# On Cardinality Constrained Optimization

**GAO, Jianjun**

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

in

Systems Engineering and Engineering Management

June 2009

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

# ABSTRACT

Although cardinality constraints naturally arise in many applications, e.g., in portfolio selection problems of choosing small number of assets from a large pool of stocks or dynamic portfolio selection problems with limited trading dates within a given time horizon and in subset selection of the regression analysis, the state-of-the-art in cardinality constrained optimization has been stagnant up to this stage, largely due to the inherent combinatorial nature of such hard problems. We focus in this research on developing efficient and implementable solution algorithms for cardinality constrained optimization by investigating prominent structures and hidden properties of such problems. More specifically, we develop solution algorithms for four specific cardinality constrained optimization problems, including (i) the cardinality constrained linear-quadratic control problem, (ii) the optimal control problem of linear switched system with limited number of switching, (iii) the time cardinality constrained dynamic mean-variance portfolio selection problem, and (iv) cardinality constrained quadratic optimization problem. Taking advantages of a linear-quadratic structure of cardinality constrained optimization problems, we strive for analytical solutions when possible. More specifically, we derive an analytical solution for problem (iii) and obtain for both problems (i) and (ii) semi-analytical expressions of the solution governed by a family of Ricatti-like equations, which still suffer an exponentially growing complexity. To achieve high-performance of the solution algorithm, we devise algorithms of a branch and bound (BnB) type with various tight and computationally-cheap lower bounds achieved by identifying suitable

SDP formulations and by exploiting geometric properties of the problem. We demonstrate efficiency of our proposed solution schemes evidenced from numerical experiments and present a firm step-forward in tackling this long-standing challenge of cardinality constrained optimization.

# 摘要

尽管现实生活中决策变量自由度受限制的问题比比皆是，例如在投资组合问题中投资者只在所有股票中选择一部分投资，或者多阶段动态投资问题中只在所有 投资周期中选择几个周期投资风险资产，又或者统计学家在回归分析中常常只考虑使用部分回归量。 遗憾的是对于这类问题的研究在过去一些年中停滞不前，这很大程度上是因为决策变量自由度这个限制实际上是个组合数。通 过对具体问题特殊性质的研究，我们在本研究中提出了一些有效的，可行的方法来处理决策变量自由度受限的问题。 具体来说，我们对下列四类问题提出了有效的算法，(i)控制次数受限的线性二次型控制问题 (ii)切换次数受限的最优切换控制问题 (iii)投资次数受限制的动态均值方差投资问题(iv)决策变量受限的二次规划问题。 我们尽力构造对于有线性二次型这一特殊结构问题的解析解。具体来讲，对于问题(iii)我们得到了解析表达式， 对于问题(i)(ii)我们得到了半解析的表达式。这些半解析的表达式是通过计算一组黎卡提(Riccati)方程迭代得到。 但是这个计算过程的复杂度仍是 成指数增长的。为寻求更高效的方法，我们采用类似分支定界的思想， 提出了不同的方法来构造计算费用低且相对较紧的下界， 比如构造合适的半定规划问题，或者利用问题自身的几何性质等。 我们的数值计算结果证明了这些方法的高效性与可靠性，为进一步的研究奠定了基础。

# Acknowledgement

I have been quite fortunate in having Professor Duan Li as my supervisor during my time of study in the Department of System Engineering and Engineering Management. His inspiring and patient guidance, continuous encouragement and support made it possible for me to complete this thesis. He is one of my beloved and respected teachers in my life.

I am also thankful to Professor Erwei Bai, Professor Shuzhong Zhang, and Professor Xiang Zhou for serving as members of my thesis committee and for their valuable comments on my thesis.

I am also grateful to Professor Xiaoling Sun for many advices on my research and to my friends Xiangyu Cui and Lan Yi for their stimulating discussions with me. I would like to take this opportunity to thank all professors, technical staffs, clerical staffs and postgraduate students in the Department of System Engineering and Engineering Management.

Finally, I wish to express my sincere thanks to my parents and my girlfriend Song Yan for their everlasting love, care, and understanding. Their support enabled me to conquer the difficulties and achieve many wonderful things that I could not accomplish myself.

# CONTENTS

# List of Tables

# List of Figures

# Notations

## Usual Notations

| Notation | Descriptions |
|---|---|
| $\mathbb{R}$ | The set of real numbers |
| $\mathbb{Z}_+$ | The set of positive integer |
| $\mathbb{S}^n_+$ | The set of positive-semidefinite matrices of order $n$ |
| $\mathbb{S}^n_{++}$ | The set of positive definite matrices of order $n$ |
| $\\|\cdot\\|_2$ | The $L_2$ norm |
| $\delta(\cdot)$ | The indicate function $\mathbb{R}^n \rightarrow \{0,1\}$ such that $\delta(a) = 1$ if $a$ is non-zero vector and $\delta(a) = 0$, otherwise |
| $\mathbf{I}_n$ | The $n \times n$ identity matrix |
| $\mathbf{0}_n$ | The $n \times n$ zero matrix |
| $\mathbf{1}_n$ | The $n$-dimensional vector with all its elements being one |
| $v(\cdot)$ | The optimal value of problem $(\cdot)$ |
| $\mathcal{C}_n^m$ | The combinatorial number $\mathcal{C}_n^m = n!/(m!(n-m)!)$ |

# Notations of Chapter 2

| Notation | Descriptions |
|---|---|
| $F^s_{n,m,T}$ | The problem defined in Problem (2.1) |
| $G^s_{n,m,T}$ | The problem defined in Problem (2.29) |
| $\hat{\mathcal{R}}(\cdot)$ | The operation defined in (2.7) |
| $\bar{\mathcal{R}}(\cdot)$ | The operation defined in (2.8) |
| $\mathcal{K}(Q)$ | The operation of eliminating dominant elements in the set $Q \subset \mathbb{S}^n_+$ |
| $\hat{\mathbb{P}}^{r_t}_t$ | The set of matrices calculated in (2.10) |
| $\bar{\mathbb{P}}^{r_t}_t$ | The set of matrices calculated in (2.11) |
| $\mathbb{P}^{r_t}_t$ | The set of matrices calculated in (2.12) |
| $\mathbb{M}^{r_t}_t$ | The action region defined in Theorem 2.3 |
| $\mathcal{T}(\mathbb{F}_{1,m,T})$ | The coefficients set from sCCLQR problem (see (2.33)) |

# Notations of Chapter 3

| Notation | Descriptions |
|---|---|
| $\mathcal{W}$ | The coefficient set defined in (3.2) |
| $\mathbf{P}_1(T)$ | The problem defined in Problem 3.1 |
| $\mathbf{P}_2(s)$ | The problem defined in Problem 3.2 |
| $\mathbf{P}_3(\infty)$ | The problem defined in Problem 3.3 |
| $\mathbf{P}_4(T,s)$ | The problem defined in Problem 3.4 |
| $\|x\|^2_S$ | The quadratic term $x'Sx$ with $x \in \mathbb{R}^n$ and $S \in \mathbb{S}^n_+$ |
| $\mathcal{R}(j,P)$ | The operation defined in (3.11) |
| $\mathcal{M}^t(i,j,r)$ | The switching region defined in Theorem 3.3 |
| $\mathbb{P}(t,i,j,r)$ | The matrices set defined in (3.12) |
| $\mathbb{C}(t,i,r)$ | The matrices set defined in (3.13) |

# Notations of Chapter 4

| Notation | Descriptions |
|---|---|
| $\mathbf{P}_t$ | The relative return vector of risky assets (4.1) |
| $\mathcal{P}_1(\sigma)$ | The problem defined in (4.2) |
| $\mathcal{P}_2(\epsilon)$ | The problem defined in (4.3) |
| $\mathcal{A}_1(\sigma, s)$ | The auxiliary problem defined in (4.6) |
| $\mathcal{A}_2(\epsilon, s)$ | The ausiliary problem defined in (4.7) |
| $\mathbf{c}_t$ | The parameter defined in (4.8) |
| $\mathbf{D}_t$ | The parameter defined in (4.9) |
| $\theta_t$ | The parameter defined in (4.10) |
| $\gamma_t$ | The parameter defined in (4.11) |

# Notations of Chapter 5

| Notation | Descriptions |
|---|---|
| $\mathcal{G}_s(D, d)$ | The cardinality constrained quadratic optimization problem |
| $\Delta(s)$ | The cardinality constrained set defined in (5.1) |
| $\mathcal{E}(P, P, \rho)$ | The ellipsoid defined in (5.10) |
| $\mathcal{B}(o, r^2)$ | The ball defined in (5.22) |
| $\theta(y)$ | The permutation of vector $y$ in ascending order |

# CHAPTER 1

# INTRODUCTION

The subject of optimization has grown by leaps and bounds, largely driven by the demand of its applications in almost all areas of engineering, science and management. The last century has witnessed numerous theoretical breakthroughs and innumerable successful applications in various fields. However, there are always many long standing and newly emerging challenges in front of the optimization community. In this thesis, we are interested in investigating one such a challenge termed cardinality constrained optimization where the freedom of the decision variables is restricted. The cardinality constraint naturally arises in many applications, e.g., statisticians are often confined themselves in identifying a subset of regressors in their regression analysis, investors always choose a small number of assets to invest from a large pool of stocks; investors never exhaust all available trading dates to modify their portfolios, and engineers only implement their control actions a few times on manufacturing lines due to a concern of the corresponding cost. Due to the combinatorial nature of the possibility in satisfying a cardinality constraint, optimization problems involving a cardinality constraint are in general very hard to solve. Up to this stage, the state-of-the-art in cardinality constrained optimization has been stagnant. The few articles addressing cardinality constraint have only proposed some seemingly naive lower bounding schemes together with a branch-and-bound framework, for example, a lower bounding scheme by dropping the cardinality constraint. We focus in this research on developing efficient and implementable solution schemes for some particular cardinality constrained optimization problems by investigating special structures and prominent features of such problems. More specifically, we focus on the following four specific cardinality constrained optimization problems.

- *The cardinality constrained linear-quadratic regulator(CCLQR)* The optimal control problem involving linear dynamics and quadratic penalty, named linear-quadratic optimal control problem, has been extensively studied in the literature since Kalman's seminar work on linear quadratic regulator (LQR). We study in this research discrete-time systems, due to a consideration of computation and applications. Implementing a control action usually incurs two types of costs, fixed (set-up) cost and variable cost associated with the magnitude or energy of the control. While the second type of cost has been extensively investigated in the literature, the first type of cost has been not yet addressed. When assuming such a fixed cost to be a constant in each time period, the LQR problem with set-up costs turns out to be equivalent to identifying the solution of a linear-quadratic control problem with limited number of control implementation, which is termed cardinality constrained linear-quadratic regulator problem(CCLQR) in this research.

- *Optimal control of linear switched systems with switching cost* Linear switched system is a particular class of hybrid systems, which consists of several sub linear-systems. Optimal control of such a system has been investigated in the literature, e.g., [75] [82] [67], and many real-world applications have been also reported in the literature, e.g., the chemical process, automotive systems and embedded systems, etc. (see [82] [75]). Although switching cost arises in many switched systems, this factor has been often neglected in most of the existing studies, except [67] in which the switched autonomous systems are studied. We focus in this research on the optimal control problem of linear switched systems with quadratic cost function and constant switching cost, where both optimal switching signal and control input are decision variables.

- *Time cardinality constrained mean-variance dynamic portfolio selection* The mean-variance formulation proposed by Markowitz [56] [55] provides the fundamental basis of portfolio selection and the results of the single-period mean-variance portfolio selection has been extended to a multi-period setting by Li and Ng [43] and to a continuous-time setting by Zhou and Li [84]. We extend further dynamic mean-variance portfolio selection by considering the management fee (cost) imposed on the non-zero position on risky assets. Due to such a set-up type of management fees charged for investing in risky assets, investors do not always invest in risky assets in all time periods. This real-world investment situation leads to a mathematical formulation in this research termed time cardinality constrained mean-variance dynamic portfolio selection, where both the investment timing and the distribution of the wealth have to be determined at the same time.

- *The cardinality constrained quadratic optimization problem.* On recognizing various applications of cardinality constrained optimization, we consider a general problem formulation of minimizing a convex quadratic function subject to a single cardinality constraint. In this research, we explore rich geometrical properties hidden behind the special structure of the cardinality constrained quadratic optimization problem. The revealed prominent features enable us to develop tight and cheap lower bounds, when adopting powerful numerical schemes of semi-definite programming (SDP). Integrating such lower bounds into a solution algorithm of a branch-and-bound type, we achieve promising numerical results for large-scale test problems.

This thesis is organized as follows. After the brief introduction presented in this chapter, we focus in Chapter 2 on the CCLQR problem. In Chapter 3, we investigate the optimal control problem of linear switched systems with an added feature of switching cost. We explore the time-cardinality constrained mean-variance portfolio selection in Chapter 4. In Chapter 5, we concentrate on the general cardinality constrained quadratic optimization problem. We emphasize here that the notations and materials of all the four technical chapters are self-contained. We finally conclude this thesis in Chapter 6 with some remarks and suggestions for future research directions.

# CHAPTER 2

# Cardinality Constrained Linear-Quadratic Control

## 2.1. Introduction and Problem Formulation

Linear-quadratic regulator (LQR) problem is, with a doubt, one of the most remarkable achievements in control theory, largely due to its mathematical elegance in tractability and a wide range of its applications. The past three decades have witnessed many extensions of the traditional LQR in the literature, see, e.g. [2], [38],[8], [5], [39], [41], [40] and [44]. We explore in this chapter another extension of the conventional discrete-time LQR problem with a finite time horizon.

We consider in this chapter the following time-varying linear system,

$$x_{t+1} \quad = \quad A_t x_t + B_t u_t, \quad t = 0, \cdots, T-1, \tag{2.1}$$

where $x_t \in \mathbb{R}^n$ is the state with given initial state $x_0$, $u_t \in \mathbb{R}^m$ is the control, $A_t \in \mathbb{R}^{n \times n}$ and $B_t \in \mathbb{R}^{n \times m}$. The performance index to be minimized is of the following quadratic form,

$$J(x, u) := \sum_{t=0}^{T-1} \left[ x'_{t+1} Q_{t+1} x_{t+1} + u'_t R_t u_t \right], \tag{2.2}$$

where $Q_t \in \mathbb{R}^{n \times n}$ and $R_t \in \mathbb{R}^{m \times m}$ are positive semi-definite and positive definite, respectively. Different from the traditional LQR problem, the cardinality constrained LQR problem imposes a limit on the number of the control implementation,

$$\sum_{t=0}^{T-1} \delta(u_t) \leq s, \tag{2.3}$$

where $s$ is a given positive integer less than or equal to $T$, $\delta(u_t) = 0$ if $u_t$ is a zero vector and $\delta(u_t) = 1$ otherwise. The discrete-time cardinality constrained linear-quadratic regulator problem (CCLQR) is stated now as follows.

**Problem 2.1.** For dynamic system (2.1), find control sequence $\{u_t\}_{t=0}^{T-1}$ that minimizes the performance index (2.2) subject to the cardinality constraint (2.3).

Denote by $(F^s_{n,m,T})$ a given specific CCLQR problem with dimension of state being $n$, dimension of control being $m$, time horizon being $T$ and number of cardinality being $s$. We use $v(\cdot)$ to denote the optimal value of problem $(\cdot)$ in this chapter.

Studying such a CCLQR problem is motivated by the consideration of the set-up cost attached to nonzero control action. Set-up costs, in many situations, prevent control actions from being implemented at every time period. Consider the following revised quadratic performance index,

$$J := \sum_{t=0}^{T-1} w\delta(u_t) + \sum_{t=0}^{T-1} [x'_{t+1} Q_{t+1} x_{t+1} + u'_t R_t u_t], \tag{2.4}$$

where $w > 0$ is the set-up cost of implementing a control action. The discrete-time LQR problem with a set-up cost is given as follows.

**Problem 2.2.** For dynamic system (2.1), find control sequence $\{u_t\}_{t=0}^{T-1}$ that minimizes the performance index (2.4).

When $w$ is set at zero, Problem 2.2 reduces to the conventional LQR problem, whereas a very large $w$ forces all $u_t$, $t = 0, \cdots, T-1$, to be zero vectors. In general, an incorporation of a set-up cost into the formulation of LQR problem is equivalent to placing an upper limit on the number of control implementation. Notice the following fact.

**Lemma 2.1.** *Monotonicity. For any $s_1$ and $s_2$ that satisfy $1 \leq s_1 < s_2 \leq T$, $v(F^{s_1}_{n,m,T}) \geq v(F^{s_2}_{n,m,T})$.*

Proof. The lemma is obvious from the fact that the feasible region of $\{u_t\}_{t=0}^{T-1}$ in problem $(F^{s_1}_{n,m,T})$ is a subset of the feasible region of $\{u_t\}_{t=0}^{T-1}$ in problem $(F^{s_2}_{n,m,T})$. □

The optimal solution for Problem 2.2 can be found by first solving problem $(F^{s}_{n,m,T})$ for $1 \leq s \leq T$, and then identifying the optimal cardinality $s^*$ such that

$$s^* = \arg\min_{0 \leq s \leq T} \{ws + v(F^{s}_{m,n,T})\}.$$

It is evident that the solution to $(F^{s^*}_{n,m,T})$ is the optimal control to Problem 2.2. Thus, an efficient solution scheme of CCLQR problem plays a key role in solving Problem 2.2.

In the remaining of this chapter, we first discuss in Section 2.2 how to use dynamic programming to solve CCLQR. For CCLQR problems with a scalar state space (sCCLQR), dynamic programming yields an analytical solution. The complexity of the exact algorithm using dynamic programming, however, grows exponentially with the dimension of the state space. Recognizing this fact, we adopt in Section 2.3 techniques from semi-definite programming to construct a corresponding sCCLQR problem, of which the solution provides an approximate solution to the primal CCLQR with a vector state space. Several illustrative examples are presented in Section 2.4. We give some conclusion remarks in Section 2.5.

Throughout this chapter, we use notion $Q \succeq 0$ $(Q \succ 0)$ to denote a positive semidefinite (positive definite) matrix $Q$, $\mathbb{S}^n_+$ ($\mathbb{S}^n_{++}$ ) the set of all $n \times n$ positive semidefinite matrices (positive definite matrices), $\mathrm{diag}(S_0, S_1, \cdots, S_{T-1})$ the block diagonal matrix with $S_t \in \mathbb{R}^{n\times m}$ for $t = 0, \cdots, T-1$, $\mathbf{I}_n$ the $n \times n$ identity matrix, and $\mathbf{0}_n$ the $n \times n$ zero matrix. For an $H \in \mathbb{H} \subset \mathbb{S}^n_+$, if there exists another $H^* \in \mathbb{H}$ such that $H \succ H^*$, $H$ is called dominated with respect to $\mathbb{H}$. We denote by $\mathcal{K}(\mathbb{H})$ the set derived from $\mathbb{H}$ by eliminating all of its dominated members. The following fact is true.

**Lemma 2.2.** *Given $x \in \mathbb{R}^n$ and $S \subset \mathbb{S}^n_+$, then*

$$\min_{H \in S} x'Hx = \min_{H \in \mathcal{K}[S]} x'Hx.$$

## 2.2. Dynamic Programming Based Solution Scheme

The CCLQR problem $(F^s_{n,m,T})$ becomes separable when we expand the state space by adding an integer-valued variable $r_t$ that satisfies the following recursive equation for $t = 0, \cdots, T-1$,

$$r_{t+1} = \begin{cases} r_t - \delta(u_t) & \text{if } r_t > 0, \\ 0 & \text{otherwize,} \end{cases}$$

where $r_0 = s$. Clearly, $r_t$ represents the remaining number of control implementation at stage $t$. We define the cost-to-go at stage $t$ for a given pair $(x_t, r_t)$ as

$$J_t(x_t, r_t) = \min_{u_t, \cdots, u_{T-1}} \left\{ x_T' Q_T x_T + \sum_{k=t}^{T-1} [x_k' Q_k x_k + u_k' R_k u_k] \mid x_t, r_t \right\}, \tag{2.5}$$

where $Q_0$ is set as a zero matrix. The following is evident from Lemma 4.1,

$$J_t(x_t, r_1) \le J_t(x_t, r_2), \quad \text{if} \;\; r_1 > r_2.$$

Thus, to attain the optimality of $(F^s_{n,m,T})$, the feasible range of $r_t$, $t = 0, \cdots, T-1$, can be confined to the following set[1],

$$F_t := \{ r_t \in \mathbb{Z}^+ \mid \max(0, s-t) \le r_t \le \min(s, T-t) \}. \tag{2.6}$$

We introduce the following notations for problem $(F^s_{n,m,T})$. At stage $t$, the Riccati operator $\hat{\mathcal{R}}_t(\cdot) : \mathbb{S}^n_+ \rightarrow \mathbb{S}^n_+$ and the degenerated Riccati operator $\bar{\mathcal{R}}_t(\cdot)$ :

[1] $\mathbb{Z}^+$ is the set of nonnegative integers

$\mathbb{S}^n_+ \rightarrow \mathbb{S}^n_+$ are defined, respectively, as,

$$\hat{\mathcal{R}}_t(P) \;:=\; A_t'(P - PB_t(B_t'PB_t + R_t)^{-1}B_t'P)A_t + Q_t, \quad P \in \mathbb{S}^n_+, \tag{2.7}$$

$$\bar{\mathcal{R}}_t(P) \;:=\; A_t'PA_t + Q_t, \quad P \in \mathbb{S}^n_+. \tag{2.8}$$

Abusing the above notaion, we further define the following for a set of semi-definite matrices, $\mathbb{P} \subseteq \mathbb{S}^n_+$,

$$\hat{\mathcal{R}}_t(\mathbb{P}) := \bigcup_{P\in\mathbb{P}} \{\hat{\mathcal{R}}_t(P)\}, \quad \bar{\mathcal{R}}_t(\mathbb{P}) := \bigcup_{P\in\mathbb{P}} \{\bar{\mathcal{R}}_t(P)\}.$$

For convenience, let $\hat{\mathcal{R}}_t(\emptyset) = \emptyset$ and $\bar{\mathcal{R}}_t(\emptyset) = \emptyset$. The solution to CCLQR problem $(F^s_{n,m,T})$ can be characterized by the following theorem.

**Theorem 2.3.** *The following control law is optimal to CCLQR problem* $(F^s_{n,m,T})$*,*

$$\delta(u_t) = \begin{cases} 1 & x_t \in \mathbb{M}^{r_t}_t, \\ 0 & x_t \notin \mathbb{M}^{r_t}_t, \end{cases}$$

*where region* $\mathbb{M}^{r_t}_t$ *is defined as,*

$$\mathbb{M}^{r_t}_t := \begin{cases} \emptyset & \text{if } \hat{\mathbb{P}}^{r_t}_t = \emptyset, \\ \mathbb{R}^n & \text{if } \bar{\mathbb{P}}^{r_t}_t = \emptyset, \\ \bigcup_{P_i\in\hat{\mathbb{P}}^{r_t}_t} \bigcap_{P_j\in\bar{\mathbb{P}}^{r_t}_t} \{x \in \mathbb{R}^n \mid x'(P_i - P_j)x \le 0\} & \text{otherwise}, \end{cases} \tag{2.9}$$

*with sets* $\bar{\mathbb{P}}^{r_t}_t \subseteq \mathbb{S}^n_+$ *and* $\hat{\mathbb{P}}^{r_t}_t \subseteq \mathbb{S}^n_+$ *being recursively calculated, for* $t = T, \cdots, 0$ *and* $r_t \in F_t$*, by*

$$\hat{\mathbb{P}}^{r_t}_t \;:=\; \hat{\mathcal{R}}_t(\mathbb{P}^{r_t-1}_{t+1}), \tag{2.10}$$

$$\bar{\mathbb{P}}^{r_t}_t \;:=\; \bar{\mathcal{R}}_t(\mathbb{P}^{r_t}_{t+1}), \tag{2.11}$$

$$\mathbb{P}^{r_t}_t \;:=\; \mathcal{K}[\hat{\mathbb{P}}^{r_t}_t \bigcup \bar{\mathbb{P}}^{r_t}_t], \tag{2.12}$$

*where* $\mathbb{P}^0_T = \{Q_T\}$ *and* $\mathbb{P}^{r_t}_t = \emptyset$ *if* $r_t \notin F_t$*. Furthermore, when* $\delta(u_t) = 1$*, the corresponding optimal control is*

$$u^*_t = -(R_t + B_t'P^*B_t)^{-1}B_t'P^*A_tx_t, \tag{2.13}$$

*where*

$$P^* := \arg \min_{P\in\mathbb{P}^{r_t-1}_{t+1}} x_t'\hat{\mathcal{R}}_t(P)x_t.$$

*Proof.* The cost-to-go in (2.5) can be calculated by the following recursion,

$$J_t(x_t, r_t) = \min_{u_t}\{x_t' Q_t x_t + u_t' R_t u_t + J_{t+1}(A_t x_t + B_t u_t, r_t - \delta(u_t))\}, \quad (2.14)$$

where $J_T(x_T, 0) = Q_T$. We claim that the the cost-to-go is of the following form at stage $t$ for any $r_t \in F_t$,

$$J_t(x_t, r_t) = \min_{P \in \mathbb{P}_t^{r_t}} x_t' P x_t, \quad (2.15)$$

where $\mathbb{P}_t^{r_t}$ is defined in (2.12). We use the induction method in the following to prove such a claim and thus the theorem.

At stage $T-1$, $F_{T-1} = \{0, 1\}$. Thus,

$$\begin{aligned} J_{T-1}(x_{T-1}, 0) &= x_{T-1}' P_{T-1}^0 x_{T-1}, \quad P_{T-1}^0 = \bar{\mathcal{R}}_{T-1}(Q_T), \\ J_{T-1}(x_{T-1}, 1) &= x_{T-1}' P_{T-1}^1 x_{T-1}, \quad P_{T-1}^1 = \hat{\mathcal{R}}_{T-1}(Q_T). \end{aligned}$$

Let $\hat{\mathbb{P}}_{T-1}^1 := \{P_{T-1}^1\}$, $\bar{\mathbb{P}}_{T-1}^1 := \emptyset$, $\mathbb{P}_{T-1}^1 := \{P_{T-1}^1\}$, $\hat{\mathbb{P}}_{T-1}^0 := \emptyset$, $\bar{\mathbb{P}}_{T-1}^0 := \{P_{T-1}^0\}$, $\mathbb{P}_{T-1}^0 := \{P_{T-1}^0\}$. Thus, $\mathbb{M}_{T-1}^1 = \mathbb{R}^n$ and $\mathbb{M}_{T-1}^0 = \emptyset$. More specifically, for $r_{T-1} = 1$, the region of $x_{T-1}$ in which the control should be implemented is $\mathbb{R}^n$ and the correspondent optimal control is $u_{T-1}^* = -(R_{T-1} + B_{T-1}' Q_T B_{T-1})^{-1} B_{T-1}' Q_T A_{T-1} x_{T-1}$. Thus, the theorem is proved to be true at stage $T-1$.

We assume now that the theorem holds at stage $k+1$ and the cost-to-go takes the following form,

$$J_{k+1}(x_{k+1}, j) = \min_{P \in \mathbb{P}_{k+1}^j} x_{k+1}' P x_{k+1},$$

for given state $x_{k+1}$ and $j \in F_{k+1}$. For convenience, let $J_{k+1}(x_{k+1}, j) = +\infty$ when $j \notin F_{k+1}$.

At stage $k$, the cost-to-go in (2.14) becomes,

$$J(x_k, j) = \min\{x_k' Q_k x_k + J(x_{k+1}, j), \ \min_{u_k}[x_k' Q_k x_k + u_k' R_k u_k + J(x_{k+1}, j-1)]\}, \quad (2.16)$$

where the first and second terms are corresponding to the cases of $\delta(u_k) = 0$ and $\delta(u_k) = 1$, respectively.

There are three different cases for different $j$. If $j = 0$, then $j - 1 \notin F_k$ and relationship (2.16) yields

$$\begin{aligned} J_k(x_k, 0) &= \min\{x_k' Q_k x_k + J_{k+1}(x_{k+1}, 0), +\infty\} \\ &= x_k' Q_k x_k + \min_{P \in \mathbb{P}_{k+1}^j} (A_k x_k)' P (A_k x_k) \\ &= \min_{P \in \mathbb{P}_k^0} x_k' P x_k, \end{aligned}$$

where $\mathbb{P}_k^0 = \bar{\mathcal{R}}_k(\mathbb{P}_{k+1}^0)$. It is clear that, when $j = 0$, no control opportunity exists for implementation and the set $\mathbb{P}_k^0$ contains one element for all $k$.

If $j = T - k$, then $j \notin F_{k+1}$ and control action must be implemented at every remaining stage, as the resulting control police will not be optimal otherwise. The cost-to-go in this situation is

$$\begin{aligned} J_k(x_k, T-k) &= \min\{+\infty, \min_{u_k}[x_k' Q_k x_k + u_k' R_k u_k + J_{k+1}(x_{k+1}, T-k-1)]\} \\ &= \min_{P \in \mathbb{P}_{k+1}^{T-k-1}} \Big\{ \min_{u_k} \big[ x_k' Q_k x_k + u_k' R_k u_k \\ &\qquad + (A_k x_k + B_k u_k)' P (A_k x_k + B_k u_k) \big] \Big\} \\ &= \min_{P \in \mathbb{P}_k^{T-k}} \{x_k' P x_k\}, \end{aligned}$$

where $\mathbb{P}_k^{T-k} = \hat{\mathcal{R}}_k(\mathbb{P}_{k+1}^{T-k-1})$ and the optimal control at stage $k$ is $u_k^* = -(R_k + B_k' P^* B_k)^{-1} B' P^* A_k x_k$, with $P^* = \arg\min_{P \in \mathbb{P}_{k+1}^{T-k-1}} x_t' \hat{\mathcal{R}}_t(P) x_t$. We should notice that when $j = T - k$, the set $\mathbb{P}_{k+1}^{j-1}$ only has a unique element.

If $0 < j < T - k$, both $j \in F_{k+1}$ and $j - 1 \in F_{k+1}$ hold and the cost-to-go in (2.16) becomes

$$J_k(x_k, j) = \min\{\bar{J}_k(x_k, j), \hat{J}_k(x_k, j)\}, \tag{2.17}$$

where

$$\bar{J}_k(x_k, j) := x_k' Q_k x_k + \min_{P \in \mathbb{P}_{k+1}^j} [(A_k x_k)' P (A_k x_k)], \tag{2.18}$$

$$\hat{J}_k(x_k, j) := \min_{u_k} [x_k' Q_k x_k + u_k' R_k u_k + \min_{P \in \mathbb{P}_{k+1}^{j-1}} [(A_k x_k + B_k u_k)' P (A_k x_k + B_k u_k)]]. \tag{2.19}$$

Note that relations in (2.18) and (2.19) are corresponding to the cases of $\delta(u_k) = 0$ and $\delta(u_k) = 1$, respectively. Note that (2.19) can be rewritten as following,

$$\begin{aligned} \hat{J}_k(x_k, j) &= \min_{P \in \mathbb{P}_{k+1}^{j-1}} \min_{u_t} [x_k' Q_k x_k + u_k' R_k u_k + (A_k x_k + B_k u_k)' P (A_k x_k + B_k u_k)] \\ &= \min_{P \in \mathbb{P}_{k+1}^{j-1}} x_k' \hat{\mathcal{R}}_k(P) x_k. \end{aligned}$$

For each $P \in \mathbb{P}_{k+1}^{j-1}$, the optimal control is

$$u_k^* = -(R_k + B_k' P B_k)^{-1} B_k' P A_k x_k. \tag{2.20}$$

It is clear that (2.18) and (2.19) can be rewritten as,

$$\bar{J}_k(x_k, j) = \min_{P \in \bar{\mathbb{P}}_k^j} x_k' P x_k, \quad \hat{J}_k(x_k, j) = \min_{P \in \hat{\mathbb{P}}_k^j} x_k' P x_k, \tag{2.21}$$

where $\bar{\mathbb{P}}_k^j = \bar{\mathcal{R}}_k(\mathbb{P}_{k+1}^j)$ and $\hat{\mathbb{P}}_k^j = \hat{\mathcal{R}}_k(\mathbb{P}_{k+1}^{j-1})$. Thus, we should implement control when $\hat{J}_k(x_k, j) \leq \bar{J}_k(x_k, j)$, which is $x_k$ dependent. Furthermore, the following set $\mathbb{M}_k^j$ specifies the region of the state $x_k$ in which relation $\hat{J}_k(x_k, j) \leq \bar{J}_k(x_k, j)$ holds,

$$\begin{aligned} \mathbb{M}_k^j &= \{x_k \in \mathbb{R}^n \mid \hat{J}_k(x_k, j) \leq \bar{J}_k(x_k, j), \} \\ &= \{x_k \in \mathbb{R}^n \mid \min_{P \in \hat{\mathbb{P}}_k^j} x_k' P x_k \leq \min_{H \in \bar{\mathbb{P}}_k^j} x_k' H x_k\} \\ &= \bigcup_{\forall P \in \hat{\mathbb{P}}_k^j} \bigcap_{\forall H \in \bar{\mathbb{P}}_k^j} \{x \in \mathbb{R}^n \mid x'(P - H)x \leq 0\}. \end{aligned}$$

Thus, we have the following result from (2.17), (2.18), (2.19), (2.20), (2.21) and Lemma 3.2,

$$J_k(x_k, j) = \min_{P \in \mathbb{P}_k^j} x_k' P x_k, \quad \mathbb{P}_k^j := \mathcal{K}[\bar{\mathbb{P}}_k^j \bigcup \hat{\mathbb{P}}_k^j].$$

The proof of the theorem is completed. □

It is interesting to note that the action region $\mathbb{M}_t^{r_t}$ defined in (2.9) is always a union of some cones, as for any $z \in \{x \in \mathbb{R}^n | x'(P_i - P_j)x \leq 0\}$, where $P_j, P_i \in \mathbb{S}_n^+$, $tz \in \{x \in \mathbb{R}^n | x'(P_i - P_j)x \leq 0\}$ holds for any $t \in \mathbb{R}$.

The optimal control law developed in Theorem 2.3 is of a feedback nature. After characterizing the action region $\mathbb{M}_t^{r_t}$ by a recursion of a Riccati type, whether the control should be implemented or not can be readily determined. Such a seemingly elegant approach is in general inefficient as the time complexity of the worst case is $\mathcal{O}(2^T s \max\{m^3, n^3\})$ (without considering the operation of $\mathcal{K}(\cdot)$), as $\mathcal{O}(\max\{m^3, n^3\})$ time is needed to compute Ricatti iteration in each step. Such an exponential growth with respect to $T$ prevents a direct implementation of Theorem 2.3 in large-scale applications. Such an algorithm becomes, however, an efficient polynomial-time algorithm for a special class of CCLQR problems with $n = 1$.

**Corollary 2.4.** *If the state space is of dimension one, i.e., $n = 1$, then the algorithm specified in Theorem 2.3 is a polynomial-time algorithm with a complexity of $\mathcal{O}(Tsm^3)$.*

Proof. When $x_t \in \mathbb{R}$, the action region defined in (2.9) is independent of state $x_t$. More specifically, the region $\mathbb{M}_t^{r_t}$ is either $\mathbb{R}$ or $\emptyset$ in such a situation. □

As mentioned before, the algorithm provided in Theorem 2.3 becomes inefficient when the size of the problem increases. We examine CCLQR in the next section from another angel. The analytical solvability of sCCLQR using dynamic programming motivates us to investigate a solution scheme that uses sCCLQR to approximate general CCLQR problems in order to identify exact solutions within a branch-and-bound framework.

## 2.3. Mathematical Programming Based Solution Scheme

### 2.3.1. Reformulation

Reformulating the LQR problem as a quadratic optimization problem has been a powerful solution technique, especially from the computational point of view [60] [71]. We adopt such a scheme to suit our purpose. It is well known that the solution of the linear system (2.1) is,

$$x_t = \Phi(t,0)x_0 + \sum_{\tau=0}^{t-1} \Phi(t,\tau+1)B_\tau u_\tau,$$

where the state transition matrix $\Phi(\cdot,\cdot) : (\mathbb{N}\times\mathbb{N})_+ \to \mathbb{R}^{n\times n}$ is given as follows[2],

$$\Phi(t,t_0) = \begin{cases} A_{t-1}A_{t-2}\cdots A_{t_0} & \text{if } \ t > t_0, \\ \mathbf{I}_n & \text{if } \ t = t_0. \end{cases} \tag{2.22}$$

We introduce the following notations for problem $(F^s_{n,m,T})$,

$$\mathcal{B}_T := \operatorname{diag}\begin{pmatrix} B_0 & B_1 & \cdots & B_{T-1} \end{pmatrix}, \tag{2.23}$$

$$\mathcal{R}_T := \operatorname{diag}\begin{pmatrix} R_0 & R_1 & \cdots & R_{T-1} \end{pmatrix}, \tag{2.24}$$

$$\mathcal{Q}_T := \operatorname{diag}\begin{pmatrix} Q_1 & Q_2 & \cdots & Q_T \end{pmatrix}, \tag{2.25}$$

$$\mathcal{F}_T : = \begin{pmatrix} \Phi(1,1) & 0 & \cdots & 0 \\ \Phi(2,1) & \Phi(2,2) & \cdots & 0 \\ \Phi(3,1) & \Phi(3,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \Phi(T,1) & \Phi(T,2) & \cdots & \Phi(T,T) \end{pmatrix}, \tag{2.26}$$

[2] $\mathbb{N}\times\mathbb{N}_+$ denotes $\{(k,k_0):\ k,k_0\in\mathbb{N},\ k\geq k_0\}$

$$\mathcal{L}_T := \begin{pmatrix} \Phi(1,1) \\ \Phi(2,1) \\ \Phi(3,1) \\ \vdots \\ \Phi(T,1) \end{pmatrix}, \; x := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix}, \; u := \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{T-1} \end{pmatrix}, \tag{2.27}$$

where $\mathcal{B}_T \in \mathbb{R}^{nT\times mT}$, $\mathcal{R}_T \in \mathbb{S}^{mT}_{++}$, $\mathcal{Q}_T \in \mathbb{S}^{nT}_{+}$, $\mathcal{F}_T \in \mathbb{R}^{nT\times nT}$, and $\mathcal{L}_T \in \mathbb{R}^{nT\times 1}$. It is clear that the state vector $x$ is a linear function of the control vector,

$$x = \mathcal{L}_T A_0 x_0 + \mathcal{F}_T \mathcal{B}_T u. \tag{2.28}$$

Problem $(F^s_{n,m,T})$ can be then reformulated as the following equivalent cardinality constrained quadratic optimization problem (CCQO) (with a constant difference in the objective function from (CCLQR)),

$$(G^s_{n,m,T}) \qquad \min \; f(u) = \frac{1}{2}u'Du + d'u, \tag{2.29}$$

$$\text{Subject to :} \sum_{t=0}^{T-1} \delta(u_t) \leq s, \tag{2.30}$$

where

$$D \;=\; 2\mathcal{B}'_T\mathcal{F}'_T\mathcal{Q}_T\mathcal{F}_T\mathcal{B}_T + 2\mathcal{R}_T, \tag{2.31}$$

$$d \;=\; 2\mathcal{B}'_T\mathcal{F}'_T\mathcal{Q}_T\mathcal{L}_T A_0 x_0. \tag{2.32}$$

Notice that $D \in \mathbb{S}^{mT}_{++}$ due to the assumptions of $\mathcal{R}_T \in \mathbb{S}^{mT}_{++}$ and $\mathcal{Q}_T \in \mathbb{S}^{mT}_{+}$.

Note that the problem $(G^s_{n,m,T})$ is readily solved by some optimization software, e.g., CPLEX solver, after introducing some artificial binary variables. Our previous numerical tests showed, however, that the computational power of such a state-of-the-art numerical solver is still limited. In this research, we exploit the structure properties hidden behind problem $(G^s_{n,m,T})$, especially these properties demonstrated in (2.31) and (2.32) for matrix $D$ and vector $d$. The main goal of this investigation is to derive a tight lower bound that plays a key role in exact solution methods of a branch and bound type. More specifically, we focus in the following how to construct a lower bound of problem $(G^s_{n,m,T})$ by identifying the best corresponding sCCLQR problem $(G^s_{1,m,T})$ under certain criterion.

### 2.3.2. Lower Bounding via sCCLQ

**Assumption 2.1.** Neither any of matrices $A_t$, $t = 0, \cdots, T-1$, nor matrix $Q_T$ is a zero matrix.

The above assumption excludes degenerate cases of problem $(F^s_{n,m,T})$ from our further consideration. If $A_\tau$ is a zero matrix for some $0 \leq \tau < T-1$, then minimizing the objective function forces $x_t$ and $u_t$ to be zero vectors for all $t > \tau$. Problem $(F^s_{n,m,T})$ reduces then to $(F^s_{n,m,\tau})$. Similarly, if $Q_T$ is a zero matrix, no penalty will be exercised on $x_T$, which leads to zero $u_{T-1}$, thus reducing problem $(F^s_{n,m,T})$ to $(F^s_{n,m,T-1})$.

As we have already revealed in Section 2.2, any sCCLQR problem can be solved analytically by using dynamic programming. We utilize this advantage to approximate the solution of CCLQR.

Denote the sets of coefficients of problems $(F^s_{n,m,T})$ and $(G^s_{n,m,T})$ as follows, respectively,

$$\begin{aligned}\Sigma(F^s_{n,m,T}) &:= \big\{A_t|_{t=0}^{T-1}, B_t|_{t=0}^{T-1}, Q_t|_{t=1}^{T}, R_t|_{t=0}^{T-1}, x_0\big\},\\ \Sigma(G^s_{n,m,T}) &:= \big\{D, d\big\}.\end{aligned}$$

Define further by $\mathbb{F}_{n,m,T} := \{\Sigma(F^s_{n,m,T})\}$, the union of all sets of coefficients of the CCLQR problem with dimensions of the state space, the control space and the time horizon being $n$, $m$ and $T$, respectively. The relationship given in (2.31) and (2.32) actually defines a mapping $\mathcal{T}(\cdot)$,

$$\mathcal{T}(\cdot): \quad \mathbb{F}_{n,m,T} \;\rightarrow\; \big\{D \in \mathbb{S}^{mT}_{++},\ d \in \mathbb{R}^{mT}\big\}, \tag{2.33}$$

and in particular, $\mathcal{T}(\Sigma(F^s_{n,m,T})) = \Sigma(G^s_{n,m,T})$. By adopting such a notation, the set $\mathcal{T}(\mathbb{F}_{1,m,T})$ includes all the sets of coefficients, $\{D, d\}$, resulted from sCCLQR problems.

The way of constructing a computable lower bound is to approximate the problem $(G^s_{n,m,T})$ by an sCCLQR problem $(\hat{G}^s_{1,m,T})$, i.e., to identify function $\hat{f}(u) = \frac{1}{2}u'Hu+d'u$, where $\{H, d\} \in \mathcal{T}(\mathbb{F}_{1,m,T})$, to best approximate the objective function $f(u) = \frac{1}{2}u'Du + d'u$ of problem $(G^s_{n,m,T})$ in a bounded area. Condition $D \succeq H$ is required to ensure $\hat{f}(u) \leq f(u)$ for all feasible $u$. The best $\hat{f}(u)$ is identified such that the following norm measure between $f(u)$ and $\hat{f}(u)$ is minimized,

$$\pi = \int_{\|u-l\|^2\leq\delta} 4|f(u) - \hat{f}(u)|^2 du,$$

where $l = -D^{-1}d$ is the center of the objective contour of $f(u)$ and $\delta$ is a given constant that controls the size of an $l$-centered ball as the region of comparison. Replacing $u$ by $y + l$ and noticing that $y'(D - H)y \leq \|D - H\|_F\|y\|^2$ and $(d + Hl)'y \leq \|d + Hl\| \cdot \|y\|$, we can get the following upper bound for $\pi$,

$$\begin{aligned} \pi &= \int_{\|y\|^2\leq\delta} [y'(D - H)y + 2(-d - Hl)'y \\ &+ (-d - Hl)'l]^2 dy \\ &\leq \int_{\|y\|^2\leq\delta} 3\{\|D - H\|_F^2\|y\|^4 + 4\|d + Hl\|^2 \cdot \|y\|^2 + \|d + Hl\|^2 \cdot \|l\|^2\}dy \\ &\leq 3\delta^2\|D - H\|_F^2 \int_{\|y\|^2\leq\delta} dy + 3(4\delta + \|l\|^2)\|d + Hl\|^2 \int_{\|y\|^2\leq\delta} dy \\ &= C_1(\delta)\|D - H\|_F^2 + C_2(\delta, l)\|d + Hl\|^2, \end{aligned} \tag{2.34}$$

where notations $\|\cdot\|$ and $\|\cdot\|_F$ are $l_2$ norms for vectors and matrices, respectively, $C_1(\delta)$ and $C_2(\delta, l)$ depend only on $\delta$ and $l$, and the arithmetic mean inequality is used to derive the first inequality above. Note that the norm $\|d + Hl\|^2$ serves as a penalty term if the center $\hat{f}(u)$ deviates from $l = -D^{-1}d$, the center of the objective contour of $f(u)$ of the primal problem. As it is difficult to minimize $\pi$ directly, we, instead, consider the following auxiliary problem $(A)$ to minimize the upper bound of $\pi$ defined in (2.34),

$$\begin{aligned} (A) \qquad & \min \ \mu\|Hl+d\|^2+\|H-D\|_F^2, \\ \text{Subject to:} \qquad & \{H,d\} \in \mathcal{T}(\mathbb{F}_{1,m,T}), && (2.35) \\ & D-H \succeq 0. && (2.36) \end{aligned}$$

Let the optimal solution of problem $(A)$ be $H^*$. Then, the optimal value of the following problem,

$$\begin{aligned} (\hat{G}^s_{1,m,T}) \qquad & \min \ \frac{1}{2}u'H^*u+d'u, \\ & \text{Subject to} : \sum_{t=0}^{T-1}\delta(u_t) \leq s, \end{aligned}$$

provides an approximation of $v(G^s_{n,m,T})$. The approximated problem $(\hat{G}^s_{1,m,T})$ can be efficiently solved by a two-step procedure: Finding an equivalent sCCLQR problem $(\hat{F}^s_{1,m,T})$ and solving it by dynamic programming.

We focus now on how to solve problem $(A)$ first. Let matrix $E_t \in \mathbb{R}^{T\times T}$ be defined for $t=0,\cdots,T-1$ with its elements $\{E_{i,j}\}_t$ being given as follows,

$$\{E_{i,j}\}_t = \begin{cases} 1 & \text{if } i=t, \ j=t,\cdots,T-1, \\ 1 & \text{if } j=t, \ i=t+1,\cdots,T-1, \\ 0 & \text{Otherwise.} \end{cases} \qquad (2.37)$$

For example, when $T=3$, $E_0$, $E_1$ and $E_2$ are given as

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

respectively. We further denote $\mathbf{1}_m$ as the $m \times m$ matrix with all its elements being 1.

**Theorem 2.5.** *For matrix $H \in \mathbb{S}^{mT}_{++}$ and vector $h \in \mathbb{R}^{mT}$, $\{H,h\} \in \mathcal{T}(\mathbb{F}_{1,m,T})$ if and only if $H$ can be expressed as*

$$H = \sum_{t=0}^{T-1}\beta_t M_t + diag(S_0,\cdots,S_{T-1}), \qquad (2.38)$$

*where* $M_t := [(E_t \otimes \boldsymbol{1}_m) \circ (hh')]$ *with* $E_t$ *being defined in (2.37),* $S_t \in \mathbb{S}^m_{++}$*, and* $\beta_t$ *being scalars satisfying* $0 < \beta_0 \leq \beta_1 \leq \beta_2 \cdots \leq \beta_{T-1}$.[3]

*Proof.* In the following proof, the vector $h$ consists of $T$ block vectors with each block in dimension $\mathbb{R}^m$, i.e., $h' = (\; h'_0, \;\; \cdots, \;\; h'_{T-1})$ with $h_t \in \mathbb{R}^m$ for $t = 0, \cdots, T-1$.

We prove the "if" part first. Given $H \in \mathbb{S}^{mT}_{++}$ and $h \in \mathbb{R}^{mT}$ and assume that $H$ can be expressed as the form in (2.38). Our target is to prove $\{H, h\} \in \mathcal{T}(\mathbb{L}_{1,m,T})$. Consider a CCLQR problem $(\hat{F}^s_{1,m,T})$ with parameters

$$\Sigma(\hat{F}^s_{1,m,T}) = \{\hat{A}_t|_{t=0}^{T-1}, \hat{B}_t|_{t=0}^{T-1}, \hat{Q}_t|_{t=1}^{T}, \hat{R}_t|_{t=0}^{T-1}, \hat{x}_0\}.$$

Let $\hat{B}_t = \beta_t h_t$, $\hat{A}_t = 1$, $\hat{R}_t = S_t$ for $t = 0, \cdots, T-1$ and $\hat{x}_0 = \frac{1}{2}$. Since $\beta_0 > 0$ and $\beta_t \geq \beta_{t-1}$, for $t = 1, \cdots, T-1$, we can further construct $\hat{Q}_t$ as follows,

$$\hat{Q}_t = \begin{cases} \frac{1}{\beta_{t-1}} & \text{if } \; t = T, \\ \frac{1}{\beta_{t-1}} - \frac{1}{\beta_t} & \text{if } \; t = 1, \cdots, T-1. \end{cases}$$

It is then straightforward to verify $\mathcal{T}(\Sigma(\hat{F}^s_{1,m,T})) = \{H, h\}$ by using (2.31) and (2.32).

We next prove the "only if" part. Given any $(F^s_{1,m,T})$ problem. From the definition of the mapping $\mathcal{T}$, we have $\mathcal{T}(\Sigma(F^s_{1,m,T})) = \{H, h\}$, where vector $h$ is given as

$$\begin{aligned} h' &= 2x'_0 A'_0 \mathcal{L}'_T \mathcal{Q}_T \mathcal{F}_T \mathcal{B}_T \\ &= \mu \begin{pmatrix} \alpha_0 B_0 & \alpha_1 B_1 & \cdots & \alpha_{T-1} B_{T-1} \end{pmatrix}, \end{aligned} \tag{2.39}$$

with $(\alpha_0, \alpha_1, \cdots, \alpha_{T-1}) = \mathcal{L}'_T \mathcal{Q}_T \mathcal{F}_T \in \mathbb{R}^{1 \times (T-1)}$ and $\mu = 2x_0 A_0$. Under Assumption 2.1, it can be verified that all $\alpha_t$'s are nonzero scalars. From (2.39), we have

$$B_t = \frac{h'_t}{\mu \alpha_t}, \;\; \text{for } t = 0, \cdots, T-1. \tag{2.40}$$

[3] Notation "$\otimes$" is the usual Kronecker product and "$\circ$" is the schur product.

As $\mathcal{L}'_T$ is the first row of $\mathcal{F}'_T$, matrix $H$ can expressed as follows from (2.31) and (2.22),

$$\begin{aligned} H \;=\; & \begin{pmatrix} \alpha_0 B'_0 B_0 & \alpha_1 B'_0 B_1 & \cdots & \alpha_{k-1} B'_0 B_{T-1} \\ \alpha_1 B'_1 B_0 & \frac{\alpha_1}{A_1} B'_1 B_1 & \cdots & \frac{\alpha_{T-1}}{A_1} B'_1 B_{T-1} \\ \alpha_2 B'_2 B_0 & \frac{\alpha_2}{A_1} B'_2 B_1 & \cdots & \frac{\alpha_{T-1}}{A_1 A_2} B'_2 B_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{T-1} B'_{T-1} B_0 & \frac{\alpha_{T-1}}{A_1} B'_{T-1} B_1 & \cdots & \frac{\alpha_{T-1}}{\prod_{i=1}^{T} A_i} B'_{T-1} B_{T-1} \end{pmatrix} \\ & + \; \mathrm{diag}\begin{pmatrix} R_0 & R_1 & R_2 & \cdots & R_{T-1} \end{pmatrix}. \end{aligned}$$

Denote

$$\beta_0 = \frac{1}{\mu^2 \alpha_0}, \quad \beta_t = \frac{1}{\mu^2 \alpha_t \prod_{i=1}^{t} A_i}, \quad \text{for } t = 1, \cdots, T-1. \tag{2.41}$$

Replacing all $B_t$ in the expression of $H$ by using (2.40) yields

$$\begin{aligned} H \;=\; & \begin{pmatrix} \beta_0 h_0 h'_0 & \beta_0 h_0 h'_1 & \cdots & \beta_0 h_0 h'_{T-1} \\ \beta_0 h_1 h'_0 & \beta_1 h_1 h'_1 & \cdots & \beta_1 h_1 h'_{T-1} \\ \beta_0 h_2 h'_0 & \beta_1 h_2 h'_1 & \cdots & \beta_2 h_2 h'_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0 h_{k-1} h'_0 & \beta_1 h_{T-1} h'_1 & \cdots & \beta_{T-1} h_{T-1} h'_{T-1} \end{pmatrix} \\ & + \; \mathrm{diag}\begin{pmatrix} R_0 & R_1 & \cdots & R_{T-1} \end{pmatrix} \\ =\; & \sum_{t=0}^{T-1} \beta_t [(E_t \otimes \mathbf{1}_m) \circ h h'] + \mathrm{diag}\begin{pmatrix} R_0 & R_1 & \cdots & R_{T-1} \end{pmatrix}. \end{aligned}$$

The following is evident from (2.39),

$$\alpha_t \prod_{i=1}^{t} A_i = \sum_{i=t}^{T} \prod_{j=1}^{i} A_j^2 Q_{i+1}, \quad \text{for} \quad t = 0, \cdots, T-1,$$

where $\prod_{j=1}^{i} A_j = 1$ if $i < j$. From Assumption 2.1 and the definition of $\beta_t$, we have $\beta_0 > 0$ and

$$\beta_0 \le \beta_1 \le \cdots \le \beta_{T-1}.$$

This completes the proof. □

Note that under the assumption that $0 < \beta_0 \leq \beta_1 \leq \ldots \leq \beta_{T-1}$, the first term in $H$, $\sum_{t=0}^{T-1} \beta_t M_t$ is positive semidefinite. The above theorem actually characterizes set $\mathcal{T}(\Sigma(F^s_{1,m,T}))$. More specifically, we can use the above theorem to simplify the formulation in problem $(A)$. We introduce the following decision vectors,

$$\hat{y} = \begin{pmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_{T-1} \end{pmatrix}, \quad \bar{y} = \begin{pmatrix} \bar{y}_0 \\ \bar{y}_1 \\ \vdots \\ \bar{y}_N \end{pmatrix}, \quad Y = \begin{pmatrix} \hat{y} \\ \bar{y} \end{pmatrix},$$

where $N := \frac{1}{2}(m+1)mT$. Matrix $H$ in problem $(A)$ can be now expressed as

$$\begin{aligned} H &= \sum_{t=0}^{T-1} \hat{y}_t \hat{M}_t + \sum_{i=0}^{N} \bar{y}_i \bar{M}_i, \\ 0 &< \hat{y}_0, \quad \hat{y}_t \leq \hat{y}_{t+1}, \quad t = 0, \cdots, T-2, \end{aligned}$$

where constant matrix $\hat{M}_t := [(E_t \otimes \mathbf{1}_m) \circ (dd')]$ and constant matrix $\bar{M}_i \in \mathbb{R}^{mT}$ can be understood from the expression of $\operatorname{diag}(S_0, S_1, \cdots, S_{T-1})$. The objective function of problem $(A)$ is actually a quadratic function with respect to $Y$, $(LY - d)'(LY-d) + c'Y$. Then problem $(A)$ can be written as the following semidefinite programming problem,

$$\begin{aligned} (A) \qquad & \min \quad \lambda, \\ \text{Subject to:} \qquad & \sum_{i=0}^{N} \bar{y}_i \bar{M}_i \succeq \varepsilon \mathbf{I}_{mT}, \\ & D - \sum_{t=0}^{T-1} \hat{y}_t \hat{M}_t - \sum_{i=0}^{N} \bar{y}_i \bar{M}_i \succeq 0, \\ & \begin{pmatrix} \mathbf{I}_N & LY - d \\ (LY-d)' & -c'Y + \lambda \end{pmatrix} \succeq 0, \\ & \hat{y}_0 > \varepsilon, \quad \hat{y}_t \leq \hat{y}_{t+1}, \quad t = 0, \cdots, T-1, \end{aligned}$$

where $\varepsilon$ is a given small positive number. After solving problem $(A)$, problem

$(\hat{G}^s_{n,m,T})$ is constructed and the equivalent $sCCLQ$ problem $(\hat{F}^s_{1,m,T})$ can be constructed by using the way specified in the if part of the proof for Theorem 2.5.

**Remark 2.1.** Note that the resulted optimal control of problem $(\hat{F}^s_{1,m,T})$ provides a good feasible control of the original problem $(F^s_{n,m,T})$. In particular, let $\{\hat{u}_t\}|_{t=0}^{T-1}$ and $\{\bar{u}_t\}|_{t=0}^{T-1}$ be the optimal control of problem $(\hat{F}^s_{1,m,T})$ and a heuristic feasible control of problem $(F^s_{1,m,T})$, respectively, then the sparsity of $\bar{u}_t$ is set according to $\hat{u}_t$, i.e., set $\delta(\bar{u}_t) = 1$ if $\delta(\hat{u}_t) = 1$ and $\delta(\bar{u}_t) = 0$, otherwise, for $t = 0, \cdots, T-1$. After fixing the sparsity of $\{\bar{u}_t\}|_{t=0}^{T-1}$, the magnitude of $\{\bar{u}_t\}_{t=0}^{T-1}$, where $\bar{u}_t \neq 0$, can be calculated as in the traditional LQR problem.

Integrating the above lower bounding scheme into a branch and bound algorithm gives rise to an efficient solution for cardinality-constrained LQR problems and thus LQR problems with set up cost (Problem Type 2). We now demonstrate our solution procedure by the following example problems.

## 2.4. Illustrative Examples

**Example 2.1.** Consider a CCLQR problem $(F^2_{2,1,4})$ with $n = 2$, $m = 1$, $T = 4$, $s = 2$, $A_t$ being identity matrix for $t = 0, 1, 2, 3$ and

$$B_0 = \begin{pmatrix} 2.42 \\ -0.83 \end{pmatrix},\ B_1 = \begin{pmatrix} 0.12 \\ 0.18 \end{pmatrix},\ B_2 = \begin{pmatrix} 0.63 \\ 2.49 \end{pmatrix},\ B_3 = \begin{pmatrix} -0.19 \\ -1.37 \end{pmatrix},$$

$$R_0 = 7.15,\ R_1 = 4.13,\ R_2 = 1.55,\ R_3 = 0.87,$$

$$Q_1 = \begin{pmatrix} 0.39 & 0.84 \\ 0.84 & 2.60 \end{pmatrix}, Q_2 = \begin{pmatrix} 0.2564 & 0.1949 \\ 0.1949 & 1.6979 \end{pmatrix},$$

$$Q_3 = \begin{pmatrix} 1.76 & -0.06 \\ -0.06 & 1.97 \end{pmatrix},\ Q_4 = \begin{pmatrix} 0.40 & -0.20 \\ -0.20 & 0.20 \end{pmatrix}.$$

Sets $\hat{\mathbb{P}}_t^{r_t}$, $\bar{\mathbb{P}}_t^{r_t}$ and $\mathbb{P}_t^{r_t}$ can be calculated recursively by using (2.10), (2.11) and (2.12),

$$\mathbb{P}_3^1 = \left\{ \begin{pmatrix} 2.11 & -0.23 \\ -0.23 & 2.13 \end{pmatrix} \right\},\ \mathbb{P}_3^0 = \left\{ \begin{pmatrix} 2.15 & -0.27 \\ -0.27 & 2.18 \end{pmatrix} \right\}$$

$$\mathbb{P}_2^2 = \left\{ \begin{pmatrix} 2.33 & -0.30 \\ -0.30 & 2.04 \end{pmatrix} \right\},\ \mathbb{P}_2^0 = \left\{ \begin{pmatrix} 2.40 & -0.07 \\ -0.07 & 3.87 \end{pmatrix} \right\},$$

$$\hat{\mathbb{P}}_2^1 = \left\{ \begin{pmatrix} 2.37 & -0.31 \\ -0.31 & 2.04 \end{pmatrix} \right\},\ \bar{\mathbb{P}}_2^1 = \left\{ \begin{pmatrix} 2.37 & -0.03 \\ -0.03 & 3.83 \end{pmatrix} \right\},\ \mathbb{P}_2^1 = \hat{\mathbb{P}}_2^1 \bigcup \bar{\mathbb{P}}_2^1,$$

$$\hat{\mathbb{P}}_1^2 = \left\{ \begin{pmatrix} 2.75 & 0.51 \\ 0.51 & 4.62 \end{pmatrix}, \begin{pmatrix} 2.74 & 0.76 \\ 0.76 & 6.32 \end{pmatrix} \right\},\ \bar{\mathbb{P}}_1^2 = \left\{ \begin{pmatrix} 2.72 & 0.54 \\ 0.54 & 4.64 \end{pmatrix} \right\},$$

$$\mathbb{P}_1^2 = \hat{\mathbb{P}}_1^2 \bigcup \bar{\mathbb{P}}_1^2,$$

$$\hat{\mathbb{P}}_1^1 = \left\{ \begin{pmatrix} 2.78 & 0.72 \\ 0.72 & 6.37 \end{pmatrix} \right\},\ \bar{\mathbb{P}}_1^1 = \left\{ \begin{pmatrix} 2.77 & 0.52 \\ 0.52 & 4.64 \end{pmatrix}, \begin{pmatrix} 2.76 & 0.80 \\ 0.81 & 6.43 \end{pmatrix} \right\},$$

$$\mathbb{P}_1^1 = \hat{\mathbb{P}}_1^1 \bigcup \bar{\mathbb{P}}_1^1,$$

$$\hat{\mathbb{P}}_0^2 = \left\{ \begin{pmatrix} 1.39 & 1.17 \\ 1.17 & 4.65 \end{pmatrix}, \begin{pmatrix} 1.51 & 1.62 \\ 1.62 & 6.23 \end{pmatrix}, \begin{pmatrix} 1.50 & 1.57 \\ 1.57 & 6.14 \end{pmatrix} \right\},$$

$$\bar{\mathbb{P}}_0^2 = \left\{ \begin{pmatrix} 2.95 & 0.53 \\ 0.53 & 4.91 \end{pmatrix}, \begin{pmatrix} 2.98 & 0.50 \\ 0.50 & 4.89 \end{pmatrix}, \begin{pmatrix} 2.97 & 0.75 \\ 0.75 & 6.59 \end{pmatrix} \right\}.$$

The action region $M_t^{r_t}$ in the state space can be then calculated. Using the polar coordinate system, we can express the action regions for stage 0 with $r_0 = 2$, stage 1 with $r_1 = 1$, stage 1 with $r_1 = 2$ and stage 2 with $r_2 = 1$ as the shadow areas in Figure 2.1. More specifically, we have

$$\begin{aligned}
\mathbb{M}_0^2 =&[0.4191\pi,\ 1.3311\pi] \cup [-0.5809\pi,\ 0.3311\pi],\\
\mathbb{M}_1^2 =&[0.1330\pi,\ 0.5847\pi] \cup [0.9375\pi, 0.9798\pi]\\
&\cup [-0.8670\pi, -0.4153\pi] \cup [-0.0625\pi, -0.0202\pi],\\
\mathbb{M}_1^1 =&[0.9425\pi, 0.9866\pi] \cup [-0.0575\pi, -0.0134\pi],\\
\mathbb{M}_2^1 =&[0.0036\pi, 0.8999\pi] \cup [-0.9964\pi, -0.1001\pi],\\
\mathbb{M}_2^2 =&[0, 2\pi],\ \mathbb{M}_2^0 = \emptyset,\ \mathbb{M}_3^1 = [0, 2\pi],\ \mathbb{M}_3^0 = \emptyset.
\end{aligned}$$

If the initial state is $x_0 = (2, 2)'$, the optimal control sequence and the state trajectory can be derived as $u_0^*(x_0) = (-0.256, 0.105)x_0$, $x_1 = (1.266, 2.550)'$, $u_1^* = 0$, $x_2 = (1.266, 2.550)'$, $u_2^*(x_2) = (-0.045, -0.349)x_2$, $x_3 = (0.737, 0.153)'$, $u_3^* = 0$, and $x_4 = (0.737, 0.153)'$.

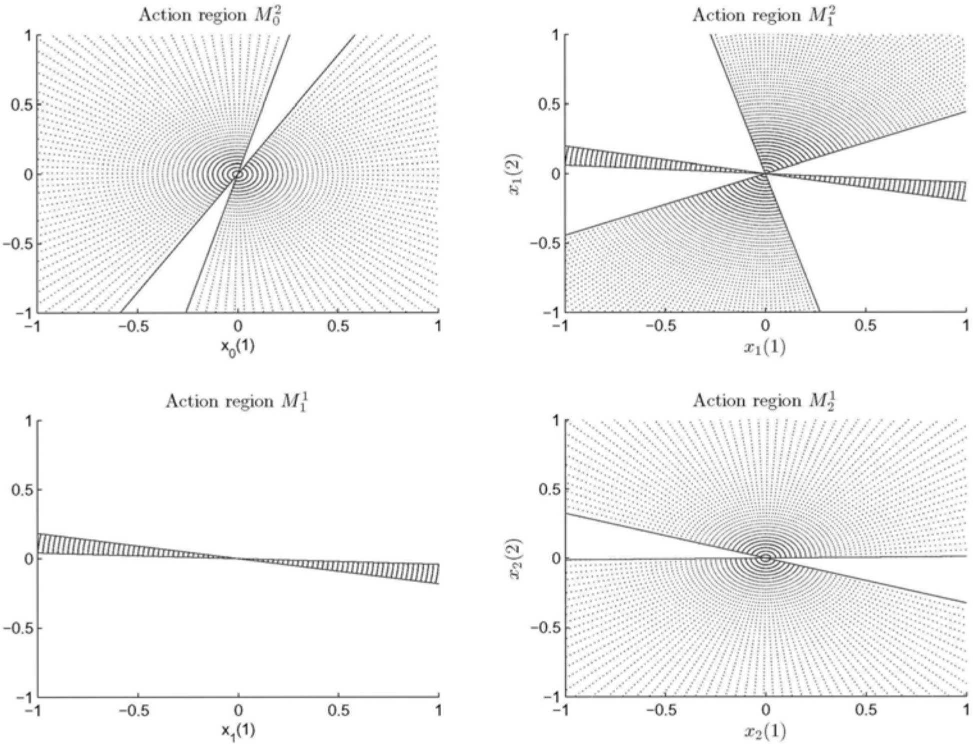

Figure 2.1: The action region of Example 2.1

**Example 2.2.** Consider the following LQR problem with $n = 3$, $m = 1$, $T = 6$ and set-up cost $w = 500$, which is imposed on non-zeros control (Problem Type 2). We let all matrices $A_t$'s be identity matrices in the systems dynamics. The other system parameters are given as follows,

$$B_0 = \begin{pmatrix} -0.145 \\ 0.026 \\ -0.108 \end{pmatrix},\ B_1 = \begin{pmatrix} -0.124 \\ -0.235 \\ 0.203 \end{pmatrix},\ B_2 = \begin{pmatrix} -0.148 \\ -0.082 \\ 0.144 \end{pmatrix},$$

$$B_3 = \begin{pmatrix} 0.297 \\ 0.080 \\ 0.265 \end{pmatrix},\ B_4 = \begin{pmatrix} 0.082 \\ 0.178 \\ -0.108 \end{pmatrix},\ B_5 = \begin{pmatrix} 0.483 \\ -0.809 \\ -0.185 \end{pmatrix},$$

$$Q_1 = \begin{pmatrix} 5.262 & 0.763 & -0.301 \\ 0.763 & 3.498 & -0.565 \\ -0.301 & -0.565 & 0.138 \end{pmatrix},\ Q_2 = \begin{pmatrix} 3.961 & -0.675 & 0.678 \\ -0.675 & 1.022 & -0.326 \\ 0.678 & -0.326 & 2.325 \end{pmatrix},$$

$$Q_3 = \begin{pmatrix} 2.264 & -0.527 & -0.019 \\ -0.527 & 2.063 & 0.876 \\ -0.019 & 0.876 & 4.945 \end{pmatrix},\ Q_4 = \begin{pmatrix} 2.548 & -0.032 & 0.207 \\ -0.032 & 2.472 & 0.005 \\ 0.207 & 0.005 & 3.089 \end{pmatrix},$$

$$Q_5 = \begin{pmatrix} 3.000 & -0.363 & -0.229 \\ -0.363 & 3.777 & -0.229 \\ -0.229 & -0.229 & 4.745 \end{pmatrix},\ Q_6 = \begin{pmatrix} 2.978 & -1.237 & 0.005 \\ -1.237 & 5.420 & -0.183 \\ 0.005 & -0.183 & 4.293 \end{pmatrix},$$

$$R_0 = 0.208,\ R_1 = 0.637,\ R_2 = 0.258,\ R_3 = 0.224,\ R_4 = 0.392,$$
$$R_5 = 0.318,\ \text{and } x_0' = \begin{pmatrix} -23.17 & -1.53 & 17.68 \end{pmatrix}'.$$

By using (2.33), we reformulate this problem to its equivalent CCQO form with coefficient matrices of

$$D = \begin{pmatrix} 1.7953 & -0.6871 & -0.1490 & -1.3635 & 0.2447 & -0.8253 \\ -0.6871 & 4.5336 & 1.6854 & 0.5347 & -1.2191 & 1.4597 \\ -0.1490 & 1.6854 & 1.7575 & 0.1492 & -0.6297 & -0.0956 \\ -1.3635 & 0.5347 & 0.1492 & 3.6098 & -0.1815 & 0.3154 \\ 0.2447 & -1.2191 & -0.6297 & -0.1815 & 1.6042 & -1.2227 \\ -0.8253 & 1.4597 & -0.0956 & 0.3154 & -1.2227 & 11.2360 \end{pmatrix},$$

and

$$d' = 10^3 \times \begin{pmatrix} 0.0595 \; 0.1925 \; 0.1540 \; 0.0001 \; -0.0531 \; -0.1206 \end{pmatrix}.$$

We identify next the sCCLQR problem which best approximates the given CCLQR problem. It is worth to mention that, to avoid numerical un-stability of the solver, we re-scale $D$ and $d$ by $10^{-2}$, which does not affect the optimal solution. We construct first the auxiliary problem $(A)$ for the re-scaled problem and then solve such a semidefinite programming problem by using SeDuMi 1.1 toolbox under Matlab 7.0 [73], e.g. We consider subproblem with $s = 2$ which yields the following solution of $(A)$,

$$\hat{y}_0 = 4.542 \times 10^{-2},\ \hat{y}_1 = 4.884 \times 10^{-2},\ \hat{y}_2 = 4.884 \times 10^{-2},\ \hat{y}_3 = 4.880 \times 10^{-2},$$
$$\hat{y}_4 = 7.306 \times 10^{-2},\ \hat{y}_5 = 7.306 \times 10^{-2}, \bar{y}_0 = 1.001 \times 10^{-4}, \bar{y}_1 = 2.707 \times 10^{-4},$$
$$\bar{y}_2 = 1.001 \times 10^{-4},\ \bar{y}_3 = 1.002 \times 10^{-4}, \bar{y}_4 = 2.823 \times 10^{\times} - 4,\ \bar{y}_5 = 1.000 \times 10^{-4}.$$

Matrix $H$ corresponding to the best sCCLQR is identified as

$$H = \begin{pmatrix} 0.2606 & 0.5197 & 0.4159 & 0.0004 & -0.1434 & -0.3258 \\ 0.5197 & 2.0801 & 1.4479 & 0.0014 & -0.4993 & -1.1342 \\ 0.4159 & 1.4479 & 1.2587 & 0.0011 & -0.3996 & -0.9076 \\ 0.0004 & 0.0014 & 0.0011 & 0.1000 & -0.0004 & -0.0009 \\ -0.1434 & -0.4993 & -0.3996 & -0.0004 & 0.4884 & 0.4682 \\ -0.3258 & -1.1342 & -0.9076 & -0.0009 & 0.4682 & 1.1635 \end{pmatrix}.$$

By implementing the branching and bound algorithm in solving problem $(F^s_{3,1,6})$, for $s = 1, \cdots, 6$, yields the results in Table 2.1, in which the optimal

Table 2.1: Solution of $F^s_{3,1,6}$

| s | Optimal index set | $v(F^s_{3,1,6})$ | $ws$ | $v(F^s_{3,1,6}) + ws$ |
|---|---|---|---|---|
| 1 | $\{2\}$ | 9737.1 | 500 | 10237.1 |
| 2 | $\{0, 2\}$ | 8262.2 | 1000 | 9262.2 |
| 3 | $\{0, 1, 2\}$ | 7448.9 | 1500 | 8948.9 |
| $4^*$ | $\{0, 1, 2, 5\}$ | 6858.5 | 2000 | 8858.5 |
| 5 | $\{0, 1, 2, 3, 5\}$ | 6557.0 | 2500 | 9057.0 |
| 6 | $\{0, 1, 2, 3, 4, 5\}$ | 6555.7 | 3000 | 9555.7 |

index set indicates the stages where a nonzero optimal control is implemented. Summing up $v(F^s_{3,1,6})$ and $ws$ and performing a comparison identifies the optimal cardinality for this example problem: $s^* = 4$ with the corresponding optimal control,$u_0 = -44.6,\ u_1 = -29.4,\ u_2 = -62.7,\ u_3 = 0,\ u_4 = 0,\ u_5 = 10.7.$

**Example 2.3.** We randomly generate 30 cases for each type of problems to check the quality of the lower bound generated by sCCLQ problems. The computational results are shown in Table 2.2. We ignore the constant $c$ in all the cases. The corresponding SDP problem $(A)$ is solved by calling Sedumi under Matlab. Since all problems are in small-scale, we are able to solve them by an enumeration method. In Table 2.2, the column "optV" denotes the optimal value, "sCCLQbound" denotes the lower bound constructed by an sCCLQ problem, and "tbound" denotes the trivial bound generated by ignoring the cardinality constraint. From Table 2.2, we can see that sCCLQ problem is tighter than the trivial bound for all these problems. However, when the size of the problem becomes larger, e.g., $T > 30$ and $m > 3$, the computation of the sCCLQ bound becomes expensive and unstable. On recognizing this difficulty, more sophisticated and stable bounding scheme will be discussed in Chapter 5.

Table 2.2: Comparison of bound

| $G^s_{n,m,T}$ | Optimal value | sCCLQBd | | tBound |
|---|---|---|---|---|
| {n, m, T, s} | OptV | value | CPUtime | vlaue |
| {4, 1, 8, 2} | -11337 | -13128 | 0.9 | -13583 |
| {8, 1,12, 4} | -32828 | -36428 | 2.5 | -37674 |
| {8, 3, 8, 2} | -33636 | -38806 | 12.0 | -39424 |
| {10, 2, 12,3} | -48264 | -56785 | 13.3 | -57173 |
| {6, 1, 15, 4} | -41564 | -44264 | 4.8 | -44310 |
| {6, 1, 20, 4} | -60055 | -63792 | 10.5 | -64921 |

## 2.5. Conclusion

Recognizing the need to incorporate set-up costs into the framework of optimal control, we present promising results of cardinality constrained linear-quadratic optimal regulator CCLQR problems. The feed-back control can be characterized via dynamic programming and the Riccati type of solution is computed in a recursive manner. However, such a procedure with exponential complexity suffers a computational difficulty while the size of problem is large. We thus choose to tackle the problem by solving an equivalent cardinality constrained quadratic optimization CCQO problem. Due to NP-hardness of such a problem, we propose a new low bounding scheme using the solution from a corresponding polynomially solvable sCCLQR problem. Combining this lower bounding scheme with a branch-and-bound algorithm, our computational results have demonstrated promising performance of the proposed method.

One interesting extension of the CCLQR problem is to considering the cardinality constraint on the change of the control, i.e., to consider the constraint $\sum_{t=1}^{T-1}\delta(u_{t-1}-u_t)\leq s$. Furthermore, since the time-invariant feedback controller is preferred in most applications, we may also consider the cardinality constraint on the change of the controller, i.e., $\sum_{t=1}^{T-1}\delta(K_{t-1}-K_t)\leq s$, where $K_t$ is the feedback gain $u_t^* = K_t x_t$.

# CHAPTER 3

# Optimal Control of Linear Switched System

## 3.1. Introduction

Linear switched systems arise naturally in many applications [6], [82] [74], such as in automotive systems, electrical circuit systems, manufacturing systems, and chemical systems. Following the categorization in [82][75], we focus in this chapter mainly on the subject termed externally forced optimal control problem of linear switched systems, which has been investigated extensively in the literature. A two-stage optimization strategy via parameterizing the switching instances is proposed in [82] for a continuous-time model of such a problem. The optimal control of a switched affine autonomous linear system is studied in [67] and an iterative algorithm is developed to find optimal switching sequence and switching instances iteratively, and the switching region is characterized by using dynamic programming (DP). A general continuous-time switching problem is investigated in [7] based on the maximum principle and an embedding method. As for discrete-time models, the computational complexity is a major issue. To cope with such a difficulty, an approximation approach is suggested in [51] [65] to obtain the value function in DP. Following the same line, a continuous-time switched homogeneous system is studied in [66]. In [50], the author considered

LQG optimal control of such a linear switched system and proposed a pruning method to reduce searching efforts. It is also worth mentioning that numerical schemes of discretizing both the state and time spaces are proposed in the literature, such as in [32], to approximate the value function.

In this chapter, different from all the previous works, we investigate the linear-quadratic optimal control problem for the discrete-time switched system with a constant switching cost. Such an addition of switching cost introduces jumps in the cost function, resulting in significant difficulties in obtaining analytical results. Using some sophisticated dynamic programming techniques, we characterize explicitly in our work the switching region, the value function and the feed-back gain of such a problem. Furthermore, some novel techniques using semidefinite programming are presented to reduce the computational burden for such a kind of problems.

## 3.2. Problem Formulation

Consider a switched system consisting of $K$ linear subsystems, where the $i$-th subsystem, $i = 1, \ldots, K$, is characterized by a quadruple $[A_i, B_i, Q_i, R_i]$ with $A_i \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times m}$, $Q_i \in \mathbb{R}^{n\times n}$ and $R_i \in \mathbb{R}^{m\times m}$, where $Q_i$ and $R_i$ are positive semidefinite and positive definite symmetric matrices, respectively. The dynamics of the $i$th subsystem is governed by

$$x(t+1) = A_i x(t) + B_i u(t),\ \ i = 1, \ldots, K. \tag{3.1}$$

Denote by

$$\mathcal{W} := \{\ [\ A_i, B_i, Q_i, R_i\ ],\ i \in \mathbb{K} := \{1, \cdots, K\}\} \tag{3.2}$$

the set of coefficient matrices of all subsystems. Let $y(t) \in \mathbb{K}$ be the active subsystem in interval $[t, t+1]$, $t = 0,\ 1,\ \cdots,\ T-1$, with $y(-1)$ being the initial active subsystem. Let $Y(T) := (y(0), y(1), \cdots, y(T-1))$ be the switching sequence of length $T$ by choosing specific $\left[A_{y(t)}, B_{y(t)}, Q_{y(t)}, R_{y(t)}\right]$ from $\mathcal{W}$ for all control instances, $t = 0,\ 1,\ \cdots,\ T-1$. Denote by $\|x\|_S^2$ the quadratic term $x'Sx$ with $S$ being symmetric and positive semidefinite. We consider now the following quadratic cost function in finite time horizon with a constant switching cost $M > 0$,

$$\sum_{t=0}^{T-1} \Big[\ \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + M \cdot \sigma(y(t-1), y(t))\ \Big] + \|x(T)\|^2_{Q(T)}, \qquad (3.3)$$

where $Q(T)$ is positive semidefinite, and the indicating function $\sigma : \mathbb{K} \times \mathbb{K} \to \{0, 1\}$ is defined such that $\sigma(a, b) = 1$ if $a \neq b$ and $\sigma(a, a) = 0$. The optimal control problem of the switched system studied in this chapter is stated now as follows.

**Problem 3.1.** ($\mathbf{P}_1(T)$) For the linear switched system given in (3.1) and (3.2), find the optimal switching sequence $\{y(t)\}_{t=0}^{T-1}$ and the optimal control $\{u(t)\}_{t=0}^{T-1}$ that minimize the cost function (3.3).

To deal with problem $\mathbf{P}_1(T)$, we consider the following auxiliary problem.

**Problem 3.2.** ($\mathbf{P}_2(s)$) For the linear switched system (3.1)-(3.2) and some $s \in \mathbb{Z}_+$ with $s \leq T$, where $\mathbb{Z}_+$ denotes the set of of nonnegative integers, find the optimal sequence $Y(T) = \{y(t)\}_{t=0}^{T-1}$ and the optimal control $\{u(t)\}_{t=0}^{T-1}$ such that

$$Y(T) \in \mathbb{Y}(s) := \left\{ Y(T) \mid \sum_{\tau=t}^{T-1} \sigma(y(\tau-1), y(\tau)) \leq s \right\} \qquad (3.4)$$

holds and the cost function

$$\sum_{t=0}^{T-1} \Big[\ \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} \Big] + \|x(T)\|^2_{Q(T)} \qquad (3.5)$$

is minimized.

We use $v(\cdot)$ to denote the optimal value of problem $(\cdot)$. The following monotonic property is evident.

**Lemma 3.1.** *$v(\boldsymbol{P}_2(s_1)) \leq v(\boldsymbol{P}_2(s_2))$, if $s_1 \geq s_2$.*

Problem $\mathbf{P}_1(T)$ can be solved by identifying the following optimal number of switching,

$$s^* = \arg \min_{s \in \{0,1,\ldots,T\}} \{M \cdot s + v(\mathbf{P}_2(s))\}. \tag{3.6}$$

It is clear that the optimal control and optimal switching sequence of problem $\mathbf{P}_1(T)$ are identical to the solution of problem $\mathbf{P}_2(s^*)$. Thus, an efficient solution method for problem $\mathbf{P}_2(s)$ plays a key role in solving problem $\mathbf{P}_1(T)$. In addition to the problem with a finite horizon, we are also interested in the infinite horizon problem as $T$ goes to infinity.

**Problem 3.3.** ($\mathbf{P}_3(\infty)$) For the linear switched system given in (3.1) and (3.2), find the optimal switching sequence $\{y(t)\}_{t=0}^{\infty}$ and the optimal control $\{u(t)\}_{t=0}^{\infty}$ that minimize the cost function

$$\sum_{t=0}^{\infty} \Big[ \, \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + M \cdot \sigma(y(t-1), y(t)) \, \Big]. \tag{3.7}$$

This chapter is organized as follows. After presenting the problem formulations in this section, we develop in Section 3.3 an exact solution procedure of solving problem $\mathbf{P}_2(s)$ by using DP. Both the switching region and feedback gain are characterized. To reduce the burden of the computation, we develop in Section 3.4 a branch and bound framework which integrates a modified DP algorithm with semidefinite programming. We extend the method developed for finite-horizon problem to the infinite-horizon problem in Section 3.5. We present some illustrative examples in Section 3.6 to demonstrate the solution procedure developed in this chapter. Finally, we give some conclusion remarks in Section 3.7.

Throughout of the chapter, we use notation $Q \succeq 0$ to denote a positive semidefinite symmetric matrix and $\mathbb{S}^n_+$ the set of all $n \times n$ positive semidefinite symmetric matrices. For given $\mathcal{A}_1$ and $\mathcal{A}_2 \subset \mathbb{S}^n_+$, the notation $\mathcal{A}_1 \succeq \mathcal{A}_2$ means that, for any $H_1 \in \mathcal{A}_1$, there exists $H_2 \in \mathcal{A}_2$ such that $H_1 \succeq H_2$. Furthermore, we introduction the following operation $\mathcal{P}(\cdot)$ for $\mathcal{A} \subset \mathbb{S}_+^n$,

$$\mathcal{P}[\mathcal{A}] := \begin{cases} \mathcal{A} \setminus \mathcal{A}^* & \text{if } \exists \mathcal{A}^* \subset \mathcal{A}, \mathcal{A}^* \succeq \mathcal{A} \setminus \mathcal{A}^*, \\ \mathcal{A} & \text{otherwise.} \end{cases}$$

Clearly, such an operation eliminates all the dominated elements in set $\mathcal{A}$. The following fact is true.

**Lemma 3.2.** *Given $x \in \mathbb{R}^n$ and $\mathcal{A} \subset \mathbb{S}_+^n$,*

$$\min_{H \in \mathcal{A}} \|x\|_H^2 = \min_{H \in \mathcal{P}[\mathcal{A}]} \|x\|_H^2.$$

## 3.3. DP-Based Solution Approach for $\mathbf{P}_2(s)$

In this section we focus on solving problem $\mathbf{P}_2(s)$. A naive way to solve problem $\mathbf{P}_2(s)$ is to enumerate all possible switching sequences in $\mathbb{Y}(s)$ defined in (3.4). We are more interested in the structure properties of the optimal control and the switching conditions in the feedback back manner. Such structure properties may benefit us in designing more efficient algorithms in solving problem $\mathbf{P}_2(s)$. More specifically, we use DP to characterize the solution of problem $\mathbf{P}_2(s)$.

We expand the state space of problem $\mathbf{P}_2(s)$ by adding $y(t-1)$ and $r(t)$ which denote, respectively, the activated subsystem and the remaining number of switching at stage $t$, where $r(t)$ satisfies the following recursion,

$$r(t+1) = \begin{cases} r(t) - \sigma(y(t-1), y(t)) & \text{if } r(t) \geq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

with $r(0) = s$. The feasible set of $r(t)$, $t = 0, \cdots, T$, is confined to

$$\mathcal{F}_t := \{r(t) \in \mathbb{Z}_+ \mid \max\{0, s-t\} \leq r(t) \leq s\}.$$

The cost-to-go function can be then expressed as follows for $t = T, \cdots, 0$, $y(t) \in \mathbb{K}$ and $r(t) \in \mathcal{F}_t$,

$$J_t(x(t), y(t-1), r(t)) := \min_{u(\tau), y(\tau), \tau \geq t} \Big[ \sum_{\tau=t}^{T-1} (\|x(\tau)\|_{Q_{y(\tau)}}^2 + \|u(\tau)\|_{R_{y(\tau)}}^2) + \|x(T)\|_{Q(T)}^2 \Big], \tag{3.9}$$

where $\{y(\tau)\}_{\tau=t}^{T}$ satisfies $\sum_{\tau=t}^{T}\sigma(y(t-1),y(t)) \leq r(t)$. The cost-to-go in (3.9) implies the following recursion,

$$\begin{aligned} & J_t(x(t),y(t-1),r(t)) \\ & = \min_{u(t),y(t)\in\mathbb{K}}[\|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + J_{t+1}(x(t+1),y(t),r(t+1))]. \end{aligned} \tag{3.10}$$

We define the following Riccati operator $\mathcal{R}(j,\cdot):\mathbb{S}_n^+ \to \mathbb{S}_n^+$ for given subsystem $j\in\mathbb{K}$ and $P\in\mathbb{S}_n^+$,

$$\mathcal{R}(j,P){:=}A_j'PA_j + Q_j - A_j'PB_j[B_j'PB_j + R_j]^{-1}B_j'PA_j. \tag{3.11}$$

We further abuse such an operation to a set operation for a set of positive semidefinite matrices $\mathcal{A}\subset\mathbb{S}_+^n$ and $j\in\mathbb{K}$,

$$\mathcal{R}(j,\mathcal{A}) := \bigcup_{\forall P\in\mathcal{A}}\{\mathcal{R}(j,P)\}.$$

Without loss of generality, we assume that $\mathcal{R}(j,\emptyset)=\emptyset$.

We define now a recursion for $t=T-1,\cdots,0$, $j\in\mathbb{K}$, and[1] for all $r\in\mathcal{F}_t$,

$$\mathbb{P}(t,i,j,r){=}\begin{cases}\mathcal{R}\left(i,\mathbb{C}(t+1,j,r-1)\right) \text{ if } j\neq i \text{ and } r>0, \\ \mathcal{R}\left(i,\mathbb{C}(t+1,i,r)\right) \qquad \text{Otherwise.}\end{cases} \tag{3.12}$$

$$\mathbb{C}(t,i,r){=}\mathcal{P}[\bigcup_{j\in\mathbb{K}}\mathbb{P}(t,i,j,r)], \tag{3.13}$$

with the boundary condition $\mathbb{C}(T,j,r)=\{Q(T)\}$ for all $r\in\mathcal{F}_T$ and $j\in\mathbb{K}$. Furthermore, based on the above recursion, we define the following switching region in $\mathbb{R}^n$ for all $i,j\in\mathbb{K}$ and $r\in\mathcal{F}_t$,

$$\mathcal{M}^t(i,j,r) := \begin{cases} \emptyset \quad \text{if } \mathbb{P}(t,i,j,r)=\emptyset, \\ \mathbb{R}^n \quad \text{if } \mathbb{P}(t,i,k,r)=\emptyset,\ \forall\, k\in\mathbb{K}, k\neq j, \\ \{\, x\in\mathbb{R}^n \mid \min_{V\in\mathbb{P}(t,i,j,r)}\|x\|_V^2 \leq \min_{U\in\mathbb{P}(t,i,k,r)}\|x\|_U^2, \forall\, k\in\mathbb{K}\}, \\ \quad \text{otherwise.}\end{cases} \tag{3.14}$$

[1] Since $t$ is given, we simplify the notation of $r(t)$ to $r$, for all $r(t)\in\mathcal{F}_t$.

**Theorem 3.3.** *At stage $t$, for given $x(t)$, $y(t-1) = i^* \in \mathbb{K}$ and $r^* \in \mathcal{F}_t$, the optimal policy of problem $\boldsymbol{P}_2(s)$ is such that subsystem $j^* \in \mathbb{K}$ is active at stage $t+1$, if and only if $x(t)$ belongs to the switching region $\mathcal{M}^t(i^*, j^*, r^*)$, i.e.,*

$$y(t) = j^* \text{ if and only if } x(t) \in \mathcal{M}^t(i^*, j^*, r^*).$$

*Furthermore, the optimal control at stage $t$ is*

$$u^*(t) = (R + B'_{j^*} P^* B_{j^*})^{-1} B_{j^*} P^* A_{j^*} x(t),$$

*where*

$$P^* := \begin{cases} \arg\min_{P \in \mathbb{C}(t+1, j^*, r^*-1)} \ \|x(t)\|^2_{\mathcal{R}(j^*, P)}, & \text{if } j^* \neq i^*, \\ \arg\min_{P \in \mathbb{C}(t+1, i^*, r^*)} \ \|x(t)\|^2_{\mathcal{R}(i^*, P)}, & \text{if } \ j^* = i^*. \end{cases}$$

*Proof.* We prove this theorem by claiming that the cost-to-go at stage $t$ takes the following form for any $j \in \mathbb{K}$ and $r \in \mathcal{F}_t$,

$$J_t(x(t), j, r) = \min_{H \in \mathbb{C}(t,j,r)} \|x(t)\|^2_H.$$

The mathematical induction starts from stage $T$,

$$J_T(x(T), j, r) = \|x(T)\|^2_{Q(T)},$$

for $j \in \mathbb{K}$ and $r \in \mathcal{F}_T$. We simply let $\mathbb{C}(T, j, r) := \{Q(T)\}$.

Suppose that the claim is true at stage $t+1$, for all $j \in \mathbb{K}$ and $r \in \mathcal{F}_{t+1}$. The following two situations are possible at stage $t$.

If $r \in \mathcal{F}_t$ and $r = 0$, we fix $y(t) = i^* \in \mathbb{K}$ for the remaining time periods as the system cannot switch anymore,

$$\begin{aligned} J_t(x(t), i^*, 0) &= \min_{u(t)} \left[ \|x(t)\|^2_{Q_{i^*}} + \|u(t)\|^2_{R_{i^*}} + \min_{H \in \mathbb{C}(t+1, i^*, 0)} \|A_{i^*} x(t) + B_{i^*} u(t)\|^2_H \right] \\ &= \min_{H \in \mathbb{C}(t, i^*, 0)} \|x(t)\|^2_H, \end{aligned} \tag{3.15}$$

where the second equality is achieved by choosing optimal control

$$u^*(t) = (R_{i^*} + B'_{i^*} H B_{i^*})^{-1} B_{i^*} H A_{i^*} x(t).$$

We let $\mathbb{P}(t,i^*,i^*,0) := \mathcal{R}(t,i^*,\mathbb{C}(t+1,i^*,0))$, $\mathbb{P}(t,i^*,j,0) := \emptyset$ for $j \neq i^*$ and $\mathbb{C}(t,i^*,0) = \mathcal{P}[\mathbb{P}(t,i^*,i^*,0)\bigcup\emptyset]$. According to the definition of the switching region, we have $\mathcal{M}^t(i^*,i^*,0) = \mathbb{R}^n$ and $\mathcal{M}^t(i^*,j,0) = \emptyset$ for all $j \neq i^*$.

If $r \in \mathcal{F}_t$ and $r > 0$ with $y(t) = i^*$, cost-to-go in (3.10) can be expressed as

$$J_t(x(t),i^*,r) {=} \min\{\hat{J}_t(x(t),i^*,r), \bar{J}_t(x(t),i^*,r)\}, \tag{3.16}$$

where

$$\hat{J}_t(x(t),i^*,r) := \min_{u(t),y(t)=i^*} \Big[\|x(t)\|^2_{Q_{i^*}} + \|u(t)\|^2_{R_{i^*}} + J_{t+1}(x(t+1),i^*,r)\Big], \tag{3.17}$$

$$\bar{J}_t(x(t),i^*,r) := \min_{u(t),y(t)\neq i^*} \Big[\|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + J_{t+1}(x(t+1),y(t),r-1)\Big]. \tag{3.18}$$

From the induction assumption, (3.17) can be further simplified as follows,

$$\begin{aligned}\hat{J}_t(x(t),i^*,r(t)) &= \min_{u(t)} \Big[\|x(t)\|^2_{Q_{i^*}} + \|u(t)\|^2_{R_{i^*}} + \min_{H\in\mathbb{C}(t+1,i^*,r(t))} \|x(t+1)\|^2_H\Big] \\ &= \min_{H\in\mathbb{P}(t,i^*,i^*,r(t))} \|x(t)\|^2_H,\end{aligned} \tag{3.19}$$

where the second equality is achieved by adopting optimal control

$$u^*(t) = (R_{i^*} + B'_{i^*}HB_{i^*})^{-1}B_{i^*}HA_{i^*}x(t)$$

and let $\mathbb{P}(t,i^*,i^*,r) := \mathcal{R}(i^*,\mathbb{C}(t+1,i^*,r))$.

On the other hand, letting $y(t+1) = j \in \mathbb{K}$ in (3.18) gives rise,

$$\begin{aligned}\bar{J}_t(x(t),i^*,r) &{=} \min_{j\neq i^*,u(t)} [\ \|x(t)\|^2_{Q_j} + \|u(t)\|^2_{R_j} + \min_{H\in\mathbb{C}(t+1,j,r-1)} \|x(t+1)\|^2_H\ ] \\ &{=}\min_{j\neq i^*}\ \min_{H\in\mathbb{C}(t+1,j,r-1)} \{\min_{u(t)}[\ \|x(t)\|^2_{Q_j} + \|u(t)\|^2_{R_j} + \|x(t+1)\|^2_H\ ]\} \\ &{=}\min_{j\neq i}\ \min_{H\in\mathbb{P}(t,i^*,j,r)} \|x(t)\|^2_H,\end{aligned} \tag{3.20}$$

where the third equality is achieved by adopting optimal control $u^*(t) = (R_j + B'_jHB_j)^{-1}B_jHA_jx(t)$ and let

$$\mathbb{P}(t,i^*,j,r) := \mathcal{R}(j,\mathbb{C}(t+1,j,r-1)).$$

Substituting both (3.20) and (3.19) into (3.16) yields

$$\begin{aligned} J_t(x(t), i^*, r) &= \min\{\min_{H\in\mathbb{P}(t,i^*,i^*,r)} \|x(t)\|_H^2, \min_{j\neq i} \min_{H\in\mathbb{P}(t,i^*,j,r)} \|x(t)\|_H^2\} \quad (3.21) \\ &= \min_{j\in\mathbb{K}}\{\min_{H\in\mathbb{P}(t,i^*,j,r(t))} \|x(t)\|_H^2,\} \\ &= \min_{H\in\mathbb{C}(t,i^*,r(t))} \|x(t)\|_H^2, \quad (3.22) \end{aligned}$$

where

$$\mathbb{C}(t, i^*, r) := \mathcal{P}[\bigcup_{j\in\mathbb{K}} \mathbb{P}(t, i^*, j, r)].$$

Thus, we know from (3.21) that system switches from subsystem $i^*$ to $j$ if and only if

$$\min_{j\in\mathbb{K}, j\neq i^*} \min_{H\in\mathbb{P}(t,i^*,j,r)} \|x(t)\|_H^2 < \min_{H\in\mathbb{P}(t,i^*,i^*,r)} \|x(t)\|_H^2. \quad (3.23)$$

We can then define the switching region for each $j$ the same as in (3.14). Note that the above conclusion is true for any $i^* \in \mathbb{K}$. □

Theorem 3.3 actually provides an algorithm in solving problem $\mathbf{P}_2(s)$. We first calculate off-line $\mathbb{P}(t, i, j, r)$ iteratively from $t = T$ to $t = 0$, for $i, j \in \mathbb{K}$ and $r \in \mathcal{F}_t$, and determine then on-line the optimal switching and control policy. Once switching region is characterized, the switching policy can be achieved for any initial subsystem and initial state. Although the solution to problem $\mathbf{P}_2(s)$ is fully characterized by Theorem 3.3, such an algorithm, however, suffers from a computational complexity of $O\left(K^T(s + 1) \cdot \max\{m^3, n^3\}\right)$, where it takes $\mathcal{O}(\max\{m^3, n^3\})$ to compute the Riccati iteration in each step. As mentioned in the literature, solving optimal control of general switching systems is NP-hard with respect to its computational complexity. When $T$ is not too large, our algorithm performs well, largely due to the pruning operation $\mathcal{P}[\cdot]$ in the algorithm. Eliminating dominated elements in $\mathbb{C}(t, j, r)$ reduces significantly computational efforts. We will show in the following that our algorithm becomes a polynomial-time algorithm when all the subsystems are of a scalar state space.

**Corollary 3.4.** *If the state spaces of all the subsystems described in (3.2) are of dimension one, i.e.,* $n = 1$*, then the algorithm provided in Theorem 3.3 is polynomial with complexity of* $O\left(TK^2(s+1)m^3\right)$.

*Proof.* Notice the fact that whether or not $\|x\|_{H_1}^2 \leq \|x\|_{H_2}^2$ is independent of $x$, if $H_1$ and $H_2$ are scalars. Thus, the cardinality of set $\mathbb{C}(t, j, r)$ is one, for all $t = T, \cdots, 0$, $j \in \mathbb{K}$ and $r \in \mathcal{F}_t$. Thus, we only need to calculate $sK^2$ times of matrix operation of (3.11), which possesses a complexity of $O\left(m^3\right)$ at most. □

In addition to the cases with a scalar state space, the algorithm in Theorem 3.3 is also efficient for the following case.

**Corollary 3.5.** *If subsystem* $i^*$ *is active at stage* $\tau$ *and is such that*

$$\begin{pmatrix} Q_{i^*} & A'_{i^*} \\ A_{i^*} & -B'_{i^*}R_{i^*}^{-1}B_{i^*} \end{pmatrix} \preceq \begin{pmatrix} Q_j & A'_j \\ A_j & -B'_jR_j^{-1}B_j \end{pmatrix}, \ \forall\, j \neq i^*,$$

*then the optimal switching strategy for problem 3.2 in the remaining stages is* $y(\tau+1) = y(\tau+2) = \cdots = y(T) = i^*$.

Corollary 3.5 actually gives a sufficient condition for an "absorbing" subsystem. Once the system switches to such a subsystem, it will never switch to others.

We invoke the following lemma from [78] for Riccati operator $\mathcal{R}(\cdot,\cdot)$ in (3.11) in order to prove the above corollary.

**Lemma 3.6.** *If* $P_i \preceq P_j$ *and*

$$\begin{pmatrix} Q_i & A'_i \\ A_i & -B'_iR_i^{-1}B_i \end{pmatrix} \preceq \begin{pmatrix} Q_j & A'_j \\ A_j & -B'_jR_j^{-1}B_j \end{pmatrix},$$

*then* $\mathcal{R}(i, P_i) \preceq \mathcal{R}(j, P_j)$.

Corollary 3.5 then follows Theorem 3.3. At stage $T$, $\mathbb{C}(T, j, r) = \{Q(T)\}$ for all $r \in \mathcal{F}_T$ and $j \in \mathbb{K}$. At stage $T-1$, because of Lemma 3.6, $\mathcal{R}(i^*, \mathbb{C}(T, j, r)) \preceq \mathcal{R}(j, \mathbb{C}(T, j, r))$, for any $j \neq i^*$ and $r \in \mathcal{F}_{T-1}$. Thus, we have $\mathcal{M}^{T-1}(i^*, i^*, r) = \mathbb{R}^n$

and $\mathcal{M}^{T-1}(i^*, j, r) = \emptyset$. That is to say, at stage $T-1$, if subsystem $i^*$ is active, the optimal switching policy is not to switch. Carrying out the similar procedure backward at $T-2, T-3, \cdots, \tau$ yields the proof.

## 3.4. Branch and Bound Algorithm

As mentioned before, the algorithm developed in Theorem 3.3 becomes inefficient when $T$ and $K$ are large. To improve our algorithm, we develop a lower-bounding scheme which can be integrated into a branch and bound solution framework. Assume that we have a feasible control of problem $\mathbf{P}_2(s)$ with an incumbent value $v^*$. To search for the real optimal solution of $\mathbf{P}_2(s)$, we gradually fix $y(t)$ from $t = 0$ to $t = T-1$. For example, fixing the first $\tau$ elements of $Y(T)$ gives rise a partially fixed switching sequence $Y(\tau, T)$,

$$Y(\tau, T) := \{i_1, \cdots, i_\tau, y(\tau+1), \cdots, y(T-1)\}. \tag{3.24}$$

We then denote $\mathbf{P}_2(Y(\tau, T), s)$ as a sub-problem of primal problem $\mathbf{P}_2(s)$ with given partially fixed switching sequence $Y(\tau, T)$. We then calculate the lower bound of problem $\mathbf{P}_2(Y(\tau, T), s)$ and denote the lower bound as $\underline{v}(\mathbf{P}_2(Y(\tau, T), s))$. If $\underline{v}(\mathbf{P}_2(Y(\tau, T), s)) > v^*$, then this branch is fathomed, since no matter what $y(\tau+1), \cdots, y(T)$ are chosen, it will never generate a better objective value than $v^*$. If $\underline{v}(\mathbf{P}_2(Y(\tau, T), s)) < v^*$, this branch will be kept alive and we continue to fix $y(\tau+1), \cdots, y(T-1)$ one by one. Of course, we need to consider the cardinality constraint (3.4) when fixing $y(t)$. Once $s$ switches are specified, we get a feasible solution and we can use such a solution to modify $v^*$ if it offers a better solution. Such a procedure can be carefully implemented by a searching process of an enumeration tree. Thus, finding a tight and cheap bound of problem $\mathbf{P}_2(Y(\tau, T), s)$ plays a key role in such a branch and bound framework.

When considering a partially fixed switching sequence, problem $\mathbf{P}_2(Y(\tau, T), s)$ is actually the same difficult as the original problem $\mathbf{P}_2(s)$.

To construct a lower bound of problem $\mathbf{P}_2(Y(\tau,T),s)$, we introduce a number $\alpha \leq T-\tau$ and denote $D(\alpha)$ as a set of stages, i.e.,

$$D(\alpha) := \{t_1, t_2, \cdots, t_\alpha\} \subseteq \{\tau+1, \tau+2, \cdots, T\}.$$

The following procedure modifies the algorithm described in Theorem 3.3 to generate a lower bound of problem $(P(Y(\tau),s))$.

**Procedure 1**:

*Input*: $T$, $\mathcal{W}$, $x_0$, $D(\alpha)$ and $y_0$.

*Output*: Lower bound of problem $P(Y(\tau),s)$: $v(\alpha)$.

S0. Let $t \leftarrow T$ and $\mathbb{C}(t,i,r) \leftarrow \{Q(T)\}$, for all $r \in \mathcal{F}_T$ and $i \in \mathbb{K}$.

S1. If $t < \tau$, go $[S2]$. Otherwise, let $t \leftarrow t-1$. For all $i,j \in \mathbb{K}$ and $r \in \mathcal{F}_t$, calculate[2]

$$\mathbb{P}(t,i,j,r) = \begin{cases} \mathcal{R}(j, \mathbb{C}(t+1,j,r-1)) & \text{if } j \neq i \text{ and } r > 0, \\ \mathcal{R}(i, \mathbb{C}(t+1,i,r)) & \text{if } j = i \\ \emptyset & \text{otherwise,} \end{cases}$$

$$\mathbb{C}(t,i,r) = \mathcal{P}[\bigcup_{j\in\mathbb{K}} \mathbb{P}(t,i,j,r)]$$

If $t \in D(\alpha)$, solve the following problem,

$$\begin{aligned} (SP) \quad \max \quad & \text{trace}(H) \\ \text{Subject to:} \quad & H \preceq P, \ \ \forall\, P \in \mathbb{C}(t,i,r), \\ & H \in \mathbb{S}_n^+, \end{aligned}$$

and $\mathbb{C}(t,i,r) \leftarrow \{H^*\}$, where $H^*$ is the solution of problem $(SP)$. Repeat this step.

S2. If $t = 0$, compute

$$v(\alpha) := \min_{P \in \mathbb{C}(0,y(0),s)} \|x_0\|_P^2 \tag{3.25}$$

[2] If $t = \tau$, we only need to calculate $i = y(\tau)$.

Otherwise, $t \leftarrow t-1$ and compute [3]

$$\mathbb{C}(t, i_t, r(t)) = \mathcal{R}(i_{t+1}, \mathbb{C}(t+1, i_{t+1}, r(t+1))), \tag{3.26}$$

Repeat this step.

**Remark 3.1.** For a given switching sequence $Y(\tau)$, if $t < \tau$ in Procedure 1, $\mathbb{C}(t, i_t, r(t))$ is calculated directly using (3.26), as $i_t$ is fixed. If $t > \tau$, we select these stages in $D(\alpha)$ to calculate in $(SP)$ the best lower-bound matrix of set $\mathbb{C}(t, i, r)$. The purpose of $(SP)$ is to force the cardinality of set $\mathbb{C}(t, i, r)$ to 1, when $t \in D(\alpha)$, in order to reduce significantly the total computational complexity. Notice that problem $(SP)$ is a semidefinite programming problem, which can be efficiently solved by the interior point method.

**Theorem 3.7.** *As $v(\alpha)$ defined in (3.25), the following facts are true:*

*(a)* $v(\alpha) = v(P(Y(\tau), s))$, *if* $\alpha = 0$.

*(b)* $v(\alpha) \le v(P(Y(\tau), s))$, *if* $\alpha > 0$.

*(c)* $v(\alpha_1) \ge v(\alpha_2)$, *if* $0 \le \alpha_1 \le \alpha_2 \le T - \tau$ *and* $D(\alpha_1) \subseteq D(\alpha_2)$.

*Proof.* If $\alpha = 0$, then there is no modification of $\mathbb{C}(t, i, r(t))$ and (a) is true. Let us prove (c) first, we assume that $D(\alpha_1) = \{t_1\}$ and $D(\alpha_2) = \{t_1, t_2\}$ and $t_2 = t_1 + 1$. In the following, we consider $i \in \mathbb{K}$ and $r \in \mathcal{F}_t$. At stage $t_2$, $\mathbb{C}(t_2, i, r(t_2))$ is modified according to Procedure 1. Note that $\hat{\mathbb{C}}(t_2, i, r(t_2)) \le \mathbb{C}(t_2, i, r(t_2))$, where $\hat{\mathbb{C}}(t_2, i, r(t_2)) := \{H^*\}$ and $H^*$ is the solution of problem (SP) in Step [S1]. At stage $t_1$, due to Lemma 3.6, $\hat{\mathbb{P}}(t_1, i, j, r(t_1)) \preceq \mathbb{P}(t_1, i, j, r(t_1))$, where $\hat{\mathbb{P}}(t_1, i, j, r(t_1))$ and $\mathbb{P}(t_1, i, j, r(t_1))$ are calculated from $\hat{\mathbb{C}}(t_2, i, r(t_2))$ and $\mathbb{C}(t_2, i, r(t_2))$, respectively. After we solve problem $(SP)$ again, $\hat{\mathbb{C}}(t_1, i, r(t_1)) \preceq \mathbb{C}(t_1, i, r(t_1))$, where $\hat{\mathbb{C}}(t_1, i, r(t_1))$ and $\mathbb{C}(t_1, i, r(t_1))$ are corresponding to $D(\alpha_2)$ and $D(\alpha_1)$, respectively. This argument can be extended to general situations

[3] $i_t$ is defined in (3.24).

with any $\alpha_1 \leq \alpha_2$ and $D(\alpha_1) \subseteq D(\alpha_2)$ by induction method. The claim (b) is true, once we let $\alpha_1 = 0$ in (c). □

The bounding procedure 1 provides us certain flexibility in finding a lower bound. Note that the bounding error $[v(P(Y(\tau), s)) - v(\alpha)]$ is monotonically increasing with respect to $\alpha$. A small $\alpha$, however, may lead to an exponential increase of the computation complexity. The number $\alpha$ actually represents the trade-off between two critical measures, the tightness and efficiency. After fixing the number $\alpha$, the stages, $t_1, \cdots, t_\alpha$, can be fixed by using some heuristics. Example 3.3 in the Section 6 will illustrate effects of different $\alpha$ and $D(\alpha)$ in such a lower bounding scheme.

## 3.5. Problem of Infinite Horizon

We now focus on solving the problem $\mathbf{P}_3(\infty)$. Assume that the following assumption is satisfied.

**Assumption 3.1.** The system matrices $(A_i, B_i)$ are stabilizable and $(A_i, U_i)$ are detectable for all $i \in \mathbb{K}$, where $U_i'U_i = Q_i$.

Due to a nonzero switching cost $M$, the total number of switching evaluated in $\sum_{t=1}^{\infty} \sigma(y(t), y(t+1))$ has to be a finite number. That is to say, the following lemma is true.

**Lemma 3.8.** *There exists $T^*$ such that the optimal switching sequence $y(t)$ remains unchanged for all $t \geq T^*$.*

It is well known that under the Assumption 3.1, there exits a unique positive semidefinite solution of the *algebraic Riccati equation* (ARE) of each subsystem [17] [36], denoted as $H_i$, i.e.,

$$H_i = A_i'H_iA_i + Q_i - A_i'H_iB_i[B_i'H_iB_i + R_i]^{-1}B_i'H_iA_i, \quad i \in \mathbb{K}. \tag{3.27}$$

We also assume the following to eliminate trivial cases where the optimal switching strategy is never to switch.

**Assumption 3.2.** The switching cost $M$ is such that $M \leq \max_{i \in \mathbb{K}}\{\|x(0)\|^2_{H_i}\}$.

Under Assumptions 3.1 and 3.2, problem $\mathbf{P}_3(\infty)$ can be tackled by a problem of a finite horizon.

**Problem 3.4.** ($\mathbf{P}_4(T, s)$) For the linear switched system specified in (3.1) and (3.2), find the optimal switching sequence $\{y(t)\}_{t=0}^{T-1}$ that satisfies (3.4) and the optimal control $\{u(t)\}_{t=0}^{T-1}$ such that the cost function,

$$\sum_{t=0}^{T-1} \left[ \ \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} \ \right] + \min_{i \in \mathbb{K}} \|x_T\|^2_{H_i}, \tag{3.28}$$

is minimized, where $H_i$ is defined in (3.27).

The relationship between Problem 3.3 and Problem 3.4 can be characterized as follows.

**Theorem 3.9.** *Suppose that* $\bar{Y}(T) := (\bar{y}(0), \bar{y}(1), \cdots, \bar{y}(T-1))$ *and* $\{\bar{u}(t)\}_{t=0}^{T-1}$ *solve Problem* $\boldsymbol{P}_4(T, s^*)$ *where* $T \geq T^*$ *and*

$$s^* := \arg\min_{i \in \mathbb{K}} \{v(\boldsymbol{P}_4(T, s)) + s \cdot M\}.$$

*Then, the optimal switching sequence of problem* $\boldsymbol{P}_3(\infty)$ *is* $\bar{Y}(\infty) = (\bar{y}(0), \cdots, \bar{y}(T-1), \bar{y}(T-1), \cdots)$ *and the optimal control is*

$$u^*(t) = \begin{cases} \bar{u}(t) & \text{if } \ t \leq T, \\ (R_{i^*} + B_{i^*} H_{i^*} B_{i^*})^{-1} B_{i^*} H_{i^*} A_{i^*} x_t & \text{if } \ t > T. \end{cases}$$

*Proof.* At any stage $T \geq T^*$, based on the principle of the optimality, the value function of Problem $\mathcal{P}_1$ is given as,

$$J_T(x_T) := \min_{y(t), u(t), t \geq T} \sum_{t=T}^{\infty} \left\{ \ \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + M \cdot \sigma(y(t-1), y(t) \right\}. \tag{3.29}$$

Because of Lemma 3.8, $y(t)$ remains unchanged for all $t \geq T$. Thus, $\sum_{t=T}^{\infty} \sigma(y(t-1), y(t)) = 0$. Furthermore, finding optimal control $\{u_t\}|_{t=T}^{\infty}$ for the value function in (3.29) is a standard LQ control problem for each subsystem. Together with Assumption 3.1 we know the Riccati difference equation converges to a unique solution, $H_i$, given in (3.27) for each subsystem. Then, we can conclude that

$$J_T(x_T) = \min_{i \in \mathbb{K}} \{\|x_T\|^2_{H_i}\}.$$

Thus, solving problem $\mathbf{P}_3(\infty)$ is equivalent to minimizing the following value function,

$$\begin{aligned} J_0(x_0) = & \min_{y(t),u(t),0 \leq t \leq T^*} \Big[ \, \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + M \cdot \sigma(y(t-1), y(t)) \, \Big] \\ & + \min_{j \in \mathbb{K}} \|x_{T^*}\|^2_{H_j}. \end{aligned} \tag{3.30}$$

Note that if we parameterize the the total number of the switching by $s \in \mathbb{N}$, then we have

$$J_0(x_0) = \min_{s \in \mathbb{N}} \left\{ \min_{y(t),u(t),0 \leq t \leq T^*} [ \, \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} ] + \min_{j \in \mathbb{K}} \|x_{T^*}\|^2_{H_j} + s \cdot M \right\},$$

which completes the proof of the theorem. □

Note that the only difference between Problem $\mathbf{P}_2(s)$ and Problem $\mathbf{P}_4(T, s)$ is the terminal cost of $x(T)$. Thus, the Algorithm developed based on Theorem 3.3 for problem $\mathbf{P}_2(s)$ is readily for problem $\mathbf{P}_4(T, s)$. However, the boundary condition defined in the recursion (3.12) and (3.13) should be changed to $\mathbb{C}(T, j, r) = \{H_j\}$ for $r \in F_T$ and $j \in \mathbb{K}$.

The above development is dependent on the assumption that $T^*$ is known, however, finding an exact $T^*$ is hard. As illustrated in Lemma 3.8 and Theorem 3.9, we only need some $T > T^*$ which can be used to convert the infinite horizon problem $\mathbf{P}_4(\infty)$ to the finite horizon problem $\mathbf{P}_2(s, T)$. A sufficient condition of finding such $T \geq T^*$ is as follows.

**Theorem 3.10.** *At stage $T$, for given $x(T)$ and $y(T-1) = i^*$, if $\|x(T)\|^2_{H_{i^*}} \leq M$, then the switching sequence $y(t) = i^*$ for all $t \geq T$ will be optimal for Problem $\boldsymbol{P}_3(\infty)$.*

*Proof.* The value function of Problem $\mathbf{P}_3(\infty)$ is

$$J_T(x(T), i^*) = \min_{u(t), y(t) \geq T} \sum_{t=T}^{\infty} \Big[ \ \|x(t)\|^2_{Q_{y(t)}} + \|u(t)\|^2_{R_{y(t)}} + M \cdot \sigma(y(t-1), y(t)) \ \Big]. \tag{3.31}$$

If the conclusion in Theorem 3.10 is not true, then there will be at least one switching in time interval $[T, \infty]$, which leads to $J_T(x(T), i^*) > M$ from the expression (3.31). As $J_T(x(T), i^*) = \|x(T)\|^2_{H_{i^*}} \leq M$ if there is no switching, switching in time interval $[T, \infty]$ can not be optimal. □

From Theorem 3.3, we know that, for all $i \in \mathbb{K}$, the corresponding closed loop system can be written as $x(t+1) = (A_i + B_i F(t))x(t)$. We can then adopt some heuristics to estimate an upper bound of $\|x(T)\|$, i.e., $\|x(T)\|^2 \leq \beta^T \|x(0)\|^2$. Since condition $\|x(T)\|^2 \rho_i \leq M$, where $\rho_i$ is the spectral norm of $H_i$, implies the condition $\|x(T)\|^2_{H_i} \leq M$, we can estimate $T$ by

$$\min_{i \in \mathbb{K}} \frac{\log(M)}{\log(\rho_i \|x(0)\|^2 \beta)}. \tag{3.32}$$

## 3.6. Illustrative Examples

**Example 3.1.** We consider the following example with $n = 2$, $m = 1$, $K = 2$, $T = 3$, and $s = 2$ to illustrate the solution process of Theorem 3.3 for problem $\mathbf{P}_2(s)$. The data of two subsystems are given as

$$\begin{aligned}\mathcal{W} = \Big\{ \big[A_1 = \begin{pmatrix} -1 & 3 \\ -1 & -2 \end{pmatrix}, \ B_1 = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \ Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, \ R_1 = 0.5\big], \\ \big[A_2 = \begin{pmatrix} 1 & -4 \\ 1 & -3 \end{pmatrix}, \ B_2 = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, \ Q_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.9 \end{pmatrix}, \ R_2 = 0.5\big] \Big\},\end{aligned}$$

and the coefficient matrix for the final state is $Q_3 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$. We follow the procedure in Theorem 3.3. At stage $t = 3$, $\mathbb{C}(3, i, r) = \{Q(3)\}$ for $i = 1, 2$ and $r = 0, 1, 2$. At stage $t = 2$, for $i = 1, 2$ and $r = 1, 2$,

$$\mathbb{P}(2, i, 1, r) = \{\begin{pmatrix} 1.50 & -0.67 \\ -0.67 & 6.94 \end{pmatrix}\},\ \mathbb{P}(2, i, 2, r) = \{\begin{pmatrix} 1.00 & -1.67 \\ -1.67 & 6.68 \end{pmatrix}\},$$

and for $r = 0$, $\mathbb{P}(2, 1, 1, 0) = \mathbb{P}(2, 1, 1, 1)$ and $\mathbb{P}(2, 2, 2, 0) = \mathbb{P}(2, 2, 2, 1)$.

At stage 1, for $i = 1, 2$, we have

$$\mathbb{P}(1, 1, 1, 1) = \mathbb{P}(1, i, 1, 2) = \{\begin{pmatrix} 1.85 & -2.09 \\ -2.09 & 26.25 \end{pmatrix}, \begin{pmatrix} 1.55 & -0.31 \\ -0.31 & 16.17 \end{pmatrix}\},$$

$$\mathbb{P}(1, 2, 2, 1) = \mathbb{P}(1, i, 2, 2) = \{\begin{pmatrix} 2.59 & -5.41 \\ -5.41 & 15.51 \end{pmatrix}, \begin{pmatrix} 2.47 & -5.24 \\ -5.24 & 15.63 \end{pmatrix}\},$$

$$\mathbb{P}(1, 1, 2, 1) = \{\begin{pmatrix} 2.59 & -5.41 \\ -5.41 & 15.51 \end{pmatrix}\}, \mathbb{P}(1, 2, 1, 1) = \{\begin{pmatrix} 1.85 & -2.09 \\ -2.09 & 26.25 \end{pmatrix}\}.$$

At stage 0, $\mathbb{P}(0, 1, 1, 2)$, $\mathbb{P}(0, 1, 2, 2)$, $\mathbb{P}(0, 2, 1, 2)$ and $\mathbb{P}(0, 2, 2, 2)$ can be calculated in the same manner. Since $x(t) \in \mathbb{R}^2$, the switching regions given in (3.14) are symmetric cones which can be easily expressed in the polar coordinate system, $x_1(t) = \cos(\theta_t)\rho_t$ and $x_2(t) = \sin(\theta_t)\rho_t$ with $\theta_t \in (0, 2\pi]$ and $\rho_t \geq 0$. Let $\Theta(x_t) := \theta_t$. The switching regions can be now characterized as follows,

$$\begin{aligned}
\mathcal{M}^2(1, 1, 2) &= \{\Theta(x_2) \in [1.54\pi, 1.92\pi] \cup [0.54\pi, 0.92\pi]\}, \\
\mathcal{M}^1(1, 1, 2) &= \{\Theta(x_1) \in [0.52\pi, 1.03\pi] \cup [-0.48\pi, 0.03\pi]\}, \\
\mathcal{M}^1(1, 1, 1) &= \{\Theta(x_1) \in [0.52\pi, 1.03\pi] \cup [-0.48\pi, 0.03\pi]\}, \\
\mathcal{M}^1(2, 1, 1) &= \{\Theta(x_1) \in [0.81\pi, 1.03\pi] \cup [-0.19\pi, 0.03\pi]\}, \\
\mathcal{M}^0(1, 1, 2) &= \{\Theta(x_1) \in [0.49\pi, 1.05\pi] \cup [-0.51\pi, 0.05\pi]\}, \\
\mathcal{M}^0(2, 1, 2) &= \{\Theta(x_1) \in [0.59\pi, 1.04\pi] \cup [-0.41\pi, 0.04\pi]\},
\end{aligned}$$

which are pictured in Figure 3.1. Table 3.1 lists the optimal switching sequences for different initial subsystems and initial states.

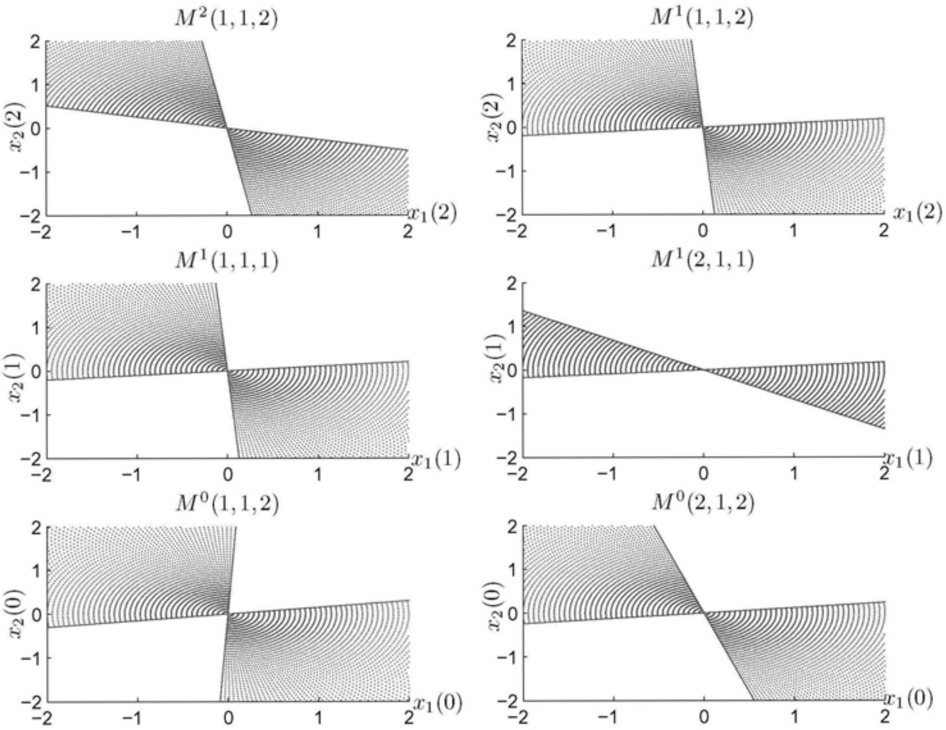

Figure 3.1: The switching regions of Example 3.1

Table 3.1: Solutions of Example 1 for various initial conditions

| Initial sys. | $x_0$ | $Y(t)$ | $v(P(s))$ |
|---|---|---|---|
| 1 | $(1,1)'$ | $\{2,2,2\}$ | 9.71 |
| 1 | $(-2,1)'$ | $\{1,2,1\}$ | 81.11 |
| 2 | $(1,2)'$ | $\{2,1,2\}$ | 63.48 |
| 2 | $(3,-1)'$ | $\{1,2,2\}$ | 123.26 |

Table 3.2: Solutions of Example 3.2 for various $s$

| $s$ | Opt switching sequence | $v(\mathbf{P}_2(s))$ | $s \cdot M$ | $v(\mathbf{P}_2(s)) + sM$ |
|---|---|---|---|---|
| 1 | $\{2,4,4,4,4,4,4,4,4,4,4\}$ | 10502875.9 | 100 | 10502975.9 |
| 2 | $\{2,4,2,2,2,2,2,2,2,2,2\}$ | 9208618.1 | 200 | 9208818.1 |
| $3^*$ | $\{2,4,1,2,2,2,2,2,2,2,2\}$ | 7933065.3 | 300 | 7933365.3 |
| 4 | $\{2,4,1,2,2,2,2,2,2,2,2\}$ | 7933065.2 | 400 | 7933465.2 |
| 5 | $\{2,4,1,2,2,2,2,2,2,2,2\}$ | 7933065.2 | 500 | 7933565.2 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

**Example 3.2.** Consider a problem $\mathbf{P}_1(T)$ with the same data of the power-train system in [77]. There are 4 subsystems ($K = 4$) for 4 different gears with $A_1 = \begin{pmatrix} -65.72 & -185.4 \\ 1.72 & 0.69 \end{pmatrix}$, $A_2 = \begin{pmatrix} -36.3 & -56.58 \\ 0.38 & 0.94 \end{pmatrix}$, $A_3 = \begin{pmatrix} -23.46 & -24.38 \\ 0.12 & 0.98 \end{pmatrix}$, $A_4 = \begin{pmatrix} -16.24 & -13.31 \\ 0.06 & 1 \end{pmatrix}$, $B_i = \begin{pmatrix} 0.2 \\ 0 \end{pmatrix}$, $Q_i = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$, $R_i = 1$, for $i = 1,2,3,4$. Let us assume the horizon to be $T = 10$, the coefficient matrix for the final state to be $Q_T = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, the initial subsystem to be the 3rd subsystem with initial state $x_0 = \begin{pmatrix} 1 \\ 50 \end{pmatrix}$, and the switching cost to be $M = 100$. Table 3.2 presents the results of the auxiliary problem $P(s)$ for different $s$. We can find that the optimal number of switching is $s^* = 3$ and the state trajectory and switching sequence are given in Figure 3.2.

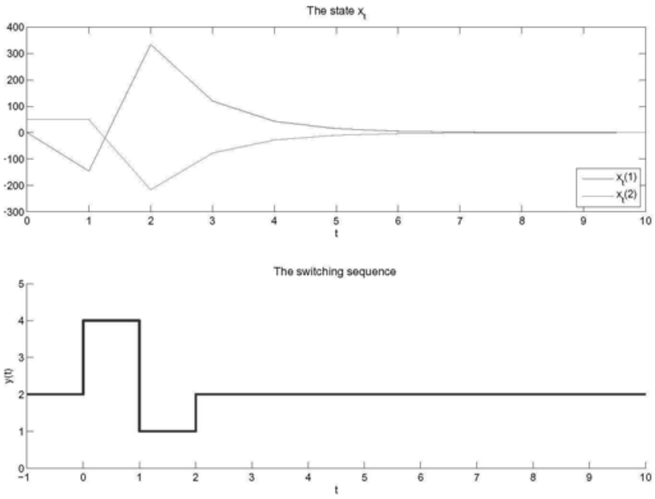

Figure 3.2: The state trajectory and switching sequence of Example 3.2

**Example 3.3.** We still use Example 3.2 to illustrate the lower bounding procedure described in Section IV. Consider the auxiliary problem $P(s)$ with $s = 3$. Consider a lower bound for a given partially fixed switching sequence,

$$Y(2) = \{2, 3, y(3), y(4), y(5), y(6), y(7), y(8), y(9), y(10)\}.$$

We perform Procedure 1 by fixing $\alpha = 2$, $\alpha = 3$ or $\alpha = 4$. The details of the lower bound and CPU time for different $D(\alpha)$ are listed in Table 3.3. All the computations are executed under Matlab 7 with Sedumi 1.1 on a PC (P4 2.6G with 1G memory). Note that the optimal value associated with sequence $Y(2)$ is $1.9814 \times 10^8$.

**Example 3.4.** We consider a randomly generated infinite horizon problem $\mathbf{P}_3(\infty)$ with 2 subsystems specified by $A_1 = \begin{pmatrix} 0.881 & -0.464 \\ 0.835 & 0.695 \end{pmatrix}$, $A_2 = \begin{pmatrix} 1.034 & -0.615 \\ 0.162 & 0.515 \end{pmatrix}$, $B_1 = \begin{pmatrix} 0.098 \\ 0.043 \end{pmatrix}$, $B_2 = \begin{pmatrix} 0.028 \\ 0.034 \end{pmatrix}$, $Q_i = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.3 \end{pmatrix}$ and $R_i = 0.1$ for $i = 1, 2$.

Table 3.3: Lower bound and CPU time for various $D(\alpha)$ in Example 3.3

| $D(\alpha)$ | LB ($10^8$) | CPU(s) | $D(\alpha)$ | LB ($10^8$) | CPU(s) |
|---|---|---|---|---|---|
| $\{3,7\}$ | 1.8591 | 50.0 | $\{4,6,9\}$ | 0.30256 | 12.0 |
| $\{4,7\}$ | 1.8695 | 28.1 | $\{5,7,9\}$ | 0.30891 | 9.6 |
| $\{5,7\}$ | 1.8648 | 20.0 | $\{5,7,8\}$ | 0.30891 | 25.1 |
| $\{6,8\}$ | 1.8648 | 8.1 | $\{2,3,5,8\}$ | 0.02952 | 12.2 |
| $\{4,6\}$ | 1.8611 | 90.0 | $\{3,4,6,8\}$ | 0.07216 | 8.9 |
| $\{4,8\}$ | 1.8690 | 30.3 | $\{2,4,5,8\}$ | 0.07030 | 13.2 |

By solving the Algebraic Riccati equation (3.27), we get

$$H_1 = \begin{pmatrix} 2.879 & -0.006 \\ -0.006 & 1.694 \end{pmatrix}, H_2 = \begin{pmatrix} 1.089 & -0.959 \\ -0.959 & 1.791 \end{pmatrix}.$$

We estimate $T$ as 15 by using the heuristic scheme described in Section 4. We then solve problem $\mathbf{P}_4(15, s)$ for different $s$. The detailed results are given in Table 3.4 and Figure 3.3. Note that in all the problems of $\mathbf{P}_4(s, 15)$ for different $s$, the condition 3.10 are satisfied, that is to say, $T$ is large enough for converting such an infinite horizon problem to one with a finite horizon.

Table 3.4: Solution of Example 3.4

| $s$ | $Y^*$ | $v(\mathbf{P}_4(15,s))$ | $sM$ | $v(\mathbf{P}_2(s)) + sM$ |
|---|---|---|---|---|
| 1 | $\{2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1\}$ | 458.5 | 10 | 468.5 |
| 2* | $\{2,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2\}$ | 204.8 | 20 | 224.8 |
| 3 | $\{2,1,1,2,2,1,1,1,1,1,1,1,1,1,1,1\}$ | 199.1 | 30 | 229.1 |
| 4 | $\{2,1,1,2,2,1,1,2,2,2,2,2,2,2,2,2\}$ | 185.1 | 40 | 225.1 |
| 5 | $\{2,1,1,2,2,1,1,1,2,2,2,1,1,1,1,1\}$ | 185.0 | 50 | 235.0 |
| 6 | $\{2,1,1,2,2,1,1,1,2,2,1,1,1,2,2,2\}$ | 184.4 | 60 | 235.0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

**Example 3.5.** We study a discrete-version of the continuous-time model in Example 2 of [82]. Since the total switching time is 1, it is of problem type $\mathbf{P}_3(s)$

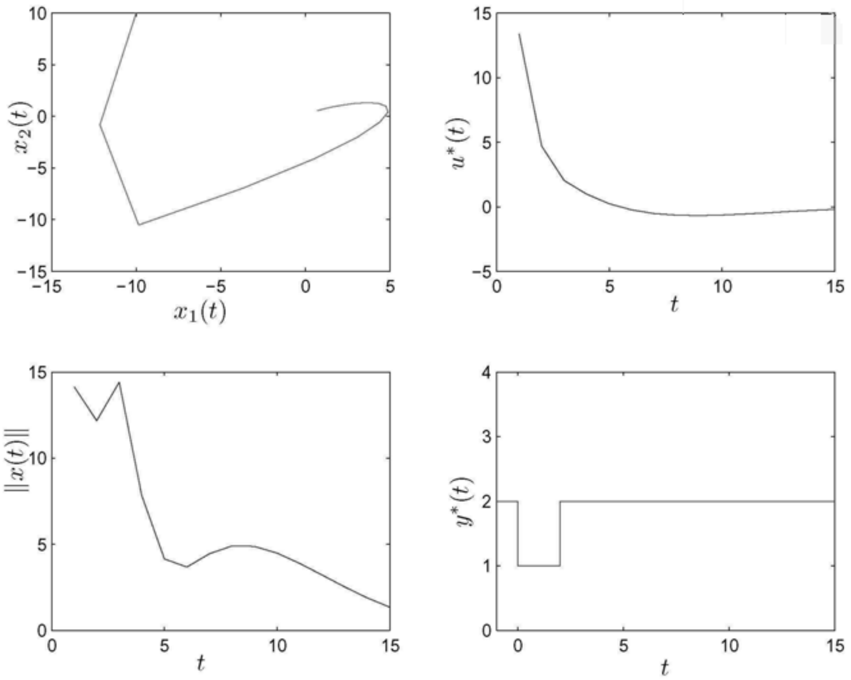

Figure 3.3: Optimal control, switching sequence and states of Example 3.4

with $s = 1$. Let the sample time period be $T_s$ and the total number of periods be $T = 2/T_s$. By expanding the state to dimension of $n = 4$, we can express the resulting problem by the following two sub-systems,

$$\left[\begin{pmatrix} \bar{A}_1(T_s) & 0_{2\times 2} \\ 0_{2\times 2} & I_{2\times 2} \end{pmatrix}, \begin{pmatrix} \bar{B}_1(T_s) \\ 0_{2\times 1} \end{pmatrix}, T_s \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}, T_s\right]$$

and

$$\left[\begin{pmatrix} \bar{A}_2(T_s) & 0_{2\times 2} \\ 0_{2\times 2} & I_{2\times 2} \end{pmatrix}, \begin{pmatrix} \bar{B}_2(T_s) \\ 0_{2\times 1} \end{pmatrix}, T_s \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}, T_s\right]$$

Table 3.5: Results for different $T_s$ of Example 3.4

| Sample Time $T_s$ | 0.1 | 0.05 | 0.02 | 0.01 | 0.005 |
|---|---|---|---|---|---|
| Total Num of Periods $T$ | 20 | 40 | 100 | 200 | 400 |
| Opt Switching Instant | 0.2 | 0.20 | 0.20 | 0.19 | 0.190 |
| CPU Time | 0.00s | 0.01s | 0.72s | 3.1s | 8.2s |

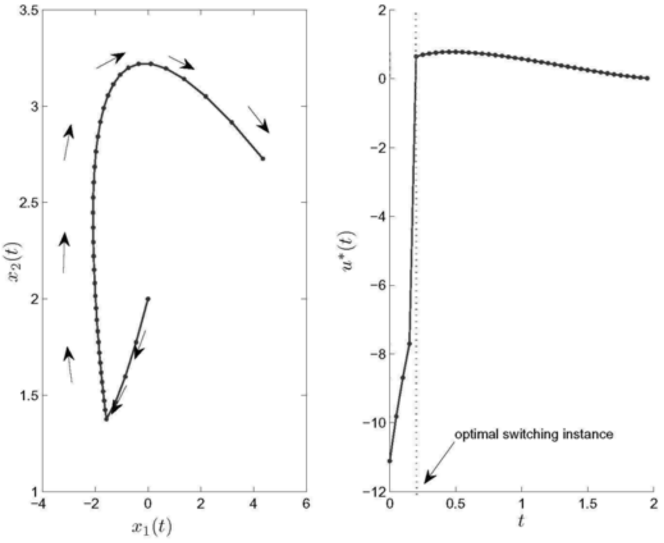

Figure 3.4: The state trajectory and the optimal control in Example 3.5

with $Q(T) = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$ and $x_0 = [0, 2, 4, 2]'$, where $\bar{A}_j(T_s)$ and $\bar{B}_j(T_s)$, $j = 1, 2$, are obtained from discretizing the original system matrices in [82] with sample period $T_s$. The state trajectory and control are shown in the Figure 3.4 ($t_s = 0.05$). The numerical results of implementing our method listed in Table 3.5 are consistent with the results in [82], and demonstrate the efficiency of our discrete-time based searching procedure when compared to the exact method in [82].

## 3.7. Conclusion

We have investigated in this chapter both finite-horizon and infinite-horizon discrete-time optimal control of linear switching systems with switching cost. We have constructed a corresponding auxiliary problem, which is a finite horizon linear-quadratic switching optimal control problem with upper limit on number of switching and have derived its switching regions. Our result on the switching regions is consistent with the result in [30] for continuous-time switched autonomous linear systems. Recognizing high computational complexity in the exact solution scheme, we have also proposed a lower-bounding method by using semidefinite programming, which can be easily integrated into a branch and bound framework. Although our preliminary computational results have already demonstrated promising results, many challenging issues remain for large-scale switching systems.

# CHAPTER 4

# Time Cardinality Constrained Mean Variance Dynamic Portfolio Selection

## 4.1. Introduction and Problem Formulation

The ground-breaking mean-variance formulation proposed by Markowitz [55, 56] initialized the study of modern portfolio selection by measuring the investment risk by the variance term of the terminal wealth. The analytical expression of the mean-variance efficient frontier in single period portfolio selection was derived by Markowitz [55] and Merton [57]. The mean-variance portfolio selection theory has been extended in the literature to dynamic settings. More specifically, the analytical optimal portfolio policies and the corresponding efficient frontiers were derived for multi-period mean-variance portfolio selection problems and for continuous-time mean-variance portfolio selection in Li and Ng [43] and Zhou and Li [84], respectively. The past eight years have witnessed numerous extensions of the mean-variance portfolio selection theory in continuous-time settings, see for examples, Li et al. [47], Lim and Zhou [49], Zhou and Yin [85], Hu and Zhou [33], Bielecki et al. [10], Li and Zhou [46], Chiu and Li [20], Xiong and Zhou [81]. Contrary to rich results in continuous-time problems, the progress of mean-varaince portfolio selection in discrete-time settings has been relatively thin, see for examples, Leippold et al. [37], Zhu et al. [86], Liang et al. [48], although discrete-time settings model the real investment world more closely than continuous-time settings. Recently, Cerny and Kallsen [1] and Cerny and Kallsen [16] studied the optimal mean-variance portfolio selection in a more general setting with a semi-martingale price process, which includes both discrete-time and continuous-time settings as its special cases. In this chapter, we extend the results of multi-period mean-variance analysis by considering the management fees for investing in risky assets.

Citing from the website of the US Securities and Exchange Commission [68], the management fees are defined as the fees that are paid out of fund assets to the fund's investment adviser for investment portfolio management. The importance of the management fees in portfolio selection has been investigated in Capon et al. [59], Golec [31], Nanda et al. [61], Brown et al. [14]. One example of the management fee can be found from the website of The American Investment Service [69]: For investors with assets under management (AUM) between US$100,000 and US$250,000, the annual fee is 0.80% of AUM or US$1,500, whichever is greater. We consider in this chapter portfolio selection problems with management fees of a nature of set-up cost. This type of situations has been often witnessed in real applications. For example, if an investor asks the American Investment Service to manage his investment with an amount less than US$187,500, he will be charged annually a fixed amount management fee of US$1,500. Due to the set-up type of management fees charged for hiring an agent in managing their investment in risky assets, investors do not always invest in risky assets in all time periods.

A related subject in the literature is the optimal investment and consumption problem with transaction costs. Transaction fees are charged when there is a transaction between riskfree account and risky assets, and transaction costs, in general, involve two types of fees: a fixed charge and a variable charge proportional to the transaction amount. Davis and Norman[25] first studied this kind of problem with proportional transaction cost in a continuous-time setting with an infinite horizon. By using the viscosity solution, Shreve and Soner [70] completely solved this infinite horizon problem. As for finite horizon problem, Liu and Loewenstein [53] carried out a study of such a problem with proportional transaction cost in a continuous-time setting with a finite horizon and derived an approach to approximate the analytical solution. Dai and Zhong [23] [24] considered the same problem, with an exception that the buying and selling boundaries are characterized by variational inequalities, and proposed some numerical solution schemes. Merton and Pliska [58] analyzed the similar problem, in which the transaction cost is a fixed fraction of the investor's portfolio. Under the framework of the impulse control, Korn [35] and Oksendal and Sulem [62] studied such a problem with both fixed and proportional transaction costs by solving the correspondent HJB equation numerically, while only one risky asset is involved in their model. Liu [52] extended the results of Korn [35] and Oksendal and Sulem [62] to a more general setting with multiple risky assets and characterized the boundaries of the transaction regions. To our knowledge, Lynch and Tan [54] represented the only work of portfolio selection problem with fixed and proportional transaction costs in a discrete-time setting, and developed some numerical solution methods.

The nature of management fee is substantially different from the transaction costs, as the management fee (for a given initial amount) is an external fee paid to investor's agent based on the time length of service, while transaction costs is an internal fee when there is a transaction between riskfree and risky assets in investor's portfolio.

We assume that the capital market consists of $n$ risky assets and one risk-free asset, all of which evolve within a time horizon of $T$ periods, $t = 0, 1, \ldots, T-1$. An investor with initial wealth $x_0$ enters the market at stage 0 and allocates his wealth among these $n+1$ assets at the beginning of each of the $T$ periods. When a non-zero amount of his wealth is allocated to some or all the $n$ risky assets at period $t$, $t = 0, \cdots, T-1$, he will be charged a constant management fee $M$. Such a cost $M$ can be understood as the cost of hiring an agent to manage his stocks. The total management fee is deducted from his terminal wealth $x_T$ at the last period $t = T$. Let the returns of the risk-free asset at different time periods be $r_t$, $t = 0, \cdots, T-1$, which are assumed to be deterministic in this chapter, although there will be no technical difficulty to extend our results to situations with random $r_t$. We denote the random return vector of the $n$ risky assets as

$$\mathbf{e}_t := \begin{pmatrix} e_t^1 & e_t^2 & \cdots & e_t^n \end{pmatrix}'$$

for $t = 0, \cdots, T-1$ and assume $\mathbf{e}_t$, $t = 0, \cdots, T-1$, are statistically independent. The mean and covariance of $\mathbf{e}_t$ are assumed to be known as

$$E[\mathbf{e}_t] = \begin{pmatrix} E[e_t^1] & E[e_t^2] & \cdots & E[e_t^n] \end{pmatrix}',$$

$$\mathrm{Cov}[\mathbf{e}_t] = \begin{pmatrix} \sigma_{1,1}^t & \cdots & \sigma_{1,n}^t \\ \vdots & \ddots & \vdots \\ \sigma_{n,1}^t & \cdots & \sigma_{n,n}^t \end{pmatrix},$$

respectively, where $\sigma_{i,j}^t := E[(e_t^i - E[e_t^i])(e_t^j - E[e_t^j])]$, for $i, j = 1, \cdots, n$. Let $u_t^i$ be the amount of dollars invested in the risky asset $i$, $i = 1, \cdots, n$, and $x_t$ be the wealth level at stage $t$. Thus, at the beginning of the $t$-th period, the amount of wealth allocated in the risk-free asset is $x_t - \sum_{i=1}^n u_t^i$. At the last period $T$, the investor's final wealth is given as follows after deducting the total management fees from his terminal wealth,

$$\hat{x}_T := x_T - \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \cdot M,$$

where $\mathbf{u}_t := \begin{pmatrix} u_t^1 & u_t^2 & \cdots & u_t^n \end{pmatrix}'$, for $t = 0, \cdots, T-1$, and $\delta(\cdot)$ is the indicate function, i.e., $\delta(a) = 0$ if $a$ is a zero vector and $\delta(a) = 1$, otherwise. Let

$$\mathbf{P}_t := \begin{pmatrix} P_t^1, & P_t^2, & \cdots, & P_t^n \end{pmatrix}' = \mathbf{e}_t - r_t \mathbf{1}_n, \tag{4.1}$$

where $\mathbf{1}_n$ is the $n$ dimensional vector with all elements being 1. The multi-period portfolio selection problem with management fees is to seek the optimal investment strategy $\mathbf{u}_t$, for $t = 0, \cdots, T-1$, such that (i) the expected final wealth, $E[\hat{x}_T]$ is maximized, subject to that the variance of the final wealth $\mathrm{Var}[\hat{x}_T]$ is not greater than a given positive level $\sigma$,

$$\begin{aligned} \mathcal{P}_1(\sigma): &\quad \max \quad E[\hat{x}_T] \\ \text{Subject to:} &\quad \mathrm{Var}[\hat{x}_T] \le \sigma, \\ &\quad \hat{x}_T = x_T - \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \cdot M, \\ &\quad x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1, \end{aligned} \tag{4.2}$$

or (ii) the variance of the final wealth, $\mathrm{Var}[\hat{x}_T]$ is minimized, subject to that the expected final wealth, $E[\hat{x}_T]$, is not smaller than a given positive level $\epsilon$,

$$\begin{aligned} \mathcal{P}_2(\epsilon): &\quad \min \quad \mathrm{Var}[\hat{x}_T] \\ \text{Subject to:} &\quad E[\hat{x}_T] \ge \epsilon, \\ &\quad \hat{x}_T = x_T - \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \cdot M, \\ &\quad x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1. \end{aligned} \tag{4.3}$$

Clearly, when $M$ is zero, problem $\mathcal{P}_1(\sigma)$ or $\mathcal{P}_2(\epsilon)$ reduces to the conventional multi-period mean-variance portfolio selection problem investigated in [43]. A large $M$, however, will prevent the investor from investing in risky assets in all time periods or even force the investor to deposit all his money into the risk-free asset for the entire time horizon.

As $E[\mathbf{e}_t\mathbf{e}_t'] = \mathrm{Cov}[\mathbf{e}_t] + E[\mathbf{e}_t]E[\mathbf{e}_t']$, it is reasonable to assume that $E[\mathbf{e}_t\mathbf{e}_t'] \succ 0$, $t = 0, \cdots, T-1$, which further implies the following ([43]),

$$E[\mathbf{P}_t\mathbf{P}_t'] \succ 0, \quad t = 0, \cdots, T-1, \tag{4.4}$$

and

$$1 - E[\mathbf{P}_t']E^{-1}[\mathbf{P}_t\mathbf{P}_t']E[\mathbf{P}_t] > 0, \quad t = 0, \cdots, T-1.$$

Let $\mathcal{I}_t$ be an information set available at time $t$ and $\mathcal{I}_{t-1} \subset \mathcal{I}_t$ for any $t$. A portfolio policy of problem $\mathcal{P}_1(\sigma)$ or $\mathcal{P}_2(\epsilon)$ is a multi-period sequence,

$$\begin{aligned}\pi :=& \{\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_{T-1}\} \\ =& \left\{ \begin{pmatrix} u_0^1 \\ \vdots \\ u_0^n \end{pmatrix}, \begin{pmatrix} u_1^1 \\ \vdots \\ u_1^n \end{pmatrix}, \cdots, \begin{pmatrix} u_{T-1}^1 \\ \vdots \\ u_{T-1}^n \end{pmatrix} \right\},\end{aligned}$$

which, for $t = 0, \dots, T-1$, maps the information set at stage $t$, $\mathcal{I}_t$, into a portfolio decision at period $t$,

$$\begin{pmatrix} u_t^1 \\ u_t^2 \\ \vdots \\ u_t^n \end{pmatrix} = \begin{pmatrix} \mu_t^1(\mathcal{I}_t) \\ \mu_t^2(\mathcal{I}_t) \\ \vdots \\ \mu_t^n(\mathcal{I}_t) \end{pmatrix}. \tag{4.5}$$

In the remaining of this chapter, we use $\pi(\cdot)$ to denote an optimal policy for problem $(\cdot)$. A multi-period portfolio policy, $\pi$, is said to be efficient if there is no other multi-period portfolio policy, $\hat{\pi}$, such that $E[\hat{x}_T]|_{\hat{\pi}} \geq E[\hat{x}_T]|_{\pi}$ and $\mathrm{Var}(\hat{x}_T)|_{\hat{\pi}} \leq \mathrm{Var}(\hat{x}_T)|_{\pi}$ with at least one strict inequality. The entire set of efficient multi-period portfolio policies can be generated by varying $\sigma$ or $\epsilon$ in problem $\mathcal{P}_1(\sigma)$ or $\mathcal{P}_2(\epsilon)$, respectively.

Problem formulations of $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$ motivate us to study the following time cardinality constrained multi-period mean-variance portfolio selection problems (TCCMV) for problem $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$, respectively,

$$\begin{aligned} \mathcal{A}_1(\sigma, s) : \max \quad & E[x_T], \\ \text{Subject to:} & \mathrm{Var}[x_T] \leq \sigma, \\ & \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \leq s, \\ & x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1, \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} \mathcal{A}_2(\epsilon, s): \quad & \min \quad \mathrm{Var}[x_T], \\ \text{Subject to:} \quad & E[x_T] \geq \epsilon, \\ & \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \leq s, \\ & x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1, \end{aligned} \tag{4.7}$$

where $s \in \{0, 1, \cdots, T\}$ is given. Different from the multi-period mean-variance formulation studied in [43], the time cardinality constrained problem imposes a limit on the number of time periods where investing in the risky assets is allowed. We use $v(\cdot)$ to denote the optimal value of problem $(\cdot)$. Note that the following fact is true.

**Lemma 4.1.** *Given $s_1, s_2 \in \{0, 1, \cdots, T\}$ and $s_1 \leq s_2$, it holds true that $v(\mathcal{A}_1(\sigma, s_1)) \leq v(\mathcal{A}_1(\sigma, s_2))$ and $v(\mathcal{A}_2(\epsilon, s_1)) \geq v(\mathcal{A}_2(\epsilon, s_2))$.*

Such results are evident from the fact that enlarging the feasible set never worsens the optimal value. Clearly, the optimal policies of problem $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$ can be obtained by identifying the best $s_1^* \in \{0, 1 \cdots, T\}$ and $s_2^* \in \{0, 1 \cdots, T\}$ as follows,

$$\begin{aligned} s_1^* &:= \arg \max_{s \in \{0, \cdots, T-1\}} v(\mathcal{A}_1(\sigma, s)) - s \cdot M, \\ s_2^* &:= \arg \min_{s \in \{0, \cdots, T-1\}} v(\mathcal{A}_2(\epsilon + s \cdot M, s)). \end{aligned}$$

It is obvious that policy $\pi(\mathcal{A}_1(\sigma, s_1^*))$ solves problem $\mathcal{P}_1(\sigma)$ and $\pi(\mathcal{A}_2(\epsilon + s_2^* \cdot M, s_2^*))$ solves $\mathcal{P}_2(\epsilon)$. Thus, deriving efficient solution procedures for time cardinality constrained portfolio selection problems $\mathcal{A}_1(\sigma, s)$ and $\mathcal{A}_2(\epsilon, s)$ plays a key role in solving multi-period portfolio selection problems with management fees, $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$.

This chapter is organized as follows. We first present in Section 4.2 the analytical solutions of problems $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$, respectively, and report then the detailed derivation of these results in Section 4.3. We give some illustrative examples in Section 4.4 to demonstrate some prominent features of both the problem formulation and the analytical solution. Finally, we conclude this chapter in Section 4.5 with a suggestion of a future research topic.

## 4.2. Analytical Solutions to Problems $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$

We state in this section the analytical optimal portfolio policies for both problems $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$. The detailed derivation of these results will be given in the next section.

Define, for $t = 0, \ldots, T-1$,

$$\mathbf{c}_t := E[\mathbf{P}_t] = E[\mathbf{e}_t] - r_t \mathbf{1}_n, \tag{4.8}$$

$$\mathbf{D}_t := E[\mathbf{P}_t\mathbf{P}_t'] = E[\mathbf{e}_t\mathbf{e}_t'] - r_t(\mathbf{1}_n E[\mathbf{e}_t]' + E[\mathbf{e}_t]\mathbf{1}_n') + r_t^2 \mathbf{1}_n\mathbf{1}_n', \tag{4.9}$$

$$\theta_t := 1 - \mathbf{c}_t'\mathbf{D}_t^{-1}\mathbf{c}_t, \tag{4.10}$$

$$\gamma_t := \prod_{\tau=t}^{T-1} r_\tau. \tag{4.11}$$

Let sequence $\{t_1, t_2, \ldots, t_T\}$ be a permutation of $\{0, 1, \ldots, T-1\}$ such that $\theta_t$, $t = 0, \ldots, T-1$, are arranged in an ascending order,

$$\theta_{t_1} \le \theta_{t_2} \le \cdots \le \theta_{t_T}. \tag{4.12}$$

For any $s \in \{0, 1, \cdots, T\}$, define

$$\xi_1(s) := (\prod_{i=1}^{s} \theta_{t_i})\gamma_0, \tag{4.13}$$

$$\xi_2(s) := (\prod_{i=1}^{s} \theta_{t_i})\gamma_0^2, \tag{4.14}$$

$$\rho(s) := 1 - \prod_{i=1}^{s} \theta_{t_i}. \tag{4.15}$$

**Theorem 4.2.** *The optimal cardinality of problem $\mathcal{P}_1(\sigma)$ is given by*

$$s^* = \arg \max_{s\in\{0,\cdots,T-1\}} (U_1(s) - s\cdot M),$$

*where*

$$U_1(s) := \sqrt{\frac{\sigma\rho(s)}{1-\rho(s)}} + x_0\gamma_0, \quad \textit{for } s \in \{0,\cdots,T\}, \tag{4.16}$$

*while the optimal portfolio policy of $\mathcal{P}_1(\sigma)$ is given as follows:*

***(i)*** *Investment to risk assets is only carried out at time instants, $t \in \{t_1, t_2, \ldots, t_{s^*}\}$, i.e.,*

$$\delta(\boldsymbol{u}_t^*) = \begin{cases} 1 & \textit{if} \quad t = t_i, \ i = 1,\cdots,s^*, \\ 0 & \textit{otherwise.} \end{cases}$$

***(ii)*** *When $\delta(u_t) = 1$,*

$$\boldsymbol{u}_t^* = \mu_t^*(x_t) = -r_t \boldsymbol{D}_t^{-1}\boldsymbol{c}_t x_t + \left(\frac{1+2\omega_1(s^*)\xi_1(s^*)x_0}{2\gamma_{t+1}\omega_1(s^*)(1-\rho(s^*))}\right)\boldsymbol{D}_t^{-1}\boldsymbol{c}_t,$$

*where*

$$\omega_1(s) := \sqrt{\frac{\rho(s)}{4\sigma(1-\rho(s))}}, \quad \textit{for } s \in \{0,\cdots,T\}. \tag{4.17}$$

**Theorem 4.3.** *The optimal cardinality of problem $\mathcal{P}_2(\epsilon)$ is given by*

$$s^* = \begin{cases} 0 & \textit{if } \epsilon \le x_0\gamma_0, \\ \arg\min_{s\in\{1,\cdots,T\}} U_2(s) & \textit{if } \epsilon > x_0\gamma_0, \end{cases}$$

*where*

$$U_2(s) := (\epsilon - x_0\gamma_0 + s\cdot M)^2(\frac{1}{\rho(s)} - 1), \quad \textit{for } s \in \{1,\cdots,T\}, \tag{4.18}$$

*while the optimal portfolio policy of $(\mathcal{P}_2(\epsilon))$ is given as follows:*

***(i)*** *Investment to risk assets is only carried out at time instants, $t \in \{t_1, t_2, \ldots, t_{s^*}\}$, i.e.,*

$$\delta(\boldsymbol{u}_t^*) = \begin{cases} 1 & \textit{if } t = t_i, \ i = 1,\cdots,s^*, \\ 0 & \textit{otherwise.} \end{cases}$$

***(ii)*** *When* $\delta(\boldsymbol{u}_t) = 1$,

$$\boldsymbol{u}_t^* = \mu_t^*(x_t) = -r_t \boldsymbol{D}_t^{-1} \boldsymbol{c}_t x_t + \left( \frac{1 + 2\omega_2(s^*)\xi_1(s^*)x_0}{2\gamma_{t+1}\omega_2(s^*)(1-\rho(s^*))} \right) \boldsymbol{D}_t^{-1} \boldsymbol{c}_t, \tag{4.19}$$

*where*

$$\omega_2(s) := \frac{\rho(s)}{2(1-\rho(s))(\epsilon - x_0\gamma_0 + s \cdot M)}, \quad \textit{for } s \in \{1, \cdots, T\}. \tag{4.20}$$

The optimal multi-period portfolio policies for problems $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$ can be easily implemented by first computing the coefficients in (4.8)-(4.15) off-line and then executing the feedback investment policy on-line. Note that problem $\mathcal{P}_1(\sigma)$ or $\mathcal{P}_2(\epsilon)$ may actually admit multiple optimal strategies and the results presented in this section specify a particular solution.

Note that parameters, $\theta_t$, $t = 0, \ldots, T-1$, essentially measure the market conditions for different time periods. The optimal investment policies in both Theorems 4.2 and 4.3 indicate that these time periods with best market conditions should be assigned priorities to seize (limited) investment opportunities. In this sense, the distribution of the cardinality is independent of the investor's wealth. On the other hand, at a time period where investment is allowed to carry out, the portfolio policy is an affine function of the investor's current wealth.

## 4.3. Derivation of the Analytical Solutions

As stated in Section 4.1, the solution to problem $\mathcal{P}_1(\sigma)$ or $\mathcal{P}_2(\epsilon)$ can be identified respectively from solving problem $\mathcal{A}_1(\sigma, s)$ or $\mathcal{A}_2(\epsilon, s)$ for different $s$. We thus focus first on the solutions of the time cardinality constrained portfolio selection problems. Obviously, when $s = 0$, investing in risky assets is not allowed for both problems, $\mathcal{A}_1(\sigma, s)$ and $\mathcal{A}_2(\epsilon, s)$. Thus, $v(\mathcal{A}_1(\sigma, 0)) = x_0\gamma_0$ and $v(\mathcal{A}_2(\epsilon, 0)) = 0$ when $x_0\gamma_0 \geq \epsilon$. Clearly, if $x_0\gamma_0 < \epsilon$, there is no solution to problem $\mathcal{A}_2(\epsilon, 0)$ and we simply let $v(\mathcal{A}_2(\epsilon, 0)) = +\infty$ in this situation. In the following, we focus on the cases with $s \in \{1, \cdots, T\}$. To solve problem $\mathcal{A}_1(\sigma, s)$ and $\mathcal{A}_2(\epsilon, s)$, we consider the following problem $\mathcal{H}(\omega, s)$ with $\omega > 0$,

$$\begin{aligned} \mathcal{H}(\omega, s): &\ \max\ E[x_T] - \omega \mathrm{Var}[x_T] \\ \text{Subject to: } & x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1, \\ & \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \leq s. \end{aligned}$$

Due to the nonseparability of the variance term mentioned in [43], problem $\mathcal{H}(\omega, s)$ is nonseparable in the sense of dynamic programming. To cope with such a difficulty, we adopt the same idea as in [43] by constructing the following auxiliary problem $\mathcal{A}(\lambda, \omega, s)$ corresponding to problem $\mathcal{H}(\omega, s)$ for a given $s \in \{1, \cdots, T\}$,

$$\begin{aligned} \mathcal{A}(\lambda, \omega, s): &\quad \max\ E[-\omega x_T^2 + \lambda x_T], \\ \text{Subject to: } &\quad x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1, \\ &\quad \sum_{t=0}^{T-1} \delta(\mathbf{u}_t) \leq s. \end{aligned}$$

Note that problem $\mathcal{A}(\lambda, \omega, s)$ is separable in the sense of dynamic programming. Using the same proofs for Theorems 1 and 2 in [43], we can have the following results which reveal the relationship between the solutions of problems $\mathcal{H}(\omega, s)$ and $\mathcal{A}(\lambda, \omega, s)$.

**Theorem 4.4.** *Let $\Pi[\mathcal{H}(\omega, s)]$ and $\Pi[\mathcal{A}(\lambda, \omega, s)]$ be the solution sets of problems $\mathcal{H}(\omega, s)$ and $\mathcal{A}(\lambda, \omega, s)$, respectively.*

*(a) For any $\pi^* \in \Pi[\mathcal{H}(\omega, s)]$, $\pi^* \in \Pi[\mathcal{A}(\lambda^*, \omega, s)]$ with $\lambda^* = 1 + 2\omega E[x_T]|_{\pi^*}$.*

*(b) Assume $\pi^* \in \Pi[\mathcal{A}(\lambda^*, \omega, s)]$. A necessary condition for $\pi^* \in \Pi[\mathcal{H}(\omega, s)]$ is $\lambda^* = 1 + 2\omega E[x_T]|_{\pi^*}$.*

Theorem 4.4 implies that the solution set of problem $\mathcal{H}(\omega, s)$ is a subset of the solution set of problem $\mathcal{A}(\lambda, \omega, s)$. Furthermore, it provides a necessary condition under which a solution of problem $\mathcal{A}(\lambda, \omega, s)$ constitutes an optimal solution of problem $\mathcal{H}(\omega, s)$. As mentioned in the proof of Theorem 2 in [43], we have an alternative way to identify $\lambda^*$ with which $\pi(\mathcal{A}(\lambda^*, \omega, s))$ solves $\mathcal{H}(\omega, s)$. Substituting $\pi(\mathcal{A}(\lambda, \omega, s))$ into the objective function of $\mathcal{H}(\omega, s)$, $v(\mathcal{H}(\omega, s)) \mid_{\pi(\mathcal{A}(\lambda,\omega,s))}$ becomes a function of $\lambda$ only. If $v(\mathcal{H}(\omega, s)) \mid_{\pi(\mathcal{A}(\lambda,\omega,s))}$ is concave with respect to $\lambda$, it is sufficient to show that $\pi(\mathcal{A}(\lambda^*, \omega, s))$ solves problem $\mathcal{H}(\omega, s)$ if

$$\lambda^* = \arg\max_{\lambda} v(\mathcal{H}(\omega, s)) \mid_{\pi(\mathcal{A}(\lambda,\omega,s))} .$$

Thus, we will find out the solution to problem $\mathcal{A}(\lambda, \omega, s)$ first in the following and characterize then the solution of problem $\mathcal{H}(\omega, s)$.

We introduce a nonnegative state variable $y_t$, $t = 0, \cdots, T$, which represents the remaining number of time periods in investing in risky assets and satisfies the following recursion,

$$y_{t+1} = \begin{cases} y_t - \delta(\mathbf{u}_t) & y_t \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the feasible set of $y_t$, $t = 0, \cdots, T$, is given as

$$F_t := \{y_t \in \{0, \cdots, T\} \mid \max\{0, s - t\} \leq y_t \leq s\}. \tag{4.21}$$

One simple fact is that $v(\mathcal{A}(\lambda, \omega, s))$ is a non-decreasing function of $s$, which implies that the cardinality constraint will be binding at least one optimal solution of $\mathcal{A}(\lambda, \omega, s)$. In other words, at least one optimal solution of $\mathcal{A}(\lambda, \omega, s)$ will have its corresponding $y_T$ equal to zero. Define the following recursions, for any $t = T-1, \ \cdots, \ 0$ and $y \in F_t$,

$$\hat{\alpha}_{(t,y)} := \theta_t r_t^2 \alpha_{(t+1,y-1)}, \quad \text{if} \quad y > 0, \tag{4.22}$$

$$\bar{\alpha}_{(t,y)} := r_t^2 \alpha_{(t+1,y)}, \tag{4.23}$$

$$\alpha_{(t,y)} := \begin{cases} \hat{\alpha}_{(t,y)} & \text{if} \quad \hat{\alpha}_{(t,y)} < \bar{\alpha}_{(t,y)} \quad \text{and} \quad y > 0, \\ \bar{\alpha}_{(t,y)} & \text{if} \quad \hat{\alpha}_{(t,y)} \geq \bar{\alpha}_{(t,y)} \quad \text{and} \quad y > 0, \\ \bar{\alpha}_{(t,y)} & \text{if} \quad y = 0, \end{cases} \tag{4.24}$$

with $\alpha_{(T,y)} = \omega$ for all $y \in F_T$.

**Lemma 4.5.** *Given $x_t$ and $y \in F_t$ at stage $t$, the following policy is optimal for problem $\mathcal{A}(\lambda, \omega, s)$.* ***(i)*** *If $y = 0$, $\delta(\boldsymbol{u}_t) = 0$. Otherwise,*

$$\delta(\boldsymbol{u}_t) = \begin{cases} 0 \ or \ 1 & if \ \ x_t = h_t \ \ or \quad \hat{\alpha}_{(t,y)} = \bar{\alpha}_{(t,y)}, \\ 1 & if \ \ x_t \neq h_t \ \ and \ \ \hat{\alpha}_{(t,y)} < \bar{\alpha}_{(t,y)}, \\ 0 & if \ \ x_t \neq h_t \ \ and \ \ \hat{\alpha}_{(t,y)} > \bar{\alpha}_{(t,y)}, \end{cases} \tag{4.25}$$

*where $h_t := \lambda/(2\omega\gamma_t)$.*

***(ii)*** *When $\delta(\boldsymbol{u}_t) = 1$,*

$$\boldsymbol{u}_t^* = \mu_t^*(x_t) = -\boldsymbol{K}_t x_t + \boldsymbol{b}_t, \tag{4.26}$$

*where*

$$\boldsymbol{K}_t := r_t \boldsymbol{D}_t^{-1} \boldsymbol{c}_t, \tag{4.27}$$

$$\boldsymbol{b}_t := \frac{\lambda}{2\omega\gamma_{t+1}} \boldsymbol{D}_t^{-1} \boldsymbol{c}_t. \tag{4.28}$$

Proof. We first define the value function of problem $\mathcal{A}(\lambda, \omega, s)$ as

$$J_t(x_t, y) = \max_{\sum_{\tau=t}^{T-1} \delta(u_\tau) \leq y} E\left[\lambda x_T - \omega x_T^2 | x_t, y\right], \tag{4.29}$$

for $t = 0, \cdots, T-1$ and $y \in F_t$. Due to the independence among $\mathbf{P}_t^\prime s$, the value function satisfies the following recursive relation,

$$J_t(x_t, y) = \max_{u_t,\ y - \delta(\mathbf{u}_t) \geq 0} E\left[J_{t+1}(r_t x_t + \mathbf{P}_t^\prime \mathbf{u}_t, y - \delta(\mathbf{u}_t)) \mid x_t, y\right] \tag{4.30}$$

with the boundary condition $J_T(x_T, y) = -\omega x_T^2 + \lambda x_T$ for all $y \in F_T$. For convenience, let $J_t(x_t, y) = -\infty$ for $y \notin F_t$. We use the mathematical induction method to prove the following claims.

**(i)** The value function takes the following quadratic form for all $t = T-1, \cdots, 0$ and $y \in F_t$,

$$J_t(x_t, y) = -\alpha_{(t,y)} x_t^2 + \beta_{(t,y)} x_t + \eta_{(t,y)}, \tag{4.31}$$

where the coefficient $\alpha_{(t,y)}$ is defined in (4.24).

**(ii)** The maximizer (with respect to $x_t$) and the maximum value of the value function at time $t$, $J_t(x_t, y)$, are both independent of $y$. More specifically, we have

$$x_t^* := \arg\max_{x_t} J_t(x_t, y) = \frac{\lambda}{2\omega\gamma_t}, \quad J_t(x_t^*, y) = \frac{\lambda^2}{4\omega}. \tag{4.32}$$

Our mathematical induction starts at stage $T$ with $J_T(x_T, y) = -\omega x_T^2 + \lambda x_T$ for all $y \in F_T$. We simply define $\alpha_{(T,y)} := \omega$, $\beta_{(T,y)} := \lambda$ and $\eta_{(T,y)} := 0$. The maximum point of $J_T(x_T, y)$ is $x_T^* = \lambda/(2\omega)$ with the corresponding value function given as $J_T(x_T^*, y) = \lambda^2/(4\omega)$.

Assume that both claims (i) and (ii) are true at stage $t = k+1$. We consider now stage $t = k$. If $y \in F_k$ and $y = 0$, then $\delta(\mathbf{u}_k) = 0$ and the value function becomes

$$J_k(x_k, 0) = -\,\alpha_{(k,0)} x_k^2 + \beta_{(k,0)} x_k + \eta_{(k,0)},$$

where $\alpha_{(k,0)} := r_k^2 \alpha_{(k+1,0)}, \beta_{(k,0)} := r_k \beta_{(k+1,0)}, \eta_{(k,0)} := \eta_{(k+1,0)}$. Together with the induction assumption, we can derive

$$\begin{aligned} x_k^* =& \frac{\beta_{(k,0)}}{2\alpha_{(k,0)}} = \frac{\beta_{(k+1,0)}}{2\alpha_{(k+1,0)} r_k} = \frac{\lambda}{2\omega\gamma_k}, \\ J_k(x_k^*, 0) =& \eta_{(k,0)} + \frac{\beta_{(k,0)}^2}{4\alpha_{(k,0)}} = \eta_{(k+1,0)} + \frac{\beta_{(k+1,0)}^2}{4\alpha_{(k+1,0)}} = \frac{\lambda^2}{4\omega}. \end{aligned}$$

If $y > 0$ and $y \in F_k$, then the value function in (4.30) can be expressed as

$$J_k(x_k, y) = \max\{\bar{J}_k(x_k, y), \hat{J}_k(x_k, y)\},$$

where

$$\bar{J}_k(x_k, y) := J_{k+1}(r_k x_k, y), \tag{4.33}$$

$$\hat{J}_k(x_k, y) := \sup_{\mathbf{u}_k \neq 0} V(x_k, y, \mathbf{u}_t), \tag{4.34}$$

$$V(x_k, y, \mathbf{u}_t) := E[J_{k+1}(r_k x_k + \mathbf{P}_k' \mathbf{u}_k, y-1)].$$

Note that the value functions (4.33) and (4.34) are corresponding to $\delta(\mathbf{u}_t) = 0$ and 1, respectively.

From the induction assumption, the expression in (4.33) can be simplified to

$$\bar{J}_k(x_k, y) = -\bar{\alpha}_{(k,y)} x_k^2 + \bar{\beta}_{(k,y)} x_k + \bar{\eta}_{(k,y)}, \tag{4.35}$$

where $\bar{\alpha}_{(k,y)} := \alpha_{(k+1,y)} r_k^2, \bar{\beta}_{(k,y)} := \beta_{(k+1,y)} r_k, \bar{\eta}_{(k,y)} := \eta_{(k+1,y)}$. We can verify that the maximizer and maximum value of $\bar{J}_k(x_k, y)$ are

$$\begin{aligned}
\bar{x}_k^* =& \frac{\bar{\beta}_{(k,y)}}{2\bar{\alpha}_{(k,y)}} = \frac{\beta_{(k+1,y)}}{2\alpha_{(k+1,y)} r_k} = \frac{\lambda}{2\omega\gamma_k}, \\
\bar{J}_k(\bar{x}_k^*, y) =& \bar{\eta}_{(k,y)} + \frac{\bar{\beta}_{(k,y)}^2}{4\bar{\alpha}_{(k,y)}} = \eta_{(k+1,y)} + \frac{\beta_{(k+1,y)}^2}{4\alpha_{(k+1,y)}} = \frac{\lambda^2}{4\omega},
\end{aligned}$$

respectively.

On the other hand, the expression in (4.34) can be written as,

$$\begin{aligned}
\hat{J}_k(x_k, y) &= \sup_{\mathbf{u}_k \neq 0} V(x_k, y, \mathbf{u}_k), \\
&= \sup_{\mathbf{u}_k \neq 0} E[-\alpha_{(k+1,y-1)}(r_k x_k + \mathbf{P}_k' \mathbf{u}_k)^2 \qquad (4.36)\\
&\quad + \beta_{(k+1,y-1)}(r_k x_k + \mathbf{P}_k' \mathbf{u}_k) + \eta_{(k+1,y-1)}] \\
&= - \hat{\alpha}_{(k,y)} x_k^2 + \hat{\beta}_{(k,y)} x_k + \hat{\eta}_{(k,y)}, \qquad (4.37)
\end{aligned}$$

where $\hat{\alpha}_{(k,y)}$, $\hat{\beta}_{(k,y)}$ and $\hat{\eta}_{(k,y)}$ are given, respectively, by,

$$\begin{aligned}
\hat{\alpha}_{(k,y)} =& \alpha_{(k+1,y-1)} r_k^2 \theta_k, \\
\hat{\beta}_{(k,y)} =& \beta_{(k+1,y-1)} r_k \theta_k, \\
\hat{\eta}_{(k,y)} =& \eta_{(k+1,y-1)} + \frac{\beta_{(k+1,y-1)}^2}{4\alpha_{(k+1,y-1)}}(1 - \theta_k).
\end{aligned}$$

There are two different cases in attaining the optimal value in (4.37). When

$$x_k \neq \frac{\beta_{(k+1,y-1)}}{2 r_k \alpha_{(k+1,y-1)}},$$

$V(x_k, y, \mathbf{u}_k)$ achieves its maximum by taking

$$u_k^* = \mu_k^*(x_k) = -\mathbf{K}_k x_k + \mathbf{b}_k, \tag{4.38}$$

where $\mathbf{K}_k$ and $\mathbf{b}_k$ are given in (4.27) and (4.28), respectively. Note that the expression of $\mathbf{b}_k$ is obtained by using the fact $\beta_{(k+1,y-1)}/(2\alpha_{(k+1,y-1)}) = \lambda/(2\omega\gamma_{k+1})$ from the induction assumption. When

$$x_k = \frac{\beta_{(k+1,y-1)}}{2r_k\alpha_{(k+1,y-1)}},$$

$V(x_k, y, \mathbf{u}_k)$ does not have a maximum in its feasible region $\{u_k \mid u_k \neq 0\}$. The supremum of $V(x_k, y, \mathbf{u}_k)$ is achieved at $u_k^* = 0$ with

$$\sup_{u_k} V(x_k, y, u_k) = \lim_{\|u_k\|\to 0} V(x_k, y, u_k) = \eta_{(k+1,y-1)} + \frac{\beta^2_{(k+1,y-1)}}{4\alpha_{(k+1,y-1)}}.$$

Thus, no matter $x_k = \beta_{(k+1,y-1)}/(2r_k\alpha_{(k+1,y-1)})$ holds or not, the maximizer and maximum value of (4.37) have the following unified expressions,

$$\begin{aligned}\hat{x}_k^* =& \frac{\hat{\beta}_{(k,y)}}{2\hat{\alpha}_{(k,y)}} = \frac{\beta_{(k+1,y-1)}}{2\alpha_{(k+1,y-1)}r_k} = \frac{\lambda}{2\omega\gamma_k},\\ \hat{J}_k(\hat{x}_k^*, y) =& \hat{\eta}_{(k,y)} + \frac{\hat{\beta}^2_{(k,y)}}{4\hat{\alpha}_{(k,y)}} = \eta_{(k+1,y-1)} + \frac{\beta^2_{(k+1,y-1)}}{4\alpha_{(k+1,y-1)}} = \frac{\lambda^2}{4\omega},\end{aligned}$$

respectively. One important fact is that both concave quadratic functions $\bar{J}_k(x_k, y)$ and $\hat{J}_k(x_k, y)$ share the same maximum point, as illustrated in Figure 4.1. Thus, for $t = T-1, \cdots, 0$ and $y \in F_t$, whether $\bar{J}_k(x_k, y) \geq \hat{J}_k(x_k, y)$ or not, only depends on a comparison between $\bar{\alpha}_{(k,y)}$ and $\hat{\alpha}_{(k,y)}$. We can conclude now that

$$J_k(x_k, y) = -\alpha_{(k,y)}x_k^2 + \beta_{(k,y)}x_k + \eta_{(k,y)}, \quad y \in F_k,$$

where

$$(\alpha_{(k,y)}, \beta_{(k,y)}, \eta_{(k,y)}) = \begin{cases} (\hat{\alpha}_{(k,y)}, \hat{\beta}_{(k,y)}, \hat{\eta}_{(k,y)}) \text{ or} & \\ (\bar{\alpha}_{(k,y)}, \bar{\beta}_{(k,y)}, \bar{\eta}_{(k,y)}) & \text{if } x_t = h_t \text{ or } \quad \hat{\alpha}_{(k,y)} = \bar{\alpha}_{(k,y)}, \\ (\hat{\alpha}_{(k,y)}, \hat{\beta}_{(k,y)}, \hat{\eta}_{(k,y)}) & \text{if } x_t \neq h_t \text{ and } \hat{\alpha}_{(k,y)} < \bar{\alpha}_{(k,y)}, \\ (\bar{\alpha}_{(k,y)}, \bar{\beta}_{(k,y)}, \bar{\eta}_{(k,y)}) & \text{if } x_t \neq h_t \text{ and } \hat{\alpha}_{(k,y)} > \bar{\alpha}_{(k,y)}. \end{cases} \tag{4.39}$$

Furthermore, the optimal policy is given in (4.26) when $(\alpha_{(k,y)}, \beta_{(k,y)}, \eta_{(k,y)}) = (\hat{\alpha}_{(k,y)}, \hat{\beta}_{(k,y)}, \hat{\eta}_{(k,y)})$. The proof is completed. □

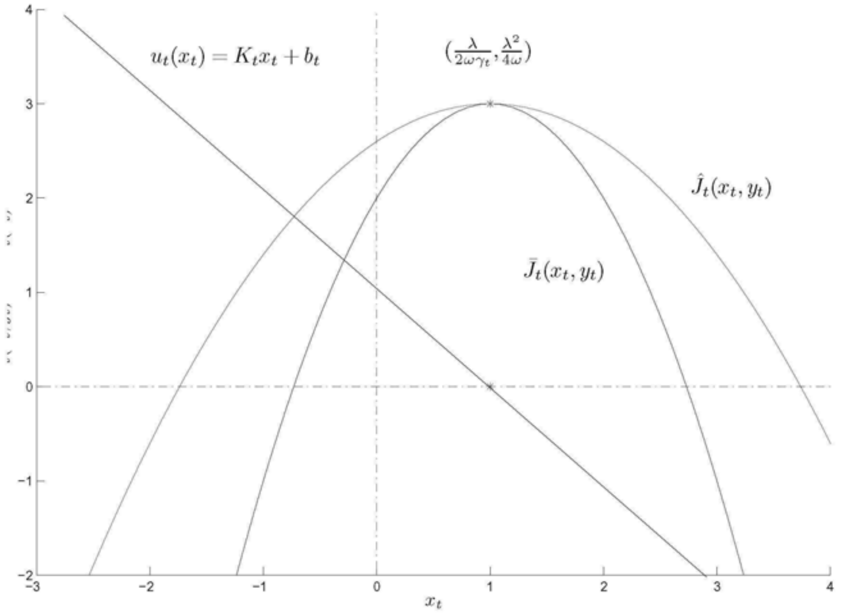

Figure 4.1: The function of $J_t(x_t, y_t)$ and $u_t(x_t)$

Clearly, at stage $t$ and for any $y \in F_t$, whether $\delta(\mathbf{u}_t) = 1$ or 0 only depends on coefficient $\alpha_{(t,y)}$, which can be computed off-line using (4.39). Most interestingly, this result is almost $x_t$-independent, except that a freedom exists at $x_t = h_t$ to take either $\delta(\mathbf{u}_t) = 1$ or 0. We emphasize here that Lemma 4.5 enables us to identify possibly multiple optimal portfolio policies for auxiliary problem $\mathcal{A}(\lambda, \omega, s)$, including solutions at which the time cardinality constraint is either binding or not. As we stated before, as $v(\mathcal{A}(\lambda, \omega, s))$ is a non-decreasing function of $s$, the cardinality constraint will be binding for at least one optimal solution to $\mathcal{A}(\lambda, \omega, s)$, i.e., the corresponding $y_T$ is equal to zero. Based on Lemma 4.5, we will specify in the following the optimal investment policy of the auxiliary problem $\mathcal{A}(\lambda, \omega, s)$ at which the time cardinality constraint is binding.

**Theorem 4.6.** ***(Solution of problem*** $\mathcal{A}(\lambda, \omega, s)$*) The following policy*

*$\pi(\mathcal{A}(\lambda,\omega,s))$ is optimal for problem $\mathcal{A}(\lambda,\omega,s)$,*

$$\delta(\boldsymbol{u}_t)=\begin{cases} 1 & \text{if} \quad t=t_i, \quad \text{for } i=1,\cdots,s, \\ 0 & \text{if} \quad t\neq t_i, \quad \text{for } i=1,\cdots,s, \end{cases} \tag{4.40}$$

*where $t_i$ is defined in (4.12), and the investment policy is given, when $\delta(\boldsymbol{u}_t)=1$, as*

$$\boldsymbol{u}_t^*=\mu_t^*(x_t)=-\boldsymbol{K}_t x_t+\boldsymbol{b}_t, \tag{4.41}$$

*where $\boldsymbol{K}_t$ and $\boldsymbol{b}_t$ are defined in (4.27) and (4.28), respectively.*

Proof. From Lemma 4.5, we know that the optimal policy of problem $\mathcal{A}(\lambda,\omega,s)$ is determined by calculation of $\alpha_{(t,y)}$, $y\in F_t$, from recursions (4.22), (4.23) and (4.24). Define $\Theta(t):=\{\theta_t,\theta_{t+1},\cdots,\theta_{T-1}\}$. As we have already recognized that at least one optimal solution to $\mathcal{A}(\lambda,\omega,s)$ consumes all time cardinality, we claim that one $\alpha_{(t,y)}$ takes the following form,

$$\alpha_{(t,y)}=\min_{\theta^{(1)},\theta^{(2)},\cdots,\theta^{(y)}\in\Theta(t)}\{\prod_{i=1}^{y}\theta^{(i)}\}\gamma_t^2\omega,$$

where $\theta^{(j)}$ is the element of $\Theta(t)$, for any $y\geq 1$, $y\in F_t$ and $t=0,\cdots,T-2$.

Note that $\alpha_{(t,y)}$ is computed in a backward fashion and the iteration stops at $\alpha_{(0,s)}$. The above claim is now evident from a comparison between the recursions in (4.22) and (4.23): $\theta_t\in\Theta(0)$, is brought into the expression of $\alpha_{(0,s)}$ only when $\sigma(\mathbf{u}_t)=1$, which leads to a conclusion that whether $\theta_t$ appears in the expression of $\alpha_{(0,s)}$ or not indicates whether or not we should do the investment in risky assets at stage $t$.

The above claim is proved by the induction method. When $t=T-2$, from (4.22) and (4.23), $\alpha_{(T-2,1)}$ and $\alpha_{(T-2,2)}$ can be computed explicitly as,

$$\begin{aligned}\alpha_{(T-2,1)}=&(\min\{\theta_{T-1},\theta_{T-2}\})r_{T-1}r_{T-2}\omega,\\ \alpha_{(T-2,2)}=&\theta_{T-1}\theta_{T-2}r_{T-1}r_{T-2}\omega.\end{aligned}$$

We assume that the claim is true for $t = k + 1$. At stage $t = k$, for $y \in F_k$ and $y > 1$, (4.22), (4.23) and (4.24) imply that

$$\begin{aligned}\alpha_{(k,y)} =& \min\{\alpha_{(k+1,y)}, \theta_k \alpha_{(k+1,y-1)}\} r_k^2 \\ =& \min\left\{ \min_{\theta^{(1)},\cdots,\theta^{(y)} \in \Theta(k+1)} \{\prod_{i=1}^{y} \theta^{(i)}\}, \theta_k ( \min_{\theta^{(1)},\cdots,\theta^{(y-1)} \in \Theta(k+1)} \{\prod_{i=1}^{y-1} \theta^{(i)}\}) \right\} \gamma_{k+1}^2 r_k^2 \omega, \\ =& \min_{\theta^{(1)},\cdots,\theta^{(y)} \in \Theta(k)} \{\prod_{i=1}^{y} \theta^{(i)}\} \gamma_k^2 \omega.\end{aligned}$$

Repeating this iterative process until $k = 0$ yields

$$\alpha_{(0,s)} = \min_{\theta^{(1)},\cdots,\theta^{(s)} \in \Theta(0)} \{\prod_{i=1}^{s} \theta^{(i)}\} \cdot (\prod_{\tau=0}^{T-1} r_\tau^2) \omega.$$

which completes the proof of Theorem 4.6. □

Using Theorem 4.6 as a solution scheme does not generate solutions of $\mathcal{A}(\lambda, \omega, s)$ at which the cardinality constraint is not binding. Fortunately, the solution algorithms for $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\epsilon)$ utilize the solution scheme specified in Theorem 4.6 for all $s = 0, \ldots, T$. Thus, no solution will be missed in the solution process.

**Theorem 4.7.** ***(Solution of problem*** $\mathcal{H}(\omega, s)$***)*** *The optimal policy* $\pi(\mathcal{A}(\lambda^*(s), \omega, s))$ *solves problem* $\mathcal{H}(\omega, s)$ *with*

$$\lambda^*(s) = \frac{1 + 2\omega\xi_1(s) x_0}{1 - \rho(s)}, \tag{4.42}$$

*the expected value and the variance of the terminal wealth under policy* $\pi(\mathcal{H}(\omega, s))$ *are given, respectively, by*

$$E[x_T(\omega, s)] = \frac{\rho(s)}{2\omega(1 - \rho(s))} + x_0\gamma_0, \tag{4.43}$$

$$Var[x_T(\omega, s)] = \frac{\rho(s)}{4\omega^2(1 - \rho(s))}, \tag{4.44}$$

*and the efficient frontier is expressed as*

$$Var[x_T(\omega, s)] = \frac{1 - \rho(s)}{\rho(s)} (E[x_T(\omega, s)] - x_0\gamma_0)^2, \ \textit{for } E[x_T(\omega, s)] > x_0\gamma_0. \tag{4.45}$$

Proof. Implied by Theorem 4.4, the optimal policy of problem $\mathcal{H}(\omega, s)$ takes the same form as $\pi(\mathcal{A}(\lambda, \omega, s))$, the policy for $\mathcal{A}(\lambda, \omega, s)$. Our target shifts now to identify an optimal $\lambda^*$ such that $\pi(\mathcal{A}(\lambda^*, \omega, s))$ solves problem $\mathcal{H}(\omega, s)$. Substituting the optimal policy $\pi^*(\mathcal{A}(\lambda, \omega, s))$ specified in Theorem 4.6 into the dynamics of the wealth yields

$$x_{t+1}(\lambda, \omega, s) = \begin{cases} r_t x_t(\lambda, \omega, s) & \text{if } t = t_i, s < i \leq T, \\ (r_t - \mathbf{P}_t' \mathbf{K}_t) x_t(\lambda, \omega, s) + \mathbf{P}_t' \mathbf{b}_t & \text{if } t = t_i, 1 \leq i \leq s, \end{cases} \tag{4.46}$$

where $t_i$ is defined in (4.12). Taking the expectation on both sides of (4.46) while noticing the independency between $\mathbf{P}_t$ and $x_t$, we have the following recursive expression for the expected wealth under the optimal policy $\pi^*(\mathcal{A}(\lambda, \omega, s))$,

$$\begin{aligned} & E[x_{t+1}(\lambda, \omega, s)] \\ & = \begin{cases} r_t E[x_t(\lambda, \omega, s)] & \text{if } t = t_i, s < i \leq T, \\ r_t \theta_t E[x_t(\lambda, \omega, s)] \lambda (1 - \theta_t)/(2\omega\gamma_{t+1}) & \text{if } t = t_i, 1 \leq i \leq s. \end{cases} \end{aligned} \tag{4.47}$$

Similarly, squaring both sides of (4.46) yields,

$$\begin{aligned} & x_{t+1}^2(\lambda, \omega, s) \\ & = \begin{cases} r_t^2 x_t^2(\lambda, \omega, s) & \text{if } t = t_i, s < i \leq T, \\ (r_t^2 - 2r_t \mathbf{P}_t' \mathbf{K}_t + \mathbf{K}_t' \mathbf{P}_t \mathbf{P}_t' \mathbf{K}_t) x_t^2(\lambda, \omega, s) & \\ \quad + 2(r_t - \mathbf{P}_t' \mathbf{K}_t) \mathbf{P}_t' \mathbf{b}_t x_t(\lambda, \omega, s) + \mathbf{b}_t' \mathbf{P}_t \mathbf{P}_t' \mathbf{b}_t & \text{if } t = t_i, 1 \leq i \leq s. \end{cases} \end{aligned} \tag{4.48}$$

Taking the expectation on both sides of (4.48) leads to the following recursive expression for the second moment of wealth $x_t$ under the optimal policy $\pi^*(\mathcal{A}(\lambda, \omega, s))$,

$$\begin{aligned} & E[x_{t+1}^2(\lambda, \omega, s)] \\ & = \begin{cases} r_t^2 E[x_t^2(\lambda, \omega, s)] & \text{if } t = t_i, s < i \leq T, \\ r_t^2 \theta_t E[x_t^2(\lambda, \omega, s)] + \lambda^2 (1 - \theta_t)/(4\omega^2 \gamma_{t+1}^2) & \text{if } t = t_i, 1 \leq i \leq s. \end{cases} \end{aligned} \tag{4.49}$$

Solving both the recursive expressions in (4.47) and (4.49) gives rise to the first and second moments of the terminal wealth under the policy $\pi(\mathcal{A}(\lambda, \omega, s))$,

$$\begin{aligned} E[x_T(\lambda, \omega, s)] &= \xi_1(s)x_0 + \frac{\lambda}{2\omega}\rho(s), && (4.50) \\ E[x_T^2(\lambda, \omega, s)] &= \xi_2(s)x_0^2 + \frac{\lambda^2}{4\omega^2}\rho(s), && (4.51) \end{aligned}$$

where $\xi_1(s)$, $\xi_2(s)$ and $\rho(s)$ are defined in (4.13), (4.14) and (4.15), respectively. Thus, the variance of the terminal wealth under the policy $\pi(\mathcal{A}(\lambda, \omega, s))$ can be expressed as

$$\begin{aligned} \text{Var}[x_T(\lambda, \omega, s)] =& E[x_T^2(\lambda, \omega, s)] - (E[x_T(\lambda, \omega, s)])^2 \\ =& (\frac{\lambda}{2\omega})^2(\rho(s) - \rho^2(s)) - \frac{\lambda}{\omega}\xi_1(s)\rho(s)x_0 + x_0^2(\xi_2(s) - \xi_1^2(s)). \end{aligned} \tag{4.52}$$

It is clear that the expected terminal wealth $E[x_T]$ is an increasing linear function of $\lambda$ and the variance in (4.52) is a quadratic function of $\lambda$. Thus, from (4.50) and (4.52), the objective function of $\mathcal{H}(\omega, s)$ under policy $\pi(\mathcal{A}(\lambda, \omega, s))$ can be written as,

$$\begin{aligned} v(H(\omega, s)) \mid_{\pi(A(\lambda,\omega,s))} =& \frac{\lambda^2}{4\omega}[\rho^2(s) - \rho(s)] + \lambda[\frac{\rho(s)}{2\omega} + \xi_1(s)\rho(s)x_0] \\ &+ \xi_1(s)x_0 - \omega x_0^2(\xi_2(s) - \xi_1^2(s)). \end{aligned} \tag{4.53}$$

The expression of $\rho(s)$ in (4.15) implies that $0 < \rho < 1$ and $(\rho^2 - \rho) < 0$, which in turn implies that $v(\mathcal{A}(\lambda, \omega, s))$ is quadratic concave function of $\lambda$. Differentiating (4.53) with respect to $\lambda$ yields,

$$\frac{dv(H(\omega, s)) \mid_{\pi(A(\lambda,\omega,s))}}{d\lambda} = \frac{\lambda}{2\omega}(\rho^2(s) - \rho(s)) + (\frac{\rho(s)}{2\omega} + \xi_1(s)\rho(s)x_0). \tag{4.54}$$

The optimal $\lambda^*(s)$ given in (4.42) is obtained then by solving $\frac{dv(\mathcal{H}(\omega,s))|_{\pi(\mathcal{A}(\lambda,\omega,s))}}{d\lambda} = 0$. Substituting $\lambda^*(s)$ into the optimal policy $\pi(\mathcal{A}(\lambda, \omega, s))$ and the expressions in (4.50) and (4.52) gives rise to the optimal policy of problem $\mathcal{H}(\omega, s)$ and the efficient pair of the expected value and the variance of the terminal wealth given in (4.43) and (4.44), respectively. Note that the relationships

$$\xi_2(s) = \frac{\xi_1(s)^2}{1 - \rho(s)}$$

and

$$\frac{\xi_1(s)}{1-\rho(s)} = \gamma_0$$

are used in the above derivation. The efficient frontier is achieved by eliminating $\omega$ from (4.43) and (4.44). □

**Theorem 4.8.** ***(Solution of problems*** $\mathcal{A}_1(\sigma, s)$ ***and*** $\mathcal{A}_2(\epsilon, s)$***)*** *For any $s \in \{1, \cdots, T\}$, policies $\pi(\mathcal{H}(\omega_1, s))$ and $\pi(H(\omega_2, s))$ solve respectively problems $\mathcal{A}_1(\sigma, s)$ and $\mathcal{A}_2(\epsilon, s)$ with $\epsilon > x_0\gamma_0$, where $\omega_1$ and $\omega_2$ are, respectively, the non-negative roots of the following equations,*

$$Var[x_T(\omega_1, s)] = \sigma, \tag{4.55}$$

$$E[x_T(\omega_2, s)] = \epsilon. \tag{4.56}$$

*Furthermore, the optimal values of problem $\mathcal{A}_1(\omega, s)$ and $\mathcal{A}_2(\omega, s)$ are given, respectively, by*

$$\begin{aligned} v(\mathcal{A}_1(\omega, s)) =& E[x_T(\omega_1, s)], \\ v(\mathcal{A}_2(\omega, s)) =& Var[x_T(\omega_2, s)]. \end{aligned}$$

*When $\epsilon \leq x_0\gamma_0$, the optimal policy of problem $\mathcal{A}_2(\epsilon, s)$ is $\delta(\boldsymbol{u}_t) = 0$ for $t = 0, \cdots, T-1$ with $v(\mathcal{A}_2(\epsilon, s)) = 0$.*

Proof. For any fixed $s \in \{1, \cdots, T\}$, we introduce Lagrangian multiplier $\omega \geq 0$ for problem $\mathcal{A}_1(\sigma, s)$,

$$\begin{aligned} \mathcal{L}(\sigma, \omega, s) \quad & \max \quad E[x_T] + \omega(\sigma - \mathrm{Var}[x_T]), \\ \text{Subject to}: \quad & \sum_{t=0}^{T-1} \delta(u_t) \leq s, \\ & x_{t+1} = r_t x_t + \mathbf{P}_t' \mathbf{u}_t, \quad t = 0, \cdots, T-1. \end{aligned}$$

By weak duality, it is clear that $v(\mathcal{L}(\sigma, \omega, s)) \geq v(\mathcal{A}_1(\sigma, s))$. On the other hand, note that solving problem $\mathcal{L}(\sigma, \omega, s)$ is equivalent to solving problem $\mathcal{H}(\omega, s)$, since $\sigma$ and $\omega$ are both constants. Thus, policy $\pi(\mathcal{H}(\omega, s))$ also solves problem

$\mathcal{L}(\sigma, \omega, s)$. Under the optimal policy $\pi(\mathcal{H}(\omega, s))$, the expected value and the variance, $E[x_T(\omega, s)]$ and $\mathrm{Var}[x_T(\omega, s)]$, are given in (4.43) and (4.44), respectively. The strong duality, $v(\mathcal{L}(\sigma, \omega, s)) = v(\mathcal{A}_1(\sigma, s))$ holds once the feasible condition $\mathrm{Var}[x_T(\omega, s)] = \sigma$ is satisfied. The existence of an optimal $\omega$ is guaranteed for all $\sigma > 0$, as evidenced from the expression in (4.44). Once $\omega_1$ solves equation $\mathrm{Var}[x_T(\omega, s)] = \sigma$, $v(\mathcal{A}_1(\sigma, s)) = E[x_T(\omega_1, s)]$.

When $\epsilon > x_0\gamma_0$, the similar argument also applies for problem $\mathcal{A}_2(\epsilon, s)$. When $\epsilon \leq x_0\gamma_0$, the constraint $E[x_T] \geq \epsilon$ will, however, not be binding. Clearly, solution $\pi^* := \{\delta(\mathbf{u}_t) = 0, t = 0, \cdots, T-1\}$ achieves optimality in this situation as $E[x_T]|_{\pi^*} = \gamma_0 x_0 \geq \epsilon$ and $\mathrm{Var}[x_T]|_{\pi^*} = 0$. □

**Proof of Theorems 4.2 and 4.3.** Clearly, the optimality of problem $\mathcal{P}_1(\sigma)$ is attained by policy $\pi^*(\mathcal{A}_1(\sigma, s_1^*))$ with

$$s_1^* := \arg \max_{s \in \{0, \cdots, T\}} v(\mathcal{A}_1(\sigma, s)) - s \cdot M,$$

where $v(\mathcal{A}_1(\sigma, s))$ is expressed as in (4.16), and, from Theorems 4.7 and 4.8, problem $\mathcal{A}_1(\sigma, s)$ is solved by policy $\pi(\mathcal{H}(\omega_1, s))$ with $\omega_1$ being given in (4.17). Similarly, the optimal policy of problem $\mathcal{P}_2(\epsilon)$ is supplied by $\pi(\mathcal{A}_2(\epsilon + s_2^* \cdot M, s_2^*))$ with

$$s_2^* := \begin{cases} \arg\min_{s \in \{1, \cdots, T\}} v(\mathcal{A}_2(\epsilon + s \cdot M, s)) & \text{if } \epsilon > x_0\gamma_0, \\ 0 & \text{if } \epsilon \leq x_0\gamma_0, \end{cases}$$

where $v(\mathcal{A}_2(\epsilon + s \cdot M, s))$ is expressed as in (4.18) and, from Theorem 4.7 and 4.8, problem $\mathcal{A}_2(\epsilon + s \cdot M, s)$ is solved by policy $\pi(\mathcal{H}(\omega_2, s))$ with $\omega_2$ being given in (4.20). □

## 4.4. Illustrative Examples

**Example 4.1.** An investor with initial wealth of $x_0 = 100$ units enters a market consisting of 4 risky assets and one risk-free asset during a time horizon of 6 consecutive time periods. The management fee for each time period is $M = 0.5$. The expected return and the covariance of the risky assets for these 6 time periods are predicted, respectively, as,

$$\begin{aligned}
&E[\mathbf{e}_0]' = (1.049,\ 1.038,\ 1.055,\ 1.032)\,,\ E[\mathbf{e}_1]' = (1.046,\ 1.038,\ 1.060,\ 1.050)\,,\\
&E[\mathbf{e}_2]' = (1.064,\ 1.039,\ 1.055,\ 1.052)\,,\ E[\mathbf{e}_3]' = (1.056,\ 1.039,\ 1.039,\ 1.043)\,,\\
&E[\mathbf{e}_4]' = (1.058,\ 1.043,\ 1.032,\ 1.036)\,,\ E[\mathbf{e}_5]' = (1.056,\ 1.048,\ 1.030,\ 1.034)
\end{aligned}$$

and

$$\mathrm{Cov}[\mathbf{e}_0] = 10^{-3} \times \begin{pmatrix} 29 & 9 & 14 & 7 \\ 9 & 14 & 15 & 8 \\ 14 & 15 & 29 & 9 \\ 7 & 8 & 9 & 9 \end{pmatrix}, \mathrm{Cov}[\mathbf{e}_1] = 10^{-3} \times \begin{pmatrix} 28 & 9 & 14 & 12 \\ 9 & 14 & 15 & 15 \\ 14 & 15 & 29 & 17 \\ 12 & 15 & 17 & 28 \end{pmatrix},$$

$$\mathrm{Cov}[\mathbf{e}_2] = 10^{-3} \times \begin{pmatrix} 32 & 9 & 13 & 13 \\ 9 & 14 & 14 & 15 \\ 13 & 14 & 24 & 15 \\ 13 & 15 & 15 & 28 \end{pmatrix}, \mathrm{Cov}[\mathbf{e}_3] = 10^{-3} \times \begin{pmatrix} 25 & 8 & 10 & 8 \\ 8 & 13 & 11 & 11 \\ 10 & 11 & 18 & 10 \\ 8 & 11 & 10 & 14 \end{pmatrix},$$

$$\mathrm{Cov}[\mathbf{e}_4] = 10^{-3} \times \begin{pmatrix} 25 & 8 & 9 & 7 \\ 8 & 13 & 11 & 9 \\ 9 & 11 & 15 & 7 \\ 7 & 9 & 7 & 10 \end{pmatrix}, \mathrm{Cov}[\mathbf{e}_5] = 10^{-3} \times \begin{pmatrix} 23 & 9 & 8 & 5 \\ 9 & 19 & 11 & 9 \\ 8 & 11 & 11 & 5 \\ 5 & 9 & 5 & 6 \end{pmatrix}.$$

The returns of the risk-free asset at different time periods are $r_0 = 1.015$, $r_1 = 1.015$, $r_2 = 1.020$, $r_3 = 1.030$, $r_4 = 1.025$, $r_5 = 1.025$. The investor considers portfolio selection problem $\mathcal{P}_1(30)$, i.e., to maximize his expected terminal wealth with a constraint that the variance of his terminal wealth does not exceed 30.

| $s$ | $\rho$ | $U_1(s)$ | $s \cdot M$ | $U_1(s) - s \cdot M$ |
|---|---|---|---|---|
| 0 | 0 | 113.714 | 0 | 113.714 |
| 1 | 0.077 | 115.297 | 0.5 | 114.797 |
| 2 | 0.145 | 115.974 | 1.0 | 114.974 |
| $3^*$ | 0.205 | 116.499 | 1.5 | 114.999 |
| 4 | 0.257 | 116.938 | 2.0 | 114.938 |
| 5 | 0.303 | 117.329 | 2.5 | 114.829 |
| 6 | 0.324 | 117.510 | 3.0 | 114.510 |

Table 4.1: Computational results of Example 4.1

It can be verified that $\theta_0 = 0.9378$, $\theta_1 = 0.9259$, $\theta_2 = 0.9230$, $\theta_3 = 0.9699$, $\theta_4 = 0.9347$ and $\theta_5 = 0.9299$. The solution to this example problem generated from implementing Theorem 4.2 is presented in Table 4.1, while the optimal cardinality is $s^* = 3$.

The optimal investment strategy is given as follows with the maximum expected return of $\hat{x}_6$ achieved at 116.499, $\mathbf{u}_0 = \mathbf{0}$, $\mathbf{u}_3 = \mathbf{0}$, $\mathbf{u}_4 = \mathbf{0}$,

$$\mathbf{u}_1 = \begin{pmatrix} -0.397 \\ 0.939 \\ -1.319 \\ -0.717 \end{pmatrix} x_0 + \begin{pmatrix} 45.089 \\ -106.666 \\ 149.836 \\ 81.458 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} -0.890 \\ 1.219 \\ -1.194 \\ -0.653 \end{pmatrix} x_2 + \begin{pmatrix} 102.600 \\ -140.565 \\ 137.716 \\ 75.339 \end{pmatrix},$$

$$\mathbf{u}_5 = \begin{pmatrix} -1.353 \\ -2.203 \\ 2.230 \\ 0.952 \end{pmatrix} x_2 + \begin{pmatrix} 168.000 \\ 273.489 \\ -276.845 \\ -118.149 \end{pmatrix}.$$

**Example 4.2.** We still use the same data as in Example 4.1 and vary now the value of $\sigma$ to trace out the efficient frontier of the terminal wealth $\hat{x}_6$. See Figure 4.2. Note that when $M = 0$ the efficient frontier is a straight line, which is identical to the efficient frontier presented in [43]. When $M > 0$, such a line is twisted because of the management fee $M$. Given the same risk level,

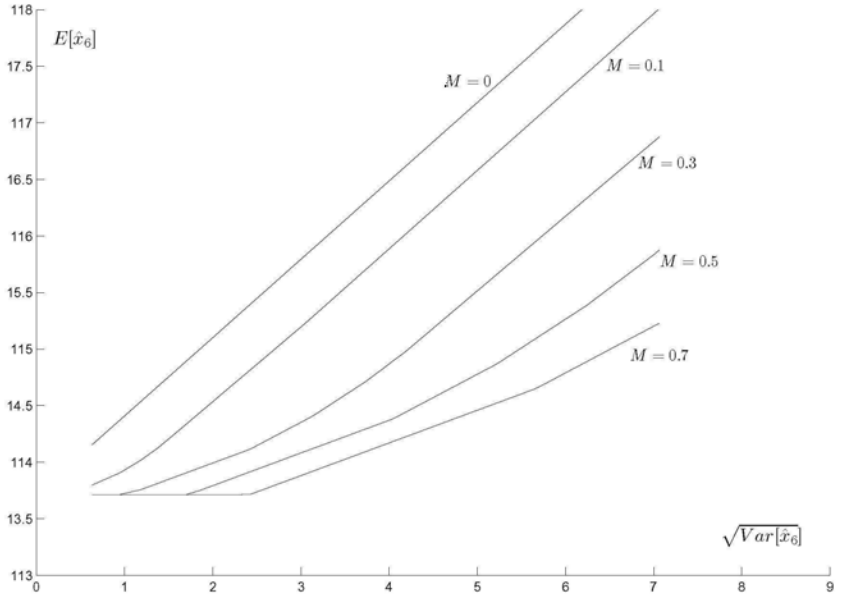

Figure 4.2: The efficient frontiers of $\hat{x}_6$

i.e., the same upper bound of the variance $\sigma$, the expected return decreases when the management fee increases. In Figure 4.3, we consider a pure TCCMV problem of Example 4.1 and different efficient frontiers are showed for different $s \in \{1, \cdots, 6\}$. The slopes of these lines are computed using (4.45) in Theorem 4.7.

**Remark 4.1.** In real implementation, we may adopt the following rolling horizon strategy. For a given $T$-period problem, if the optimal policy at time 0 advices you to invest in the first $l$ periods, you follow. Otherwise, the market situation is not good enough for you to plunge into at time 0, you wait, perform the analysis of a $(T-1)$-period problem at time $t = 1$ with updated information, and decide accordingly. In summary, when you reevaluate your investment strategy at time $t$, you should invest at time $t$ for a time horizon of $l$ periods, only if the analytical solution of the remaining $(T-t)$-period problem suggests you to invest in the first $l$ periods starting from time $t$ and advices you not to invest in the $(l+1)$th

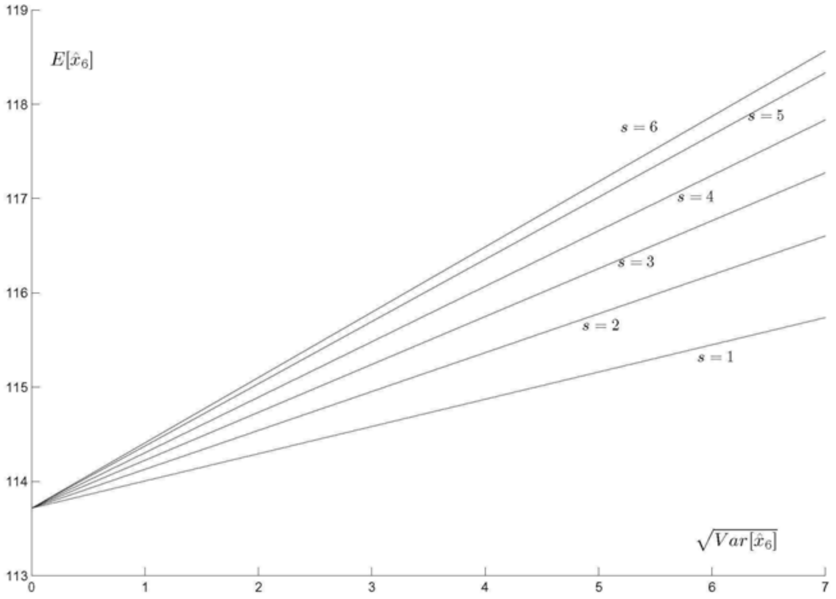

Figure 4.3: The efficient frontier of $\hat{x}_6$ for different $s$

period starting from time $t$.

## 4.5. Conclusion

Motivated by a common observation from financial markets: The set-up type of management fees in asset management often induces situations where investors do not invest in risky assets in all time periods, we have presented an analytical solution approach for the time cardinality constrained mean-variance dynamic portfolio selection problem (TCCMV) and the dynamic mean-variance portfolio selection problem with management fees. Interestingly, the analytical solution of (TCCMV) reveals that the best distribution of the cardinality of the investment time periods is entirely decided by the market parameter $\theta_t$. Varying the cardinality $s$, the solution of the dynamic mean-variance portfolio selection problem with management fees can be generated.

One interesting future research topic is to investigate a portfolio selection problem where investment policies are confined to be buy-and-hold policies and a management fee will be charged only when policy is recalculated.

# CHAPTER 5

# On Geometric Approach of Solving CCQO

## 5.1. Introduction

### 5.1.1. Problem Formulation

In this chapter, we consider exact solution approaches for the following cardinality constrained quadratic optimization problem (CCQO),

$$\begin{aligned} \mathcal{G}_s(D,d): \ & \min_{u} \quad f(x) = \frac{1}{2}u'Du + d'u, \\ \text{Subject to: } & u \in \Delta(s) := \left\{ u \in \mathbb{R}^T \mid \sum_{t=1}^{T} \delta(u_i) \leq s < T \right\}, \end{aligned} \tag{5.1}$$

where $u' := (u_1, u_2, \cdots, u_T)$, $d \in \mathbb{R}^T$ and $D \in \mathbb{S}^T_+$. The cardinality constraint (5.1) is also known as the $\mathbf{L}_0$ norm constraint, i.e., $\sum_{i=1}^{T} |\text{sign}(u_i)| \leq s$. However, we still prefer the notation $\delta(\cdot)$ to make a consistency with the previous chapters.

To ensure that $v(\mathcal{G}_s(D,d)) > -\infty$, we assume the following condition for problem $\mathcal{G}_s(D,d)$.

**Assumption 5.1.** [4] There exists $\hat{u} \in \mathbb{R}^T$such that $D\hat{u} + d = 0$.

Clearly, problem $\mathcal{G}_s(D,d)$ is NP-hard in general. This can be seen from the basic reduction method [29] [21]. We construct the following problem,

$$\hat{\mathcal{G}}_s: \min_{u \in \mathbb{R}^T} \left\{ \hat{f} := \|u - \mathbf{1}_T\|_2^2 + \|Au\|_2^2 \mid u \in \Delta(s) \right\},$$

where $A \in \mathbb{R}^{l \times T}$ and $l \leq T$. Any instance of problem $\hat{\mathcal{G}}_s$ is polynomially reducible to an instance of problem $\mathcal{G}_s(D, d)$. That is to say, problem $\hat{\mathcal{G}}_s$ is no more difficult than $\mathcal{G}_s(D, d)$. Since $u \in \Delta(s)$, minimizing the first term of $\hat{\mathcal{G}}_s$ enforces $u_i$ taking either 0 or 1 for $i = 1, \cdots, T$. More specifically, at least $T - s$ of $u_i$ are zero. Thus, the optimal value of problem $\hat{\mathcal{G}}_s$ is lower bounded, i.e., $v(\hat{\mathcal{G}}_s) \geq T - s$. Answering the question "whether equality $v(\hat{\mathcal{G}}_s) = T - s$ holds or not" turns out to find the integer (binary) solution of linear systems $Au = 0$ such that $u \in \{0, 1\}^T$ and $\sum_{i=1} u_i \leq s$ which is a well known NP-complete decision problem [29].

Although problem $\mathcal{G}_s(D, d)$ is generally hard to solve, it becomes an easy problem when the rank of matrix $D$ is low.

**Lemma 5.1.** *If* ***rank***$(D) = r$ *with* $r \leq s$, *then* $v(\mathcal{G}_s(D, d)) = -\frac{1}{2} d' D^\dagger d$.

Proof. Under the assumption (5.1), problem $\mathcal{G}_s(D, d)$ is lower bounded, i.e., $v(\mathcal{G}_s(D, d)) \geq -\frac{1}{2} d' D^\dagger d$. The equality holds when $u^*$ solves the equation $Du^* + d = 0$. Note that any $u^* \in \{u | Du + d = 0\}$ can be expressed as $u^* = \sum_{i=1}^{T-r} \zeta_i h_i - D^\dagger d$, where $h_i$, $i = 1, \cdots, T - r$, span the null space of $D$. Note that $u \in \Delta(s)$ if at least $T - s$ elements of $u$ are zeros. Thus, when $r \leq s$, it holds that $T - r \geq T - s$ and we can always find some $\zeta_i^*$, for $i = 1, \cdots, T - r$ such that $T - r$ elements of $u^*$ are zero and hence $u^* \in \Delta(s)$. □

In the following, we mainly consider the case where **rank**$(D) = T$. Without loss of generality, we also assume that the eigenvalues of $D$ are arranged in an ascending order,

$$0 < \lambda_1^D \leq \lambda_2^D \leq \cdots \leq \lambda_T^D. \tag{5.2}$$

Furthermore, let

$$l := D^{-1} d \text{ and } C := -\frac{1}{2} l' d. \tag{5.3}$$

In this chapter, we use the following notions. Let $A \in \mathbb{S}^T_{++}$, $b \in \mathbb{R}^T$ and $z \in \mathcal{Z}(s)$, where

$$\mathcal{Z}(s) := \{z \in \{0,1\}^T \mid \mathbf{1}'_T z = s\}.$$

For any $z^* \in \mathcal{Z}(s)$, the notation $A[z^*] \in \mathbb{S}^s_{++}$ denotes the principle sub matrix of $A$ constructed according to $z^*$, i.e., the $i$-th row and column are taken from $A$ to construct $A[z^*]$ when $z^*_i = 1$. Similarly, $b[z^*] \in \mathbb{R}^s$ denotes the truncated vector of $b$ according to $z^*$.

**Lemma 5.2.** *Given $A \in \mathbb{S}^T_{++}$ and $b \in \mathbb{R}^T$, $((ZAZ)^\dagger b)[z] = A[z]^{-1}b[z]$ and $b'(ZAZ)^\dagger b = b[z]'A[z]^{-1}b[z]$, where $Z := diag\{z\}$ and $z \in \mathcal{Z}(s)$.*

Proof. This lemma is obvious based on the property of the pseudo-inverse: If $A_1 \succ 0$, then $\begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix}^\dagger = \begin{pmatrix} A_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}$. □

The cardinality constraint (5.1) indicates the total possible number of the sparsity to be $\sum_{i=1}^s \mathcal{C}^i_T$. Clearly, enumerating all the sparsities yields the solution of $\mathcal{G}_s(D,d)$, i.e., finding

$$z^* = \arg\min_{z\in\mathcal{Z}} \big\{ -\frac{1}{2}d[z]'(D[z])^{-1}d[z]\big\}, \tag{5.4}$$

where $\mathcal{Z} = \bigcup_{t=1,\cdots,s}\mathcal{Z}(t)$. Then, $u^* = (Z^*DZ^*)^\dagger b$ solves problem $\mathcal{G}_s(D,d)$, where $Z^* = \text{diag}\{z^*\}$. Under the Assumption $D \succ 0$, the following fact is true.

**Lemma 5.3.** *The solution of problem $\mathcal{G}_s(D,d)$ can be found by identifying*

$$z^* = \arg\min_{z\in\mathcal{Z}(s)} \big\{ -\frac{1}{2}d[z]'(D[z])^{-1}d[z]\big\}, \tag{5.5}$$

*and $(Z^*DZ^*)^\dagger b$ solves problem $\mathcal{G}_s(D,d)$, where $Z^* = diag\{z^*\}$.*

Proof . Let $u^*$ be the optimal solution of problem $\mathcal{G}_s(D,d)$. If $\sum_{i=1}^T \delta(u^*_i) = s$, then the lemma is true. We only need to consider the case where $\sum_{i=1}^T \delta(u^*_i) = \hat{s} < s$. From (5.4), we know $u^* = (Z^*DZ^*)^\dagger d$ with $z^* \in \mathcal{Z}(\hat{s})$ and

$$-d[z^*]'(D[z^*])^{-1}d[z] \le -d[z]'(D[z])^{-1}d[z],\ \forall\ z \in \mathcal{Z}(s), \tag{5.6}$$

On the other hand, there exists $\bar{z} \in \mathcal{Z}(s)$, such that $D[\bar{z}] = \begin{pmatrix} D[z^*] & \bar{D}_{12} \\ \bar{D}'_{12} & \bar{D}_{22} \end{pmatrix}$. Note that the basic property of a partitioned matrix and the assumption of $D \succ 0$ imply

$$-\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}' \begin{pmatrix} D[z^*] & \bar{D}_{12} \\ \bar{D}'_{12} & \bar{D}_{22} \end{pmatrix}^{-1} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \leq -d'_1 D[z^*]^{-1} d_1. \tag{5.7}$$

Thus,

$$\min_{z \in \mathcal{Z}(s)} \{-\frac{1}{2} d[z]'(D[z])^{-1} d[z]\} \leq \min_{z \in \mathcal{Z}(\hat{s})} \{-\frac{1}{2} d[z]'(D[z])^{-1} d[z]\}.$$

In conclusion, there exists $\bar{z} \in \mathcal{Z}(s)$, such that $d[\bar{z}]'D[\bar{z}]^{-1}d[\bar{z}] = d[z^*]'D[z^*]^{-1}d[z^*]$ and $u^* = (\bar{Z}D\bar{Z})^{\dagger}d$ with $\bar{Z} = \text{diag}(\bar{z})$. □

Lemma 5.3 suggests a $\mathcal{C}_T^s$ enumeration method in solving problem $\mathcal{G}_s(D,d)$.

**Assumption 5.2.** We assume that $l \notin \Delta(s)$.

The above assumption rules out the degenerate cases. Since if $l \in \Delta(s)$ then $l$ solves problem $\mathcal{G}_s(D,d)$ and optimal value is $C$.

### 5.1.2. Literature Review

The quadratic optimization problems involving a cardinality constraint have been studied in many literatures. In [11], the author considers such a problem with polyhedra constraints $Au \leq b$ and upper-lower bound constraints of $u$, i.e., $0 \leq u \leq lb$. To deal with the cardinality constraint (5.1), the author proposes to use surrogate constraint $\sum_i (u_i/lb_i) \leq s$. Furthermore, they consider an algorithm of a branch and bound type. Following this line, recently, Bertsimas and Shioda consider the exact solution approach of such a kind of problem in [9]. Unlike [11], the author construct the lower bound in [9] by ignoring the cardinality constraint (5.1) and the resulting relaxation problem becomes a convex quadratic programming problem. Such a relaxation leads to a form ready for the Lemek's pivoting method, which is the "simplex"-type of method to solve the quadratic programming and linear complementary problems. While integrating such a bounding method to branch and bound algorithm, the "warm-start" strategy is achieved at each step, which accelerates the whole algorithm. In [18], the author considers the mean-variance portfolio selection problem with a cardinality constraint and uses a heuristic method. i.e., genetic search, tabu search to solve such a problem. Blog [12] also proposes a dynamic programming heuristic to solve such a portfolio selection problem in his early work. Xie, He and Zhang adopt a randomized algorithm to solve such a mean-variance investment problem with cardinality constraint in [80]. In [22], Li, Sun and Wang proposed efficient numerical solution approach to find the optimal lot solution of such a portfolio selection problem with cardinality constraint.

It is also worth mentioning the recent progress on the sparse signal reconstruction via $\mathbf{L}_1$ norm. The problem of sparse signal reconstruction can be stated as following. Given $y \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times m}$, find $\omega^*$ that solves one of the following types of questions,

$$\begin{aligned} &\text{(maximum likehood)} \quad \min \ \left\{\|A\omega - y\|_2^2 \mid \|\omega\|_0 \le k\right\}, \\ &\text{(minium sparsity)} \quad \min \ \{\|\omega\|_0 | \ y = A\omega\}, \end{aligned}$$

where $p << m$ and $k < p$. One of the heuristic methods of solving above problems is to replace the $\mathbf{L}_0$ norm by the $\mathbf{L}_1$ norm, which generates a convex optimization problem. It has been proved in [26] and [15] that under some conditions (e.g., the condition CS1-CS3 in [26]) the $\mathbf{L}_1$ norm heuristic also solves the minimum sparsity problem with overwhelming probability. Note that problem $\mathcal{G}_s(D,d)$ is different from the maximum likehood reconstruction problem, since matrix $D$ is of full rank.

### 5.1.3. Mean-Variance Portfolio Selection with Transaction Cost

One direct application of problem $\mathcal{G}_s(D,d)$ is the mean-variance portfolio selection model with a cardinality constraint on the total number of risk-assets selection. Different from [12] and [18], we assume that the short selling is allowed. Suppose that there are one risk-free asset with return $r_f$ and $T$ risky assets with random return $X_i$, with known expected values and covariance,

$$r_i = E(X_i), \quad \text{and } \sigma_{ij} = Cov(X_i, X_j), \ i,j = 1,\cdots,T.$$

Let $u_i$ be the liquid share of $i$-th security. The investor enters the market with initial wealth $W_0$ and seeks the optimal portfolio $u' := (u_1,\cdots,u_T)$ on the $T$ risky assets and $u_f \in \mathbb{R}$ on the risk-free asset. Let the current price of each liquid share of security be $a_i$. Initially, the wealth balance is

$$W_0 = u_f + a'u,$$

where $a' := (a_1,\cdots,a_T)$. At the end of investment period, the return of the holding security is the random term $R_p := \sum_{i=1}^{T} u_i X_i$. The mean and variance of $R_p$ are, respectively,

$$\begin{aligned} E(R_p) =& E\left[\sum_{i=1}^{T} u_i X_i\right] = r'u, \\ \text{Var}(R_p) =& \text{Var}\left[\sum_{i=1}^{T} u_i X_i\right] = u'\Sigma u, \end{aligned}$$

where $r' := (r_1,\cdots,r_T)$ and $\Sigma := \{\sigma_{ij}\}|_{i,j=1}^{T}$. Let the transaction cost be a positive constant $M$. Then, the total expected return of the portfolio is

$$r'u + r_f u_f - M\sum_{i=1}^{T} \delta(u_i).$$

Given a desire level of return $\bar{r}$, investor is seeking the minimum variance solution of following problem,

$$\begin{aligned} \mathcal{P}_{mv}: \quad & \min_{u,u_f} \;\; u'\Sigma u, \\ \text{Subject to: } & r'u + r_f u_f - M\sum_{i=1}^{T}\delta(u_i) = \bar{r} \\ & a'u + u_f = W_0. \end{aligned}$$

To solve problem $\mathcal{P}_{mv}$, we construct the auxiliary problem,

$$\begin{aligned} \mathcal{P}_{mv}(s): \quad & \min_{u,u_f} \;\; u'\Sigma u, \\ \text{Subject to: } & r'u + r_f u_f - Ms = \bar{r}, \\ & a'u + u_f = W_0, \\ & u \in \Delta(s). \end{aligned}$$

The solution of problem $\mathcal{P}_{mv}$ can be identified by finding the optimal cardinality $s^*$,

$$s^* := \arg\min_{1\le s\le T}(v(\mathcal{P}_{mv}(s))).$$

The optimal solution of problem $\mathcal{P}_{mv}(s^*)$ also solves problem $\mathcal{P}_{mv}$.

**Theorem 5.4.** *If $u^*$ solves problem $\mathcal{P}_{mv}(s)$ and $\hat{u}^*$ solves the following problem,*

$$\hat{\mathcal{P}}_{mv}(s): \quad \min_{u\in\mathbb{R}^T} \;\; \left\{\; u'\Sigma u + \hat{a}'u \;|\;\; u \in \Delta(s) \;\right\},$$

*where $\hat{a} := r - r_f a$, then $u^*$ and $\hat{u}^*$ have the same sparsity and $u^* = \rho\hat{u}^*$ with $\rho \in \mathbb{R}$.*

Proof. Clearly, problem $\mathcal{P}_{mv}(s)$ can be solved by enumerating $\sum_{j=1}^{s}\mathcal{C}_T^j$ sparsity patterns. In particular, we can solve problem $\mathcal{P}_{mv}(s)$ explicitly for any given $z \in \mathcal{Z}(s)$ by considering the following truncated problem,

$$\mathcal{P}_{mv}(s,z): \;\min_{u[z]}\Big\{\; u[z]'\Sigma[z]u[z] \mid r[z]'u[z] + r_f u_f = \bar{r} + Ms,\;\; a[z]'u[z] + u_f = W_0\Big\}.$$

As $\Sigma[z] \succ 0$, problem $\mathcal{P}_{mv}(s,z)$ is a standard convex problem. Note that if $r[z] - r_f a[z] = 0$, problem $\mathcal{P}_{mv}(s,z)$ is infeasible. We simply let $v(\mathcal{P}_{mv}(s,z)) = +\infty$ under this case. When $\hat{a} \neq 0$, the solution of problem $\mathcal{P}_{mv}(s,z)$ is $u[z]^* = \rho\Sigma[z]^{-1}\hat{a}[z]$ with $\rho := (Ms + \bar{r} - r_f W_0)/(\hat{a}[z]'\Sigma[z]^{-1}\hat{a}[z])$ and the optimal value is

$$v(\mathcal{P}_{mv}(s,z)) = \begin{cases} (Ms + \bar{r} - r_f W_0)^2/(\hat{a}[z]'\Sigma[z]^{-1}\hat{a}[z]) & \text{if } \hat{a} \neq 0, \\ +\infty & \text{if } \hat{a} = 0. \end{cases} \tag{5.8}$$

The expression of $v(\mathcal{P}_{mv}(s,z))$ in (5.8) reveals that finding the optimal $z^* \in \mathcal{Z}(s)$ of problem $\mathcal{P}_{mv}(s)$ is equivalent to maximizing $\hat{a}[z]'\Sigma[z]^{-1}\hat{a}[z]$. Thus, the optimal $z^*$ can be identified as

$$\begin{aligned} z^* &= \arg\min_{z\in\mathcal{Z}} v(\mathcal{P}_{mv}(s,z)) = \arg\max_{z\in\mathcal{Z}} \hat{a}[z]'\Sigma[z]^{-1}\hat{a}[z] \\ &= \arg\min_{z\in\mathcal{Z}} -\hat{a}[z]'\Sigma[z]^{-1}\hat{a}[z]. \end{aligned} \tag{5.9}$$

Then, the optimal solution of problem $\mathcal{P}_{mv}(s)$ is $u^* = \rho^*\Sigma[z^*]^{-1}\hat{a}[z^*]$ with $\rho^* := (Ms^* + \bar{r} - r_f W_0)/(\hat{a}[z^*]'\Sigma[z^*]^{-1}\hat{a}[z^*])$. On the other hand, from (5.5), the optimal sparsity of problem $\hat{\mathcal{P}}_{mv}(s)$ can also be specified by (5.9) with optimal solution being $\hat{u}^* = (Z^*\Sigma Z^*)^\dagger \hat{a}$. Comparing the optimal solution of problem $\mathcal{P}_{mv}(s)$ and $\hat{\mathcal{P}}_{mv}(s)$, we have $u^* = \rho^*\hat{u}^*$ which completes the proof of the theorem. □

Although problem $\mathcal{P}_{mv}(s)$ involves both equality constraints and cardinality constraint, from Theorem 5.4, we can solve it by considering a CCQO problem $\hat{\mathcal{P}}_{mv}(s)$, in which only cardinality constraint is included.

## 5.2. Lower Bounding Schemes

The branch-and-bound(BnB) type of algorithms is ready to be used in solving problem $\mathcal{G}_s(D,d)$ exactly. The main framework of using a BnB algorithm to solve problem $\mathcal{G}_s(D,d)$ is similar to our previous work [42]. As a tight and cheap bound plays a key role in such an algorithm, we focus in the following on constructing efficient lower bounds of problem $\mathcal{G}_s(D,d)$.

For any $P \in \mathbb{S}^T_{++}$ and $p \in \mathbb{R}^T$, we denote following ellipsoid in $\mathbb{R}^T$ as

$$\mathcal{E}(P,p,\rho) := \left\{ y \in \mathbb{R}^T \mid (y+p)'P(y+p) \le \rho \right\}, \tag{5.10}$$

where $\rho \ge 0$. Clearly, the objective contour of $\mathcal{G}_s(D,d)$ is an ellipsoid in $\mathbb{R}^T$ space, for any $\tau \ge C$,

$$\begin{aligned}\mathcal{E}(D,l,\rho(\tau)) :=& \{u \in \mathbb{R}^T \mid f(u) \le \tau\} \\ =& \{u \in \mathbb{R}^T \mid (u+l)'D(u+l) \le 2\tau - 2C\},\end{aligned}$$

where $\rho(\tau) := 2\tau - 2C$. Geometrically, minimizing $f(u)$ under the constraint (5.1) is equivalent to find the minimum ellipsoid that touches the set $\Delta(s)$, i.e.,

$$\begin{aligned} \mathcal{G}_s(D,d): \quad & \min \quad \tau, \\ \text{Subject to:} \quad & u \in \mathcal{E}(D,l,\rho(\tau)), \\ & u \in \Delta(s). \end{aligned} \tag{5.11}$$

For a vector $y \in \mathbb{R}^T$, we define a corresponding vector $\theta(y) \in \mathbb{R}^T$ with $(\theta_1(y), \theta_2(y), \cdots, \theta_T(y))'$ being a permutation of $y' = (y_1, y_2, \cdots, y_T)$ and $\theta_1(y)$ $\le \theta_2(y) \cdots \le \theta_T(y)$.

Ignoring the cardinality constraint (5.1) in problem $\mathcal{G}_s(D,d)$ generates the trivial lower bound, $v(\mathcal{D}_{cont}) = C$. Regarding $v(\mathcal{D}_{cont})$ as benchmark, we are interested in finding a lower bound $J$ with $J \ge v(\mathcal{D}_{cont})$ in the worst case.

### 5.2.1. Lower Bound via Hyper-box

In this Section, we use the hyper-box to construct the lower bound of problem $\mathcal{G}_s(D,d)$. Define the hyper-box by two vectors $\upsilon \in \mathbb{R}^T$ and $\vartheta \in \mathbb{R}^T$,

$$[\upsilon, \vartheta] := \left\{ y \in \mathbb{R}^T \mid \upsilon_i \le y_i \le \vartheta_i, \ i = 1, \cdots, T \right\},$$

where $\upsilon_i$ are $\vartheta_i$ are the $i$-th component of $\upsilon$, and $\vartheta$, respectively. The following Lemma is true [45].

**Lemma 5.5.** *The circumscribed hyper-box of ellipsoid* $\mathcal{E}(D, l, \rho(\tau))$ *is given by* $\big[\upsilon(\tau), \vartheta(\tau)\big] := \big\{y \in \mathbb{R}^T \mid \upsilon(\tau) \le y \le \vartheta(\tau)\big\}$, *where*

$$\upsilon(\tau) = -\,l - \sqrt{\rho(\tau)}\ vec\{\sqrt{\hat{D}_i}\},$$
$$\vartheta(\tau) = -\,l + \sqrt{\rho(\tau)}\ vec\{\sqrt{\hat{D}_i}\},$$

*and* $vec\{\sqrt{\hat{D}_i}\} : = \Big(\ \sqrt{\hat{D}_1}\ \ \sqrt{\hat{D}_2}\ \ \cdots\ \ \sqrt{\hat{D}_T}\ \Big)'$ *with* $\hat{D}_i$ *being the i-th component of the diagonal of* $D^{-1}$.

The above explicit expression of such a circumscribed hyper-rectangle motivates us to relax the original problem, $\mathcal{G}_s(D, d) : \min_u \big\{\tau \mid u \in \mathcal{E}(D, l, \rho(\tau)),\ u \in \Delta(s)\big\}$ to $\min_u \big\{\tau \mid \upsilon(\tau) \le u \le \vartheta(\tau),\ u \in \Delta(s)\big\}$, which can be expressed in the following form,

$$\begin{aligned} \mathcal{D}_{box} :\quad & \min \quad \tau \\ \text{Subject to:}\quad & \frac{(u_t + l_t)^2}{\hat{D}_t} \le \rho(\tau), \quad t = 1, \cdots, T, \\ & u \in \Delta(s). \end{aligned} \tag{5.12}$$

Clearly, since the feasible region is enlarged, we have $\upsilon(\mathcal{D}_{box}) \le \upsilon(\mathcal{G}_s(D, d))$. Furthermore, problem $\mathcal{D}_{box}$ can be solved explicitly.

**Theorem 5.6.** *The optimal value of problem* $\mathcal{D}_{box}$ *is* $\upsilon(\mathcal{D}_{box}) = \frac{1}{2}\theta_{T-s}(\varrho) + C$, *where* $\varrho := (\varrho_1, \cdots, \varrho_T)$ *and* $\varrho_t := (l_t)^2/\hat{D}_t$.

Proof. When $u_t = 0$, the constraints indexed by $t$ in (5.12) give rise $(l_t)^2/\hat{D}_t \le \rho(\tau) = 2\tau - 2C$. Thus, the minimum $\tau$ corresponding to this set of inequalities is the maximum among $\frac{1}{2}(l_t)^2/\hat{D}_t + C$, for $i = 1, \cdots, T$. The cardinality constraint (5.1) sets at least $(T - s)$ $u_t$'s equal to 0. Thus, the minimum of $\tau$ is the $(s+1)$st largest of $\frac{1}{2}(l_t)^2/\hat{D}_t + C$ among all $t = 1, \cdots, T$. □

**Remark 5.1.** Clearly, $\upsilon(\mathcal{D}_{box}) \ge \upsilon(\mathcal{D}_{cont})$ and the improvement is $\theta_{s+1}(\varrho)/2$ when compared with $\mathcal{D}_{cont}$. Under Assumption 5.2, the (s+1)-th largest of $l_t^2$ is not zero. Thus, $\upsilon(\mathcal{D}_{box}) > \upsilon(\mathcal{D}_{cont})$. Computation of $\mathcal{D}_{box}$ only involves inverse calculation of a $(k \times k)$ matrix with $k \leq T$ (in iterations of BnB algorithm, the dimension of the matrix is reduced), which can be completed in at most $\mathcal{O}(k^3)$ time. The computational efforts can be further reduced by adopting a "warm-start" strategy, e.g., if an inverse matrix is stored at a parent node, the inverse matrices at its children nodes can be computed by directly modifying the correspondent columns and rows of such a matrix at the parent node. Such a strategy greatly speeds up the branch and bound procedure, as evidenced from our numerical experiments.

Besides the way of constructing a lower bound $\mathcal{D}_{box}$ directly from the minimum circumscribed box of $\mathcal{E}(D, l, \rho(\tau))$, we could use such a box to estimate the upper and lower bound of each $u_i$. Furthermore, we could use the idea of surrogation to relax the cardinality constraint (5.1). Suppose that an incumbent $f(\hat{u})$ is known, then the optimal solution of problem $\mathcal{G}_s(D, d)$ is in the ellipsoid $u \in \mathcal{E}(D, l, 2f(\hat{u}) - 2C)$. Denote the circumscribed box of $\mathcal{E}(D, l, 2f(\hat{u}) - 2C)$ as $\left[\upsilon^*, \vartheta^*\right]$, which imposes bounds on each $u_i$, $\upsilon_i^* \leq u_i \leq \vartheta_i^*$, for $i = 1, \cdots, T$. Define the following index set for $\varpi \in \{-1, 0, 1\}$,

$$\kappa(\varpi) := \left\{ \, i \mid \operatorname{sign}(\upsilon_i^*) \cdot \operatorname{sign}(\vartheta_i^*) = \varpi, \text{ for } i = 1, \cdots, T \right\} \subseteq \{1, \cdots, T\},$$

Let $u^* = (u_1^*, \cdots, u_T^*)'$ be the optimal solution of problem $\mathcal{G}_s(D, d)$. As $u^* \in \left[\vartheta^*, \upsilon^*\right]$, there exist different cases.

- If $\operatorname{sign}(\upsilon_i^*) \cdot \operatorname{sign}(\vartheta_i^*) = 1$, then $u_i^*$ will not be zero.
- If $\operatorname{sign}(\upsilon_i^*) \cdot \operatorname{sign}(\vartheta_i^*) = -1$, then $u_i^*$ could be zero.
- If $\operatorname{sign}(\upsilon_i^*) \cdot \operatorname{sign}(\vartheta_i^*) = 0$, then either $\upsilon_i^* = 0$ or $\vartheta_i^* = 0$ is zero. If $\upsilon_i^* = 0$ and $\vartheta_i^* = 0$, then $u_i^* = 0$.

Thus, we can consider the following relaxed problem,

$$\begin{aligned} \mathcal{D}_{surgt}: \quad & \min \ \frac{1}{2}u'Du + d'u, \\ \text{Subject to:} \quad & \sum_{i\in\kappa(-1)} g_i^1(u_i) + \sum_{i\in\kappa(0)} g_i^2(u_i) \leq s - |\kappa(1)|, && (5.13) \\ & \upsilon_i^* \leq u_i \leq \vartheta_i^*, \quad \text{for } i = 1, \cdots, T, && (5.14) \end{aligned}$$

where $g_i^1(\cdot) : \mathbb{R} \to \mathbb{R}$ and $g_i^2(\cdot) : \mathbb{R} \to \mathbb{R}$, in particular,

$$g_i^1(y) := \frac{y^+}{\vartheta_i^*} + \frac{y^-}{\upsilon_i^*}, y \in \mathbb{R}, \qquad (5.15)$$

$$g_i^2(y) := \begin{cases} y/\vartheta_i^* & \text{if } \ \upsilon_i^* = 0, \\ y/\upsilon_i^* & \text{if } \ \vartheta_i^* = 0, \end{cases} \qquad (5.16)$$

and $y^+ := \max\{y, 0\}$, $y^- := \min\{y, 0\}$. Since $g_i^1(u_i) \leq |\text{sign}(u_i)|$ and $g_i^2(u_i) \leq |\text{sign}(u_i)|$, (5.13) is a relaxation of the cardinality constraint (5.1). Indeed, the constraints (5.13) and (5.14) define a polyhedra and problem $\mathcal{D}_{surgt}$ turns out to be a convex quadratic programming problem. However, explicitly writing out such polyhedron constraints is not cheap, e.g., if $T = 3$, $|\kappa(1)| = 0$, $|\kappa(0)| = 0$, $\kappa(-1) = 3$, constraint (5.13) is unfolded as

$$\begin{aligned} &\frac{u_1}{\vartheta_1^*} + \frac{u_2}{\vartheta_2^*} + \frac{u_3}{\vartheta_3^*} \leq s, \ \frac{u_1}{\upsilon_1^*} + \frac{u_2}{\vartheta_2^*} + \frac{u_3}{\vartheta_3^*} \leq s, \frac{u_1}{\vartheta_1^*} + \frac{u_2}{\upsilon_2^*} + \frac{u_3}{\vartheta_3^*} \leq s, \\ &\frac{u_1}{\vartheta_1^*} + \frac{u_2}{\vartheta_2^*} + \frac{u_3}{\upsilon_3^*} \leq s, \ \frac{u_1}{\upsilon_1^*} + \frac{u_2}{\upsilon_2^*} + \frac{u_3}{\vartheta_3^*} \leq s, \frac{u_1}{\upsilon_1^*} + \frac{u_2}{\vartheta_2^*} + \frac{u_3}{\upsilon_3^*} \leq s, \\ &\frac{u_1}{\vartheta_1^*} + \frac{u_2}{\upsilon_2^*} + \frac{u_3}{\upsilon_3^*} \leq s, \ \frac{u_1}{\upsilon_1^*} + \frac{u_2}{\upsilon_2^*} + \frac{u_3}{\upsilon_3^*} \leq s. \end{aligned}$$

The number of constraints are in the order of $\mathcal{O}(2^{|\sigma(-1)|})$, which is usually very large. Historically, quadratic minimization problem with $\mathbf{L}_1$ norm constraint has been known as *Least Absolute Selection and Shrinkage Operator* problem (LASSO) [76] and several algorithms have been designed, e.g., based on active set methods, [63] [76] and based on interior point methods [19].

To lower the computational burden, we do further relaxation of problem

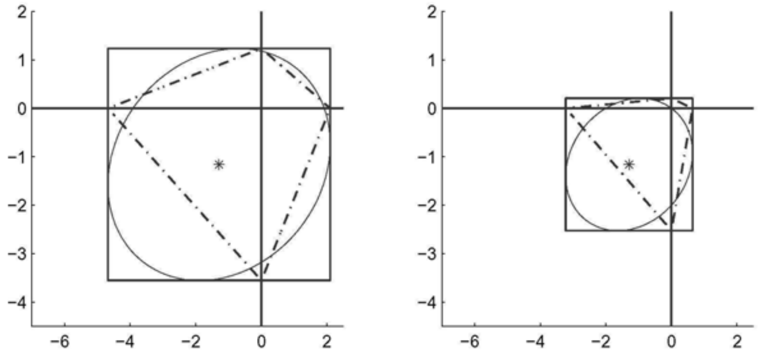

Figure 5.1: Box and surrogate constraints in case of $T = 2$

$\mathcal{D}_{surgt}$ by considering the following problem for given $\rho > 0$,

$$\begin{aligned}\hat{\mathcal{D}}_{surgt}: \ \min \ & \frac{1}{2}u'Du + d'u \\ & + \rho(\sum_{i\in\kappa(-1)} g_i^1(u_i) + \sum_{i\in\kappa(0)} g_i^2(u_i) - s + |\kappa(1)|),\end{aligned}$$

$$\text{Subject to: } v_i^* \leq u_i \leq \vartheta_i^*, \quad \text{for } i = 1, \cdots, T.$$

Note that constraint (5.13) is actually $\mathcal{L}_1$ norm constraint, i.e., for $a < 0$ and $b > 0$, we have

$$(\frac{y^-}{a} + \frac{y^+}{b}) = \frac{1}{2}\left[(\frac{1}{b} + \frac{1}{a})|y| + (\frac{1}{b} - \frac{1}{a})y\right],$$

leading to $g_i^1(u_i) = H_i|u_i| + h_i u_i$ with

$$H_i := \frac{1}{v_i} + \frac{1}{\vartheta_i}, \quad h_i := \frac{1}{v_i} - \frac{1}{\vartheta_i}.$$

Thus, problem $\hat{\mathcal{D}}_{surgt}$ is readily reformulated as a second-order cone programming problem [13].

**Remark 5.2.** From Theorem 5.6, we can see that $v(\mathcal{D}_{box}) \geq v(\mathcal{D}_{cont})$. Under Assumption 5.2, the inequality is always held. However, it is not guaranteed that $v(\mathcal{D}_{surgt}) > v(\mathcal{D}_{cont})$. When $l$ is a feasible point of constraints (5.13) and (5.14), $v(\mathcal{D}_{surgt}) = v(\mathcal{D}_{cont})$. See an illustrative example in $\mathbb{R}^2$ in Figure 5.1.

### 5.2.2. Lower Bound via Axis-aligned Ellipsoid

To identify a computable lower bound, we approximate problem $\mathcal{G}_s(D,d)$ by considering a special class of problems $\hat{\mathcal{G}}_s(H,h,\xi)$,

$$\hat{\mathcal{G}}_s(H,h,\xi): \quad \min \quad \left\{ \hat{f}(u) := \frac{1}{2}u'Hu + h'u + \xi \mid u \in \Delta(s) \right\}, \tag{5.17}$$

where $H \in \mathbb{S}^T_{++}$, $h \in \mathbb{R}^T$, $\xi \in \mathbb{R}$ and $\{H,h\}$ is of a special structure such that problem $\hat{\mathcal{G}}_s(H,h,\xi)$ can be efficiently solved in polynomial time. Define a class of $(T \times T)$ diagonal matrices by

$$\Lambda := \operatorname{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_T\},$$

where $\lambda_t \geq 0$, for $t = 1, \cdots, T$. We now consider a class of problem $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$ and develop a solution scheme to identify the optimal member within this class that best bounds $\mathcal{G}_s(D,d)$ from below in the sense of minimizing the duality gap. Once $\Lambda$ is given, problem $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$ can be solved explicitly.

**Lemma 5.7.** *The optimal value of problem* $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$ *is* $\frac{1}{2}\sum_{j=1}^{T-s} \theta_j(\eta) - \frac{1}{2}l'\Lambda l + \xi$*, where* $\eta := (\eta_1, \cdots, \eta_T)$ *and* $\eta_t := \lambda_t (l_t)^2$*, for* $t = 1, \cdots, T$.

Proof. To minimize $\hat{f}(u) = \frac{1}{2}\sum_{t=1}^{T} \lambda_t (u_t + l_t)^2 - \frac{1}{2}l'\Lambda l + \xi$, the objective function of $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$, under the cardinality constraint in (5.1), the best strategy is to distribute the cardinality to $s$ controls according to the $s$ largest $\eta_t$, such that the $s$ corresponding terms of $\lambda_t(u_t + l_t)^2$ are set to 0. The lemma then follows. □

Given incumbent $\hat{u}$, better feasible solutions must be within the bounded region $(u+l)'D(u+l) \leq 2f(\hat{u}) - l'd$. Thus, theoretically, we only need to consider a bounded region, $(u+l)'D(u+l) \leq \omega$, with $\omega = 2f(\hat{u}) - l'd$, within which the objective function of $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$, $\hat{f}(u) = \frac{1}{2}u'\Lambda u + (\Lambda l)'u + \xi$, bounds the objective function of $\mathcal{G}_s(D,d)$, $f(u) = \frac{1}{2}u'Du + d'u$, from below. However, when we minimize $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$ using Lemma 5.7, we do not take into account constraint $(u+l)'D(u+l) \leq \omega$. From the properties of any optimal solution (assumed to be $\tilde{u}$) of $\hat{\mathcal{G}}_s(\Lambda, \Lambda l, \xi)$, we have $(\tilde{u}+l)'D(\tilde{u}+l) \leq l'Dl \leq \lambda_T^D \|l\|^2$, where $\lambda_T^D$ is the maximum eigenvalue of $D$. Thus, in our real implementation, we set $\omega$ equal to $\lambda_T^D \|l\|^2$.

**Lemma 5.8.** *The condition that $\hat{f}(u) \leq f(u)$ for all $u \in \{u \in \mathbb{R}^T \mid (u+l)'D(u+l) \leq \omega\}$ is satisfied if and only if there exists $\mu \geq 0$ such that*

$$\begin{pmatrix} (\frac{1}{2}+\mu)D - \frac{1}{2}\Lambda & 0 \\ 0 & \xi_0 - \hat{\xi} - \omega\mu \end{pmatrix} \succeq 0, \tag{5.18}$$

*where $\hat{\xi} := \xi - \frac{1}{2}l'\Lambda l$ and $\xi_0 := -\frac{1}{2}l'd$.*

Proof. Note that $f(u) - \hat{f}(u) = \frac{1}{2}(u+l)'(D-\Lambda)(u+l) + \xi_0 - \hat{\xi}$. Let $z := u+l$ and define the following set of feasible $(\Lambda, \hat{\xi})$ such that $f(u) - \hat{f}(u) \geq 0$ in the region of $(u+l)'D(u+l) \leq \omega$,

$$\Pi_1 := \left\{ (\Lambda, \hat{\xi}) \mid \frac{1}{2}z'(D-\Lambda)z + \xi_0 - \hat{\xi} \geq 0, \ \forall\ z:\ z'Dz \leq \omega \right\}. \tag{5.19}$$

We claim that $\Pi_1$ is equivalent to the following set,

$$\Pi_2 := \left\{ (\Lambda, \hat{\xi}) \mid \frac{1}{2}z'(D-\Lambda)z + (\xi_0 - \hat{\xi})\tau^2 \geq 0, \ \forall\ z, \tau:\ z'Dz \leq \tau^2\omega \right\}. \tag{5.20}$$

For any $(\Lambda, \hat{\xi}) \in \Pi_2$, we also have $(\Lambda, \hat{\xi}) \in \Pi_1$ by letting $\tau = 1$. Thus, $\Pi_2 \subseteq \Pi_1$. Next we consider any $(\Lambda^*, \hat{\xi}^*) \in \Pi_1$. Letting $\hat{z} := z\tau$ gives rise $\hat{z}D\hat{z} \leq \tau^2\omega$ and

$$\frac{1}{2}\hat{z}'(D-\Lambda^*)\hat{z} + \tau^2(\xi_0 - \hat{\xi}) = \frac{1}{2}\tau^2 z'(D-\Lambda^*)z + \tau^2(\xi_0 - \hat{\xi}) \geq 0. \tag{5.21}$$

Thus, $(\Lambda^*, \hat{\xi}^*) \in \Pi_2$ and $\Pi_1 \subseteq \Pi_2$.

Let $y' := (z', \tau)$ and apply $\mathcal{S}$-Lemma [64] to set $\Pi_2$, we conclude that $y' \begin{pmatrix} \frac{1}{2}(D-\Lambda) & 0 \\ 0 & \xi_0 - \hat{\xi} \end{pmatrix} y \geq 0$ for all $y' \begin{pmatrix} -D & 0 \\ 0 & \omega \end{pmatrix} y \geq 0$ if and only if there exists a $\mu \geq 0$ such that $y' \begin{pmatrix} (\frac{1}{2}+\mu)D - \frac{1}{2}\Lambda & 0 \\ 0 & \xi_0 - \hat{\xi} - \omega\mu \end{pmatrix} y \geq 0$ for all $y \in \mathbb{R}^T$. □

To minimize the duality gap between $\mathcal{G}_s(D,d)$ and $\hat{\mathcal{G}}_s(\Lambda,\Lambda l,\xi)$ is equivalent to finding the best $\Lambda$ and $\xi$ of the following problem,

$$\begin{aligned}\mathcal{D}_{diag}:\ & \max_{\Lambda\succeq 0,\xi}\min_{u}\ f(u)=\frac{1}{2}u'\Lambda u+(\Lambda l)'u+\xi,\\ \text{Subject to:}\ & u\in\Delta(s),\\ & \{\Lambda,\xi\}\ \text{satisfies}\ \ (5.18).\end{aligned}$$

Due to the explicit expression of the optimal value of problem $\hat{\mathcal{G}}_s(\Lambda,\Lambda l,\xi)$ given in Lemma 5.7, problem $\mathcal{D}_{diag}$ can be simplified to

$$\begin{aligned}\mathcal{D}_{diag}:\ & \max_{\Lambda\succeq 0,\xi}\frac{1}{2}\sum_{j=1}^{T-s}\theta_j(\eta)+\hat{\xi},\\ \text{Subject to}:\ & \{\Lambda,\hat{\xi}\}\ \text{satisfies}\ \ (5.18).\end{aligned}$$

As maximizing the summation of $T-s$ smallest $\theta_i(\eta)$ can be cast into a linear representation [79], problem $\mathcal{D}_{diag}$ is equivalent to the following SDP problem with $a_t$ and $z$ being auxiliary variables,

$$\begin{aligned}\mathcal{D}_{diag}:\ & \max_{\Lambda\succeq 0,\xi}\quad -(T-s)z+\sum_{t=1}^{T}a_t+\hat{\xi}\\ \text{Subject to:}\ & \{\Lambda,\hat{\xi}\}\ \text{satisfies}\ \ (5.18),\\ & a_t\le\frac{1}{2}(l_t)^2\lambda_t+z,\quad t=1,\cdots,T,\\ & a_t\le 0,\quad t=1,\cdots,T.\end{aligned}$$

Geometrically, problem $\mathcal{D}_{diag}$ constructs the "best" axis-aligned ellipsoid that contains the objective contour of $\mathcal{G}_s(D,d)$ for any given $\tau$.

### 5.2.3. Lower Bounding via Balls

We denote a ball with center $(-o)\in\mathbb{R}^T$ and radius $r>0$ as

$$\mathcal{B}(o,r^2):=\left\{u\in\mathbb{R}^T\mid\|u+o\|_2^2\le r^2\right\}. \tag{5.22}$$

To construct a lower bound of the problem $\mathcal{G}_s(D,d)$, one simple idea is to relax the ellipsoid (5.11) to its circumscribed ball with minimum radius. Consider the following problem,

$$\begin{aligned} \mathcal{D}_{ball}: &\ \min \ \ \tau, \\ \text{Subject to}: &\ u \in \mathcal{B}(l, \rho(\tau)/\lambda_1^D), \\ &\ u \in \Delta(s). \end{aligned}$$

Since $\lambda_1^D \|u+l\|_2^2 \le (u+l)'D(u+l)$ for all $u$, then $\mathcal{E}(D,l,\rho(\tau)) \subseteq \mathcal{B}(l,\rho(\tau)/\lambda_1^D)$. Thus, $v(\mathcal{D}_{ball}) \le v(\mathcal{G}_s(D,d))$. Solving problem $\mathcal{D}_{ball}$ is equivalent to

$$\min_u \ \left\{ \lambda_1^D \|u+l\|_2^2 \mid u \in \Delta(s) \ \right\}.$$

From the proof of Theorem 5.13 in Appendix, $v(\mathcal{D}_{ball}) = C + \lambda_1^D \sum_{i=1}^{T-s} \theta_i(l^2)$, where $(l^2)' := ((l_1)^2, (l_2)^2, \cdots, (l_T)^2)$. Compared with the trivial bound $\mathcal{D}_{cont}$, the improvement is $\lambda_1^D \sum_{i=1}^{T-s} \theta_i(l^2)$. Under Assumption 5.2, we have $v(\mathcal{D}_{ball}) > v(\mathcal{D}_{cont})$. Furthermore, note that $(\lambda_1^D \mathbf{I}_T, C)$ is a feasible solution of problem $\mathcal{D}_{diag}$. Then, we have $v(\mathcal{D}_{diag}) \ge v(\mathcal{D}_{ball}) > v(\mathcal{D}_{cont})$.

We now improve further this ball bound. Assume that there exists $k \in \{1, \cdots, T\}$, such that $\lambda_1^D \le \lambda_2^D \cdots \le \lambda_k^D < \lambda_{k+1}^D \cdots \le \lambda_T^D$. Let the spectral decomposition of $D$ be $D = \Gamma' \Lambda_D \Gamma$, where $\Lambda_D := \operatorname{diag}\{\lambda_i^D\}|_{i=1}^T$ and $\Gamma'\Gamma = \mathbf{I}$. We construct a matrix $H := \Gamma' \Lambda_D^k \Gamma$ where

$$\Lambda_D^k := \operatorname{diag}\big\{\lambda_1^D, \cdots, \lambda_k^D, \lambda_{k+1}^D, \lambda_{k+1}^D, \cdots, \lambda_{k+1}^D\big\}. \tag{5.23}$$

Then, we have $D \succeq H$ which implies $(u+l)'D(u+l) \succeq (u+l)'H(u+l)$. Thus, $\mathcal{E}(D,l,\rho(\tau)) \subseteq \mathcal{E}(H,l,\rho(\tau))$ for any $\tau > 0$.

The following problem provides a lower bound of problem $\mathcal{G}_s(D,d)$,

$$\begin{aligned} \mathcal{D}_{ellp}^k: &\ \min \ \tau, \\ \text{Subject to:} &\ u \in \mathcal{E}(H, l, \rho(\tau)), \\ &\ u \in \Delta(s). \end{aligned}$$

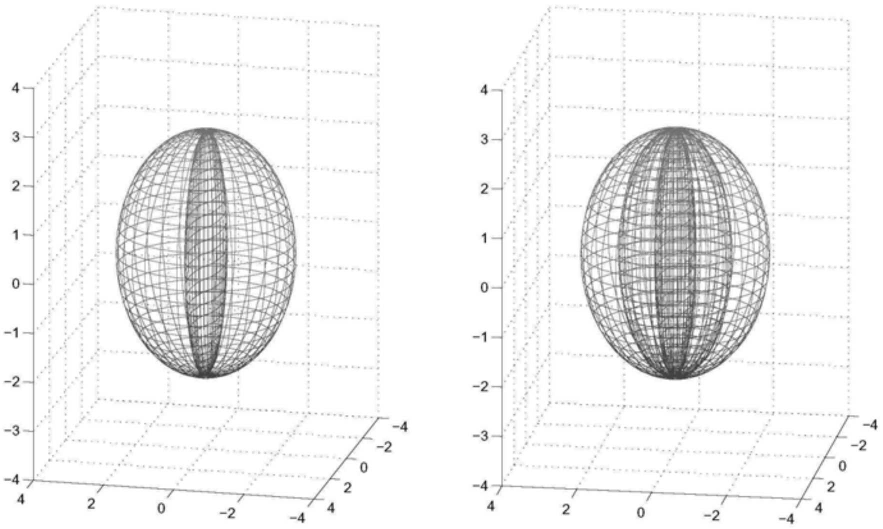

Figure 5.2: The ball bound and ellipsoid bound

Compared with the ball bound $\mathcal{D}_{ball}$, the feasible region $\mathcal{E}(H, l, \rho(\tau))$ is smaller than $\mathcal{B}(l, \rho(\tau)/\lambda_1^D)$, i.e.,

$$\mathcal{E}(D, l, \rho(\tau)) \subseteq \mathcal{E}(H, l, \rho(\tau)) \subseteq \mathcal{B}(l, \rho(\tau)/\lambda_1^D).$$

An illustration in $\mathbb{R}^3$ is shown in Figure 5.2. In the left sub figure of Figure 5.2, $\mathcal{E}(D, l, \rho(\tau))$ is covered by a ball $\mathcal{B}(l, \rho(\tau)/\lambda_1^D)$. In the right sub figure of Figure 5.2, $\mathcal{E}(D, l, \rho(\tau))$ is covered by both $\mathcal{E}(H, l, \rho(\tau))$ and $\mathcal{B}(l, \rho(\tau)/\lambda_1^D)$. Since $H$ is of a special structure, problem $\mathcal{D}_{ellp}^k$ can be solved efficiently.

**Theorem 5.9.** *Consider an ellipsoid $\mathcal{E}(\Lambda_k, 0, \rho)$ with $\rho > 0$, where $\Lambda_k := diag\{\lambda_i\}|_{i=1}^T$ and $0 < \lambda_1 \leq \cdots \leq \lambda_k < \lambda_{k+1} = \cdots = \lambda_T$. Then ellipsoid $\mathcal{E}(\Lambda_k, 0, \rho)$ can be decomposed as*

$$\mathcal{E}(\Lambda_k, 0, \rho) = \bigcup_{\beta \in E(k)} \mathcal{B}(\beta, r^2(\beta)),$$

*where*

$$E(k) :=\{\beta \in \mathbb{R}^T \mid \sum_{i=1}^{k} \frac{\lambda_i \lambda_{k+1}^2 \beta_i^2}{(\lambda_{k+1} - \lambda_i)^2} \leq \rho,\ \beta_j = 0, j = k+1, \cdots, T\}, \tag{5.24}$$

$$r^2(\beta) :=\frac{\rho}{\lambda_{k+1}} - \sum_{i=1}^{k} \frac{\lambda_i \beta_i^2}{\lambda_{k+1} - \lambda_i}. \tag{5.25}$$

Proof. For any $y^* \in \mathcal{E}(\Lambda_k, 0, \rho)$,

$$\sum_{i=1}^{k} \lambda_i (y_i^*)^2 + \lambda_{k+1} \sum_{j=k+1}^{T} y_j^* \leq \rho.$$

Let $\beta^* = (\beta_1^*, \beta_2^*, \cdots, \beta_k^*, 0, \cdots, 0)$, where $\beta_i^* = y_i^*(\lambda_{k+1} - \lambda_i)/\lambda_{k+1}$ for $i = 1, \cdots, k$ and $\beta_j = 0$ for $j = k+1, \cdots, T$. Then, we have

$$\sum_{i=1}^{k} \frac{(\beta_i^*)^2 \lambda_i \lambda_{k+1}^2}{(\lambda_{k+1} - \lambda_i)^2} = \sum_{i=1}^{k} (y_i^*)^2 \lambda_i \leq \rho - \lambda_{k+1} \sum_{i=k+1}^{T} (y_i^*)^2 \leq \rho.$$

Let

$$r^2(\beta^*) = \frac{\rho}{\lambda_{k+1}} - \sum_{i=1}^{k} \frac{\lambda_i (\beta^*)^2}{\lambda_{k+1} - \lambda_i}.$$

Since $\lambda_{k+1} - \lambda_i \neq 0$, we show that $\|y^* - \beta^*\|_2^2 \leq r^2(\beta^*)$ by replacing $\beta_i^*$ with $y_i^*(\lambda_{k+1} - \lambda_i)(\lambda_{k+1})$,

$$\begin{aligned}\|y^* - \beta^*\|_2^2 - r^2(\beta^*) =& \sum_{i=1}^{k} (y_i^* - y_i^* \frac{\lambda_{k+1} - \lambda_i}{\lambda_{k+1}})^2 + \sum_{j=k+1}^{T} y_j^* - (\frac{\rho}{\lambda_{k+1}} \\ & - \sum_{i=1}^{k} \frac{(y_i^*)^2 \lambda_i (\lambda_{k+1} - \lambda_i)}{\lambda_{k+1}^2}), \\ =& \sum_{i=1}^{k} \frac{(y_i^*)^2 \lambda_i}{\lambda_{k+1}} + \sum_{i=k+1}^{T} (y_i^*)^2 - (\frac{\rho}{\lambda_{k+1}}) \leq 0,\end{aligned}$$

which shows that for any $y^* \in \mathcal{E}(\Lambda_k, 0, \rho)$, there exists $\beta^*$ such that $y^* \in \mathcal{B}(\beta^*, r^2(\beta^*))$, implying $\mathcal{E}(\Lambda_k, 0, \rho) \subseteq \bigcup_{\beta \in E(k)} \mathcal{B}(\beta, r^2(\beta^*))$.

On the other hand, for any $\bar{y} \in \mathcal{B}(\bar{\beta}, \bar{r}^2)$ with

$$\bar{r}^2 = (\rho/\lambda_{k+1}) - \sum_{i=1}^{k} \frac{\bar{\beta}_i^2 \lambda_i}{\lambda_{k+1} - \lambda_i},$$

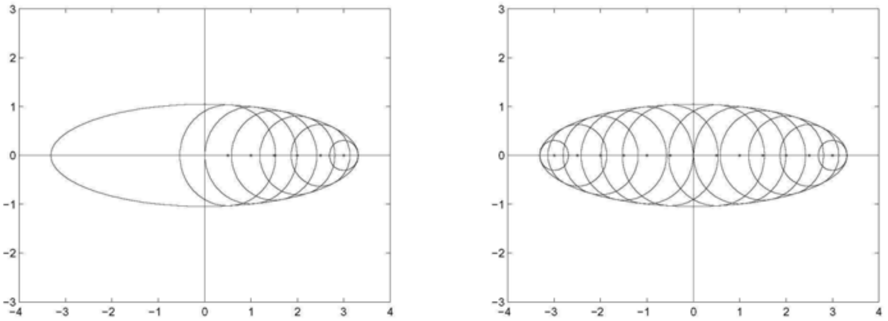

Figure 5.3: Decomposition of ellipsoid in $\mathbb{R}^2$

the following inequality is held, for $i = 1, \cdots, k$,

$$\frac{\bar{y}_i^2 \lambda_i}{\lambda_{k+1}} \leq (\bar{y}_i - \bar{\beta}_i)^2 + \frac{\bar{\beta}_i^2 \lambda_i}{\lambda_{k+1} - \lambda_i}, \tag{5.26}$$

due to the fact that for $i = 1, \cdots, k$,

$$\frac{\lambda_{k+1} - \lambda_i}{\lambda_{k+1}} (\bar{y}_i - \frac{\bar{\beta}_i \lambda_{k+1}}{\lambda_{k+1} - \lambda_i})^2 \geq 0$$

implies

$$-\frac{\bar{y}_i^2 \lambda_i}{\lambda_{k+1}} + \bar{y}_i^2 - 2\bar{\beta}_i \bar{y}_i + \bar{\beta}_i^2 + \frac{\bar{\beta}_i^2 \lambda_i}{\lambda_{k+1} - \lambda_i} \geq 0.$$

From the inequality (5.26), we have

$$\sum_{i=1}^{k} \frac{\bar{y}_i^2 \lambda_i}{\lambda_{k+1}} + \sum_{j=k+1}^{T} \bar{y}_j^2 \leq \sum_{i=1}^{k} [(\bar{y}_i - \bar{\beta}_i)^2 + \frac{\bar{\beta}_i^2 \lambda_i}{\lambda_{k+1} - \lambda_i}] + \sum_{j=k+1}^{T} \bar{y}_j^2 \leq \frac{\rho}{\lambda_{k+1}}, \tag{5.27}$$

where the first inequality is due to the fact of $\bar{y} \in \mathcal{B}(\bar{\beta}, \bar{r}^2)$. The inequality (5.27) implies $\bar{y} \in \mathcal{E}(\Lambda_k, 0, \rho)$. That is to say, $\bigcup_{\beta \in E(k)} \mathcal{B}(\beta, r^2) \subseteq \mathcal{E}(\Lambda_k, 0, \rho)$. □

The decomposition of the ellipsoid in $\mathbb{R}^2$ is shown in Figure 5.3. The center of the balls are along the longest radius of the ellipsoid. The union of the infinite balls is nothing but the ellipsoid itself.

Applying the result in Theorem 5.9 to the ellipsoid $\mathcal{E}(\Lambda_D^k, 0, \rho(\tau))$ yields the following decomposition,

$$\mathcal{E}(\Lambda_D^k, 0, \rho(\tau)) = \bigcup_{\beta \in E(k)} \mathcal{B}(\beta, r^2(\beta)),$$

where

$$
\begin{aligned}
E(k) :=& \{\beta \in \mathbb{R}^T \mid \sum_{i=1}^{k} \kappa_i \beta_i^2 \leq \rho(\tau),\ \beta_j = 0, j = k+1, \cdots, T\}, && (5.28)\\
r^2(\beta) :=& (\rho(\tau) - \sum_{i=1}^{k} (\iota_i \beta_i^2))/\lambda_{k+1}^D, && (5.29)\\
\kappa_i :=& (\lambda_i^D (\lambda_{k+1}^D)^2)/(\lambda_{k+1}^D - \lambda_i^D)^2, \text{ for } i = 1, \cdots, k, && (5.30)\\
\iota_i :=& (\lambda_i^D \lambda_{k+1}^D)/(\lambda_{k+1}^D - \lambda_i^D), \text{ for } i = 1, \cdots, k. && (5.31)
\end{aligned}
$$

Since the shape and size of ellipsoid $\mathcal{E}(\Lambda_D^k, 0, \rho(\tau))$ are coordinate independent, the affine mapping $u = \Gamma' y - l$ maps $y \in \mathcal{E}(\Lambda_D^k, 0, \rho(\tau))$ to $u \in \mathcal{E}(H, l, \rho(\tau))$. Then ellipsoid $\mathcal{E}(H, l, \rho(\tau))$ is decomposed as,

$$
\mathcal{E}(H, l, \rho(\tau)) = \bigcup_{\beta \in E(k)} \mathcal{B}(\Gamma'\beta - l, r^2(\beta)),
$$

where $E(k)$ and $r^2(\beta)$ are defined by (5.28) and (5.29), respectively. The problem $\mathcal{D}_{ellp}^k$ becomes,

$$
\begin{aligned}
\mathcal{D}_{ellp}^k : \quad & \min\ \tau = \frac{1}{2}\rho(\tau) + C,\\
\text{Subject to: } \quad & u \in \bigcup_{\beta \in E(k)} \mathcal{B}(\Gamma'\beta - l, r^2(\beta)), && (5.32)\\
& u \in \Delta(s).
\end{aligned}
$$

Note that constraint (5.32) can be expressed more explicitly as,

$$
u \in \bigcup_{\beta \in E(k)} \{u \mid \lambda_{k+1}^D \|u - \Gamma'\beta + l\|_2^2 + \sum_{i=1}^{k} \iota_i \beta_i^2 \leq \rho(\tau)\}. \qquad (5.33)
$$

**Theorem 5.10.** *It holds that $v(\mathcal{D}_{ellp}^k) = v(\hat{\mathcal{D}}_{ellp}^k)$ where problem $\hat{\mathcal{D}}_{ellp}^k$ is given by*

$$
\begin{aligned}
\hat{\mathcal{D}}_{ellp}^k : \quad & \min\ \tau = \frac{1}{2}\rho(\tau) + C,\\
\textit{Subject to: } \quad & \lambda_{k+1}^D\, dis(\beta) + \sum_{i=1}^{k} \iota_i \beta_i^2 \leq \rho(\tau), && (5.34)\\
& \sum_{i=1}^{k} \kappa_i \beta_i^2 \leq \rho(\tau) \textit{ and } \beta_j = 0, \quad \textit{for } j = k+1, \cdots, T, && (5.35)
\end{aligned}
$$

*where*

$$dis(\beta) := \min_{u\in\mathbb{R}^T} \left\{ \ \|u - \Gamma'\beta + l\|_2^2 \mid u \in \Delta(s) \right\}. \tag{5.36}$$

Proof. We consider $(u, \beta, \rho(\tau))$ and $(\beta, \rho(\tau))$ to be the decision variables for problem $\mathcal{D}_{ellp}$ and $\hat{\mathcal{D}}_{ellp}$, respectively. Let the feasible regions of $\mathcal{D}_{ellp}$ and $\hat{\mathcal{D}}_{ellp}$ be $\Pi$ and $\hat{\Pi}$, respectively. Clearly, for any $(\bar{\beta}, \bar{\rho}(\tau)) \in \hat{\Pi}$, it holds that $(\bar{u}, \bar{\beta}, \bar{\rho}(\tau)) \in \Pi$, where $\bar{u} := \arg\min_{u\in\Delta(s)} \|u - \Gamma'\beta + l\|_2^2$. Then, $v(\hat{\mathcal{D}}_{ellp}^k) \geq v(\mathcal{D}_{ellp}^k)$. On the other hand, (5.33) and (5.34) give rise

$$\lambda_{k+1}^D(\min_u \|u - \Gamma'\beta + l\|_2^2) + \sum_{i=1}^k \iota_i\beta_i^2 \leq \lambda_{k+1}^D\|u - \Gamma'\beta + l\|_2^2 + \sum_{i=1}^k \iota_i\beta_i^2,$$

for any $u \in \Delta(s)$ and $\beta \in E(k)$, which implies that $v(\hat{\mathcal{D}}_{ellp}^k) \leq v(\mathcal{D}_{ellp}^k)$. We complete the proof of the theorem. □

From Theorem 5.10, we know that solving problem $\hat{\mathcal{D}}_{ellp}^k$ gives rise the optimal value of problem $\mathcal{D}_{ellp}^k$ and hence the lower bound of problem $\mathcal{G}_s(D, d)$. Note that the function dis$(\beta)$ defined in (5.36)of problem $\hat{\mathcal{D}}_{ellp}^k$ is a key issue of solving such a problem. Clearly, Function dis$(\beta)$ measures the distance between the affine space, $\{y \in \mathbb{R}^T \mid y = \Gamma'\beta - l\}$ and the set $\Delta(s)$. Various properties of such a function dis$(\beta)$ are discussed in Appendix 5.4.1.

We prove in Theorem 5.13 (Appendix 5.4.1) that function dis$(\beta)$ is a piecewise convex quadratic function. Problem $\hat{\mathcal{D}}_{ellp}$ is thus not convex in general. From Theorem 5.13, for any given $k$, such dis$(\beta)$ can be divided into at most $N$ (N is bounded by $\mathcal{O}\big((T(T-1))^k\big)$) pieces of convex quadratic functions and each function is defined by a polyhedra set, i.e.,

$$\text{dis}(\beta) = \begin{cases} \beta'Q_1\beta + q_1'\beta + c_1 & A_1\beta \leq b_1, \\ \vdots & \vdots \\ \beta'Q_N\beta + q_N'\beta + c_N & A_N\beta \leq b_N. \end{cases}$$

Theoretically, problem $\hat{\mathcal{D}}_{ellp}^k$ can be solved by comparing $N$ sub-problems $\hat{\mathcal{D}}_{ellp}^k(t)$, for $t = 1, \cdots, N$,

$$\hat{\mathcal{D}}^k_{ellp}(t): \quad \min\ \tau = \frac{1}{2}\rho(\tau) + C,$$

$$\text{Subject to: } \lambda^D_{k+1}(\beta' Q_t \beta + q_t'\beta + c_t) + \sum_{i=1}^{k} \iota_i \beta_i^2 \le \rho(\tau), \tag{5.37}$$

$$\sum_{i=1}^{k} \kappa_i \beta_i^2 \le \rho(\tau) \text{ and } \beta_j = 0, \quad \text{for } j = k+1, \cdots, T, \tag{5.38}$$

$$A_t \beta \le b_t. \tag{5.39}$$

Clearly, problem $\hat{\mathcal{D}}^k_{ellp}(t)$ is a convex problem. However, from the computational point of view, we are more interested in the case $k = 1$. Suppose $\lambda^D_1 < \lambda^D_2 \le \lambda^D_3 \cdots \le \lambda^D_T$. Then the lower bounding problem becomes

$$\hat{\mathcal{D}}^1_{ellp}: \quad \min \quad \frac{1}{2}\rho(\tau) + C,$$

$$\text{Subject to: } \lambda^D_2 \text{dis}(\beta_1) + \iota_1 \beta_1^2 \le \rho(\tau), \tag{5.40}$$

$$\kappa_1 \beta_1^2 \le \rho(\tau), \tag{5.41}$$

where $h$ is the first column of $\Gamma'$, $\beta_1 \in \mathbb{R}$, and

$$\begin{aligned} \text{dis}(\beta_1) &:= \min_{u \in \mathbb{R}^T} \left\{ \|u - h\beta_1 + l\|_2^2 \mid u \in \Delta(s) \right\}, \\ \kappa_1 &:= \lambda^D_1 (\lambda^D_2)^2 / (\lambda^D_2 - \lambda^D_1), \\ \iota_1 &:= \lambda^D_1 \lambda^D_2 / (\lambda^D_2 - \lambda^D_1). \end{aligned}$$

**Remark 5.3.** From Corollary 5.14 in Appendix, when $k = 1$, the total number of the quadratic pieces of $\text{dis}(\beta_1)$ is bounded by $T(T-1)$. In each piece, the corresponding sub problem $\hat{\mathcal{D}}^1_{ellp}(t)$, for $t = 1, \cdots, N$, is

$$\hat{\mathcal{D}}^1_{ellp}(t): \quad \min\ \tau = \frac{1}{2}\rho(\tau) + C,$$

$$\text{Subject to: } \lambda^D_2(a_t\beta_1^2 + b_t\beta_1 + c_t) + \iota_1\beta_1^2 \le \rho(\tau), \tag{5.42}$$

$$\kappa_1 \beta_1^2 \le \rho(\tau), \tag{5.43}$$

$$I_t \le \beta_1 \le I_{t+1}, \tag{5.44}$$

Table 5.1: Optimal value and various bounds of Example 5.1

| $v(\mathcal{G}_2(D,d))$ | $v(\mathcal{D}_{cont})$ | $v(\mathcal{D}_{ball})$ | $v(\mathcal{D}_{box})$ | $v(\mathcal{D}_{diag})$ |
|---|---|---|---|---|
| $-168.9$ | $-749.4$ | $-526.2$ | $-254.8$ | $-328.1$ |

where scalars $a_t$, $b_t$, $c_t$, $I_t$ and $I_{t+1}$ are given in Corollary 5.14. Note that problem $\hat{\mathcal{D}}^1_{ellp}(t)$ can be explicitly solved. In the following, we use an example to illustrate the solution procedure in details.

**Example 5.1.** Let us consider an example of problem $\mathcal{G}_s(D,d)$ with $T=6$ and $s=2$. The matrix $D$ and vector $d$ are given, respectively, as

$$D = \begin{pmatrix} 27.171 & -5.738 & 2.479 & -2.768 & 4.931 & 1.725 \\ -5.738 & 18.358 & 5.030 & -5.615 & 10.004 & 3.500 \\ 2.479 & 5.030 & 27.827 & 2.426 & -4.322 & -1.512 \\ -2.768 & -5.615 & 2.426 & 27.292 & 4.825 & 1.688 \\ 4.931 & 10.004 & -4.322 & 4.825 & 21.404 & -3.007 \\ 1.725 & 3.500 & -1.512 & 1.688 & -3.007 & 28.948 \end{pmatrix},$$

$$d' = \begin{pmatrix} 37.745 & -26.329 & -80.284 & 34.905 & 7.296 & -51.002 \end{pmatrix}.$$

Problem $(\mathcal{G}_2(D,d))$ can be solved by pure enumeration. Various bounds are shown in Table 5.1. We now compute the bound $(\hat{\mathcal{D}}^1_{ellp})$. It can be verified that $\Lambda_D = \operatorname{diag}\{1,30,30,30,30,30\}$ and

$$l' = (\ 11.369,\ 19.636,\ -11.539,\ 11.057,\ -17.384,\ -7.867),$$
$$h' = (\ 0.312,\ 0.634,\ -0.274,\ 0.306,\ -0.544,\ -0.190\ ).$$

Furthermore, constant factors are computed as $\kappa_1 = 0.7134$, and $\iota_1 = 0.6897$. Applying Algorithm 1 to identify the distance function $\operatorname{dis}(\beta_1)$ in the interval [0,40], we can explicitly write out the constraints (5.40) and (5.41). To avoid large number, we scale both sides of constraints (5.40) and (5.41) by $\lambda_2 = 30$ as

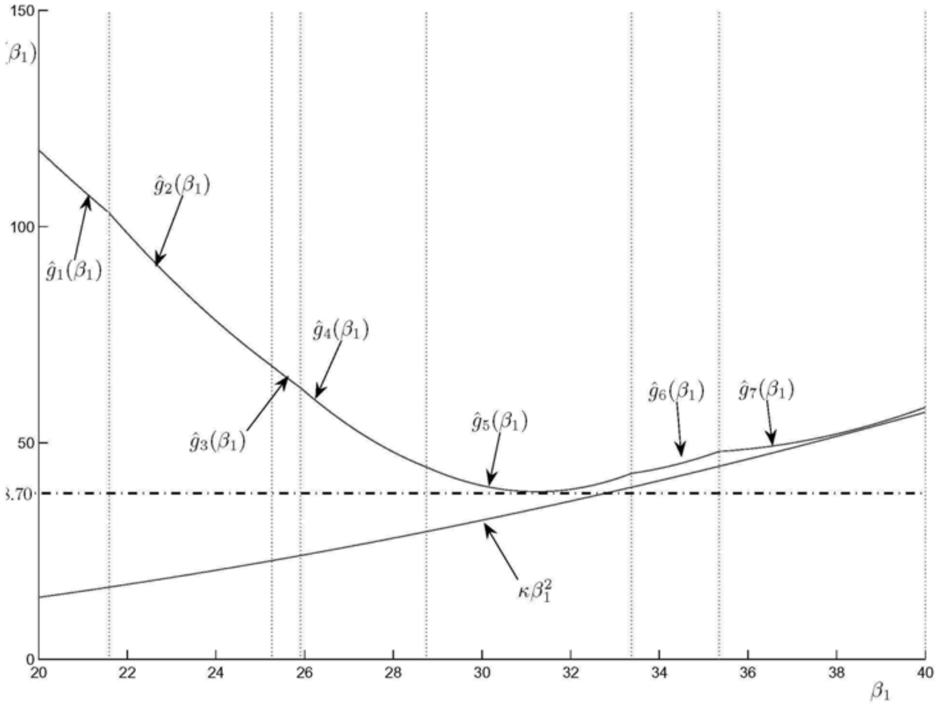

Figure 5.4: The constraints (5.40) and (5.41) of Example 5.1

follows,

$$(\lambda_2\mathrm{dis}(\beta_1)+\iota\beta_1^2)/\lambda_2=\begin{cases}\hat{g}_1=0.30\beta_1^2-23.17\beta_1+446.56 & \beta_1\in[0,21.6],\\ \hat{g}_2=0.52\beta_1^2-35.78\beta_1+615.61 & \beta_1\in[21.6,25.3],\\ \hat{g}_3=0.63\beta_1^2-41.73\beta_1+698.98 & \beta_1\in[25.3,25.9],\\ \hat{g}_4=0.83\beta_1^2-53.56\beta_1+871.92 & \beta_1\in[25.9,28.7],\\ \hat{g}_5=0.89\beta_1^2-57.67\beta_1+939.30 & \beta_1\in[28.7,33.4],\\ \hat{g}_6=0.52\beta_1^2-35.79\beta_1+615.61 & \beta_1\in[33.4,35.3],\\ \hat{g}_7=0.30\beta_1^2-23.17\beta_1+446.56 & \beta_1\in[35.3,40.0].\end{cases}$$

The constraint (5.40) becomes $\hat{g}_i\le\rho(\tau)/\lambda_2$, for $i=1,\cdots,7$. The scaled functions $\hat{g}_i\le\rho(\tau)/\lambda_2$ and $\kappa\beta_1^2/\lambda_2\le\rho(\tau)/\lambda_2$ are shown in Figure 5.4. Then, we can find the minimum $\rho(\tau)^*/\lambda_2=38.7$. Note that lower bound $v(\hat{\mathcal{D}}^1_{ellp})=-168.9$ is the same as $v(\mathcal{G}_2(D,d))$.

### 5.2.4. Lower Bound in Block-wise CCQO Problem

In Section 2.3, the resulting CCQO problem $(G^s_{n,m,T})$ is a block-wise cardinality constrained optimization problem. The hyper-box bound and axis-aligned ellipsoid bound can be easily extended to such a situation. Consider the block-wise CCQO problem,

$$\begin{aligned} \mathcal{G}^m_s(D,d): \ & \min \ f^m(u) := \frac{1}{2}u'Du + d'u, \\ \text{Subject to: } & \sum_{t=0}^{T-1} \delta(u_t) \le s, \end{aligned}$$

where $D \in \mathbb{S}^{mT}_{++}$, $d \in \mathbb{R}^{mT}$, $u \in \mathbb{R}^{mT}$ and $m \ge 1$. Variable $u$ consists of $T$ blocks of vectors, $u' = (u'_1, u'_2, \cdots, u'_T)$ and $u'_t = (u^1_t, \cdots, u^m_t)$. Similarly, let $l' := D^{-1}d$ and $l$ consists of $T$ blocks, $l' = (l'_1, l'_2, \cdots, l'_T)$ with $l'_t := (l^1_t, \cdots, l^m_t)$.

The hyper-box bound $\mathcal{D}^m_{box}$ of problem $\mathcal{G}^m_s(D,d)$ can be constructed as follows,

$$\begin{aligned} \mathcal{D}^m_{box}: \ & \min \quad \tau \\ \text{Subject to: } & \frac{(u^i_t + l^i_t)^2}{\hat{D}^i_t} \le 2\tau - l'd, \ \text{for } t = 1, \cdots, T, \ i = 1, \cdots, m, \\ & \sum_{t=1}^{T} \delta(u_t) \le s. \end{aligned}$$

Let $\hat{D} \in \mathbb{R}^{mT}$ be the diagonal of $D^{-1}$ with $\hat{D}' := \left(\hat{D}'_1, \cdots, \hat{D}'_T\right)$ with $\hat{D}'_t := \left(\hat{D}^1_t, \cdots, \hat{D}^m_t\right)$. Similar to Theorem 5.6, we have the following result.

**Theorem 5.11.** *The optimal value of problem* $\mathcal{D}^m_{box}$ *is* $v(\mathcal{D}^m_{box}) = \frac{1}{2}\theta_{T-s}(\varrho) - \frac{1}{2}l'd$, *where* $\varrho := (\varrho_1, \cdots, \varrho_T)$ *and* $\varrho_t := \max_{i\in\{1,\cdots,m\}}\{(l^i_t)^2/\hat{D}^i_t\}$.

The proof of Theorem 5.11 is similar to that of Theorem 5.6, except that $\varrho_t$ is replaced by the maximum element of $(l^i_t)^2/\hat{D}^i_t$ for $i = 1, \cdots, m$.

The bound $\mathcal{D}^m_{diag}$ can be also extended for problem $\mathcal{G}^m_s(D,d)$. Let $\Lambda \in \mathbb{S}^{mT}_{+}$ be a block diagonal matrix,

$$\Lambda := \text{diag}\{\lambda^1_1, \cdots, \lambda^m_1, \cdots, \lambda^1_T, \cdots, \lambda^m_T\}.$$

Similar to (5.17), we consider the following class of problems,

$$\hat{\mathcal{G}}_s^m(\Lambda, \Lambda l, \xi): \quad \min \quad \left\{ \hat{f}^m(u) := \frac{1}{2}u'\Lambda u + (\Lambda l)'u + \xi \mid \sum_{t=1}^{T} \delta(u_t) \le s, \right\}. \quad (5.45)$$

Given $\Lambda$, problem $\hat{\mathcal{G}}_s^m(\Lambda, \Lambda l, \xi)$ can be solved explicitly.

**Lemma 5.12.** *The optimal value of problem $\hat{\mathcal{G}}_s^m(\Lambda, \Lambda l, \xi)$ is $\frac{1}{2}\sum_{j=1}^{T-s}\theta_j(\eta) - \frac{1}{2}l'\Lambda l + \xi$, where $\eta := (\eta_1, \cdots, \eta_T)$ and $\eta_t := \sum_{i=1}^{m}\lambda_t^i(l_t^i)^2$, for $t = 1, \cdots, T$.*

The condition (5.18) in Lemma 5.8 guarantees that the objective function of $\mathcal{G}_s^m(\Lambda, \Lambda l, \xi)$ is smaller than $\mathcal{G}_s^m(D, d)$, i.e., $\hat{f}^m(u) \le f^m(u)$ in the region $(u + l)'D(u + l) \le \omega$. Minimizing the duality gap between $\mathcal{G}_s^m(D, d)$ and $\mathcal{G}_s^m(\Lambda, \Lambda l, \xi)$ is equivalent to finding the optimal $\Lambda$ and $\xi$ of the following problem,

$$\begin{aligned} \mathcal{D}_{diag}^m \ &: \max_{\Lambda \succeq 0,\ \xi} \min_{u} f^m(u) = \frac{1}{2}u'\Lambda u + (\Lambda l)'u + \xi, \\ \text{Subject to}: &\ \sum_{i=1}^{T}\delta(u_i) \le s, \\ &\ \{\Lambda, \xi\} \text{ satisfies } \ (5.18). \end{aligned}$$

We then use Lemma 5.12 and the linear representation of the summation of $T - s$ smallest $\theta_i(\eta)$ in [79] to simplify the problem $\mathcal{D}_{diag}^m$ to the following form,

$$\begin{aligned} \mathcal{D}_{diag}^m \ &: \ \max_{\Lambda \succeq 0, \hat{\xi}} \ -(T - s)z + \sum_{t=1}^{T} a_t + \hat{\xi} \\ \text{Subject to}: &\ \{\Lambda, \xi\} \text{ satisfies } \ (5.18), \\ &\ a_t \le \frac{1}{2}\sum_{i=1}^{m}(l_t^i)^2\lambda_t^i + z, \ t = 1, \cdots, T, \\ &\ a_t \le 0, t = 1, \cdots, T, \end{aligned}$$

where $\hat{\xi} = \xi - \frac{1}{2}l'\Lambda l$, $a_t$, $t = 1, \cdots, T$ and $z$ are auxiliary variables.

## 5.3. Numerical Test

### 5.3.1. Random Tests for CCLQR Problem

The lower-bounding schemes proposed in previous sections are implemented on PC with 2.2GHz processor and 1.0GB RAM. We combine such lower bounding methods with a BnB algorithm. Our BnB routine is coded by C++ with SDPA package [28].

The parameters $n$, $m$, $T$, $A_t$, $B_t$, $R_t$ for $t = 0, \cdots, T-1$ and $Q_t$ for $t = 1 \cdots, T$ are defined in Chapter 2. The parameter $A_t$, $B_t$, $R_t$ and $Q_t$ are time variant. Without loss of generality, we generate the following class of time-variant linear systems by the following random generation procedure.

- Fix $Q_t = 0.5\mathbf{I}_n$ and $R_t = \mathbf{I}_m$
- Generate $B_t$ with its elements being uniformly distributed in $[-1, 1]$
- Generate $A_t = \alpha W_t^1 + (1-\alpha)W_t^2 + W$, where the norms of the eigenvalues of $W_t^1$ and $W_t^2$ are uniformly distributed in $[0, 0.8]$ and $[1, 1.6]$, respectively, and the elements of $W_t$ are normally distributed (following $N(0, \xi^2)$). Parameter $\alpha$ is set in $[0, 1]$ and $\xi$ is set at different values for different test problems to avoid data explosion of matrix $D$.

We compare the computational results of our BnB algorithm with the results by the standard mixed-integer solver in CPLEX [34]. Note that the block-wise CCQO problem $\mathcal{G}_s^m(D, d, c)$ can be reformulated as a mixed-integer programming problem.

$$\begin{aligned}
\mathcal{G}_s^m(D, d): \quad & \min \ f(u) := \frac{1}{2}u'Du + d'u, \\
\text{Subject to:} \quad & \|u_t\|_2^2 \le \zeta_t M_t, \quad \text{for} \quad t = 0, \cdots, T-1, \\
& \sum_{t=0}^{T-1} \delta(\zeta_t) \le s,
\end{aligned}$$

where $M_t$ is the upper bound of $\|u_t\|_2^2$ estimated by incumbent.

For each type of problems listed in Table 5.2, Table 5.3, Table 5.4 and Table 5.5, 50 randomly generated problems are examined, while the limit of CPU time is set at 60s, i.e., we stop the algorithm when the running time reaches 60s and record the incumbent. In the tables, Columns "Node", "Time", "Succ" and "optV" represent the average number of nodes explored, average CPU times, the number of problems solved successfully and the average optimal value, respectively. Both CPLEX solver and our BnB solver successfully solve all problems of Types 1 - 5 in Table 5.2 and Table 5.4 and our BnB solver performs much better. For problems of Types 6-10 in Table 5.3 and Table 5.5 , neither CPLEX solver nor our BnB solver solves all randomly generated problems successfully within 60 seconds. Compared with CPLEX, however, our BnB solver solves more instances successfully and obtains a better average objective value. Evidenced from Table5.3 and Table 5.5 (Types 6 - 10), our BnB algorithm visits much more nodes than the CPLEX solver within the same time limit, demonstrating much higher efficiency. In Table 5.6, we compare at the root node 4 different lower bounding schemes for their relative gaps defined by $(v^* - \underline{v})/|v^*|$, where $v^*$ is the optimal value and $\underline{v}$ is the lower bound value. Generally speaking, the bound $\mathcal{D}_{diag}$ offers the tightest bound for CCLQR problems, followed by "$\mathcal{D}_{surgt}$" or $\mathcal{D}_{box}$. Note that calculating $\mathcal{D}_{box}$ is much cheaper than calculating $\mathcal{D}_{diag}$ or "$\mathcal{D}_{surgt}$". Thus, in our BnB algorithm, we calculate bound $\mathcal{D}_{diag}$ at the root node to obtain a tight lower bound and calculate $\mathcal{D}_{box}$ at all children nodes to achieve a high speed.

We then check time-invariant cases of CCLQR problem, i.e., $A_t$, $B_t$, $Q_t$ and $R_t$ are kept unchange in whole horizon $T$. All the parameter are generated in the following way.

- Generate $B$ with its elements being uniformly distributed in $[-1, 1]$. Let $B_t = B$ for $t = 0, \cdots, T-1$. Fixed $Q_t = 0.5\mathbf{I}_n$ and $R_t = \mathbf{I}_m$.

- Generate $A = \alpha W^1 + W$, where the norms of the eigenvalues of $W^1$ is

Table 5.2: Computational results of CPLEX solver for time-variant cases, I

| Type | | CPLEX | | | |
|---|---|---|---|---|---|
| No | {n,m,T,s} | Node | CPU Times | Succ | Optv |
| 1 | {20,1,40,10} | 11871 | 3.8 | 50 | 5188 |
| 2 | {40,1,40,10} | 11359 | 4.2 | 50 | 30070 |
| 3 | {40,2,40,10} | 20849 | 20.8 | 50 | 5747 |
| 4 | {50,3,30,8} | 2374 | 4.1 | 50 | 12722 |
| 5 | {30,1,50,10} | 16481 | 8.7 | 50 | 8367 |

Table 5.3: Computational results of CPLEX solver for time-variant cases, II

| Type | | CPLEX | | | |
|---|---|---|---|---|---|
| No | {n,m,T,s} | Node | CPU Times | Succ | Optv |
| 6 | {50,1,50,10} | 73902 | 35.0 | 36 | 36865 |
| 7 | {30,1,60,15} | 47231 | 25.0 | 47 | 47090 |
| 8 | {50,1,60,15} | 81609 | 44.9 | 21 | 24571 |
| 9 | {60,1,70,15} | 63908 | 59.3 | 8 | 51618 |
| 10 | {60,1,80,20} | 65474 | 60.0 | 0 | 47650 |

uniformly distributed in $[0, 1.1]$, respectively, and the elements of $W$ are normally distributed (following $N(0, \xi^2)$). Set $\xi$ at a reasonable value to avoid data explosion of matrix $D$. Then let $A_t = A$ for $t = 0, \cdots, T-1$.

The results are shown in Table 5.7 and Table 5.8. Similarly, 50 trials are examined for each type of problem. The results shows that our BnB algorithm performs better than the CPLEX solver. Compared with the time-variant cases(see Table 5.4, Table 5.5, Table 5.2, Table 5.3), both solver use less time to solve the time-invariant problem for the same type of problem. We can conclude that the time invariant problems are in general easier to solve.

Table 5.4: Computational results of BnB solver for time-variant cases, I

| Type | | BnB | | | |
|---|---|---|---|---|---|
| No | {n,m,T,s} | Node | CPU Times | Succ | Optv |
| 1 | {20,1,40,10} | 8398 | 0.3 | 50 | 5118 |
| 2 | {40,1,40,10} | 8985 | 0.4 | 50 | 30070 |
| 3 | {40,2,40,10} | 14933 | 1.7 | 50 | 5747 |
| 4 | {50,3,30,8} | 1654 | 0.7 | 50 | 12722 |
| 5 | {30,1,50,10} | 11933 | 0.9 | 50 | 8367 |

Table 5.5: Computational results of BnB solver for time-variant cases, II

| Type | | BnB | | | |
|---|---|---|---|---|---|
| No | {n,m,T,s} | Node | CPU Times | Succ | Optv |
| 6 | {50,1,50,10} | 103427 | 4.9 | 49 | 36809 |
| 7 | {30,1,60,15} | 63893 | 5.7 | 50 | 47034 |
| 8 | {50,1,60,15} | 154232 | 14.3 | 48 | 24509 |
| 9 | {60,1,70,15} | 385344 | 40.8 | 25 | 50900 |
| 10 | {60,1,80,20} | 374651 | 53.1 | 10 | 47150 |

### 5.3.2. Random Test Problems for CCQO Problem

We compare various lower bounding schemes developed in the previous section for CCQO problem $\mathcal{G}_s(D,d)$. All the tested problems are randomly generated under the Matlab platform. The bounding scheme are coded under the Matlab and corresponding SDP and SOC problems are solved by using Sedumi [73]. In order to control the conditional number of matrix $D$, we generate the problem by following procedure:

- Generate $\chi := \lambda_{min}/\lambda_{max}$ uniformly in a pre-given interval $[\underline{\chi}, \bar{\chi}]$. Let $\lambda_{min} = 1$ and $\lambda_{max} = 1/\chi$. Generate $\lambda_i$ uniformly from $[\lambda_{min}, \lambda_{max}]$, for $i = 1, \cdots, T-2$. Let $\Lambda := \text{diag}\{\lambda_{min}, \lambda_1, \cdots, \lambda_{max}\}$. Generate the unitary

Table 5.6: Relative ratios of different lower bounding schemes for time variant cases

| {n,m,T,s} | $v(\mathcal{D}_{diag})$ | $v(\mathcal{D}_{box})$ | $v(\mathcal{D}_{surgt})$ | $v(\mathcal{D}_{cont})$ |
|---|---|---|---|---|
| {20,1,40,10} | 19.0% | 24.2% | 24.2% | 27.9% |
| {40,1,40,10} | 20.7% | 25.9% | 25.2% | 29.0% |
| {40,2,40,10} | 14.1% | 18.5% | 17.5% | 22.4% |
| {50,3,30,8} | 24.2% | 32.1% | 31.1% | 35.4% |
| {30,1,50,10} | 12.5% | 16.2% | 17.3% | 19.1% |
| {50,1,50,10} | 27.5% | 35.1% | 34.3% | 38.3% |
| {30,1,60,15} | 12.2% | 15.2% | 15.7% | 18.2% |
| {50,1,60,15} | 13.2% | 16.7% | 15.6% | 20.1% |
| {60,1,70,15} | 22.1% | 27.1% | 25.6% | 31.0% |
| {60,1,80,20} | 14.5% | 18.0% | 17.2% | 21.2% |

matrix $\Gamma$ and let $D = \Gamma'\Lambda\Gamma$.

- Generate $d$ with each of its element being uniformly distributed in $[-10, 10]$.

In Tables 5.9, 5.10, and 5.11, the column "$\zeta$" is the relative ratio defined as $\zeta = (v^* - \underline{v})/|v^*|$, where $\underline{v}$ is the lower bound and $v^*$ is the upper bound of problem $\mathcal{G}_s(D, d)$, respectively. The column "Time" records the cpu time of computation. We test 8 types of problems, i.e., problem with $\{T, s\}$ being set as $\{50, 10\}$, $\{50, 20\}$, $\{60, 10\}$, $\{60, 20\}$, $\{70, 10\}$,$\{70, 30\}$, $\{80, 10\}$ and $\{80, 30\}$. For each of these problems, the conditional number $\chi$ of $D$ is controlled in three intervals $[0.01, 0.1]$, $[0.2, 0.5]$ and $[0.6, 0.9]$. For all of the cases $\chi \in [0.6, 0.9]$, $\chi \in [0.2, 0.5]$ and $\chi \in [0.01, 0.1]$, the bound $(\mathcal{D}_{diag})$ is the tightest. However, it is expensive to compute such a bound. Although the bounds $\mathcal{D}_{ball}$, $\mathcal{D}_{ellp}$, $\mathcal{D}_{surgt}$ are not as tight as $\mathcal{D}_{diag}$, they can be computed more efficiently.

Table 5.7: Computational results of CPLEX solver for time-invariant cases

| Type | | CPLEX | | |
|---|---|---|---|---|
| No | {n,m,T,s} | Node | CPU Times | Succ |
| 1 | {20,1,40,10} | 6871 | 1.8 | 50 |
| 2 | {40,1,40,10} | 9359 | 3.2 | 50 |
| 3 | {40,2,40,10} | 10849 | 5.8 | 50 |
| 4 | {50,3,30,8} | 1074 | 0.5 | 50 |
| 5 | {30,1,50,10} | 12481 | 3.7 | 50 |
| 6 | {50,1,50,10} | 73402 | 8.0 | 50 |
| 7 | {30,1,60,15} | 43231 | 8.5 | 50 |
| 8 | {50,1,60,15} | 61609 | 9.9 | 50 |

## 5.4. Appendix

### 5.4.1. Distance Between Affine Apace and $\Delta(s)$

Given $H \in \mathbb{R}^{T\times k}$ and $h \in \mathbb{R}^T$ with $\text{rank}(H) = k$, define the following affine space,

$$\mathcal{Y}_k(H,h) := \left\{ y \in \mathbb{R}^T \mid y = H\beta + h, \beta \in \mathbb{R}^k \right\}.$$

We are interested in characterizing the following distance function $\text{dis}(\beta) := \min \|x-y\|_2^2$, where $x \in \Delta(s)$ and $y \in \mathcal{Y}_k(H,h)$,

$$\text{dis}(\beta) := \min_{x\in\mathbb{R}^T} \left\{ \|y-x\|_2^2 \mid y \in \mathcal{Y}_k(H,h), x \in \Delta(s) \right\}.$$

Let $H_i' \in \mathbb{R}^{1\times k}$ be the $i$-th row of matrix $H$ and $h_i$ be the $i$-th element of $h$. Let $g(\beta) \in \mathbb{R}^T$ with $g(\beta)' = (g_1(\beta), g_2(\beta), \cdots, g_T(\beta))$, where

$$g_i(\beta) := |H_i'\beta + h_i|, \text{ for } i = 1, \cdots, T. \tag{5.46}$$

Then, the following is true.

Table 5.8: Computational results of BnB solver for time-invariant cases

| Type | | BnB | | |
|---|---|---|---|---|
| No | {n,m,T,s} | Node | CPU Times | Succ |
| 1 | {20,1,40,10} | 4523 | 0.1 | 50 |
| 2 | {40,1,40,10} | 5110 | 0.2 | 50 |
| 3 | {40,2,40,10} | 1181 | 1.1 | 50 |
| 4 | {50,3,30,8} | 1532 | 0.6 | 50 |
| 5 | {30,1,50,10} | 11012 | 0.9 | 50 |
| 6 | {50,1,50,10} | 91980 | 3.2 | 50 |
| 7 | {30,1,60,15} | 41218 | 4.2 | 50 |
| 8 | {50,1,60,15} | 140309 | 9.9 | 50 |

**Theorem 5.13.** *The distance function dis($\beta$) is a piece-wise continuous quadratic function with respect to $\beta$, i.e.,*

$$dis(\beta) = \begin{cases} \beta' Q_1 \beta + q_1' \beta + c_1, & \text{if } A_1\beta \le b_1, \\ \quad \vdots & \quad \vdots \\ \beta' Q_N \beta + q_N' \beta + c_N, & \text{if } A_N\beta \le b_N, \end{cases}$$

*where the number $N$ is upper bounded by $\mathcal{O}\left((T(T-1))^k\right)$.*

Proof. For any fixed $\beta^*$, we have

$$\text{dis}(\beta^*) := \min_{x \in \Delta(s)} \|H\beta^* + h - x\|_2^2 = \sum_{i=1}^{T} (H_i'\beta^* + h_i - x_i)^2.$$

Since $x \in \Delta(s)$, at least $T - s$ of $x_i$ are zeros. To minimize $\|H\beta^* + h - x\|_2^2$, we must choose $x$ with its $s$ nonzero components to cancel out $s$ largest $|H_i'\beta^* + h_i|$ and the summation of the $(T-s)$ smallest $(H_i'\beta^* + h_i)^2$ determines the minimum distance dis($\beta^*$). We define an index set $\mathcal{I}(\beta) \subset \{1, 2, \cdots, T\}$ which consists of the first smallest $T - s$ indices of $g_i(\beta)$. Then

$$\text{dis}(\beta^*) = \sum_{i \in \mathcal{I}(\beta^*)} g_i(\beta)^2 = \beta' Q^* \beta + (q^*)'\beta + c^*, \tag{5.47}$$

Table 5.9: Comparison results of various bounds with $\chi \in [0.6, 0.9]$

| | Tested Problem with $0.6 \leq \gamma \leq 0.9$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\{T, s\}$ | $\{50, 10\}$ | | $\{50, 20\}$ | | $\{60, 10\}$ | | $\{60, 20\}$ | |
| | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) |
| $\mathcal{D}_{cont}$ | 0.61 | 0.00 | 0.14 | 0.00 | 0.72 | 0.00 | 0.19 | 0.00 |
| $\mathcal{D}_{ball}$ | 0.07 | 0.00 | 0.02 | 0.00 | 0.08 | 0.00 | 0.02 | 0.00 |
| $\mathcal{D}_{box}$ | 0.45 | 0.00 | 0.12 | 0.00 | 0.65 | 0.00 | 0.16 | 0.00 |
| $\mathcal{D}_{surgt}$ | 0.21 | 0.79 | 0.04 | 0.72 | 0.35 | 0.72 | 0.03 | 0.71 |
| $\mathcal{D}_{diag}$ | 0.05 | 13.03 | 0.01 | 14.01 | 0.05 | 17.21 | 0.01 | 19.02 |
| $\mathcal{D}_{ellp}$ | 0.06 | 0.00 | 0.02 | 0.01 | 0.07 | 0.01 | 0.02 | 0.01 |
| $\{T, s\}$ | $\{70, 10\}$ | | $\{70, 30\}$ | | $\{80, 10\}$ | | $\{80, 30\}$ | |
| | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) |
| $\mathcal{D}_{cont}$ | 0.87 | 0.00 | 0.13 | 0.00 | 1.04 | 0.00 | 0.19 | 0.00 |
| $\mathcal{D}_{ball}$ | 0.11 | 0.00 | 0.02 | 0.00 | 0.16 | 0.01 | 0.03 | 0.01 |
| $\mathcal{D}_{box}$ | 0.78 | 0.00 | 0.10 | 0.00 | 0.86 | 0.00 | 0.16 | 0.00 |
| $\mathcal{D}_{surgt}$ | 0.56 | 0.82 | 0.01 | 0.83 | 0.80 | 1.13 | 0.08 | 0.90 |
| $\mathcal{D}_{diag}$ | 0.07 | 23.01 | 0.08 | 25.12 | 0.10 | 49.12 | 0.02 | 52.23 |
| $\mathcal{D}_{ellp}$ | 0.10 | 0.00 | 0.01 | 0.01 | 0.15 | 0.01 | 0.03 | 0.01 |

where

$$Q^* := \sum_{i \in \mathcal{I}(\beta^*)} H_i H_i', \; q^* := 2 \sum_{i \in \mathcal{I}(\beta^*)} h_i H_i', \;\; c^* = \sum_{i \in \mathcal{I}(\beta^*)} h_i^2. \tag{5.48}$$

Clearly, the order of $g_i(\beta)$ changes when $\beta$ changes. Furthermore, the order of $g_i(\beta)$ changes only when $g_i(\beta) - g_j(\beta)$ switches sign for any pair $i$ and $j$, $i = 1, \cdots, T-1$ and $j = i, \cdots, T$. Note that

$$(g_i(\beta))^2 - (g_j(\beta))^2 = ((H_i + H_j)\beta + (h_i + h_j))((H_i - H_j)\beta + (h_i - h_j)). \tag{5.49}$$

Table 5.10: Comparison results of various bounds with $\chi \in [0.2, 0.5]$

| | Tested Problem with $0.2 \leq \chi \leq 0.5$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\{T, s\}$ | $\{50, 10\}$ | | $\{50, 20\}$ | | $\{60, 10\}$ | | $\{60, 20\}$ | |
| | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) |
| $\mathcal{D}_{cont}$ | 0.63 | 0 | 0.19 | 0.00 | 0.82 | 0.00 | 0.31 | 0.00 |
| $\mathcal{D}_{ball}$ | 0.28 | 0.00 | 0.08 | 0.00 | 0.33 | 0.00 | 0.12 | 0.00 |
| $\mathcal{D}_{box}$ | 0.56 | 0.00 | 0.15 | 0.00 | 0.70 | 0.00 | 0.28 | 0.00 |
| $\mathcal{D}_{surgt}$ | 0.51 | 0.57 | 0.11 | 0.60 | 0.56 | 0.63 | 0.24 | 0.90 |
| $\mathcal{D}_{diag}$ | 0.18 | 9.98 | 0.05 | 9.10 | 0.21 | 16.1 | 0.07 | 17.88 |
| $\mathcal{D}_{ellp}$ | 0.24 | 0.00 | 0.07 | 0.01 | 0.31 | 2.10 | 0.10 | 0.01 |
| $\{T, s\}$ | $\{70, 10\}$ | | $\{70, 30\}$ | | $\{80, 10\}$ | | $\{80, 30\}$ | |
| | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) |
| $\mathcal{D}_{cont}$ | 1.06 | 0.00 | 0.13 | 0.00 | 1.26 | 0.00 | 0.20 | 0.00 |
| $\mathcal{D}_{ball}$ | 0.47 | 0.00 | 0.06 | 0.00 | 0.60 | 0.01 | 0.10 | 0.01 |
| $\mathcal{D}_{box}$ | 0.90 | 0.00 | 0.11 | 0.00 | 1.13 | 0.00 | 0.18 | 0.00 |
| $\mathcal{D}_{surgt}$ | 0.79 | 0.79 | 0.89 | 0.78 | 0.80 | 1.13 | 0.12 | 0.90 |
| $\mathcal{D}_{diag}$ | 0.33 | 26.01 | 0.03 | 27.12 | 0.40 | 49.12 | 0.06 | 52.23 |
| $\mathcal{D}_{ellp}$ | 0.42 | 0.01 | 0.05 | 0.00 | 0.52 | 0.01 | 0.09 | 0.01 |

We borrow some concepts from the discrete geometry by considering the hyperplane arrangement generated by the following hyplanes in $\mathbb{R}^k$,

$$p_{i,j}^1 = \{\beta \in \mathbb{R}^k \mid (H_i + H_j)\beta + (h_i + h_j) = 0\}, \tag{5.50}$$

$$p_{i,j}^2 = \{\beta \in \mathbb{R}^k \mid (H_i - H_j)\beta + (h_i - h_j) = 0\}, \tag{5.51}$$

for $i = 1, \cdots, T-1$ and $j = i, \cdots, T$. The total number of these hyperplanes is $(T-1)T$. Note that a cell $E$ of the hyperplane arrangement corresponding to $p_{i,j}^t$'s, $t = 1, 2$, is a $(k)$-dimensional polyhedral set formed by the half spaces induced by $p_{i,j}^t$'s hyperplanes. Such a cell can be characterized by a $(T-1)T$

Table 5.11: Comparison results of various bounds with $\chi \in [0.01, 0.1]$

| | Tested Problem with $0.01 \le \chi \le 0.1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\{T, s\}$ | $\{50, 10\}$ | | $\{50, 20\}$ | | $\{60, 10\}$ | | $\{60, 20\}$ | |
| | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) |
| $\mathcal{D}_{cont}$ | 1.63 | 0 | 0.67 | 0.00 | 1.52 | 0.00 | 0.77 | 0.00 |
| $\mathcal{D}_{ball}$ | 1.04 | 0.00 | 0.52 | 0.00 | 1.26 | 0.00 | 0.62 | 0.00 |
| $\mathcal{D}_{box}$ | 1.20 | 0.00 | 0.51 | 0.00 | 1.30 | 0.00 | 0.65 | 0.00 |
| $\mathcal{D}_{surgt}$ | 1.09 | 0.71 | 0.53 | 0.56 | 1.22 | 0.57 | 0.62 | 0.60 |
| $\mathcal{D}_{diag}$ | 0.59 | 9.55 | 0.41 | 7.23 | 0.71 | 6.23 | 0.42 | 6.11 |
| $\mathcal{D}_{ellp}$ | 0.92 | 0.01 | 0.47 | 0.01 | 1.10 | 0.01 | 0.57 | 0.01 |
| $\{T, s\}$ | $\{70, 10\}$ | | $\{70, 30\}$ | | $\{80, 10\}$ | | $\{80, 30\}$ | |
| | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) | $\zeta$ | Time(s) |
| $\mathcal{D}_{cont}$ | 1.83 | 0 | 0.36 | 0.00 | 2.84 | 0.00 | 0.76 | 0.00 |
| $\mathcal{D}_{ball}$ | 1.21 | 0.00 | 0.31 | 0.00 | 2.00 | 0.00 | 0.63 | 0.00 |
| $\mathcal{D}_{box}$ | 1.64 | 0.00 | 0.30 | 0.00 | 2.50 | 0.00 | 0.69 | 0.00 |
| $\mathcal{D}_{surgt}$ | 1.32 | 0.71 | 0.29 | 0.96 | 1.40 | 0.87 | 0.50 | 0.90 |
| $\mathcal{D}_{diag}$ | 0.25 | 15.55 | 0.11 | 15.23 | 0.63 | 16.23 | 0.43 | 15.00 |
| $\mathcal{D}_{ellp}$ | 1.02 | 0.02 | 0.28 | 0.01 | 1.29 | 0.01 | 0.52 | 0.02 |

dimensional sign vector $w$, $\text{sign}(E) := (w_1, \cdots, w_{T-1})$, where

$$w_i := ((w^1_{i,i+1} w^2_{i,i+1}), (w^1_{i,i+2} w^2_{i,i+2}), \cdots, w^1_{i,T} w^2_{i,T})). \tag{5.52}$$

The sign of the hyperplane $w^1_{i,j}$ or $w^2_{i,j}$, for $j = i+1, \cdots, T$, is specified by

$$w^1_{i,j} = \begin{cases} + & \text{if } \ p^1_{i,j}(\beta) \ge 0 \\ - & \text{if } \ p^1_{i,j}(\beta) < 0 \end{cases}, \text{ for } j = i+1, \cdots, T, \tag{5.53}$$

$$w^2_{i,j} = \begin{cases} + & \text{if } \ p^2_{i,j}(\beta) \ge 0 \\ - & \text{if } \ p^2_{i,j}(\beta) < 0 \end{cases}, \text{ for } j = i+1, \cdots, T. \tag{5.54}$$

The order of $g_i(\beta)$ is determined by the sign vectors of all the cells of the hyperplane arrangement. It has been known that the number of cells of the hyperplane arrangement generated by (5.50) and (5.51) is upper bounded by $\mathcal{O}\left((T(T-1))^k\right)$ (see [83] and [27]). Since each cell of a hyperplane arrangement is a polyhedra, we could unify the expression of such polyhedra as $A_i\beta \leq b_i$. Taking the summation of $T-s$ smallest $g_i(\beta)$ yields the quadratic function in (5.47) and (5.48). Moreover, on some boundary $p^1_{i^*,j^*}$ ( or $p^2_{i^*,j^*}$ ) of an individual cell, it may hold that $g_{i^*}(\beta) = g_{j^*}(\beta)$ for some $i^*$ and $j^*$. On the two sides of this boundary, $g_{i^*}(\beta)$ and $g_{j^*}(\beta)$ change order. If $i^* \in \mathcal{I}(\beta)$ and $j^* \in \mathcal{I}(\beta)$ or $i^* \notin \mathcal{I}(\beta)$ and $j^* \notin \mathcal{I}(\beta)$, dis$(\beta)$ keeps the same form on both sides of the boundary. If $i^* \in \mathcal{I}(\beta)$ and $j^* \notin \mathcal{I}(\beta)$ or $i^* \notin \mathcal{I}(\beta)$ and $j^* \in \mathcal{I}(\beta)$, due to $g_{i^*}(\beta) = g_{j^*}(\beta)$, dis$(\beta)$ is still continuous on the boundary. □

In the following, we further distinguish the discussion of the distance function dis$(\beta)$ for $k = 1$ and $k > 1$.

**dis$(\beta)$ with $k = 1$**

**Corollary 5.14.** *When $k = 1$, the distance function dis$(\beta)$ is a piece-wise quadratic function,*

$$dis(\beta) = a_j\beta^2 + b_j\beta + c_j, \quad I_j \leq \beta \leq I_{j+1},$$

*for $j = 1, \cdots, N$, where $N \leq T(T-1)$.*

Proof. The proof of the corollary follows Theorem 5.13. The cell of the hyperplane arrangement degenerates to an interval on a real line when $k = 1$. The upper bound of $N$ can be exactly calculated. Clearly, the set $\mathcal{I}(\beta)$ changes only when some function $g_i(\beta)$ intersects with $g_j(\beta)$, with $i \in I(\beta)$ and $j \notin I(\beta)$. That is to say, $N \leq S_T$, where $N$ is the number of total changes of $I(\beta)$ as $\beta$ varies from $-\infty$ to $\infty$ and $S_T$ is the total number of intersection points between the functions $g_i(\beta)$ and $g_j(\beta)$ for $i \neq j$, and $i, j = 1, \cdots, T$. The number $S_T$ can be computed in a recursive way, e.g., $S_2 = 2$ and $S_t = S_{t-1} + 2(t-1)$ for $t \geq 3$. Solving such a recursion yields $S_t = (t-1)t - 2$ for $t = 2, \cdots, T$. Thus, theoretically, we can divide the interval $[-\infty, \infty]$ into at most $N \leq T(T-1)$ consecutive intervals $[I_j, I_{j+1}]$, for $i = 1, \cdots, N$. □

From Corollary 5.14, although the total number of intervals $N$ is bounded by $T(T-1)$, it is expensive and unnecessary to compute $N$ quadratic functions directly, while identifying distance function $\text{dis}(\beta)$ in a pre-given interval. Suppose that we identify the distance function $\text{dis}(\beta)$ with $\beta \in [\omega_l, \omega_u]$. Without loss of generality, we assume

$$-\infty < -\frac{h_1}{H_1} \leq -\frac{h_2}{H_2} \leq \cdots \leq -\frac{h_T}{H_T} < \infty.$$

In each of the intervals, $[-\infty, -\frac{l_1}{h_1}]$, $\cdots$, $[\frac{l_n}{h_n}, \infty]$, the function, $g_i(\beta)$, is monotone with respect to $\beta$. Thus, the interval $[\omega_l, \omega_u]$ can be partitioned by some monotone intervals. For any one of these monotone intervals, we could use Algorithm 1 to identify the distance function $\text{dis}(\beta)$. In Algorithm 1, we sequentially check the intersection point between function $g_i(\beta)$, $i \in \mathcal{I}(\beta)$ and $g_j(\beta)$, $j \notin \mathcal{I}(\beta)$. We introduce the following Table $\mathbb{T}$ to store the data,

| $\mathbb{T}$ | $i_1$ | $i_2$ | $\cdots$ | $i_s$ |
|---|---|---|---|---|
| $i_1$ | $\mathbb{T}(1,1)$ | $\mathbb{T}(1,2)$ | $\cdots$ | $\mathbb{T}(1,s)$ |
| $i_2$ | $\mathbb{T}(2,1)$ | $\mathbb{T}(2,2)$ | $\cdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| $i_{T-s}$ | $\mathbb{T}(T-s,1)$ | $\cdots$ | $\cdots$ | $\mathbb{T}(T-s,s)$ |

The column 0 and row 0 are filled with the index set $\mathcal{I}_\beta$ and its complementary set $\bar{\mathcal{I}}_\beta := \{1, \cdots, T\} \setminus \mathcal{I}(\beta)$. In Algorithm 1, since all $g_i(\beta)$ are monotone, we assume that each $g_i(\beta)$ takes linear form $g_i(\beta) = H_i\beta + h_i$, for $i = 1, \cdots, T$. Since there is only one index getting out and one index getting into the index $\mathcal{I}(\beta)$ in each consecutive interval, we only modify one column and one row in each step of operation, which leads to a linear time operation $\mathcal{O}(T)$.

**Example 5.2.** Consider an example with $T = 6$, $k = 1$ and $s = 3$. The initial searching interval of $\beta$ is $[-1, 1.3]$. We want to find the distance function $\text{dis}(\beta)$

**Algorithm 1** Monotone interval searching:

Input : Interval $[I_1, I_2]$, $H$, $h$.

S0 : Let $lb \leftarrow I_1$, $ub \leftarrow I_1$, $J^* \leftarrow +\infty$ and $endflag \leftarrow 0$. Sort $g_i(lb)$ for $i = 1, \cdots, T$ in an ascending order. Choose the first $T - s$ minimum of them to construct $\mathcal{I}(\beta_1)$. Place remaining indices in $\bar{\mathcal{I}}(\beta_1)$. Initialize Table $\mathbb{T}$ by filling column 0 and row 0 by index set $\mathcal{I}(lb)$ and $\bar{\mathcal{I}}(lb)$, respectively. Let

$$\mathbb{T}(i,j) = \begin{cases} \frac{h_{ti}-h_{tj}}{H_{tj}-h_{ti}} \text{ if } ub \leq \frac{H_{ti}-H_{tj}}{H_{tj}-H_{ti}} \leq lb, \\ +\infty \qquad \qquad otherwise, \end{cases} \tag{5.55}$$

where $ti = \mathbb{T}(i, 0)$ and $tj = \mathbb{T}(0, j)$ for $i = 1, \cdots, T - s$ and $j = 1, \cdots, s$. Go to [S1].

S1 : If $\mathbb{T}(i, j) \neq \infty$, for $i = 1, \cdots, T - s$, $j = 1, \cdots, s$, go to [S3]. Otherwise, go to [S2].

S2 **:** Find $\mathbb{T}(i^*, j^*) = \min_{1 \leq i, 1 \leq j} \mathbb{T}(i, j)$. Let $lb \leftarrow ub$, $ub \leftarrow \mathbb{T}(i^*, j^*)$ and $\mathbb{T}(i^*, j^*) \leftarrow \infty$. If $ub = I_2$, let $endflag = 1$. Go to [S4].

S3 : Use $\mathbb{T}(i, 0)$, $i = 1, \cdots, T - s$ to construct $\mathcal{I}(\beta_1)$. Identify $\text{dis}(\beta_1)$ as (5.47) and (5.48). Record $lb$ and $ub$. If $endflag = 1$, stop. Otherwise, Go to [S4].

S4 : Exchange $\mathbb{T}(0, j^*)$ and $\mathbb{T}(i^*, 0)$. Update $j^*$-th column and $i^*$-th row of Table $\mathbb{T}$ by 5.55. Go to [S1].

between $\mathcal{Y}_1(H,h)$ and $\Delta(s)$, where

$$H' = \begin{pmatrix} -0.5 & 0.5 & 1 & -0.75 & -2 & -4 \end{pmatrix}',$$
$$h' = \begin{pmatrix} 1 & 2.5 & 4 & 3.4 & 3.5 & 5.2 \end{pmatrix}'.$$

Functions $g_i(\beta)$ are specified as (see the Figure 5.5),

$$g_1(\beta) = |-0.5\beta + 1|, \quad g_2(\beta) = |0.5\beta + 2.5|,$$
$$g_3(\beta) = |\beta + 4|, \quad g_4(\beta) = |-0.75\beta + 3.4|,$$
$$g_5(\beta) = |-2\beta + 3.5|, \quad g_6(\beta) = |-4\beta + 5.2|.$$

In each of the following sub intervals, $g_i(\beta)$, $i = 1, \cdots, 6$, are monotone,

$$[-5, 5] = \cup[-5, -4] \cup [-4, 1.3] \cup [1.3, 1.75] \cup [1.75, 2] \cup [2, 4.53] \cup [4.53, 5]$$

Thus, in the interval $[-1, 1.3]$, we have

$$g_1(\beta) = -0.5\beta + 1, \quad g_2(\beta) = 0.5\beta + 2.5,$$
$$g_3(\beta) = \beta + 4, \quad g_4(\beta) = -0.75\beta + 3.4,$$
$$g_5(\beta) = -2\beta + 3.5, \quad g_6(\beta) = -4\beta + 5.2.$$

We then use Algorithm 1 to identify the distance function in the interval $[-1, 1.3]$. At the first step, let $lb = ub = -1$ and we compute $g_i(-1)$ for $i = 1, \cdots, 6$. Initialize the index set as $\mathcal{I}(-1) = \{1, 2, 3\}$ and $\bar{\mathcal{I}}(-1) = \{4, 5, 6\}$ according to the order of $g_i(-1)$. We construct Table 5.12. The last column records the minimum in each row. We find the minimum in Table 5.12 to be

Table 5.12: Table of Step 1 in Example 5.2

| Table $\mathbb{T}$ | col 0 | col 1 | col 2 | col 3 | col4 |
|---|---|---|---|---|---|
| row 0 | $\mathcal{I}$ | 6 | 4 | 5 | min |
| row 1 | 1 | 1.200 | $+\infty$ | $+\infty$ | 1.200 |
| row 2 | 2 | 0.600 | 0.720 | 0.400 | 0.400 |
| row 3 | 3 | 0.240 | $-0.343$ | $-0.167$ | $-0.342$ |

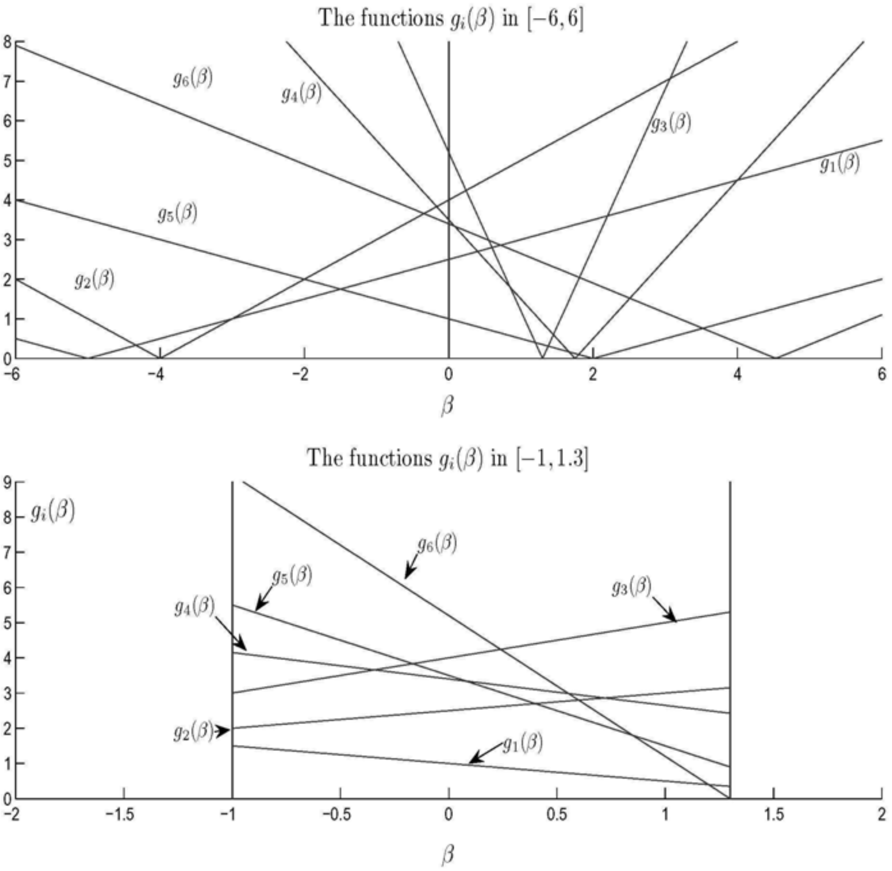

Figure 5.5: Functions $g_i(\beta)$ in Example 5.2

$-0.342$. Then we identify the interval of $\mathrm{dis}(\beta)$ as $[lb, ub] = [-1, -0.342]$. The distance function is

$$\mathrm{dis}(\beta) = 1.5\beta^2 + 9.5\beta + 23.25, \quad \beta \in [-1, -0.342].$$

Since $-0.342$ is in row 3 and column 2, we update these row and column of Table 5.12, leading to Table 5.13. The minimum in Table 5.13 is 0.08, the interval is $[lb, ub] = [-0.342, 0.08]$ and $\mathcal{I}([-0.342, 0.080]) = \{1, 2, 4\}$. The distance function in this interval is

$$\mathrm{dis}(\beta) = 1.062\beta^2 - 36\beta + 18.51, \quad \beta \in [-0.342, 0.080]. \tag{5.56}$$

Then we update Table 5.13 to Table 5.14. The minimum of Table 5.14 is 0.60,

Table 5.13: Table of Step 2 in Example 5.2

| Table $\mathbb{T}$ | col 0 | col 1 | col 2 | col 3 | col 4 |
|---|---|---|---|---|---|
| row0 | $\mathcal{I}$ | 6 | 3 | 5 | min in row |
| row1 | 1 | 1.200 | $+\infty$ | $+\infty$ | 1.200 |
| row2 | 2 | 0.600 | $+\infty$ | 0.400 | 0.400 |
| row3 | 4 | 0.554 | $+\infty$ | 0.008 | 0.08 |

Table 5.14: Table of Step 3 in Example 5.2

| Table $\mathbb{T}$ | col0 | col1 | col 2 | col3 | col4 |
|---|---|---|---|---|---|
| row0 | $\mathcal{I}$ | 6 | 3 | 4 | min in row |
| row1 | 1 | 1.200 | $+\infty$ | $+\infty$ | 1.200 |
| row2 | 2 | 0.600 | $+\infty$ | 0.720 | 0.600 |
| row3 | 5 | 0.850 | $+\infty$ | $+\infty$ | 0.85 |

the interval is $[lb, ub] = [0.08, 0.60]$ and $\mathcal{I}([0.08, 0.60]) = \{1, 2, 3\}$. The distance function in this interval is

$$\text{dis}(\beta) = 4.5\beta^2 - 12.5\beta + 19.5, \quad \beta \in [0.08, 0.60]. \tag{5.57}$$

Then we update table as Table 5.15. The minimum of Table 5.15 is 1.3, the interval is $[lb, ub] = [0.60, 1.3]$ and $\mathcal{I}([0.60, 1.3]) = \{1, 2, 3\}$. The distance function in this interval is

Table 5.15: Table of Step 4 in Example 5.2

| Table $\mathbb{T}$ | col0 | col1 | col 2 | col3 | col4 |
|---|---|---|---|---|---|
| row0 | $\mathcal{I}$ | 2 | 3 | 4 | min in row |
| row1 | 1 | $+\infty$ | $+\infty$ | $+\infty$ | $+\infty$ |
| row2 | 6 | $+\infty$ | $+\infty$ | $+\infty$ | $+\infty$ |
| row3 | 5 | $+\infty$ | $+\infty$ | $+\infty$ | $+\infty$ |

$$\text{dis}(\beta) = 20.25\beta^2 - 56.6\beta + 40.29, \quad \beta \in [0.60, 1.3]. \tag{5.58}$$

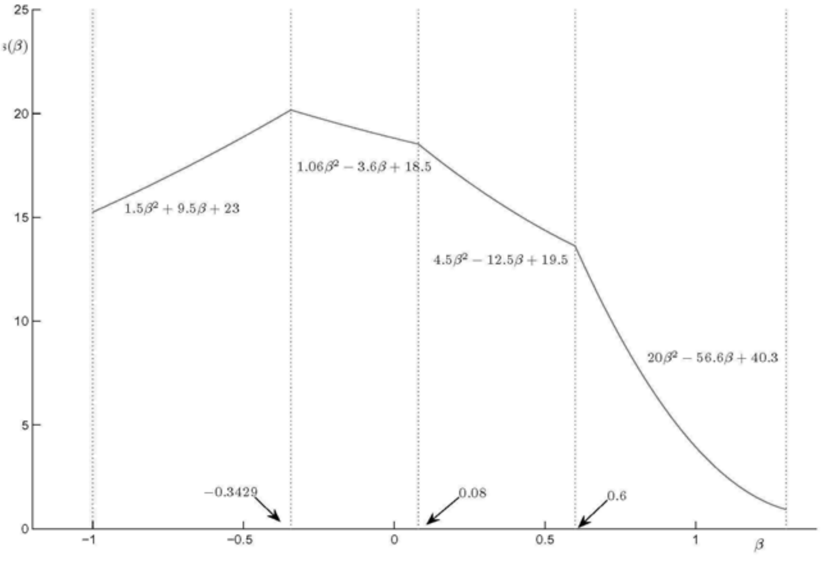

Figure 5.6: Distance function $\mathrm{dis}(\beta)$ in $[-1.3, 1]$

All elements of Table 5.15 is $\infty$. We complete characterization of the distance function $\mathrm{dis}(\beta)$. (See Figure 5.6).

### dis($\beta$) with $k > 1$

From Theorem 5.13, we know that the distance function $\mathrm{dis}(\beta)$ can be identified by checking the cells of hyperplane arrangement. Finding the cells of the hyperplane arrangement has been investigated in the literature. For example, the authors proposed in [3] and [72] a cell enumeration method by reverse searching method. Such a method consumes $\mathcal{O}((T(T-1))^k C_{lp})$ time to enumerate all the cells, where $C_{lp}$ is the time for a linear programming. Note that the cells are searched in the whole $\mathbb{R}^T$ space. Here we are interested in enumerating all the cells in a bounded region. Suppose that $\beta$ is bounded by $v_i \leq \beta_i \leq \vartheta_i$ for $i = 1, \cdots, k$. The algorithm listed in [72] is ready to solve such a cell enumeration problem with type of box boundary except for a modification of the routine

*AllAjc* in [72].

As we have illustrated in Theorem 5.13, the sign of each cell is a one-to-one mapping to the order of all $g_i(\beta)$. Here we specify such a mapping. We follow the notation used in Theorem 5.13. The sign vector $w = \text{sign}(E)$ is settled in a $(T-1)\times T$ sign matrix $\Omega(E)$ with $\Omega_{i,j}(E)$ being its element in the $i$-th row and the $j$-th column. Let "$\circ$" be an operator, i.e., $(+\circ+) = +$, $(+\circ-) = -$, $(-\circ+) = -$ and $(-\circ-) = +$. We make the following arrangement.

- let $\Omega_{i,i}(E) = 0$.
- The upper-triangle of $\Omega(E)$ is set as, $\Omega(E)_{i,j} = w^1_{i,j}\circ w^2_{i,j}$, for $i = 1,\cdots,T-1$ and $j = i+1,\cdots,T$.
- The lower-triangle of $\Omega(E)$ takes the opposite sign of the upper -triangle.

From (5.49), (5.52) and (5.53), we can conclude that $g_i(\beta)$ is the $t$-th smallest in cell $E$, if there is $t$ "+$''$ in $i$-th row of $\Omega(E)$, i.e., the $i$-th row of $\Omega(E)$ is given as follows,

$$\underbrace{(+,-,+,\cdots,-)}_{\text{there are } t \text{ "+"}} \longleftrightarrow \quad g_i(\beta) \text{ is } t\text{-th smallest elemetnt.}$$

**Example 5.3.** Consider an example with $T = 4$, $s = 2$ and

$$H = \begin{pmatrix} 4 & 2 \\ 5 & 1 \\ 1 & -1 \\ 2 & -0.5 \end{pmatrix}, \quad h = \begin{pmatrix} -2 \\ -6 \\ -1 \\ 2 \end{pmatrix}.$$

We consider in this example the box region $-1 \le \beta_1 \le 4$ and $-1.5 \le \beta_2 \le 1$ and introduce the hyperplanes as,

$$\begin{aligned}
&p^1_{1,2} = -\beta_1 + \beta_2 + 4 = 0, \;\; p^2_{1,2} = 9\beta_1 + 3\beta_2 - 8 = 0,\\
&p^1_{1,3} = 3\beta_1 + 3\beta_2 - 1 = 0, \;\; p^2_{1,3} = 5\beta_1 + \beta_2 - 3 = 0,\\
&p^1_{1,4} = 2\beta_1 + 2.5\beta_2 - 4 = 0, \;\; p^4_{1,2} = 6\beta_1 + 1.5\beta_2 = 0,\\
&p^1_{2,3} = 4\beta_1 + 2\beta_2 - 5 = 0, \;\; p^2_{2,3} = 6\beta_1 - 7 = 0,\\
&p^1_{2,4} = 3\beta_1 + 1.5\beta_2 - 8 = 0, \;\; p^2_{2,4} = 7\beta_1 + 0.5\beta_2 - 4 = 0,\\
&p^1_{3,4} = -\beta_1 - 0.5\beta_2 - 3 = 0, \;\; p^2_{1,2} = 3\beta_1 - 1.5\beta_2 + 1 = 0.
\end{aligned}$$

It can be verified that the box region $-1 \leq \beta_1 \leq 4$ and $-1.5 \leq \beta_2 \leq 1$ is on one side of the following hyperplane,

$$p^2_{1,3} > 0, \; p^2_{1,4} > 0, \; p^2_{2,4} > 0, \; p^1_{3,4} < 0, \; p^2_{3,4} > 0.$$

Thus, $\omega^2_{1,3} = 1$, $\omega^2_{1,4} = 1$, $\omega^2_{2,4} = 1$, $\omega^1_{3,4} = -1$, $\omega^2_{3,4} = 1$. We can enumerate the cells of the arrangements generated by these hyperplanes in the box region $\beta_1 \in [1, 4]$ and $\beta_2 \in [-1.5, 1]$. By using the algorithm of cell enumeration, the sign vector of the hyperplane arrangement are listed in Table 5.16. All the hyperplanes and the arrangement are illustrated in Figure 5.7. Then we can write out the order of the $g_i$ in each cell. For example, let us consider $cell_2$ in Table 5.16 with

$$\mathrm{sign}(cell_2) = ((-+, ++, ++), (++, ++), (+-)).$$

The corresponding sign matrix is

$$\Omega(cell_2) = \begin{pmatrix} 0 & - & + & + \\ + & 0 & + & + \\ - & - & 0 & - \end{pmatrix},$$

leading to $g_3(\beta) < g_4(\beta) < g_1(\beta) < g_2(\beta)$. Once the order of $g_i(\beta)$ is known, the distance function can be expressed by using (5.47) and (5.48),

$$\mathrm{dis}(\beta) = \beta' \begin{pmatrix} 5 & -2 \\ -2 & 1.25 \end{pmatrix} \beta + \begin{pmatrix} 6 & 0 \end{pmatrix} \beta + 5.$$

The other pieces of the distance function can be expressed in the similar fashion.

Table 5.16: The cells of hyperplanes in Example 5.3

| No | $(w_{1,2}, w_{1,3}, w_{1,4}), (w_{2,3}, w_{2,4}), (w_{3,4})$ |
|---|---|
| 1 | $(++, ++, ++), (++, ++), (-+)$ |
| 2 | $(-+, ++, ++), (++, ++), (-+)$ |
| 3 | $(-+, ++, -+), (++, ++), (-+)$ |
| 4 | $(-+, ++, -+), (++, +-), (-+)$ |
| 5 | $(++, ++, -+), (++, +-), (-+)$ |
| 6 | $(++, ++, -+), (-+, +-), (-+)$ |
| 7 | $(++, +-, -+), (-+, +-), (-+)$ |
| 8 | $(+-, +-, -+), (-+, +-), (-+)$ |
| 9 | $(+-, +-, -+), (--, +-), (-+)$ |
| 10 | $(+-, --, -+), (--, +-), (-+)$ |
| 11 | $(+-, --, -+), (--, ++), (-+)$ |
| 12 | $(+-, --, -+), (-+, ++), (-+)$ |
| 13 | $(+-, --, -+), (+-, ++), (-+)$ |
| 14 | $(+-, --, ++), (+-, ++), (-+)$ |
| 15 | $(+-, --, ++), (++, ++), (-+)$ |

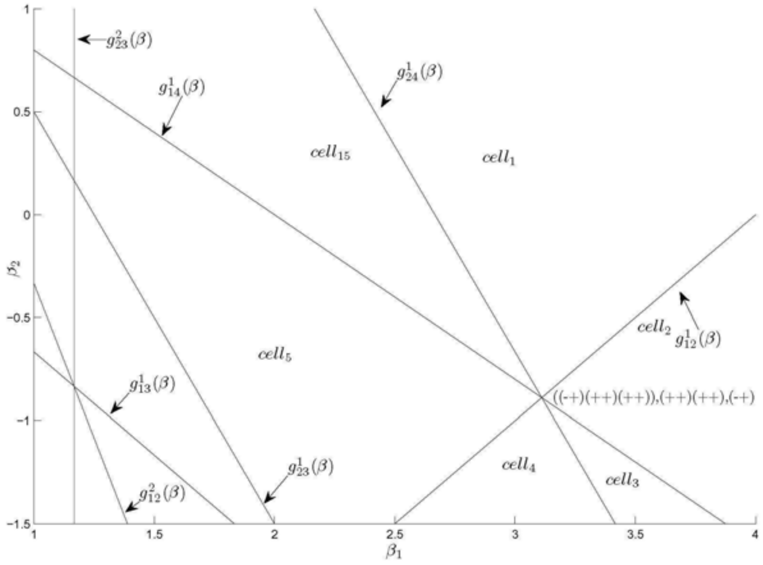

Figure 5.7: The cells of hyperplane arrangement in Example 5.3

# CHAPTER 6

# Conclusion

Stimulated by the urgent need of developing efficient solution algorithms for solving challenging cardinality constrained optimization problems arisen in real-world applications, we focus in this research on four specific cardinality constrained optimization problems, i.e., (i) the cardinality constrained linear-quadratic control problem, (ii) the cardinality constrained optimal control of linear switched systems, (iii) time cardinality constrained dynamic mean-variance portfolio selection and (iv) the cardinality constrained quadratic optimization problem. We give some conclusion remarks in this chapter and discuss some possible directions for further research.

As decision makers always prefer solutions of a feedback type for dynamic optimization (control) problems, we strive for the analytical solutions for both Problems (i) and (ii) by using sophisticated dynamic programming (DP). Due to the combinatorial nature of the possibilities in satisfying the cardinality constraint, even the elegant linear-quadratic setting does not help much this time under the cardinality constrained settings. Although the solution procedures of a Ricciti-type are derived in Chapters 2 and 3 for Problems (i) and (ii), respectively, the computation of the feedback gain is not a polynomial procedure in general, except for the case of $n = 1$. We thus switch our efforts to solution algorithms of a branch and bound type in order to alleviate and reduce the computational burden. For Problem (i), we reformulate it as a block-wise cardinality constrained quadratic programming problem(CCQO) and develop a lower bound from a corresponding sCCLQR formulation. For Problem (ii), we revise the DP procedure directly and combine it with semidefinite programming to construct a lower bound. Based on our results for Problems (i) and (ii), we may consider some possible extensions, including finding a polynomial approximation schemes (PTAS) of Problem (i) or (ii), enhancing the lower bound of problem (ii) by investigating its mathematical programming formulation, and generalizing the results of Problems (i) and (ii) to corresponding problem formulations with Gaussian-noise.

For problem (iii), due to the non-separability of the variance term and the cardinality constraint, it seems, at first glance, hopeless to find the exact analytical solution of such a problem. Fortunately, due to the embedding scheme developed in [43], such a problem can be tackled by a two-step procedure: 1) constructing and solving an auxiliary problem, which is a sparable stochastic linear-quadratic control problem; and 2) finding the exact auxiliary problem which offers the solution of the primary problem. Luckily, the analytical solution is achievable for the auxiliary problem of TCCMV (Problem (iii)) and hence the TCCMV problem can be solved explicitly. As we have mentioned, one interesting extension of Problem (iii) is to consider only investment policies which are confined to be buy-and-hold policies, where the management fee is charged when the policy is revised.

We developed various lower-bounding schemes for CCQO (Problem (iv)) in Chapter 5 by exploring geometric features hidden behind the cardinality constrained quadratic optimization problems. Compared to the existing literature where a lower bound is obtained by simply relaxing the cardinality constraint, our main idea in constructing lower bounds is more innovative. More specifically, we modify the objective function while keeping the cardinality constraint such that an analytical solution can be achieved under such settings. Utilizing the prominent geometric features of Problem (iv), we purposely identify certain subclasses of cardinality constrained optimization problems which can be solved in polynomial time and use them to lower bound the primary problem. Our newly proposed novel lower bounding schemes have been proven in our numerical tests to perform much better than the outcomes based on the previous thinking.

# Bibliography

[1] A.Cerny and J. Kallsen. On the structure of general mean-variance hedging strgtegies. *Annals of Probability*, 35:1479–1531, 2007.

[2] B. O. D. Anderson. *Optimal Control: Linear Quadratic Methods.* Prentice Hall, Englewood Cliff, NJ, 1990.

[3] D. Avis and K. Fukuda. Reverse search for enumeration. *Discrete Applied Mathematics*, 65:21–46, 1996.

[4] A. Beck and M. Teboulle. Global optimality conditions for quadratic optimization problems with binary constraints. *SIAM Journal on Optimization*, 11:179–188, 2000.

[5] A. Bemporad, M.Morari, V. Dua, and EN. Pistikopoulos. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002.

[6] A. Bemporad and M. Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–427, 1999.

[7] S. C. Bengea and R. A. DeCarlo. Optimal control of switching systems. *Automatica*, 41(1):11–27, 2005.

[8] D. Bertsimas and D. B. Brown. Constrained stochastic lqc: a tractable approach. *IEEE Transactioin on Automatic Control*, 52(10):1826–1814, 2007.

[9] D. Bertsimas and R. Shioda. Algorithm for cardinality-constrained quadratic optimization. *Computational Optimization and Applications*, 2007.

[10] T. R. Bielecki, H. Q. Jin, S. R. Pliska, and X. Y. Zhou. Continuous-time mean-variance portfolio selection with bankruptcy prohibition. *Mathematical Finance*, 15:213–244, 2005.

[11] D. Bienstock. A computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74:121–140, 1996.

[12] B. Blog, G. van der Hoek, A. H. G. Rinnooy Kan, and G. T. Timmer. The optimal selelction of portfolios. *Management Science*, 29(792–798):7, 1983.

[13] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University, 2004.

[14] S. J. Brown, W. N. Goetzmann, and B. Liang. Fees on fees in funds of funds. *Yale International Center for Finance*, pages Yale ICF Working Paper No. 02–33, June 14, 2004.

[15] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete freguency information. *IEEE Transcation on Information Theory*, 52(2):489–509, 2006.

[16] A. Cerny and J. Kallsen. Hedging by sequential regressions revisited. *To appear in Mathematical Finance*, 2009.

[17] S. W. Chan and G. C. Goodwin. Convergence properties of the riccati difference equation in optimal filtering of nonstabilizable system. *IEEE Transaction on Automatic Control*, 29(2):110–118, 1984.

[18] T. J. Chang, J. E. Beasley, and Y. M. Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers and Operations Research*, pages 1271–1302, 2000.

[19] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[20] M. C. Chiu and D. Li. Asset and liability management under a continuous-time mean variance optimization framework. *Insurance: Mathematics and Economics*, 39:330–355, 2006.

[21] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction To Algorithms*. The MIT Press, 2001.

[22] X. L. Sun D. Li and J. Wang. Optimal lot solution to cardinality constrained mean-variance formulation for portfolio selection. *Mathematical Finance*, 16(1):83–101, 2006.

[23] M. Dai and Y. Zhong. Finite horizon optimal investment with transaction costs: a parabolic double obstacle problem. *Working paper, Social Science Research Network*, SSRN-id875225, 2006.

[24] M. Dai and Y. Zhong. Penalty methods for continuous-time portfolio selection with proportional transaction costs. *Working paper, Social Science Research Network*, SSRN-id121015, 2006.

[25] M. H. A. Davis and A. R. Norman. Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15:676–713, 1990.

[26] D. L. Donoho. Compressed sensing. *IEEE Transaction Information Theory*, 52(4):1289, 2006.

[27] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer, 1979.

[28] K. Fujisawa, M. Kojima, K. Nakata, and M. Yamashita. Sdpa (semidefinite programming algorithm) user's manual — version 6.2.0. *Research Report B-308, Dept. Math and Comp. Sciences, Tokyo Institute of Technology*, 2004.

[29] M. R. Garey and D. S. Johnson. *Computer and Intractability, A Guide to The Thoery of NP-Completeness.* W. H. Freeman Co, Francisco, 1979.

[30] A. Giua, C. Seatzu, and C. Van Der Mee. Optimal control of switched autonomus linear systems. In *Proceeding of 40th IEEE Conference on Decision and Control*, pages 2472–2477, Orlando, Florida, USA, 2001.

[31] J. H. Golec. The effects of mutual fund managers' characteristics on their portfolio performance, risk and fees. *Financial Services Review*, 5:133–148, 1996.

[32] S. Hedlund and A. Rantzer. Convex dynamic programming for hybrid systems. *IEEE transaction on Automatic Control*, 47(9):1536–1540, 2002.

[33] Y. Hu and X. Y. Zhou. Constrained stochastic lq control with random coefficients, and application to portfolio selection. *SIAM Journal on Control and Optimization*, 44:444–466, 2005.

[34] ILOG. *ILOG CPLEX 9.0 User Manual.* ILOG CPLEX Division, Incline Village,NV, 2003.

[35] R. Korn. Portfolio optimization with strictly positive transaction costs and impulse control. *Finance and Stochastics*, 2:85–114, 1998.

[36] P. Lancaster and L. Rodman. *Algebraic Riccati equations.* Clearendon Press, Oxford, 1995.

[37] M. Leippold, F. Trojani, and P. Vanini. A geometric approach to multiperiod mean variance optimization of assets and liabilities. *Journal of Economic Dynamics and Control*, 28:1079–1113, 2004.

[38] F. Lewis and V. Syrmos. *Optimal Control.* John Wiley and Sons, New York, 1995.

[39] D. Li. On the minimax solution of multiple linear-qudaratic problems. *IEEE Transaction Automatic Control*, 35:1153–1156, 1990.

[40] D. Li. Hierarchical control for large-scale systems with general multiple linear-quadratic structure. *Automatica*, 29:1451–1461, 1993.

[41] D. Li. On general multiple linear-quadratic control problems. *IEEE Transaction on Automatic Control*, 38:1722–1727, 1993.

[42] D. Li and J. J. Gao. Cardinality constrained linear-quadratic control in discrete-time. *Special Issue on Control Applications of Optimisation - Control and Aeronautics, Optimal Control, Control of Partial Differential Equations, International Journal of Tomography and Statistics*, 5:103–108, 2007.

[43] D. Li and W. L. Ng. Optimal dynamic portfolio selection: multiperiod mean-variance formulation. *Mathematical Finance*, 10:387–406, 2000.

[44] D. Li and C. W. Schmidt. Cost smoothing in discrete-time linear-quadratic control. *Automatica*, 33:447–452, 1997.

[45] D. Li and X. L. Sun. *Nonlinear Integer Programming*. Springer, Boston, 2006.

[46] X. Li and X. Y. Zhou. Continuous-time mean-variance efficiency: the 80% rule. *Annals of Applied Probability*, 16:1751–1763, 2006.

[47] X. Li, X. Y. Zhou, and A. E. B. Lim. Dynamic mean-variance portfolio selection with no-shorting constraints. *Annals of Applied Probability*, 40:1540–1555, 2001.

[48] J. F. Liang, S. Z. Zhang, and D. Li. Optioned portfolio selection: models and analysis. *To Appear in Mathematical Finance*, 18:569–593, 2008.

[49] A. E. B. Lim and X. Y. Zhou. Mean-variance portfolio selection with random parameters in a complete market. *Mathematics of Operations Research*, 27:101–120, 2002.

[50] B. Lincoln and B. Bernhardsson. Lqr optimization of linear system switching. *IEEE Transactions On Automatic Control*, 47(10):1701–1705, 2002.

[51] B. Lincoln and A. Rantzer. Relaxing dynamic programming. *IEEE Transactions on Automatica Control*, 51(8):1249–1260, 2006.

[52] H. Liu. Optimal consumption and investment with transaction costs and multiple risky assets. *Jounral of Finance*, 59:289–338, 2004.

[53] H. Liu and M. Loewenstein. Optimal portfolio selection with transaction costs and finite horizon. *The Review of Financial Studies*, 15:805–835, 2002.

[54] A. W. Lynch and S. Tan. Multiple risky assets, transaction costs and return predictablility: implications for portfolio chioce. *Wroking paper, new York Universiy*, 2002.

[55] H. M. Markowitz. *Portfolio Selection: Efficient Diversification of Investment*. John Wiley and Sons, New York, 1959.

[56] H. M. Markowitz. *Mean-variance analysis in portfolio choice and capital markets*. MA: Basil Blackwell, Cambridge, 1989.

[57] R. C. Merton. An analytical derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis*, pages 1851–1872, 1972.

[58] A. Morton and S. R. Pliska. Optimal portfolio management with fixed transaction costs. *Mathematical Finance*, 5:337–356, 1995.

[59] G. J. Fitzsimons N. Capon and R. A. Prince. An individual level analysis of the mutual fund investment decision. *Journal of Financial Services Research*, 10:59–82, 1996.

[60] M. L. Nagurka and V. Yen. Development of linear quadratic control laws via control parametriztion. *International Jouranl of Systems Scinece*, 23:2125–2139, 1992.

[61] V. Nanda, M. P. Narayanan, and V. A. Warther. Liquidity, investment ability, and mutual fund structure. *Journal of Financial Economics*, 57:417–443, 2000.

[62] B. Oksendal and A. Sulem. Optimal consumption and portfolio with both fixed and proportional transaction costs. *SIAM Journal on Control and Optimization*, 40:1765–1790, 2002.

[63] M. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in lest squares problems. *IMA Jouranl of Numerical Analysis*, 20(3):389–403, 2000.

[64] I. Polik and T. Terlaky. A survey of the s-lemma. *SIAM Review*, 49(3):371–418, 2007.

[65] A. Rantzer. On approxiamtimate dynamic programming in switching systems. In *Proceeding of 44th IEEE Conference on Decision and Control*, pages 1391–1396, 2005.

[66] M. Rinehart, M. Dahleh, and I. Kolmonovsky. Optimal control of swtched homogeneous systems. In *Proceeding of the American Control Conference*, pages 1377–1382, 2007.

[67] C. Seatzu, D. Corona, A. Giua, and A. Bemorad. Optimal control of continuous-time switched affine systems. *IEEE Transactions on Automatic Control*, 51(5):726–741, 2006.

[68] US Securities and Exchange Commission. *Management fee*. http://www.sec.gov/investor/pubs/inwsmf.htm#factors, Washington D.C., US, 2008.

[69] The American Investment Service. *Management fee.* http://www.americaninvestment.com/, US, 2008.

[70] S. E. Shreve and H. M. Soner. Optimal investment and consumption with transaction costs. *Annals of Applied Probability*, 4:609–692, 1994.

[71] L. Silverman. Discrete riccati equations: alternative algorithms, asymptotic properties, and system theory interpretations. *Control and Dynamic systems*, pages 313–386, 1976.

[72] N. Sleumer. Output-sensitive cell enumeration in hyperplane arrangements. *Nordic journal of computing*, 6:137–161, 1999.

[73] J. Sturm. Using sedumi 1.02 a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11:625–653, 1999.

[74] Z. Sun and S. Ge. *Switched linear systems: Control and design.* Springer, 2005.

[75] Z. Sun and S. S. Ge. Analysis and synthesis of switched linear control systems. *Automatica*, 41(2):181–195, 2005.

[76] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(267–288), 1996.

[77] L. Y. Wang, A. Beydoun, J. Sun, and I. Kolmansovsky. Optimal hybrid control with application to automotive powertrain systems. *Lecture Notes in Control and Information Sciences*, 222:190–200, 1997.

[78] H. K. Wimmer and M. Pavon. A comparsion theorem for matrix riccati difference equations. *System and Control Letters*, 19(3):233–239, 1992.

[79] W.Ogryczak and A. Tamir. Minimizing the sum of the k largest functions in linear time. *Information Processing Letter*, 85:117–122, 2003.

[80] J. Xie, S. He, and S. Zhang. Randomized portfolio selection with constraints. *Pacific Journal of Optimization*, 4:89–112, 2008.

[81] J. Xiong and X. Y. Zhou. Mean-variance portfolio selection under partial information. *SIAM Journal on Control and Optimization*, 46:156–175, 2007.

[82] X. Xu and P. J. Antsaklis. Optimal control of switched systems based on parameterization of the switching instans. *IEEE Transactions on Automatic Control*, 49(1):2–16, 2004.

[83] T. Zaslavsky. Facing up to arrangements: face-count formulas for partitions of space by hyperplanes. *American mathematical society*, 1(1–101), 1975.

[84] X. Y. Zhou and D. Li. Continuous time mean-variance portfolio selection: a stochastic lq framework. *Applied Mathematics and Optimization*, 42:19–33, 2000.

[85] X. Y. Zhou and G. Yin. Markowitz's mean-variance portfolio selection with regime switching: a continuous-time model. *SIAM Journal on Control and Optimization*, 42:1466–1482, 2003.

[86] S. S. Zhu, D. Li, and S. Y. Wang. Risk control over bankruptcy in dynamic portfolio selection: a generalized mean-variance formulation. *IEEE transactions on Automatic Control*, 49:447–457, 2004.