol. 98, pp. 189–213, 2000.

[49] L. S. Jennings, M. E. Fisher, K. L. Teo, and C. J. Goh, *MISER 3.3 Optimal Control Software: Theory and User Manual*. University of Western Australia, Perth, Australia, 1996.

[50] L. S. Jennings and K. L. Teo, "A computational algorithm for functional inequality constrained optimization problems," *Automatica*, vol. 26, pp. 371–375, 1990.

[51] ——, *A numerical algorithm for constrained optimal control problems with applications to harvesting*. in Dynamicas of Complex Interconnected Biological Systems, Birkhauser Boston, Boston, 1990.

[52] C. H. Jiang, Q. Lin, C. J. Yu, K. L. Teo, and G. R. Duan, "An exact penalty method for free terminal time optimal control problem with continuous inequality constraints," *Journal of Optimization Theory and Applications*, DOI 10.1007/s10957-012-0006-9 2012.

[53] M. Jnger, T. Liebling, G. N. D. Naddef, W. Pulleyblank, G. Reinelt, G. Rinaldi, and L. Wolsey., *50 Years of Integer Programming 1958C2008: The Early Years and State-of-the-Art Surveys*. Springer-Verlag, 2009.

[54] J. Y. Kaakinen and T. Saramaki, "An algorithm for the optimization of adjustable fractional delay all-pass filters," *Proc. IEEE International Symposium on Circuits and Systems*, vol. 3, pp. 153–156, May 2004.

[55] N. karmarkar, "A new polynomial-time algorithm for linear programming," *Combinatorics*, vol. 4, pp. 373–395, 1984.

[56] C. Kaya and J. Martínez, "Euler discretization and inexact restoration for optimal control," *Journal of Optimization Theory and Applications*, vol. 134, no. 2, pp. 191–206, 2007.

[57] D. E. Kirk, *Optimal control theory: An introduction.* New York: Dover, 2004.

[58] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 481–492, 1951.

[59] M. López and G. Still, "Semi-infinite programming," *European Journal of Operational Research*, vol. 180, pp. 491–518, 2006.

[60] H. W. J. Lee, *Computational studies of optimal controls.* Ph.D thesis, University of Western Australia, Perth, Australia.

[61] H. W. J. Lee, K. L. Teo, V. Rehbock, and L. S. Jennings, "Control parametrization enhancing technique for time optimal control problems," *Dynamic Systems and Applications*, vol. 6, pp. 243–262, 1997.

[62] ——, "Control parametrization enhancing technique for optimal discrete-valued control problems," *Automatica*, vol. 35, no. 8, pp. 1401–1407, 1999.

[63] B. Li, C. J. Yu, K. L. Teo, and G. R. Duan, "An exact penalty function method for continuous inequality constrained optimal control problem," *Journal of Optimization Theory and Applications*, DOI 10.1007/s10957-011-9904-5 2011.

[64] D. H. Li, L. Q. Qi, J. Tam, and S. Y. Wu, "A smoothing newton method for semi-infinite programming," *Journal of Global Optimization*, vol. 3, pp. 169–194, 2004.

[65] H. L. Li and C. T. Chou, "A global approach for nonlinear mixed discrete programming in design optimization," *Engineering Optimization*, vol. 22, no. 2, pp. 109–122, 1993.

[66] R. Li, K. L. Teo, K. H. Wong, and G. R. Duan, "Control parameterization enhancing transform for optimal control of switched systems," *Mathematical and Computer Modelling*, vol. 43, no. 11-12, pp. 1393–1403, 2006.

[67] Y. C. Lim, R. Yang, D. N. Li, and J. J. Song, "Signed power-of-two term allocation scheme for the design of digital filters," *IEEE Transactions on Circuits and Systems II: Analog and Digital signal processing*, vol. 46, no. 5, May 1999.

[68] J. T. Linderoth and M. W. P. Savelsbergh, "A computational study of search strategies for mixed integer programming," *INFORMS Journal on Computing*, vol. 11, no. 2, pp. 173–188, 1999.

[69] Y. Liu, S. Ito, H. W. J. Lee, and K. L. Teo, "Semi-infinite programming approach to continuously-constrained linear-quadratic optimal control problems," *Journal Of Optimization Theory and Applications*, vol. 108, no. 3, pp. 617–632, March 2001.

[70] Y. Liu and K. L. Teo, "An adaptive dual parametrization algorithm for quadratic semi-infinite programming problems," *Journal of Global Optimization*, vol. 24, pp. 205–217, 2002.

[71] Y. Liu, K. L. Teo, and S. Ito, "Global optimization in quadratic semi-infinite programming," *Computing Supplementa*, vol. 15, pp. 119–132, 2001.

[72] Y. Liu, K. L. Teo, L. S. Jennings, and S. Wang, "On a class of optimal control problems with state jumps," *Journal of Optimization Theory and Applications*, vol. 98, no. 1, pp. 65–82, 1998.

[73] Y. Liu, K. L. Teo, and S. Y. Wu, "A new quadratic semi-infinite programming algorithm based on dual parametrization," *Journal of Global Optimization*, vol. 29, pp. 401–413, 2004.

[74] H. T. Loh and P. Y. Papalambros, "Computational implementation and tests of a sequential linearization algorithm for mixed-discrete nonlinear design optimization," *Journal of Mechanical Design*, vol. 113, no. 3, pp. 325–344, 1991.

[75] ——, "A sequential linearization approach for solving mixed-discrete nonlinear design optimization problems," *Journal of Mechanical Design*, vol. 113, pp. 325–334, 1991.

[76] L. Luk̆san and J. Vlček, "Indefinitely preconditioned inexact newton method for large sparse equality constrained nonlinear programming problems," *Numerical Linear Algebra with Applications*, vol. 5, pp. 219–247, 1998.

[77] S. Lucidi and F. Rinaldi, "Exact penalty functions for nonlinear integer programming problems," *Journal of Optimization Theory and Applications*, vol. 145, no. 3, pp. 479–488, 2010.

[78] M. A. L. M. A. Goberna, *Linear Semi-Infinite Optimization*. Wiley, Chichester, 1998.

[79] M. Makundi, T. I. Laakso, and V. Valimaki., "Efficient tunable iir and allpass structures," *Electron. Lett.*, vol. 37, no. 6, pp. 344–345, March 2001.

[80] W. Murray and K. M. Ng, "An algorithm for nonlinear optimization problems with binary variables," *Computational Optimization and Applications*, vol. 47, no. 2, pp. 257–288, 2010.

[81] C. NG and L. Zhang, "Discrete filled function method for discrete global optimization," *Computational Optimization and Applications*, vol. 31, no. 1, pp. 87–115, 2005.

[82] Q. Ni, C. Ling, L. Q. Qi, and K. L. Teo, "A truncated projected newton-type algorithm for large-scale semi-infinite programming," *SIAM Journal on Optimization*, vol. 16, pp. 1137–1154, 2006.

[83] J. Nocedal and S. J. Wright, *Numerical Optimization, second ed.* Springer, 2006.

[84] P. M. Pardalos, "Continuous approaches to discrete optimization problems," *in Nonlinear Optimization and Applications*, pp. 313–328, 1996.

[85] P. M. Pardalos, O. Prokopyev, and S. Busygin, "Continuous approaches for solving discrete optimization problems," *in Handbook on Modelling for Discrete Optimization*, pp. 39–60, 2006.

[86] P. M. Pardalos and J. B. Rosen, "Constrained global optimization: algorithms and applications," vol. 268, 1987.

[87] P. M. Pardalos and V. Yatsenko, "Optimization and control of bilinear systems," 2009.

[88] E. Polak, "On the mathematical foundations of nondifferentiable optimization in engineering design," *SIAM Review*, vol. 29, pp. 21–89, 1987.

[89] E. Polak and D. Q. Mayne, "An algorithm for optimization problems with functional inequality constraints," *IEEE Transactions on Automatic Control*, vol. 21, pp. 184–193, 1976.

[90] E. Polak, D. Q. Mayne, and D. M. Stimler, "Control system design via semi-infinite optimization: a review," *Proceedings of the IEEE*, vol. 72, no. 12, pp. 1777–1794, 1984.

[91] E. Polak, K. S. Pister, and D. Ray, "Optimal design framed structures subjected to earthquak," *Engineering Optimization*, vol. 12, pp. 65–71, 1976.

[92] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," *in Optimization, R. Fletcher, ed., Academic Press, New York*, pp. 283–298, 1969.

[93] ——, "On search directions for minimization algorithms," *MathematicalProgramming*, vol. 4, pp. 193–201, 1973.

[94] ——, "A fast algorithm for nonlinearly constrained optimization calculations," *Numerical Analysis Dundee 1977, Springer Verlag*, pp. 144–157, 1977.

[95] ——, "Algorithms for nonlinear constraints that use lagrangian functions," *Mathematical Programming*, vol. 14, pp. 224–248, 1978.

[96] ——, "The convergence of variable metric methods for nonlinearly constrained optimization calculations," *Nonlinear Programming 3, Academic Press, New York and London,*, pp. 27–63, 1978.

[97] P. Z. R. Hettich, *Numerische Methoden der Approximation und der semi-infiniten Optimierung.* Stuttgart: Teubner, 1982.

[98] R. Rabenstein, "Minimization of transient signals in recursive time-varying filters," *Circuits Syst. Signal Process.*, vol. 7, no. 3, pp. 345–359, 1988.

[99] V. Rehbock and L. Caccetta, "Two defence applications involving discrete valued optimal control," *ANZIAM Journal*, vol. 44, pp. 33–54, 2002.

[100] U. T. Ringertz, "On methods for discrete structural optimization," *Engineering Optimization*, vol. 13, no. 1, pp. 47–64, 1988.

[101] R. T. Rockafellar, "The multiplier method of hestenes and powell applied to convex programming," *Journal of Optimization Theory and Applications*, vol. 12, pp. 555–562, 1973.

[102] J. B. Rosen and J. Kreuser, *Iterative Methods for Sparse Linear Systems*. SIAM Publications, Philadelphia, PA, second ed., 2003.

[103] H. L. Royden, *Real analysis*. Prentice Hall, 1988.

[104] T. Ruby and V. Rehbock, "Numerical solutions of optimal switching control problems," in *Optimization and control with applications*, L. Qi, K. L. Teo, and X. Yang, Eds. New York: Springer, 2005, pp. 447–459.

[105] E. Sandgren, "Nonlinear integer and discrete programming in mechanical design optimization," *Journal of Mechanical Design*, vol. 112, no. 2, pp. 223–229, 1988.

[106] D. F. Shanno and E. M. Simantiraki, "Interior point methods for linear and nonlinear programming," RAL, Chilton, Tech. Rep. RAL-TR-96-042, Jun 1996.

[107] D. K. Shin, Z. Gürdal, and O. H. Griffin JR, "A penalty approach for nonlinear optimization with discrete design variables," *Engineering Optimization*, vol. 16, no. 1, pp. 29–42, 1990.

[108] J. J. Shyu, S. C. Pei, and C. H. Chan, "Minimax phase error design of allpass variable frational-delay digital filters by iterative weighted least-squares method," *Signal Processing.*, vol. 89, pp. 1774–1781, 2009.

[109] A. Siburian and V. Rehbock, "Numerical procedure for solving a class of singular optimal control problems," *Optimization Methods and Software*, vol. 19, no. 3-4, pp. 413–426, 2004.

[110] D. E. Stewart, "A numerical algorithm for optimal control problems with switching costs," *Journal of the Australian Mathematical Society, Series B*, vol. 34, no. 2, pp. 212–228, 1996.

[111] X. L. Sun, J. L. Li, and H. Z. Luo, "Convex relaxation and lagrangian decomposition for indefinite integer quadratic programming," *Optimization*, vol. 59, no. 5, pp. 627–641, 2010.

[112] K. L. Teo, "Control parametrization enhancing transform to optimal control problems," *Nonlinear Analysis*, vol. 63, no. 5-7, pp. 2223–2236, 2005.

[113] K. L. Teo and C. J. Goh, "A simple computational procedure for optimization problems with functional inequality constraints," *IEEE Transactions on Automatic Control*, vol. 32, pp. 940–941, 1987.

[114] K. L. Teo, C. J. Goh, and K. H. Wong, *A unified computational approach to optimal control problems*. Longman Scientific & Technical, 1991.

[115] K. L. Teo, L. S. Jennings, H. W. J. Lee, and V. Rehbock, "The control parameterization enhancing transform for constrained optimal control problems," *Journal of the Australian Mathematical Society—Series B*, vol. 40, no. 3, pp. 314–335, 1999.

[116] K. L. Teo, V. Rehbock, and L. S. Jennings, "A new computational algorithm for functional inequality constrained optimization problems," *Automatica*, vol. 29, pp. 789–792, 1993.

[117] K. L. Teo, X. Q. Yang, and L. S. Jennings, "Computational discretization algorithms for functional inequality constrained optimization," *Annals of Operations Research*, vol. 98, pp. 215–234, 2000.

[118] R. Tiryono, V. Rehbock, and W. B. Lawrance, "Optimal control of hybrid power systems," *Dynamics of Continuous, Discrete and Impulsive Systems Series B,*, vol. 10, no. 3, pp. 429–440, 2003.

[119] X. J. Tong and S. Z. Zhou, "A smoothing projected newton-type method for semismooth equations with bound constraints," *Journal of Industrial and Management Optimization*, vol. 2, pp. 235–250, 2005.

[120] L. Trefethen and D. Bau, *Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.

[121] C. Tseng and J. Arora, "On implementation of computational algorithms for optimization design," *International Journal for Numerical Methods in Engineering*, vol. 26, no. 6, pp. 1365–1402, 1988.

[122] C. C. Tseng, "Design of 1-d and 2-d variable fractional delay allpass filters using weighted least square methods," *IEEE Trans. Circuits Syst. I*, vol. 49, no. 10, pp. 1413–1422, October 2002.

[123] S. Wang, F. Gao, and K. L. Teo, "An upwind finite-difference method for the approximation of viscosity solutions to Hamilton-Jacobi-Bellman equations," *IMA Journal of Mathematical Control and Information*, vol. 17, no. 2, pp. 167–178, 2000.

[124] S. Wang, L. S. Jennings, and K. L. Teo, "Numerical solution of Hamilton-Jacobi-Bellman equations by an upwind finite volume method," *Journal of Global Optimization*, vol. 27, no. 2-3, pp. 177–192, 2003.

[125] S. Wang, K. L. Teo, and H. W. J. Lee, "A new approach to nonlinear mixed discrete programming problems," *Engineering Optimization*, vol. 30, no. 3, pp. 249–262, 1998.

[126] S. F. Woon, V. Rehbock, and R. Loxton, "Global optimization method for continuous-time sensor scheduling," *Nonlinear Dynamics and Systems Theory*, vol. 10, no. 2, pp. 175–188, 2010.

[127] ——, "Towards global solutions of optimal discrete-valued control problems," *Optimal Control Applications and Methods*, DOI: 10.1002/oca.1015 2011.

[128] C. Z. Wu and K. L. Teo, "Global impulsive optimal control computation," *Journal of Industrial and Management Optimization*, vol. 2, no. 4, pp. 435–450, 2006.

[129] S. Y. Wu, D. H. Li, L. Q. Qi, and G. L. Zhou, "An iterative method for solving kkt system of the semi-infinite programming," *Optimization Methods and Software*, vol. 20, pp. 629–643, 2005.

[130] X. Xu and P. J. Antsaklis, "Optimal control of switched systems based on parameterization of the switching instants," *IEEE Transactions on Automatic Control*, vol. 49, no. 1, pp. 2–16, 2004.

[131] X. Q. Yang and K. L. Teo, "Nonlinear lagrangian functions and applications to semi-infinite programs," *Annals of Operations Research*, vol. 103, pp. 235–250, 2001.

[132] C. J. Yu, B. Li, R. Loxton, and K. L. Teo, "Optimal discrete-valued control computation," *Journal of Global Optimization*, DOI 10.1007/s10898-012-9858-7 2012.

[133] C. J. Yu, K. L. Teo, and Y. Q. Bai, "An exact penalty function method for nonlinear mixed discrete programming problems," *Optimization Letters*, DOI 10.1007/s11590-011-0391-2 2011.

[134] C. J. Yu, K. L. Teo, and H. H. Dam, "Design of allpass variable fractional delay filter with signed powers-of-two coefficients," 2012, submitted to IEEE Transactions on Signal Processing.

[135] C. J. Yu, K. L. Teo, L. S. Zhang, and Y. Q. Bai, "A new exact penalty function method for continuous inequality constrained optimization problems," *Journal of Industrial and Management Optimization*, vol. 6, no. 4, pp. 895–910, 2010.

[136] ——, "On a refinement of the convergence analysis for the new exact penalty function method for continuous inequality constrained optimization problem," *Journal of Industrial and Management Optimization*, 2012, to appear.

[137] J. Zabczyk, *Mathematical control theory: An introduction*. Boston: Birkhäuser, 2008.

[138] L. S. Zhang, "New simple exact penalty function for constrained minimization on $\mathbb{R}^n$," *The 4^{th} Australia-China workshop on optimization: Theory, Methods and Applications*, December 2009.

[139] X. J. Zheng, X. L. Sun, and D. Li, "Separable relaxation for nonconvex quadratic integer programming: integer diagonalization approach," *Journal of Optimization Theory and Applications*, vol. 146, no. 2, pp. 463–489, 2010.