ANALYSIS OF CONVEX RELAXATIONS FOR NONCONVEX OPTIMIZATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yingjie Bi

May 2020

© 2020 Yingjie Bi
ALL RIGHTS RESERVED

ANALYSIS OF CONVEX RELAXATIONS FOR NONCONVEX OPTIMIZATION

Yingjie Bi, Ph.D.

Cornell University 2020

Nonconvex optimizations are ubiquitous in many application fields. One important aspect of dealing with a nonconvex optimization problem is to study its convex relaxation, which is often part of some approximation algorithms that can find quality solutions to the original nonconvex problem. In this dissertation, we analyze two prominent ways of obtaining convex relaxations: Lagrangian duality and semidefinite programming. The difference between the optimal value of the convex relaxation and that of the original problem is measured by the integrality gap or duality gap, which is of central importance in convex relaxation including providing a performance limitation that one can prove for the corresponding approximation algorithm. This dissertation focuses on how to estimate such integrality gap or duality gap more accurately.

For convex relaxation with Lagrangian duality, we propose a refinement of the Shapley-Folkman lemma and derive a new estimate for the duality gap of problems with separable objective and linear constraints. The improvement over the existing results is attributed to two sources. First, instead of using a single number measurement, a series of numbers are introduced to characterize the nonconvexity of a function in a potentially much finer manner. Second, we consider all subproblems jointly instead of approximating each subproblem individually as people had done before. We apply our result to the network utility maximization problem in networking and the dynamic spectrum management problem in communication as

examples to demonstrate that the new bound can be qualitatively tighter than the existing ones. The idea is also generalized to cases with separable nonlinear constraints, which is illustrated by an application to the network utility maximization problem with traffic split granularity constraints.

For convex relaxation with semidefinite programming, we examine two topics: the maximum cut problem and the Shannon capacity of graph. They are likely the two most well-known triumphs that semidefinite programming has in combinatorial optimization. Many semidefinite programming relaxations can be deduced from a systematic approach called the sum-of-squares hierarchy. Naturally, one wonders whether higher-degree sum-of-squares relaxations will provide relaxations with tighter integrality gap, and we study this question for the above two problems in this dissertation. For the maximum cut problem, an instance of integrality gap 0.96 is given first for the degree-4 sum-of-squares relaxation, and we further construct instances as candidates for even looser integrality gap. For the Shannon capacity of graph, we develop general conic programming upper bounds for the Shannon capacity of graph, which include the previous attempts based on sum-of-squares relaxations as special cases, and show that it is impossible to find better upper bounds for the Shannon capacity than the Lovász number along this way.

BIOGRAPHICAL SKETCH

Yingjie Bi received the B.S. degree in microelectronics from Peking University, Beijing, China in 2014. Since August 2014, Yingjie has been a Ph.D. student in the School of Electrical and Computer Engineering at Cornell University. His research interest includes optimization theory and control theory with applications to computer networks.

ACKNOWLEDGEMENTS

First, I would like to express my deep gratitude to my advisor, Prof. Kevin Tang. His friendly guidance and expert advice were invaluable to me all the time during my Ph.D. study.

Moreover, I would like to thank the rest of my committee: Prof. Aaron Wagner and Prof. Madeleine Udell, for their helpful suggestions and insightful comments.

My sincere thank also goes to my fellow group members in Cornell Networks Group: Andrey Gushchin, Shih-Hao Tseng, Ning Wu and Jiangnan Cheng. It was a great pleasure to work together with you.

Finally, I would like to thank my wife and my parents for their continuous support and encouragement in my life.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Optimization is a mathematical formulation aiming to model the decision making that identifies the best choice among a set of alternatives. The general form of an optimization problem is:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{1.1}$$

Here the vector $x \in \mathbb{R}^n$ is the *decision variable*, representing a choice. The function f is the *objective function*, where $f(x)$ is smaller means that x is a better choice. The *constraints* $g_i(x) \leq 0$ describe the requirements or limitations on the possible choices we are allowed to select. A vector x is called a *feasible solution* if it satisfies all the constraints. Furthermore, a vector $\hat{x}$ is called an *optimal solution* if $f(\hat{x}) \leq f(x)$ for any feasible solution x. Variants of optimization problems, such as problems of maximization or with equality constraints, can be easily converted into an equivalent problem of the form (1.1).

Due to its generality, optimization has found wide applications in engineering, business and many other fields. Most of the applications involve solving the corresponding optimization problems, i.e., finding an optimal solution. The available techniques of solving an optimization problem depend on the shapes of the objective function and constraints. Accordingly, optimization problems can be categorized as following:

- The problem (1.1) is called a *linear program* if the functions f and g_i are all linear.
- The problem (1.1) is called a *convex optimization problem* if the functions f and g_i are all convex, which includes linear programs as special cases.

1.1 Nonconvex Optimization and Its Convex Relaxation

In the history of optimization theory, there are two major transitions with great impact. Around 1950s, the theory of linear programming was established, and at that time the linearity was regarded as the distinction between easy and hard problems. During 1980s, the development of interior point methods enabled people to tackle convex optimization problems as efficiently as linear programs, and the boundary of hardness had moved to convexity versus nonconvexity. Thirty years later, people now once again revisit this boundary between easy and difficulty optimization problems. This is primarily because nonconvexity is ubiquitous in practice. Here are a few examples illustrating how nonconvexity is raised in application problems:

Machine learning [1, Chapter 4]. The training of a neural network is to determine the weights in the network such that the network can accurately map the inputs in the training set to their designated outputs. The input-output relationship of a neural network is nonconvex even if the input-output relationship of single neuron can be convex, because the composition of convex functions is generally no longer convex. As the result, the nonconvex input-output relationship of the network induces the nonconvexity in the objective function when we write down the optimization problem minimizing the training error.

Power systems [2]. The optimal power flow is a fundamental problem in the power system operation. The goal of the optimal power flow problem is to optimize the distribution of power in the power network subject to Kirchhoff's law and other physical limitations. The decision variables here include the power injection and bus voltage at each bus, and by Kirchhoff's law they are related by a quadratic equality, which introduces nonconvexity into the constraints of the problem.

Moreover, many decision making problems involve binary or discrete decisions. In general, there are two different ways to integrate the integer constraints into the general problem (1.1). As an example, if we assume x_i can only take 0 or 1, either we can let the corresponding function g_i in (1.1) be

$$g_i(x_i) = \begin{cases} 0, & \text{if } x_i = 0 \text{ or } x_i = 1, \\ +\infty, & \text{otherwise,} \end{cases}$$

or the rewrite the binary constraint as a polynomial constraint such as

$$x_i(1 - x_i) = 0.$$

In Chapter 5 we will adopt the first approach and in Chapter 7 we will adopt the second. In both approaches, the corresponding constraint functions g_i in (1.1) will be nonconvex.

As shown above, nonconvex optimizations are general enough that they can conveniently model varieties of decision making problems. Unfortunately, their generality also allows reductions from NP-hard problems. Therefore, it is unlikely to find an efficient algorithm that can accurately solve a general nonconvex optimization problem. Like all NP-hard problems, there are three ways to tackle with nonconvex optimization problems: First, we can give up finding an algorithm that solves all instances of the problem and only focus on certain cases. Second, if the problem size is small, then exponential searching algorithms with good optimization may well suffice. Finally, one can design approximation algorithms that find near-optimal solutions in polynomial time. The work in this dissertation belongs to this category.

Various techniques have been used in the design of approximation algorithms including linear programming, semidefinite programming, randomized algorithms

and clever uses of many standard algorithm design techniques like greedy and dynamic programming, which are extensively studied in books such as [3, 4]. In general, the field of approximation algorithms is highly problem-specific. Techniques tailored for one particular problem cannot usually be generalized to other situations. However, some guiding frameworks exist, one of which is based on convex relaxations and usually contains three steps:

The first step is to find some convex optimization problem whose optimal value is less than or equal to the optimal value of the original problem, assuming the original is a minimization problem. The new problem is called a *relaxation* of the original problem, and it can be obtained by replacing the original objective function f with another convex function f' satisfying $f' \leq f$, or modifying the constraints such that any feasible solution to the original problem is still feasible to the relaxed one. The second step is to solve the relaxed problem by efficient convex optimization algorithms. The last step is to modify the obtained optimal solution to the relaxation into an approximate solution to the original problem. In the simplest case when the objective function is the only changed part in the relaxation, this step can be skipped. As a little more complex example, consider a linear integer-programming problem relaxed by a linear program with all integer constraints ignored. The optimal solution to the relaxed linear program is usually fractional, and we have to round all the fractional values into integers in order to obtain an approximate solution to the original integer program. Generally, the last step that makes the relaxed solution feasible to the original problem is much more complicated than simply rounding a fractional number, but this step is often collectively called the rounding step.

Among the three steps in this framework, the second step of solving the relax-

ation is usually rather straightforward. The last rounding step tends to be the most difficult and needs deep insight on individual problems. In contrast, the first step of relaxing the original problem is more routine, with some standard approaches to follow. This dissertation is concerned with two widely used methods of relaxations: convex relaxation with Lagrangian duality and with semidefinite programming.

As we discussed above, finding the relaxation is only one step in the design of approximation algorithms, and in general studying the relaxation will not necessarily lead to a good approximation algorithm for the original problem. However, in the following we will see that even having the relaxation alone can lead to useful results for both engineering problems and theoretical computer science.

1.2 Convex Relaxation and Engineering Problems

The optimal value of the convex relaxation is a lower bound for the optimal value of the original problem. This lower bound can play important roles in solving and analyzing nonconvex optimization problems, in cases we need an actual solution to the original problem. For example, the optimal solution to the convex relaxation can be used as an initial guess in the local searching for the solution to the original nonconvex problem, and many global optimization algorithms for nonconvex optimizations can also benefit from having an lower bound which can be efficiently computed (see [5] for examples).

To obtain a tight lower bound from the relaxation, the relaxation itself has to be as close to the original problem as possible. As an illustration, let us consider the nonconvex problem (1.1) in which the objective function f is nonconvex and the constrains g_i are all convex. In the relaxation, we can use a convex objective

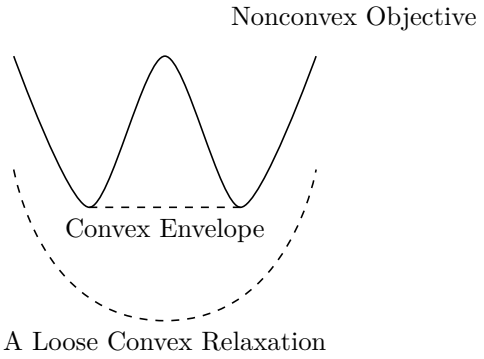


Figure 1.1: The convex relaxation of a nonconvex objective.

f' with $f' \leq f$. To make the convex relaxation tight, the new convex function f' should be the one that is closest to the original objective function f. With some technical details ignored, this function f' is called the *convex envelope* of the function f (see Figure 1.1). In Chapter 2, we will see that such convex envelope can be constructed by convex conjugates.

Furthermore, the convex relaxation can possibly suggest architecture decisions for the engineering problems. Many nonconvex optimization problems raised in fields such as communication [6] and machine learning [7, 8] have a separable structure. For these problems, their Lagrange dual problems do not have any coupling constraints and thus they can be solved by a distributed architecture. In Chapter 3 and Chapter 5, we will study the convex relaxation based on the Lagrangian duality for separable problems. Naturally, a rounding algorithm is still necessary to recover a feasible solution to the original problem from the dual solution. Some techniques for recovering feasible primal solutions are given in [9].

In some situations, we are given an iteration process that is already used in practice. The question, sometimes called the reverse engineering problem, is to find what optimization problem the process is trying to solve. If the obtained optimization problem is convex, then we can immediately show that this process

has desired properties such as global convergence and the stability of the equilibrium. Unfortunately, without the convexity the situation is much more complex to deal with. For example, for many years TCP congestion control has been studied from the perspective of reverse engineering [10]. In this model, the control rules of window size in the congestion control protocols are interpreted as a distributed algorithm to maximize the network utility. While being an elegant theory, it does not capture many important factors. For example, the network utility maximization model assumes that every user reacts to the same congestion signal on the link, which is not accurate in practice due to the propagation delay. However, if we generalize the model to include multiple congestion signals, it becomes very hard to analyze because of lacking of the joint convexity. Similar issues also occur in economics such as general equilibrium theory [11]. The tatonnement process is a system of differential equations relating the change of prices with the excess demand of commodities, which is commonly used in the study of the stability of general equilibrium. With some global convex assumption, it is not hard to establish the uniqueness, the global convergence and the stability of the equilibrium. However, people have not understood much about the behavior of the dynamical system without the convexity assumption [12]. By studying the properties of the convex relaxation, we can hope to answer these convergence and stability questions for the process.

1.3 Convex Relaxation and Theoretical Computer Science

Approximation algorithms based on convex relaxations can yield heuristic results for nonconvex optimizations. The further step is to seek provable performance guarantees for theses approximation algorithms. In this section, we assume the

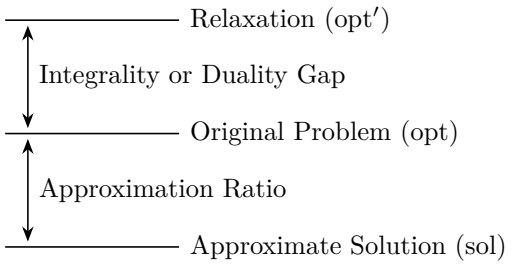


Figure 1.2: The relationship among the values of the original problem, the relaxed problem and the found approximate solution for a maximization problem.

original problem is a maximization problem for the convenience of later chapters. For a fixed instance of the problem, denote the optimal value of the original problem as opt, the optimal value of the relaxation as opt′, and the value of the approximate solution as sol. Then we have the inequality

$$\text{sol} \leq \text{opt} \leq \text{opt}',$$

and there are two ratios involved in the above inequality. The first ratio, sol / opt, is the *approximation ratio*. The second ratio, opt / opt′, is usually called the *integrality gap*. In different scenarios, this gap is also characterized in other forms. For example, when the convex relaxation is obtained through Lagrangian duality, the difference $\text{opt}' - \text{opt}$ is exactly the *duality gap*. The relationship among the values of the original problem, the relaxed problem and the found approximate solution is summarized in Figure 1.2.

The reader should be careful about the notations and names here. The value of the integrality gap is larger actually means that we have a tighter relaxation. Similarly, the value of the approximation ratio is larger means that we have a better approximation algorithm.

It is also worth pointing out that the approximation ratio and the integrality gap defined above rely on the given instance of the problem. To identify the worst-

case performance of the approximation algorithm, we say that the approximation ratio of the algorithm is at least α if for all instances of the problem

$$\text{sol} \geq \alpha \, \text{opt} \, .$$

In simple words, such algorithm is called an α-approximation algorithm. On the other hand, we say that the approximation ratio of the algorithm is at most α if there exists an instance with

$$\text{sol} \leq \alpha \, \text{opt} \, .$$

Similar terminologies can also be applied to the integrality gap.

For example, problems like the knapsack problem [13] and the Euclidean traveling salesman problem [14] have *polynomial-time approximation scheme* (PTAS), which means the existence of a $(1-\epsilon)$-approximation algorithm for any $\epsilon > 0$. On the other hand, some problems are extremely hard to approximate. One example is the maximum clique problem finding the largest subset of vertices in a graph with n vertices such that all of them are adjacent with each other. Then a trivial $1/n$-approximation algorithm can simply return any single vertex as the approximate maximum clique. However, there does not exist an $O(n^{\epsilon-1})$-approximation algorithm for any $\epsilon > 0$ unless $P = NP$ [15], which essentially says that there does not exist any approximation algorithm that is qualitatively better than the trivial strategy.

The value opt of the original problem is generally unknown and hard to find, while opt$'$, being the value of a convex optimization problem, can often be analyzed. As a result, the standard approach to prove the performance guarantee for the approximation algorithm is to bound the maximum performance loss during the rounding step by showing that for all instances of the problem we have

$$\text{sol} \geq \alpha \, \text{opt}' \, . \tag{1.2}$$

The above inequality (1.2) implies

$$\mathrm{sol} \geq \alpha \, \mathrm{opt}' \geq \alpha \, \mathrm{opt}$$

and thus the algorithm is an α-approximation algorithm. At the same time, (1.2) also implies

$$\mathrm{opt} \geq \mathrm{sol} \geq \alpha \, \mathrm{opt}',$$

so the integrality gap of the relaxation is also at least α.

To demonstrate the above approach, in the following we will present a 1/2-approximation algorithm for the knapsack problem based on the linear programming relaxation. Although it is not the best known approximation algorithm for the knapsack problem, it is a good illustration of the essence of how to design an approximation algorithm using convex relaxations and how to prove its performance guarantee.

Example 1.1 (1/2-approximation algorithm for the knapsack problem)**.** In the knapsack problem, we are given a set of n items, each with weight w_i and value a_i, and our goal is to select a subset of them such that their total weight is not exceeding W and their total value is maximized. In this problem, we can introduce binary variables x_i, where $x_i = 1$ means selecting the item i and $x_i = 0$ means not selecting. Then the knapsack problem can be formulated as the following integer program:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} a_i x_i \\ \text{s. t.} \quad & \sum_{i=1}^{n} w_i x_i \leq W, \\ & x_i \in \{0, 1\}, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{1.3}$$

Excluding the trivial cases, we assume that the weight of each item $w_i \leq W$ and

the total weight of all items

$$\sum_{i=1}^{n} w_i > W.$$

First, we write down the relaxation

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} a_i x_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i x_i \le W, \\ & 0 \le x_i \le 1, \quad \forall i = 1, \ldots, n \end{aligned} \tag{1.4}$$

of the original integer program (1.3). The next step is to solve the linear program (1.4). If the items are ordered by their value per unit weight, i.e.,

$$\frac{a_1}{w_1} \ge \frac{a_2}{w_2} \ge \cdots \ge \frac{a_n}{w_n},$$

then

$$x_1 = 1, \ldots, x_k = 1, x_{k+1} = \frac{W - x_1 - \cdots - x_k}{w_{k+1}}, x_{k+2} = 0, \ldots, x_n = 0$$

is the optimal solution to (1.4), where k is the largest integer such that

$$w_1 + \cdots + w_k \le W.$$

The corresponding optimal value of the relaxation is

$$\text{opt}' = a_1 + \cdots + a_k + \frac{W - x_1 - \cdots - x_k}{w_{k+1}} a_{k+1}.$$

Now we are going to round the above fractional solution into an integer solution. Consider the following two possible rounded solutions

$$y_1 = 1, \ldots, y_k = 1, y_{k+1} = 0, y_{k+2} = 0, \ldots, y_n = 0$$

with value $a_1 + \cdots + a_k$ and

$$y'_1 = 0, \ldots, y'_k = 0, y'_{k+1} = 1, y'_{k+2} = 0, \ldots, y'_n = 0$$

with value a_{k+1}. Since

$$a_1 + \cdots + a_k + a_{k+1} \geq \mathrm{opt}',$$

at least the value of one of the solution is at least $\mathrm{opt}'/2$. Returning that solution gives us a 1/2-approximation algorithm.

In this dissertation, we will study how to estimate the integrality gap for the relaxation, either measured in ratio or in difference. From the perspective of proving performance guarantees for approximation algorithms, there are two main reasons to study the convex relaxation even without a corresponding rounding step.

First, the integrality gap provides a limitation of what performance guarantee we can prove for the approximation algorithm based on this convex relaxation. As we demonstrated above, the common way to establish that the approximation ratio of the algorithm is at least α is through proving the inequality (1.2), which simultaneously implies that the integrality gap of the relaxation is also at least α. Therefore, it is impossible to prove that the resulted approximation algorithm is an α-approximation algorithm if we have already found an instance of integrality gap

$$\mathrm{opt} \,/\, \mathrm{opt}' < \alpha.$$

This is the reason why we are interested in finding an instance of loose integrality gap for the maximum cut problem in Chapter 7.

Second, many unsolved questions on the inapproximability of hard problems, including the well-known unique games conjecture [16], can be formulated in the form of the NP-hardness of distinguishing two types of instances: Given an instance whose value is either greater than a or less than b, we are asked to determine which case actually happens. Such decision problems can be directly solved by a

relaxation without the rounding algorithm. If the integrality gap of the convex relaxation satisfies

$$\mathrm{opt} \,/\, \mathrm{opt}' \geq b/a,$$

then the value opt of the original problem is greater than a if and only if the value opt$'$ of the convex relaxation is greater than a. Hence the inapproximability result can be refuted by showing the integrality gap of the corresponding convex relaxation is at least b/a. For a good overview of this area, see [17].

1.4 Main Contributions

In Part I of the dissertation, we will look at convex relaxations through Lagrangian duality. After giving a background introduction to the Lagrangian duality in Chapter 2, in Chapter 3 we propose a refinement of the Shapley-Folkman lemma and derive a new estimate for the duality gap of nonconvex optimization problems with separable objective functions and linear constraints based on concepts like kth convex hull and finer characterization of nonconvexity of a function. We apply our result to the network utility maximization problem in networking and the dynamic spectrum management problem in communication as examples in Chapter 4 to demonstrate that the new bound can be qualitatively tighter than the existing ones. The idea is also generalized to cases with separable nonlinear constraints in Chapter 5, which will be illustrated by an application to the network utility maximization problem with traffic split granularity constraints.

In Part II we turn to convex relaxations through semidefinite programming. Chapter 6 gives a brief introduction to the semidefinite relaxations and the sum-of-squares hierarchy. Then we study the integrality gap for the sum-of-squares

relaxations of the maximum cut problem in Chapter 7. An instance of integrality gap 0.96 will be given first for the degree-4 sum-of-squares relaxation, and we will further construct instances as candidates for even looser integrality gap. In Chapter 8, we consider the possibility of developing general conic programming upper bounds for the Shannon capacity of graph, which include the previous attempts based on sum-of-squares relaxations as special cases, and show that it is impossible to find better upper bounds for the Shannon capacity than well-known Lovász number along this way.

1.5 Notations and Conventions

In this dissertation, we use superscript to index vectors and use subscript to refer to a particular component of a vector. For instance, both x^i and x^{ij} are vectors, but x^i_s is the sth component of the vector x^i. For two vectors x and y, $x \leq y$ means $x_s \leq y_s$ holds for all components. $\mathbb{R}^n_+$ is the set of $n \times 1$ nonnegative column vectors.

For a function $f : \mathbb{R}^n \to [-\infty, +\infty]$:

- The function f is *proper* if it does not always equal to $+\infty$ and never equals to $-\infty$.
- The *domain* of f is the subset

$$\operatorname{dom} f = \{x \in \mathbb{R}^n | f(x) \neq +\infty\}.$$

- The *epigraph* of the function f is the set

$$\operatorname{epi} f = \{(x, y) \in \mathbb{R}^n \times \mathbb{R} | y \geq f(x)\}.$$

- The function f is *lower-semicontinuous* at point x_0 if

$$f(x_0) \leq \liminf_{x \to x_0} f(x).$$

  The function f is lower-semicontinuous if it is lower-semicontinuous at every point, which is equivalent to the closeness of its epigraph $\operatorname{epi} f$.

For a subset $S \in \mathbb{R}^n$:

- $\operatorname{conv} S$ is the convex hull of S.
- $\operatorname{cl} S$ is the closure of S.

The following are several matrix cones that will be frequently used in this dissertation:

- $\mathcal{S}_n$ is the cone of $n \times n$ symmetric matrices.
- $\mathcal{P}_n$ is the cone of $n \times n$ positive semidefinite matrices.
- $\mathcal{N}_n$ is the cone of $n \times n$ nonnegative symmetric matrices.
- $\mathcal{C}_n$ is the cone of $n \times n$ copositive matrices, i.e., all symmetric matrices $Q \in \mathcal{S}_n$ such that $x^T Q x \geq 0$ for any $x \in \mathbb{R}^n_+$.

Part I

Convex Relaxation with Lagrangian Duality

CHAPTER 2

PRELIMINARIES

In this chapter, we briefly review the theory of Lagrangian duality, which will form the foundation for the convex relaxation studied in this part. For a more comprehensive treatment of this material, see [18].

2.1 Lagrange Dual Problem and Convex Conjugate

Consider a general optimization problem:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0, \end{aligned} \tag{2.1}$$

where the decision variable $x \in \mathbb{R}^n$ and functions $f : \mathbb{R}^n \to (-\infty, +\infty]$, $g : \mathbb{R}^n \to (-\infty, +\infty]^m$. The basic idea of the Lagrangian duality is to move the constraints into the objective by adding Lagrangian multipliers. Define the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m_+ \to (-\infty, +\infty]$ as

$$L(x, y) = f(x) + y^T g(x).$$

Let p be the optimal value of the original problem (2.1), then for any given x

$$\sup_{y \geq 0} L(x, y) = \begin{cases} f(x), & \text{if } g(x) \leq 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

and thus the original problem (2.1) can be rewritten as

$$p = \inf_x \sup_{y \geq 0} L(x, y).$$

If we exchange the order of the infimum and the supremum in the above equality and define

$$d = \sup_{y \geq 0} \inf_x L(x, y),$$

then we have $d \leq p$, which can be seen by the following simple argument: Since

$$\inf_x L(x, y) \leq L(z, y), \quad \forall z \in \mathbb{R}^n, \forall y \in \mathbb{R}^m_+,$$

then

$$\sup_{y \geq 0} \inf_x L(x, y) \leq \sup_{y \geq 0} L(z, y), \quad \forall z \in \mathbb{R}^n,$$

which further implies

$$d = \sup_{y \geq 0} \inf_x L(x, y) \leq \inf_z \sup_{y \geq 0} L(z, y) = p.$$

For convenience, we introduce the Lagrange dual function as

$$h(y) = \inf_x L(x, y) \tag{2.2}$$

and rewrite the definition of d as the following optimization problem:

$$\begin{aligned} \max \quad & h(y) \\ \text{s.t.} \quad & y \geq 0. \end{aligned} \tag{2.3}$$

The original problem (2.1) will be called the primal problem and the problem (2.3) will be called the dual problem. The above argument shows that the dual problem can be regarded as a convex relaxation of the primal problem. This fact is usually referred to as the *weak duality*.

Theorem 2.1 (weak duality)**.** *If p is the optimal value of the primal problem* (2.1) *and d is the optimal value of the dual problem* (2.3)*, then $d \leq p$.*

As an example, we consider the special case in which $g(x)$ in the primal problem (2.1) is linear:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & Ax \leq b, \end{aligned} \tag{2.4}$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. In this case, the Lagrange dual function $h(y)$ in (2.2) becomes

$$h(y) = \inf_x (f(x) + y^T(Ax - b)) = \inf_x (f(x) + (A^T y)^T x) - b^T y.$$

If we introduce the conjugate function $f^* : \mathbb{R}^n \to [-\infty, +\infty)$ for the original objective function f defined as

$$f^*(x^*) = \sup_x (x^{*T} x - f(x)),$$

then $h(y)$ can be rewritten as

$$h(y) = -f^*(-A^T y) - b^T y$$

and the Lagrange dual problem (2.3) becomes

$$\begin{aligned} \max \quad & -f^*(-A^T y) - b^T y \\ \text{s. t.} \quad & y \geq 0. \end{aligned} \tag{2.5}$$

The concept of the conjugate function is widely used in the field of convex analysis. See [19] for an extensive introduction to the theory of conjugate functions. In this dissertation, the following property of the conjugate function will be used:

Theorem 2.2. *If a function f is bounded below by some affine function, then*

$$\operatorname{epi} f^{**} = \operatorname{cl} \operatorname{conv} \operatorname{epi} f.$$

Proof. See [20, Theorem X.1.3.5]. □

The double conjugate f^{**} is exactly the convex envelope introduced in Section 1.2. By the above theorem, f^{**} satisfies the following two properties:

- f^{**} is a convex and lower-semicontinuous function satisfying $f^{**} \leq f$.

- For any convex and lower-semicontinuous function h satisfying $h \leq f$, we have $h \leq f^{**}$.

Obviously, if f itself is convex and lower-semicontinuous, then $f^{**} = f$. More precisely, we have the following result:

Corollary 2.1. *For a convex function f, if f is finite and lower-semicontinuous at a given point x, then $f^{**}(x) = f(x)$.*

Proof. It is easy to see that f is proper, because if $f(y) = -\infty$ at some point y then f equals to $-\infty$ on the open segment between y and x, which is contradicted with the lower-semicontinuity of f at x. Since f is proper and convex, it is bounded below by some affine function and $\operatorname{epi} f$ is convex. By Theorem 2.2,

$$\operatorname{epi} f^{**} = \operatorname{cl} \operatorname{epi} f,$$

which implies $f^{**}(x)$ is also finite and $(x, f^{**}(x)) \in \operatorname{cl} \operatorname{epi} f$. Choose a sequence $(x^k, y_k) \in \operatorname{epi} f$ such that

$$\lim_{k\to\infty} x^k = x, \quad \lim_{k\to\infty} y_k = f^{**}(x).$$

Then by the lower-semicontinuity of f at x,

$$f(x) \leq \liminf_{k\to\infty} f(x^k) \leq \lim_{k\to\infty} y_k = f^{**}(x),$$

while the reverse direction of the above inequality is evident. □

2.2 The Duality Gap

The weak duality establishes the inequality $p \geq d$ for the optimal primal value p of (2.1) and optimal dual value d of (2.3). It is natural to ask in which cases these

two values are equal. This property is called the strong duality, and it holds if the functions f and g in (2.1) are convex in addition with some other technical conditions. If the strong duality does not hold, then the difference $p - d$ is called the duality gap for the problem (2.1).

To study the duality gap, we introduce the perturbation function $v : \mathbb{R}^m \to [-\infty, +\infty]$ by letting $v(z)$ be the optimal value of the perturbed problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s. t.} \quad & g(x) \leq z. \end{aligned}$$

The reason why we are interested in the perturbation function v is its relationship with both the primal value p and the dual value d:

Lemma 2.1. *For the perturbation function v defined above, $p = v(0)$ and $d = v^{**}(0)$.*

Proof. The first equality is obvious. To prove the second one, observe that

$$\begin{aligned} -v^*(-z^*) &= \inf_z (z^{*T} z + v(z)) \\ &= \inf_z \inf_{x: g(x) \leq z} (z^{*T} z + f(x)) \\ &= \inf_x \inf_{z: z \geq g(x)} (z^{*T} z + f(x)) \\ &= \inf_x \inf_{z: z \geq 0} (z^{*T}(z + g(x)) + f(x)) \\ &= \inf_{z: z \geq 0} \inf_x (z^{*T} z + L(x, z^*)) \\ &= \begin{cases} \inf_x L(x, z^*) = h(z^*), & \text{if } z^{*T} \geq 0, \\ -\infty, & \text{otherwise,} \end{cases} \end{aligned}$$

where h is the Lagrange dual function given in (2.2). Then the dual problem

$$d = \sup_{y \geq 0} h(y) = \sup_y (-v^*(-y)) = v^{**}(0).$$

□

The above observation suggests that the strong duality can be proven by showing $v(0) = v^{**}(0)$. Furthermore, the perturbation function v inherits nice properties from the functions f and g in the optimization problem (2.1).

Lemma 2.2. *If both the functions f and g in (2.1) are convex, then the corresponding perturbation function v is also convex.*

Proof. We need to prove that the set

$$\operatorname{epi} v = \{(z, w) \in \mathbb{R}^m \times \mathbb{R} | w \geq v(z)\}$$

is convex. Pick up two points $(z^i, w_i) \in \operatorname{epi} v$, for $i = 1, 2$, and $0 \leq \lambda \leq 1$. Then for any $\epsilon > 0$, there exists $x^i \in \mathbb{R}^n$ with

$$f(x^i) < w_i + \epsilon, \quad g(x^i) \leq z^i.$$

By the convexity of f and g,

$$(\lambda x^1 + (1-\lambda)x^2, \lambda w_1 + (1-\lambda)w_2 + \epsilon) \in \operatorname{epi} f,$$
$$g(\lambda x^1 + (1-\lambda)x^2) \leq \lambda z^1 + (1-\lambda)z^2,$$

which implies

$$v(\lambda z^1 + (1-\lambda)z^2) \leq f(\lambda x^1 + (1-\lambda)x^2) \leq \lambda w_1 + (1-\lambda)w_2 + \epsilon.$$

Let $\epsilon \to 0$ and we get

$$(\lambda z^1 + (1-\lambda)z^2, \lambda w_1 + (1-\lambda)w_2) \in \operatorname{epi} f,$$

which shows that the function v is convex. □

Lemma 2.3. *If both the functions f and g in (2.1) are lower semi-continuous and the domain of f is bounded, then the corresponding perturbation function v is also lower semi-continuous.*

Proof. For any $z \in \mathbb{R}^m$, we want to show that if $z^k \to z$ as $k \to \infty$

$$l = \liminf_{k\to\infty} v(z^k) \geq v(z).$$

The above inequality clearly holds when $l = +\infty$. If $l < +\infty$, by considering a subsequence of $\{v(z^k)\}_{k=1}^{\infty}$, without loss of generality we can assume $v(z^k) < +\infty$ for each k. Since the domain of f is bounded, we can find $\hat{x}^k \in \operatorname{dom} f$ attaining the optimal value of the perturbed problem related to $v(z^k)$, i.e.,

$$v(z^k) = f(\hat{x}^k), \quad g(\hat{x}^k) \leq z^k.$$

By extracting a convergent subsequence for $\{\hat{x}^k\}_{k=1}^{\infty}$, we can assume $\{\hat{x}^k\}_{k=1}^{\infty}$ has a limit $\hat{x}$. Then by the lower-semicontinuity of g,

$$g(\hat{x}) \leq \liminf_{k\to\infty} g(\hat{x}^k) \leq z,$$

which implies that $f(\hat{x}) \geq v(z)$. Now

$$l = \liminf_{k\to\infty} v(z^k) = \liminf_{k\to\infty} f(\hat{x}^k) \geq f(\hat{x}) \geq v(z),$$

because f is also lower semi-continuous. □

Based the above two properties of the perturbation function v, we have the two following conditions guaranteeing the strong duality:

Theorem 2.3 (Slater's condition)**.** *Assume the functions f and g in* (2.1) *are convex. If there exists a feasible solution x_0 satisfying*

$$f(x_0) < +\infty, \quad g(x_0) < 0,$$

then $p = d$.

Proof. By Lemma 2.2, the perturbation function v is also convex. If $p = v(0) = -\infty$, then $p = d$ by the weak duality. Otherwise, v is both finite and continuous at the origin since

$$0 \in \operatorname{int} \operatorname{dom} v.$$

By Theorem 2.2, $v(0) = v^{**}(0)$. □

Theorem 2.4. *Assume the functions f and g in (2.1) are convex and lower semi-continuous. If (2.1) is feasible and the domain of f is bounded, then $p = d$.*

Proof. By Lemma 2.2 and Lemma 2.3, the perturbation function v is both convex and lower semi-continuous. Since $v(0)$ is further finite, $v(0) = v^{**}(0)$ by Theorem 2.2. □

In cases when either f or g is not convex, generally there will be a positive duality gap $p - d > 0$. For problems with linear constraints, the duality gap can be directly interpreted by considering the optimization problem in which the original objective function f is replaced by its convex envelope f^{**}:

$$\begin{aligned} \min \quad & f^{**}(x) \\ \text{s. t.} \quad & g(x) \leq 0. \end{aligned} \tag{2.6}$$

The above problem (2.6) is a convex relaxation of the original problem. Furthermore, its dual problem is also the same as (2.5) except that f^* should be replaced by f^{***}. Let p' and d' be the primal and dual optimal value of (2.6), respectively. Since f^* is convex and lower-semicontinuous for any function f, $f^{***} = f^*$ and thus the two dual values $d' = d$. If the strong duality $p' = d'$ holds for the convex relaxation (2.6), then the duality gap $p - d = p - p'$ can be regarded as the difference of optimal values between the original problem and the relaxed problem (2.6).

CHAPTER 3

SEPARABLE NONCONVEX PROBLEMS WITH LINEAR CONSTRAINTS

In this chapter, we consider nonconvex optimization problems with separable objective and linear constraints. By proposing a refinement of the Shapley-Folkman lemma, we derive a new estimate for the duality gap of these problems. The problem formulation and pervious results on its duality gap are given in Section 3.1. The Shapley-Folkman lemma and its refined version are stated and proved in Section 3.2. Our new bound for the duality gap depends on some finer characterization of the nonconvexity of a function, which is introduced in Section 3.3, and some examples of computing the nonconvexity are provided in Section 3.4. Section 3.5 establishes our new bound for the duality gap.

In Chapter 4, we will apply our result to two examples, a network utility maximization problem in networking and the dynamic spectrum management problem in communication, to demonstrate that the new bound can be qualitatively tighter than the previous ones. Chapter 5 extends the major idea in this chapter to the cases with general convex or even nonconvex constraints.

3.1 Problem Formulation

The general formulation for a nonconvex problem with separable objective and linear constraints is as following:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} f_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{n} A_i x^i \leq b. \end{aligned} \tag{3.1}$$

Here $x^i \in \mathbb{R}^{n_i}$ are the decision variables. The function $f_i : \mathbb{R}^{n_i} \to (-\infty, +\infty]$ is lower semi-continuous, and its domain is bounded. A_i is a matrix of size $m \times n_i$, so there are m linear constraints in total. This problem is the special case of the problem (2.4) discussed in Section 2.1. Since the convex conjugate is additive, i.e., if

$$f(x^1, \ldots, x^n) = \sum_{i=1}^{n} f_i(x^i),$$

then

$$\begin{aligned} f^*(x^{1*}, \ldots, x^{n*}) &= \sup_{x^1,\ldots,x^n} \left(\sum_{i=1}^{n} x^{i*T} x^i - \sum_{i=1}^{n} f_i(x^i) \right) \\ &= \sum_{i=1}^{n} \sup_{x^i} (x^{i*T} x^i - f_i(x^i)) \\ &= \sum_{i=1}^{n} f_i^*(x^{i*}). \end{aligned}$$

Specializing the dual problem given by (2.5), we can write down the dual problem of (3.1) as following:

$$\begin{aligned} \max \quad & -\sum_{i=1}^{n} f_i^*(-A_i^T y) - b^T y \\ \text{s.t.} \quad & y \geq 0, \end{aligned} \tag{3.2}$$

In this chapter, we always assume the feasibility on the primal problem (3.1). Furthermore, based on our assumptions of f_i, the conjugate function f_i^* will always take finite value. As a result, the optimal value of the dual problem (3.2) is guaranteed to be finite. Following the notations in Chapter 2, we denote the optimal value of the primal problem (3.1) and dual problem (3.2) as p and d, respectively.

The authors of [21] presented the following upper bound for the duality gap of (3.1):

$$p - d \leq \min\{m + 1, n\} \max_{i=1,\ldots,n} \rho(f_i). \tag{3.3}$$

Here $\rho(f)$ is the nonconvexity of a proper function f defined by

$$\rho(f) = \sup \left\{ f\left(\sum_j \alpha_j x^j\right) - \sum_j \alpha_j f(x^j) \right\} \tag{3.4}$$

over all finite convex combinations of points $x^j \in \operatorname{dom} f$, i.e., $f(x^j) < +\infty$, $\alpha_j \geq 0$ with $\sum_j \alpha_j = 1$.

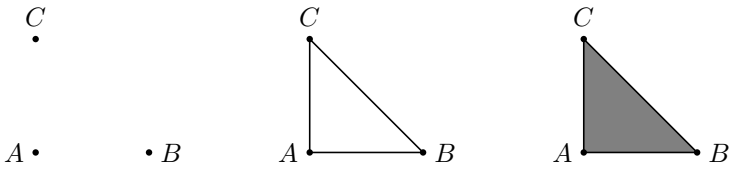
In [9], an improved bound for the duality gap was given by

$$p - d \leq \sum_{i=1}^{\min\{m,n\}} \rho(f_i), \tag{3.5}$$

where we assume that $\rho(f_1) \geq \cdots \geq \rho(f_n)$. Although the bound (3.5) is only a slight improvement over the original bound (3.3) by a factor of $m/(m+1)$, it nevertheless shows that (3.3) can never be tight except for some trivial situations. But as will be demonstrated by the examples in Chapter 4, the bound (3.5) can still be very conservative.

The foundation of the previous bounds is the Shapley-Folkman lemma, which was originally stated and used to establish the existence of approximate equilibria in economy with nonconvex preferences [22]. It roughly says that the sum of a large number of sets is close to be convex and thus can be used to generalize results on convex objects to nonconvex ones. The Shapley-Folkman lemma has found applications in many fields including economics and optimization theory. It is of particular use for estimating the duality gap of a general nonconvex optimization problem, which provides an indication of the nonconvexity of such a problem [18, 23, 24]. In this chapter, we aim at providing a tighter duality gap estimation via refining the original Shapley-Folkman lemma.

If the domain of some function f_i in (3.1) is not convex, by definition $\rho(f_i) = +\infty$. In this case, all the above bounds and our new bound in Section 3.5 will be



(a) $\mathrm{conv}_1 S$ (b) $\mathrm{conv}_2 S$ (c) $\mathrm{conv}_3 S$

Figure 3.1: The kth convex hull of a three-point set $S = \{A, B, C\}$.

useless. To handle this issue, we can replace the nonconvexity of the domain by appropriate nonconvex constraints that will be covered by the techniques introduced in Chapter 5 later.

3.2 Shapley-Folkman Lemma and Its Refinement

The original Shapley-Folkman lemma (Theorem 3.1) is given below. We omit its proof here since it will be a simple corollary of our refined Shapley-Folkman lemma.

Theorem 3.1. *Let $S_1, S_2, \ldots, S_n$ be subsets of $\mathbb{R}^m$. For each $z \in \mathrm{conv} \sum_{i=1}^n S_i = \sum_{i=1}^n \mathrm{conv}\, S_i$, there exist points $z^i \in \mathrm{conv}\, S_i$ such that $z = \sum_{i=1}^n z^i$ and $z^i \in S_i$ except for at most m values of i.*

To write down our refined version, we need to first introduce the concept of kth convex hull.

Definition 3.1. The *kth convex hull* of a set S, denoted by $\mathrm{conv}_k S$, is the set of convex combinations of k points in S, i.e.,

$$\mathrm{conv}_k S = \left\{ \sum_{j=1}^k \alpha_j v^j \middle| v^j \in S, \alpha_j \geq 0, \forall j = 1, \ldots, k, \sum_{j=1}^k \alpha_j = 1 \right\}.$$

Figure 3.1 gives a simple example to illustrate the definition of kth convex hull. In Figure 3.1, the set $S = \{A, B, C\}$, $\mathrm{conv}_1 S = S$, $\mathrm{conv}_2 S$ are the segments

AB, BC and CA, while $\operatorname{conv}_3 S$ is the full triangle which is also the convex hull of set S. In general, Carathéodory's theorem implies that $\operatorname{conv}_{m+1} S = \operatorname{conv} S$ for any set $S \subseteq \mathbb{R}^m$. However, for a particular set, the minimum k such that $\operatorname{conv}_k S = \operatorname{conv} S$ can be smaller than $m+1$, and this number intuitively reflects how the set is closer to being convex. For instance, if we start from $T = \operatorname{conv}_2 S$, the set in Figure 3.1(b), then $\operatorname{conv}_k T = \operatorname{conv} T$ for $k = 2$.

Next, we recall the concept of k-extreme points of a convex set, which is a generalization of extreme points.

Definition 3.2. A point z in a convex set S is called a *k-extreme* point of S if we cannot find $(k+1)$ independent vectors $d^1, d^2, \ldots, d^{k+1}$ such that $z \pm d^i \in S$.

According to our definition, if a point is k-extreme, then it is also k'-extreme for $k' \geq k$. For a convex set in $\mathbb{R}^m$, a point is an extreme point if and only if it is 0-extreme, a point is on the boundary if and only if it is $(m-1)$-extreme, and every point is m-extreme. For example, in Figure 3.1(c), the vertices A, B, C are 0-extreme, the points on segments AB, BC and CA are 1-extreme, and all the points are 2-extreme.

Now we can state our refined Shapley-Folkman lemma:

Theorem 3.2. *Let $S_1, S_2, \ldots, S_n$ be subsets of $\mathbb{R}^m$. Assume z is a k-extreme point of $\operatorname{conv} \sum_{i=1}^n S_i$, then there exist integers $1 \leq k_i \leq k+1$ with $\sum_{i=1}^n k_i \leq k+n$ and points $z^i \in \operatorname{conv}_{k_i} S_i$ such that $z = \sum_{i=1}^n z^i$.*

The original Shapley-Folkman lemma (Theorem 3.1) now becomes a direct corollary of Theorem 3.2. Since any point $z \in \operatorname{conv} \sum_{i=1}^n S_i$ is an m-extreme point. Applying Theorem 3.2 on this point gives a decomposition $z = \sum_{i=1}^n z^i$ with

$z^i \in \text{conv}_{k_i} S_i \subseteq \text{conv}\, S_i$ and $\sum_{i=1}^n k_i \leq m+n$. Then the conclusion in Theorem 3.1 follows because $z^i \in S_i$ if $k_i = 1$, while the number of indices i with $k_i \geq 2$ is bounded by m.

Our Theorem 3.2 is similar to the refined version of Shapley-Folkman lemma proposed in [25]. However, the result in [25] does not take extremeness of the point into account, which can be regarded as a special case of Theorem 3.2 for $k = m$.

To prove Theorem 3.2, we need the following property of k-extreme points in a polyhedron:

Lemma 3.1. *Let $P \subseteq \mathbb{R}^m$ be a polyhedron and z be a k-extreme point of P, then there exists a vector $a \in \mathbb{R}^m$ such that the set $\{y \in P | a^T y \leq a^T z\}$ is in a k-dimensional affine subspace.*

Proof. Assume that the polyhedron P is represented by $Ax \geq b$. Let $A_=$ be the submatrix of A containing the rows of active constraints for the point z, and $b_=$ be the vector containing the corresponding constants in b. The dimension of the kernel of $A_=$ is at most k. Otherwise, we can find independent and sufficiently small vectors $d^1, \ldots, d^{k+1}$ such that $A_= d^i = 0$ and $A(z \pm d^i) \geq b$ for $i = 1, \ldots, k+1$. This implies $z \pm d^i \in P$, which contradicts with the k-extremeness of point z.

Let a be the vector such that a^T is the sum of all rows in $A_=$. Consider a point y satisfying $Ay \geq b$ and $a^T y \leq a^T z$. Since adding all inequalities together in $A_= y \geq b_= = A_= z$ gives $a^T y \geq a^T z$, we must have $A_= y = b_=$. Therefore, y is in the affine subspace defined by $A_= x = b_=$ whose dimension is at most k. □

In the literature, the point satisfying the conclusion of Lemma 3.1 is called a *k-exposed point*. For a general convex set S, a k-extreme point may fail to be a

k-exposed point, although it must be in the closure of the set of k-exposed points if S is compact [26]. For the special case of polyhedra, these two concepts are equivalent, and Lemma 3.1 is a generalization of the well-known result that an extreme point of a polyhedron is the unique minimizer of some linear function.

Proof of Theorem 3.2. Since z is in the convex hull of $\sum_{i=1}^{n} S_i$, there exists some integer l such that z can be written as

$$z = \sum_{j=1}^{l} \alpha_j \sum_{i=1}^{n} v^{ij}, \tag{3.6}$$

in which $v^{ij} \in S_i$, $\alpha_j \geq 0$, $j = 1, \ldots, l$ and $\sum_{j=1}^{l} \alpha_j = 1$.

Define

$$S_i' = \{v^{i1}, \ldots, v^{il}\} \subseteq S_i,$$

then (3.6) actually tells us that $z \in \operatorname{conv} \sum_{i=1}^{n} S_i'$, so z must be k-extreme in this polytope that lies in $\operatorname{conv} \sum_{i=1}^{n} S_i$. By Lemma 3.1, there exists a vector $a \in \mathbb{R}^m$ such that the set

$$\left\{ y \in \operatorname{conv} \sum_{i=1}^{n} S_i' \middle| a^T y \leq a^T z \right\}$$

is in a k-dimensional affine subspace L of $\mathbb{R}^m$. Without loss of generality, we assume that the subspace

$$L = \{y \in \mathbb{R}^m | y_{k+1} = y_{k+2} = \cdots = y_m = 0\}.$$

Next, consider the following linear program in which β_{ij} are the decision vari-

ables:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{n}\sum_{j=1}^{l} \beta_{ij} a^T v^{ij} \\
\text{s. t.} \quad & \sum_{i=1}^{n}\sum_{j=1}^{l} \beta_{ij} v_s^{ij} = z_s, \quad \forall s = 1, \ldots, k, \\
& \sum_{j=1}^{l} \beta_{ij} = 1, \quad \forall i = 1, \ldots, n, \\
& \beta_{ij} \geq 0, \quad \forall i = 1, \ldots, n, \ \forall j = 1, \ldots, l.
\end{aligned}$$

Setting $\beta_{ij} = \alpha_j$ gives a feasible solution to the above problem with objective value $a^T z$. Among all the optimal solutions, pick up a particular vertex solution β_{ij}^*, which should have at least nl active constraints. We already have $k + n$ active constraints, so the number of nonzero β_{ij}^* entries is at most $k + n$. Define

$$z^i = \sum_{j=1}^{l} \beta_{ij}^* v^{ij}, \quad z' = \sum_{i=1}^{n} z^i,$$

and let k_i be the number of nonzero entries in $\beta_{i1}^*, \ldots, \beta_{il}^*$. Since $\sum_{j=1}^{l} \beta_{ij}^* = 1$, there must be a nonzero one and thus $k_i \geq 1$. Now we know that $z^i \in \text{conv}_{k_i} S_i$, and $\sum_{i=1}^{n} k_i \leq k+n$ implies that each k_i cannot exceed $k+1$. The remaining thing to show is $z_s = z'_s$ for $s = k + 1, \ldots, m$. Because

$$z' \in \sum_{i=1}^{n} \text{conv}\, S'_i = \text{conv} \sum_{i=1}^{n} S'_i$$

and

$$a^T z' = \sum_{i=1}^{n}\sum_{j=1}^{l} \beta_{ij}^* a^T v^{ij} \leq a^T z,$$

$z' \in L$. Since $z \in L$, the last $m - k$ components of both z and z' are all zeros, so $z = z'$. □

In Section 3.5, when proving the bound for the duality gap, we will not directly apply Theorem 3.2 but a special case of it given by the following Corollary 3.1. At that time, we will see how Corollary 3.1 will improve the bound compared with the existing result such as [25] without the consideration of extremeness.

Corollary 3.1. *Let $S_1, S_2, \ldots, S_n$ be subsets of $\mathbb{R}^m$. If $z \in \operatorname{conv} \sum_{i=1}^n S_i$, then there exist integers $1 \le k_i \le m$ with $\sum_{i=1}^n k_i \le m - 1 + n$ and points $z^i \in \operatorname{conv}_{k_i} S_i$ such that $z_s = \sum_{i=1}^n z_s^i$ for $s = 1, \ldots, m-1$ and $z_m \ge \sum_{i=1}^n z_m^i$.*

Proof. Using the same argument in the proof of Theorem 3.2, choose $S_i' \subseteq S_i$ containing finite points such that $z \in \operatorname{conv} \sum_{i=1}^n S_i'$. Since $\operatorname{conv} \sum_{i=1}^n S_i'$ is a compact set,

$$\inf \left\{ w_m \middle| w \in \operatorname{conv} \sum_{i=1}^n S_i', w_1 = z_1, \ldots, w_{m-1} = z_{m-1} \right\}$$

can be achieved by some point w^*. w^* is an $(m-1)$-extreme point of $\operatorname{conv} \sum_{i=1}^n S_i'$, and applying Theorem 3.2 on the point w^* gives the desired result. □

3.3 Characterization of Nonconvexity

To improve the bound (3.5), some finer characterization of the nonconvexity of a function has to be introduced. In parallel to the definition of kth convex hull of a set, define the *kth nonconvexity* $\rho^k(f)$ of a proper function f to be the supremum in (3.4) taken over the convex combinations of k points instead of arbitrary number of points. Obviously,

$$0 = \rho^1(f) \le \rho^2(f) \le \cdots \le \rho(f).$$

In fact, we have the following property:

Proposition 3.1. *For any proper function $f : \mathbb{R}^n \to (-\infty, +\infty]$, $\rho^{n+1}(f) = \rho(f)$.*

Proof. We only need to show that $\rho(f) \le \rho^{n+1}(f)$. Choose any convex combination $x = \sum_{j=1}^l \alpha_j x^j$ with all points $x^j \in \operatorname{dom} f$, $\alpha_j \ge 0$, and $\sum_{j=1}^l \alpha_j = 1$. Since

$(x^j, f(x^j)) \in \operatorname{epi} f$, the point

$$\left(\sum_{j=1}^{l} \alpha_j x^j, \sum_{j=1}^{l} \alpha_j f(x^j)\right) \in \operatorname{conv} \operatorname{epi} f.$$

Using Corollary 3.1 on a single set $S_1 = \operatorname{epi} f$, we can find $(y^i, t_i) \in \operatorname{epi} f$, $\beta_i \geq 0$, $i = 1, \ldots, n+1$, and $\sum_{i=1}^{n+1} \beta_i = 1$ such that

$$x = \sum_{j=1}^{l} \alpha_j x^j = \sum_{i=1}^{n+1} \beta_i y^i, \quad \sum_{j=1}^{l} \alpha_j f(x^j) \geq \sum_{i=1}^{n+1} \beta_i t_i.$$

Now

$$f(x) - \sum_{j=1}^{l} \alpha_j f(x^j) \leq f\left(\sum_{i=1}^{n+1} \beta_i y^i\right) - \sum_{i=1}^{n+1} \beta_i t_i \leq f\left(\sum_{i=1}^{n+1} \beta_i y^i\right) - \sum_{i=1}^{n+1} \beta_i f(y^i),$$

which implies $\rho(f) \leq \rho^{n+1}(f)$. □

For lower semi-continuous functions, the following proposition provides an equivalent definition for the kth nonconvexity, which sheds light on the connection between the concepts of kth nonconvexity and kth convex hull.

Proposition 3.2. *Assume a proper function f is lower semi-continuous and it is bounded below by some affine function. Let $f^{(k)}$ be the function whose epigraph is the closure of the kth convex hull of the epigraph of f, i.e.,*

$$\operatorname{epi} f^{(k)} = \operatorname{cl} \operatorname{conv}_k \operatorname{epi} f.$$

Then

$$\rho^k(f) = \sup_x \{f(x) - f^{(k)}(x)\}, \tag{3.7}$$

where we interpret $(+\infty) - (+\infty) = 0$.

Proof. The assumption on the function f implies that $f^{(k)}$ is also a proper function. Consider an arbitrary k-point convex combination of points $x^j \in \operatorname{dom} f$, for $j = 1, \ldots, k$. Following the first step in the proof of Proposition 3.1, we have

$$\left(\sum_{j=1}^{k} \alpha_j x^j, \sum_{j=1}^{k} \alpha_j f(x^j)\right) \in \operatorname{conv}_k \operatorname{epi} f \subseteq \operatorname{cl} \operatorname{conv}_k \operatorname{epi} f = \operatorname{epi} f^{(k)}.$$

Therefore,

$$f\left(\sum_{j=1}^{k}\alpha_j x^j\right)-\sum_{j=1}^{k}\alpha_j f(x^j)\le f\left(\sum_{j=1}^{k}\alpha_j x^j\right)-f^{(k)}\left(\sum_{j=1}^{k}\alpha_j x^j\right),$$

which implies

$$\rho^k(f)\le\sup_x\{f(x)-f^{(k)}(x)\}.$$

Now we prove the reverse direction. For any $x\in\operatorname{dom} f^{(k)}$,

$$(x,f^{(k)}(x))\in\operatorname{epi} f^{(k)}=\operatorname{cl}\operatorname{conv}_k\operatorname{epi} f.$$

There exists a sequence $(\kappa^l,\eta_l)\in\operatorname{conv}_k\operatorname{epi} f$ such that

$$\lim_{l\to\infty}\kappa^l=x,\quad\lim_{l\to\infty}\eta_l=f^{(k)}(x).$$

For each l, since $(\kappa^l,\eta_l)\in\operatorname{conv}_k\operatorname{epi} f$, there exists $\alpha_j\ge 0$ for $j=1,\dots,k$ such that $\sum_{j=1}^{k}\alpha_j=1$ and

$$\kappa^l=\sum_{j=1}^{k}\alpha_j x^j,\quad\eta_l\ge\sum_{j=1}^{k}\alpha_j f(x^j)$$

in which $x^j\in\operatorname{dom} f$. Thus

$$f(\kappa^l)-\eta_l\le f\left(\sum_{j=1}^{k}\alpha_j x^j\right)-\sum_{j=1}^{k}\alpha_j f(x^j)\le\rho^k(f).$$

On the other hand, by the lower semi-continuity of f,

$$f(x)\le\liminf_{l\to\infty}f(\kappa^l)\le\lim_{l\to\infty}\eta_l+\rho^k(f)=f^{(k)}(x)+\rho^k(f).$$

The case $x\notin\operatorname{dom} f^{(k)}$ is trivial since in this case $f(x)=f^{(k)}(x)=+\infty$. □

As a remark, the equality (3.7) can be regarded as a generalization for the alternative definition of nonconvexity

$$\rho(f)=\sup_x\{f(x)-f^{**}(x)\}$$

used in [9].

3.4 Examples of Computing Nonconvexity

In this section, three examples will be given to illustrate how to calculate the kth nonconvexity of a particular function. The results in Example 3.1 and Example 3.2 will be used by the network utility maximization problem in Section 4.1, and Example 3.3 will be used by the dynamic spectrum management problem in Section 4.2.

Example 3.1. Consider the function

$$f(x) = f(x_1, \ldots, x_n) = \min_{s=1,\ldots,n} x_s$$

defined on the box $0 \le x \le 1$, $x \in \mathbb{R}^n$. It is already known that $\rho(f) = (n-1)/n$ (see [9, Table 1]). By Proposition 3.1, $\rho^k(f) = \rho(f) = (n-1)/n$ for $k \ge n+1$.

For $k = 1, \ldots, n$, as in the proof of Proposition 3.2, pick up any k-point convex combination of points $0 \le x^j \le 1$, $j = 1, \ldots, k$. For a given $i \in \{1, \ldots, k\}$, let $s(i)$ be the index such that $x^i_{s(i)}$ is the minimum among $x^i_1, \ldots, x^i_n$, then

$$\begin{aligned} f(x) &= \min_{s=1,\ldots,n} \left\{ \sum_{j=1}^{k} \alpha_j x^j_s \right\} \le \sum_{j=1}^{k} \alpha_j x^j_{s(i)} \\ &\le \alpha_i x^i_{s(i)} + 1 - \alpha_i = \alpha_i f(x^i) + 1 - \alpha_i, \end{aligned}$$

where we use the fact that all x^j are within the box $0 \le x \le 1$. Summing up among $i = 1, \ldots, k$, we have

$$k f(x) \le \sum_{i=1}^{k} \alpha_i f(x^i) + k - 1,$$

which implies

$$f(x) - \sum_{i=1}^{k} \alpha_i f(x^i) \le \frac{k-1}{k} \left(1 - \sum_{i=1}^{k} \alpha_i f(x^i) \right) \le \frac{k-1}{k}.$$

The above argument shows that $\rho^k(f) \leq (k-1)/k$. In fact, the equality holds, which can be easily seen by considering the average of first k points of

$$x^1 = (0, 1, \ldots, 1),$$
$$x^2 = (1, 0, \ldots, 1),$$
$$\ldots,$$
$$x^n = (1, 1, \ldots, 0).$$

In conclusion,

$$\rho^k(f) = \begin{cases} \dfrac{k-1}{k}, & \text{if } k = 1, \ldots, n, \\ \dfrac{n-1}{n}, & \text{if } k \geq n+1. \end{cases}$$

Example 3.2. Consider the function

$$g(x) = g(x_1, \ldots, x_n) = -\log \max_{s=1,\ldots,n} x_s$$

defined on the region $x \geq 0$ except $x = 0$.

For $k = 1, \ldots, n$, pick up any k-point convex combination. Without loss of generality, assume the coefficients $\alpha_j > 0$ for $j = 1, \ldots, k$. For a given $i \in \{1, \ldots, k\}$, let $s(i)$ be the index such that $x^i_{s(i)}$ is the maximum among $x^i_1, \ldots, x^i_n$, then

$$g(x) = -\log \max_{s=1,\ldots,n} \left\{ \sum_{j=1}^{k} \alpha_j x^j_s \right\} \leq -\log \sum_{j=1}^{k} \alpha_j x^j_{s(i)}$$
$$\leq -\log(\alpha_i x^i_{s(i)}) = -\log \alpha_i + g(x^i).$$

Summing up among $i = 1, \ldots, k$ with weight α_i, we have

$$g(x) \leq -\sum_{i=1}^{k} \alpha_i \log \alpha_i + \sum_{i=1}^{k} \alpha_i g(x^i)$$
$$\leq \log k + \sum_{i=1}^{k} \alpha_i g(x^i).$$

The above argument shows that $\rho^k(g) \leq \log k$. In fact, the equality holds, which can be easily seen by considering the average of first k points of

$$x^1 = (1, 0, \ldots, 0),$$
$$x^2 = (0, 1, \ldots, 0),$$
$$\ldots,$$
$$x^n = (0, 0, \ldots, 1).$$

To calculate $\rho^{n+1}(g)$, define $h(x) = -\log \sum_{s=1}^n x_s$. Then $h(x)$ is convex and $g(x) - \log n \leq h(x) \leq g(x)$. Thus, for any $(n+1)$-point convex combination,

$$\begin{aligned} g\left(\sum_{j=1}^{n+1} \alpha_j x^j\right) &\leq h\left(\sum_{j=1}^{n+1} \alpha_j x^j\right) + \log n \\ &\leq \sum_{j=1}^{n+1} \alpha_j h(x^j) + \log n \\ &\leq \sum_{j=1}^{n+1} \alpha_j g(x^j) + \log n. \end{aligned}$$

Therefore, $\rho^{n+1}(g) \leq \log n$. On the other hand, $\rho^{n+1}(g) \geq \rho^n(g) = \log n$.

In conclusion,

$$\rho^k(g) = \begin{cases} \log k, & \text{if } k = 1, \ldots, n, \\ \log n, & \text{if } k \geq n+1. \end{cases}$$

Example 3.3. Consider the function

$$h_\sigma(x) = h_\sigma(x_1, \ldots, x_n) = \sum_{s=1}^n \log \frac{\|x\|_1 - x_s + \sigma}{\|x\|_1 + \sigma}$$

defined on the box $0 \leq x \leq 1$, $x \in \mathbb{R}^n$. Here σ is a parameter in the range $0 < \sigma \leq 1$.

For complicated functions such as this one, it is usually hard to compute their kth nonconvexity exactly. However, sometimes we can approximate the kth nonconvexity of a function by reducing it to another function whose nonconvexity is

already known. Using this technique, we are able to show that

$$\rho^k(h_\sigma) \leq \log(k/\sigma).$$

Define an auxiliary function

$$H(x;\sigma) = \prod_{s=1}^{n} \frac{\|x\|_1 - x_s + \sigma}{\|x\|_1 + \sigma},$$

then $h_\sigma(x) = \log H(x;\sigma)$. To compute the kth nonconvexity for the function h_σ, we first prove some elementary properties for the function $H(x;\sigma)$.

Lemma 3.2. *The function $H(x;\sigma)$ has the following properties:*

(a) For any vectors x and y in the region $0 \leq x, y \leq \sigma$, if $y \leq x$, then $H(y;\sigma) \geq H(x;\sigma)$.

(b) $\sigma H(x;1) \leq H(x;\sigma) \leq H(x;1)$.

Proof. For any x in the region $0 \leq x \leq \sigma$, the partial derivatives

$$\begin{aligned}
\frac{\partial H(x;\sigma)}{\partial x_i} &= H(x;\sigma)\left(\sum_{s=1}^{n} \frac{1}{\|x\|_1 - x_s + \sigma} - \frac{1}{\|x\|_1 - x_i + \sigma} - \frac{n}{\|x\|_1 + \sigma}\right) \\
&= H(x;\sigma)\left(\sum_{s=1}^{n} \frac{x_s}{(\|x\|_1 - x_s + \sigma)(\|x\|_1 + \sigma)} - \frac{1}{\|x\|_1 - x_i + \sigma}\right) \\
&\leq H(x;\sigma)\left(\sum_{s=1}^{n} \frac{x_s}{\|x\|_1(\|x\|_1 + \sigma)} - \frac{1}{\|x\|_1 + \sigma}\right) = 0,
\end{aligned}$$

which gives the first property.

For the second property, it is obvious to see that $H(x;\sigma) \leq H(x;1)$. The other inequality is equivalent to

$$p(\sigma) = \frac{1}{\sigma} H(x;\sigma) - H(x;1) \geq 0.$$

The partial derivative

$$\begin{aligned}\frac{\partial H(x;\sigma)}{\partial \sigma} &= H(x;\sigma)\left(\sum_{s=1}^{n}\frac{1}{\|x\|_1 - x_s + \sigma} - \frac{n}{\|x\|_1 + \sigma}\right)\\ &= H(x;\sigma)\sum_{s=1}^{n}\frac{x_s}{(\|x\|_1 - x_s + \sigma)(\|x\|_1 + \sigma)}\\ &\le H(x;\sigma)\sum_{s=1}^{n}\frac{x_s}{\sigma(\|x\|_1 + \sigma)}\\ &= \frac{1}{\sigma}H(x;\sigma)\frac{\|x\|_1}{\|x\|_1 + \sigma} \le \frac{1}{\sigma}H(x;\sigma)\end{aligned}$$

implies that

$$p'(\sigma) = -\frac{1}{\sigma^2}H(x;\sigma) + \frac{1}{\sigma}\frac{\partial H(x;\sigma)}{\partial \sigma} \le 0.$$

Therefore, the function $p(\sigma)$ is nonincreasing. Together with $p(1) = 0$, we have proved the nonnegativity of $p(\sigma)$. □

To upper bound the kth nonconvexity of the function h_σ, consider arbitrary points x^j for $j = 1, \ldots, k$ with corresponding combination weights $\alpha_j > 0$. Define k vectors $y^1, \ldots, y^k$ in $\mathbb{R}^k$ by

$$\begin{aligned} y^1 &= (1/H(x^1;1), 0, \ldots, 0),\\ y^2 &= (0, 1/H(x^2;1), \ldots, 0),\\ &\ldots,\\ y^k &= (0, 0, \ldots, 1/H(x^k;1)).\end{aligned}$$

Using the result of nonconvexity for the function g given in Example 3.2 and the properties proved in Lemma 3.2, we have

$$\begin{aligned} h_\sigma\left(\sum_{j=1}^{k}\alpha_j x^j\right) &= \log H\left(\sum_{j=1}^{k}\alpha_j x^j;\sigma\right) \le \log H\left(\sum_{j=1}^{k}\alpha_j x^j;1\right) && \text{by property (b)}\\ &\le \log H(\alpha_j x^j;1), \quad \forall j = 1,\ldots,k, && \text{by property (a)}\end{aligned}$$

and then

$$
\begin{aligned}
h_\sigma\left(\sum_{j=1}^k \alpha_j x^j\right) &\le \log \min_{j=1,\dots,k} H(\alpha_j x^j; 1) \\
&\le \log \min_{j=1,\dots,k} \frac{1}{\alpha_j} H(\alpha_j x^j; \alpha_j) = \log \min_{j=1,\dots,k} \frac{1}{\alpha_j} H(x^j; 1) && \text{by property (b)} \\
&= g\left(\sum_{j=1}^k \alpha_j y^j\right) \le \sum_{j=1}^k \alpha_j g(y^j) + \log k && \text{by the nonconvexity of } g \\
&= \sum_{j=1}^k \alpha_j \log H(x^j; 1) + \log k \le \sum_{j=1}^k \alpha_j \log \frac{1}{\sigma} H(x^j; \sigma) + \log k && \text{by property (b)} \\
&= \sum_{j=1}^k \alpha_j h_\sigma(x^j) + \log \frac{k}{\sigma}.
\end{aligned}
$$

The above argument shows that the kth nonconvexity $\rho^k(h_\sigma) \le \log(k/\sigma)$.

In the above example, an upper bound for the kth nonconvexity of function h_σ is obtained by a reduction from the nonconvexity of g in Example 3.2. Along this line of thoughts, it is conceivable to find the exact value for the kth nonconvexity of h_σ if we are able to reduce h_σ to itself (but with just k variables).

3.5 Bounding Duality Gap

Now we can state the main result on the duality gap between the primal problem (3.1) and the dual problem (3.2).

Theorem 3.3. *Assume that the primal problem* (3.1) *is feasible, i.e.,* $p < +\infty$. *Then there exist integers* $1 \le k_i \le m+1$ *such that* $\sum_{i=1}^n k_i \le m+n$ *and the duality gap*

$$p - d \le \sum_{i=1}^n \rho_i^{k_i}.$$

Here $\rho_i^k = \rho^k(f_i)$ *is the* k*th nonconvexity of function* f_i.

First, following the proof of the strong duality in Section 2.2, we similarly define the perturbation function $v : \mathbb{R}^m \to [-\infty, +\infty]$ by letting $v(z)$ be the optimal value of the perturbed problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} f_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{n} A_i x^i \leq b + z. \end{aligned}$$

By Lemma 2.1, $p = v(0)$ and $d = v^{**}(0)$.

Proof of Theorem 3.3. Since (3.1) is feasible, $v(0) = p < +\infty$. Let

$$\xi = \sum_{i=1}^{n} \inf_{x^i} f_i(x^i),$$

then by our assumption of f_i, ξ is finite. $v(z) \geq \xi$ for all $z \in \mathbb{R}^m$. As a consequence, $v(z)$ is bounded below by some affine function, so

$$-\infty < v^{**}(0) \leq v(0) < +\infty, \quad \operatorname{epi} v^{**} = \operatorname{cl} \operatorname{conv} \operatorname{epi} v.$$

By Lemma 2.3, v is lower semi-continuous. Since

$$(0, v^{**}(0)) \in \operatorname{epi} v^{**} = \operatorname{cl} \operatorname{conv} \operatorname{epi} v,$$

for every $\epsilon > 0$, there exists $(\kappa, \eta) \in \operatorname{conv} \operatorname{epi} v$ which is sufficiently close to $(0, v^{**}(0))$ such that

$$v(\kappa) \geq v(0) - \epsilon, \quad \eta \leq v^{**}(0) + \epsilon. \tag{3.8}$$

Because $(\kappa, \eta) \in \operatorname{conv} \operatorname{epi} v$, there exists some integer l and $\alpha_j \geq 0$ for $j = 1, \ldots, l$ such that $\sum_{j=1}^{l} \alpha_j = 1$ and

$$\kappa = \sum_{j=1}^{l} \alpha_j z^j, \quad \eta \geq \sum_{j=1}^{l} \alpha_j v(z^j)$$

in which $z^j \in \operatorname{dom} v$.

For each $j = 1, \ldots, l$, find $(\hat{x}^{1j}, \ldots, \hat{x}^{nj})$ attaining the optimal value of the perturbed problem related to $v(z^j)$, i.e.,

$$v(z^j) = \sum_{i=1}^{n} f_i(\hat{x}^{ij}), \quad \sum_{i=1}^{n} A_i \hat{x}^{ij} \leq b + z^j,$$

which means there exists some vector $w^j \in \mathbb{R}^m_+$ such that

$$(b + z^j - w^j, v(z^j)) \in \sum_{i=1}^{n} C_i,$$

where

$$C_i = \{(A_i x^i, f_i(x^i)) | f_i(x^i) < +\infty, x^i \in \mathbb{R}^{n_i}\}.$$

Taking convex combination of the points above, we have

$$\left(b + \kappa - \sum_{j=1}^{l} \alpha_j w^j, \sum_{j=1}^{l} \alpha_j v(z^j) \right) \in \operatorname{conv} \sum_{i=1}^{n} C_i.$$

Now we can apply Corollary 3.1[1], which gives points $(r^i, s_i) \in \operatorname{conv}_{k_i} C_i$ with $1 \leq k_i \leq m + 1$ such that

$$b + \kappa \geq b + \kappa - \sum_{j=1}^{l} \alpha_j w^j = \sum_{i=1}^{n} r^i, \quad \eta \geq \sum_{j=1}^{l} \alpha_j v(z^j) \geq \sum_{i=1}^{n} s_i$$

and $\sum_{i=1}^{n} k_i \leq m + n$. Since $(r^i, s_i) \in \operatorname{conv}_{k_i} C_i$, there exists $\tilde{x}^{ij} \in \mathbb{R}^{n_i}$, $\beta_{ij} \geq 0$ for $j = 1, \ldots, k_i$ such that $f_i(\tilde{x}^{ij}) < +\infty$, $\sum_{j=1}^{k_i} \beta_{ij} = 1$ and

$$r^i = \sum_{j=1}^{k_i} \beta_{ij} A_i \tilde{x}^{ij}, \quad s_i = \sum_{j=1}^{k_i} \beta_{ij} f_i(\tilde{x}^{ij}).$$

Thus,

$$\kappa \geq \sum_{i=1}^{n} \sum_{j=1}^{k_i} \beta_{ij} A_i \tilde{x}^{ij} - b = \sum_{i=1}^{n} A_i \sum_{j=1}^{k_i} \beta_{ij} \tilde{x}^{ij} - b, \tag{3.9}$$

and

$$\sum_{i=1}^{n} \rho_i^{k_i} + \eta \geq \sum_{i=1}^{n} \left(\rho_i^{k_i} + \sum_{j=1}^{k_i} \beta_{ij} f_i(\tilde{x}^{ij}) \right) \geq \sum_{i=1}^{n} f_i \left(\sum_{j=1}^{k_i} \beta_{ij} \tilde{x}^{ij} \right). \tag{3.10}$$

[1]Here if we apply Theorem 3.2 with dimension $m + 1$ instead of using Corollary 3.1, all the remaining argument still works but the bound for $\sum_{i=1}^{n} k_i$ has to be weakened to $m + n + 1$ from $m + n$. Therefore, the consideration of extremeness in Corollary 3.1 provides the exact improvement parallel to how (3.5) improves from the earliest bound (3.3).

From (3.9) we know that

$$\left(\sum_{j=1}^{k_1}\beta_{1j}\tilde{x}^{1j},\ldots,\sum_{j=1}^{k_n}\beta_{nj}\tilde{x}^{nj}\right)$$

is feasible to the perturbed problem related to $v(\kappa)$, so the corresponding objective value satisfies

$$\sum_{i=1}^{n} f_i\left(\sum_{j=1}^{k_i}\beta_{ij}\tilde{x}^{ij}\right)\geq v(\kappa).$$

The above inequality, together with (3.10) and (3.8), implies

$$v^{**}(0)+\epsilon+\sum_{i=1}^{n}\rho_i^{k_i}\geq v(0)-\epsilon.$$

We finish the proof by letting $\epsilon\to 0$. Because all the k_i depend on ϵ, we have to choose the worst case of $\sum_{i=1}^{n}\rho_i^{k_i}$ encountered in this process. □

From a computational viewpoint, since we do not know the k_i that appeared in Theorem 3.3, in order to find a number for the bound, we have to find the worst case k_i by solving the following optimization problem

$$\begin{aligned}\max\quad & \sum_{i=1}^{n}\rho_i^{k_i}\\ \text{s.t.}\quad & 1\leq k_i\leq m+1,\ k_i\in\mathbb{Z},\quad \forall i=1,\ldots,n,\\ & \sum_{i=1}^{n}k_i\leq m+n.\end{aligned}\tag{3.11}$$

Let B be the optimal value of (3.11), then

$$B\leq\sum_{i=1}^{n}\rho(f_i).$$

On the other hand, since for any feasible solution of (3.11), the number of k_i with $k_i\geq 2$ is bounded by m, so

$$B=\sum_{i:k_i\geq 2}\rho_i^{k_i}\leq\sum_{i=1}^{m}\rho(f_i)$$

if $\rho(f_1) \geq \cdots \geq \rho(f_n)$. The above argument shows that the bound B given by the optimization problem (3.11) is at least as tight as the bound (3.5) in [9].

To illustrate the procedure to calculate the bound B, consider the simple case where all the x^i in the primal problem (3.1) are one-dimensional and all the functions f_i equal to the same function f. In this case, $\rho_i^{k_i} = \rho(f)$ if $k_i \geq 2$. The optimal value to (3.11) is attained when the number of k_i that equal to 2 is maximized, so the optimal value is $\min\{m, n\}\rho(f)$, which is the same as the result given by (3.5). The Example 1 used in [9] belongs to this category. It hence explains why the bound (3.5) is tight for that example. However, if the dimension of x^i in the primal problem can be arbitrarily large, the bound (3.5) can be very loose. As will be shown in Chapter 4, the difference between the bound (3.5) and the exact duality gap tends to infinity for a series of problems.

CHAPTER 4

APPLICATIONS OF DUALITY GAP ESTIMATION

In this chapter, we will first apply the bound of the duality gap proposed in Chapter 3 to the network utility maximization problem in networking in Section 4.1 and then to the dynamic spectrum management problem in communication in Section 4.2. In both cases, we are able to obtain tighter results with less effort than the ones based on the previous bounds.

4.1 Joint Routing and Congestion Control in Networking

Consider a network with N users and L links. Let a strictly positive vector $c \in \mathbb{R}^L$ contain the capacity of each link. Each user i has K^i available paths to send its commodity. We assume that the users are sorted such that $K^1 \geq \cdots \geq K^N$. The routing matrix of user i, denoted by R^i, is a $L \times K^i$ matrix defined by

$$R^i_{lk} = \begin{cases} 1, & \text{if the } k\text{th path of user } i \text{ passes through link } l, \\ 0, & \text{otherwise.} \end{cases}$$

For example, as shown in Figure 4.1, a five-link network supports two users, each of which has two available paths. The corresponding routing matrices are given by

$$R^1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad R^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

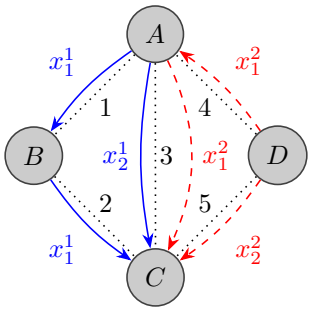


Figure 4.1: A sample network with five links (dotted line) and two users. The first user, whose source is A and destination is C, has rate x_1^1 and x_2^1 on its two paths (solid line). The second user, whose source is D and destination is C, has rate x_1^2 and x_2^2 on its two paths (dashed line).

Let $x^i \in \mathbb{R}^{K^i}$ be the vector in which x_k^i is the amount of commodity sent by user i on its kth path. Assume that each user i has a utility function $U_i(\cdot)$ depending on the vector x^i, then the network utility maximization problem can be written as

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} U_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{N} R^i x^i \le c, \\ & x^i \ge 0, \quad \forall i = 1, \ldots, N. \end{aligned} \tag{4.1}$$

If all the utility functions $U^i(\cdot)$ are concave, then the above problem (4.1) can be solved by standard convex optimization techniques. Difficulty arises when $U^i(\cdot)$ is not concave. For example, if we restrict each user to choose only one path (single-path routing) and want to maximize the total throughput of the network, then the corresponding utility function is

$$U_i(x^i) = \max_{s=1,\ldots,K^i} x_s^i.$$

Define

$$f_i(x^i) = \begin{cases} \min\limits_{s=1,\ldots,K^i} (-x_s^i), & \text{if } 0 \le x^i \le \|c\|_\infty, \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.2}$$

Here $\|c\|_\infty$ is the maximum link capacity in the network. Now the original network utility maximization problem (4.1) is equivalent to the following problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} f_i(x^i) \\ \text{s.t.} \quad & \sum_{i=1}^{N} R^i x^i \le c. \end{aligned} \tag{4.3}$$

The above problem is a particular case of the general optimization problem with separable objectives (3.1) studied in Chapter 3.5. Using the same technique as shown in Example 3.1, we can prove that

$$\rho^k(f_i) = \frac{k-1}{k}\|c\|_\infty, \quad k = 1, \ldots, K^i,$$
$$\rho^{K^i+1}(f_i) = \rho(f_i) = \frac{K^i - 1}{K^i}\|c\|_\infty.$$

In the following, suppose each user has a large number of paths to select. More explicitly, $K^i \ge L+1$ is assumed for user i. Based on the bound (3.5), the duality gap is bounded by

$$\sum_{i=1}^{\min\{N,L\}} \frac{K^i - 1}{K^i}\|c\|_\infty,$$

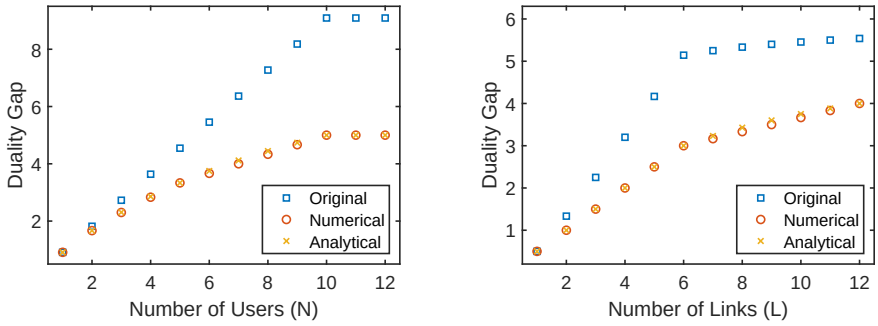
which is at least

$$\min\{N, L\}\frac{L}{L+1}\|c\|_\infty. \tag{4.4}$$

In contrast, by Theorem 3.3, the duality gap is bounded by the optimal value of the following optimization problem:

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \frac{k_i - 1}{k_i}\|c\|_\infty \\ \text{s.t.} \quad & 1 \le k_i \le L+1,\ k_i \in \mathbb{Z}, \quad \forall i = 1, \ldots, N, \\ & \sum_{i=1}^{N} k_i \le N + L. \end{aligned} \tag{4.5}$$

Let N' be the number of users whose $k_i \ge 2$, then $0 \le N' \le \min\{N, L\}$. If $N' > 0$,



(a) $L = 10$ (b) $N = 6$

Figure 4.2: The comparison among the original bound (4.4), the numerical result from directly solving (4.5) and the analytical result for the linear utility case.

using the inequality between arithmetic mean and harmonic mean,

$$\begin{aligned}\sum_{i=1}^{N}\frac{k_i-1}{k_i} &= \sum_{i:k_i\geq 2}\frac{k_i-1}{k_i} = N' - \sum_{i:k_i\geq 2}\frac{1}{k_i}\\ &\leq N' - \frac{N'^2}{\sum_{i:k_i\geq 2} k_i} \leq N' - \frac{N'^2}{N'+L}\\ &= \frac{L}{1+L/N'} \leq \min\{N,L\}\frac{L}{L+\min\{N,L\}}.\end{aligned}$$

The above analysis provides an upper bound for the problem (4.5), which in turn is an upper bound for the duality gap. Taking the $N \geq L$ case as an example, by the above inequality, we can bound the duality gap by $L\|c\|_\infty/2$, essentially half of the bound given by (3.5).

In principle, we can directly solve (4.5) to yield better upper bounds, and this is particularly practical for small instances. In Figure 4.2, we compare the original bound (4.4), the numerical result from directly solving (4.5) and the above analytical result, for fixed N or fixed L with the assumption that $\|c\|_\infty = 1$. Figure 4.2 shows that our analytical result is much better than the original bound, and in all cases it is almost tight compared with the numerical result. In fact, the above analysis can be regarded as solving the problem (4.5) exactly without considering all integer constraints on k_i.

Next, we consider another case in which each user has logarithmic utility but still must choose only one path. The utility function of user i can be written as

$$U_i(x^i) = \log \max_{s=1,\dots,K^i} x_s^i.$$

Define

$$g_i(x^i) = \begin{cases} -\log \max_{s=1,\dots,K^i} x_s^i, & \text{if } 0 \le x^i \le \|c\|_\infty,\ x^i \neq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then the network utility maximization problem (4.1) is equivalent to the problem obtained by replacing f_i with g_i in (4.3). Using the result in Example 3.2,

$$\rho^k(g_i) = \log k, \quad k = 1, \dots, K^i,$$
$$\rho^{K^i+1}(g_i) = \rho(g_i) = \log K^i.$$

Applying the bound (3.5) to this case, we can bound the duality gap by

$$\sum_{i=1}^{\min\{N,L\}} \log K^i, \tag{4.6}$$

which is at least $\min\{N, L\} \log(L+1)$. On the other hand, by Theorem 3.3, the duality gap is bounded by the optimal value of the following optimization problem

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \log k_i \\ \text{s.t.} \quad & 1 \le k_i \le L+1,\ k_i \in \mathbb{Z}, \quad \forall i = 1, \dots, N, \\ & \sum_{i=1}^{N} k_i \le N + L. \end{aligned} \tag{4.7}$$

If we still let N' be the number of users whose $k_i \ge 2$, then $0 \le N' \le \min\{N, L\}$ and the above bound

$$\begin{aligned} \sum_{i=1}^{N} \log k_i = \sum_{i:k_i \ge 2} \log k_i = \log \prod_{i:k_i \ge 2} k_i &\le \log \left(\frac{\sum_{i:k_i \ge 2} k_i}{N'} \right)^{N'} \\ &\le \log \left(\frac{N' + L}{N'} \right)^{N'} \le \min\{N, L\} \log \left(1 + \frac{L}{\min\{N, L\}} \right), \end{aligned}$$

where in the last step the monotonicity of the function $(1 + 1/x)^x$ is used. Note that the new bound is qualitatively tighter than the bound (4.6) provided by (3.5) by removing a logarithm factor of $O(\log L)$ when $N \geq L$.

4.2 Dynamic Spectrum Management in Communication

Consider a communication system consisting of L users sharing a common band. The band is divided equally into N tones. Each user l has a power budget p_l which can be allocated across all the tones. Let x_l^i be the power of user l allocated on tone i. Due to the crosstalk interference between users, the total noise for a user on tone i is the sum of a background noise σ_i and the power of all other users on the same tone. Therefore, the achievable transmission rate of user l on tone i is given by

$$u_l^i = \frac{1}{N} \log \left(1 + \frac{x_l^i}{\|x^i\|_1 - x_l^i + \sigma_i} \right).$$

The dynamic spectrum management problem is to maximize the total throughput of all users under the power budget constraints, which can be formulated as the following nonconvex optimization problem:

$$\begin{array}{ll} \max & \displaystyle\sum_{l=1}^{L} \sum_{i=1}^{N} u_l^i \\ \text{s. t.} & \displaystyle\sum_{i=1}^{N} x_l^i \leq p_l, \quad \forall l = 1, \ldots, L, \\ & x_l^i \geq 0, \quad \forall i = 1, \ldots, N, \ \forall l = 1, \ldots, L. \end{array} \tag{4.8}$$

For simplicity, we assume that the noises $\sigma_i \leq 1$ and the power budgets $p_l \leq 1$ (if not, then scale all the σ_i and p_l simultaneously). The latter requires all the variables $x_l^i \leq 1$. Using the function h_σ introduced in Example 3.3, the objective

function of (4.8) can be rewritten as a sum of separable objectives:

$$\sum_{l=1}^{L}\sum_{i=1}^{N} u_l^i = -\frac{1}{N}\sum_{i=1}^{N} h_{\sigma_i}(x^i).$$

For the purpose of designing dual algorithms, it is of great interest to estimate the duality gap for the problem (4.8). In [27], the authors showed that the duality gap will tend to zero if the number of users L is fixed and the number of tones N goes to infinity. [6] further determined the convergence rate of the duality gap to be $O(1/\sqrt{N})$. Using the bound (3.5), we now demonstrate how to improve the convergence rate estimation to $O(1/N)$, which can be only achieved by the method in [6] in the special case where all the noises σ_i are the same.[1]

Example 3.3 proves that the nonconvexity

$$\rho^k(h_{\sigma_i}) \le \log\frac{k}{\sigma_i} \le \log\frac{k}{\sigma}, \quad k = 1, \ldots, L+1.$$

where σ is the minimum among all the noises σ_i, so (3.5) implies that the duality gap is upper bounded by

$$\frac{\min\{N, L\}}{N}\log\frac{L+1}{\sigma}, \tag{4.9}$$

which is in the order of $O(1/N)$ if L is fixed and N increases.

In order to further improve the estimation (4.9) for the duality gap, we can resort to Theorem 3.3 and follow the exact same steps for solving (4.7), which shows that the duality gap is upper bounded by

$$\frac{\min\{N, L\}}{N}\log\frac{1 + L/\min\{N, L\}}{\sigma}.$$

Like the previous example, our bound is still tighter than the one (4.9) from (3.5) by removing a logarithm factor.

[1] The paper [6] actually studied the generalization of problem (4.8) under the existence of path loss coefficient between different users. However, the argument for $O(1/N)$ provided here can also be adapted to the general problem.

CHAPTER 5

SEPARABLE NONCONVEX PROBLEMS WITH NONLINEAR CONSTRAINTS

In this chapter, we generalize the bound in Chapter 3 to problems with separable nonlinear constraints (not necessarily convex) in Section 5.1. The result will be used in bounding the duality gap for the network utility maximization problem with granularity constraints in Section 5.2.

5.1 Generalization with Separable Nonlinear Constraints

We consider general separable problems with nonlinear constraints such as

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} f_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{n} g_i(x^i) \leq b. \end{aligned} \tag{5.1}$$

Here each $g_i : \mathbb{R}^{n_i} \to (-\infty, +\infty]^m$ is lower semi-continuous. Note that the previous problem (3.1) we studied is a special case of the optimization problem (5.1) if we choose $g_i(x^i) = A_i x^i$.

If the functions g_i are not convex, the duality gap should not only depend on the nonconvexity of functions f_i but also somehow relate to the functions g_i. Like [28], we define the kth order nonconvexity of a proper function $f : \mathbb{R}^n \to (-\infty, +\infty]$ with respect to another proper function $g : \mathbb{R}^n \to (-\infty, +\infty]^m$, denoted by $\rho^k(f, g)$. To do this, for each $x \in \mathbb{R}^n$, we introduce a set $G^k(x; g) \subseteq \mathbb{R}^m$ such that $y \in G^k(x; g)$ if and only if there exist $x^j \in \mathbb{R}^n$ and $\beta_j \in \mathbb{R}$, for $j = 1, \ldots, k$,

satisfying

$$\begin{cases} x = \sum_{j=1}^{k} \beta_j x^j, \\ y = \sum_{j=1}^{k} \beta_j g(x^j), \quad g(x^j) < +\infty, \ \forall j = 1, \ldots, k, \\ \sum_{j=1}^{k} \beta_j = 1, \quad \beta_j \geq 0, \ \forall j = 1, \ldots, k. \end{cases} \tag{5.2}$$

Define the auxiliary function $h^k(x; f, g) : \mathbb{R}^n \to (-\infty, +\infty]$ by

$$h^k(x; f, g) = \inf_{z \in \mathbb{R}^n} \{f(z) | g(z) \leq y, \ \forall y \in G^k(x; g)\}. \tag{5.3}$$

Then $\rho^k(f, g)$ is defined by

$$\rho^k(f, g) = \sup \left\{ h^k \left(\sum_{j=1}^{k} \alpha_j x^j; f, g \right) - \sum_{j=1}^{k} \alpha_j f(x^j) \right\}.$$

over all possible convex combinations $\alpha_j \geq 0$, $j = 1, \ldots, k$, with $\sum_{j=1}^{k} \alpha_j = 1$ of points x^j satisfying $f(x^j) < +\infty$. If the function g is convex, then in the infimum of (5.3) we can choose $z = x$, which gives $h^k(x; f, g) \leq f(x)$ and $\rho^k(f, g) \leq \rho^k(f)$. However, the last inequality may not hold if g is not convex.

Theorem 3.3 can be modified accordingly to the case with nonlinear constraints by replacing $\rho^{k_i}(f_i)$ with $\rho^{k_i}(f_i, g_i)$. In the case when all g_i are convex, $\rho^{k_i}(f_i, g_i) \leq \rho^{k_i}(f_i)$, which implies that the original conclusion in Theorem 3.3 remains true. However, in general, considering g_i explicitly in Theorem 5.1 has the potential to provide tighter bound even for convex constraints including the linear cases.

Theorem 5.1. *Assume that the primal problem* (5.1) *is feasible, i.e.,* $p < +\infty$. *Then there exist integers* $1 \leq k_i \leq m+1$ *such that* $\sum_{i=1}^{n} k_i \leq m+n$ *and the duality gap*

$$p - d \leq \sum_{i=1}^{n} \rho_i^{k_i}.$$

Here $\rho_i^k = \rho^k(f_i, g_i)$ *is the* k*th nonconvexity of function* f_i *with respect to function* g_i.

Proof. Like the proof of Theorem 3.3, we define the perturbation function $v : \mathbb{R}^m \to [-\infty, +\infty]$ by letting $v(z)$ be the optimal value of the perturbed problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} f_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{n} g_i(x^i) \leq b + z. \end{aligned}$$

By the same argument as in the proof of Theorem 3.3, $-\infty < v^{**}(0) = d \leq v(0) = p < +\infty$, and for $\epsilon > 0$ there exists $(\kappa, \eta) \in \operatorname{conv} \operatorname{epi} v$ which is sufficiently close to $(0, v^{**}(0))$ such that

$$v(\kappa) \geq v(0) - \epsilon, \quad \eta \leq v^{**}(0) + \epsilon.$$

Proceed exactly the same as the proof of Theorem 3.3. We decompose κ into z^j such that

$$\kappa = \sum_{j=1}^{l} \alpha_j z^j, \quad \eta \geq \sum_{j=1}^{l} \alpha_j v(z^j),$$

and introduce $w^j \in \mathbb{R}^m_+$ with

$$\left(b + \kappa - \sum_{j=1}^{l} \alpha_j w^j, \sum_{j=1}^{l} \alpha_j v(z^j) \right) \in \operatorname{conv} \sum_{i=1}^{n} C_i,$$

where C_i is defined by

$$C_i = \{(g_i(x^i), f_i(x^i)) | f_i(x^i) < +\infty, g_i(x^i) < +\infty, x^i \in \mathbb{R}^{n_i}\}.$$

Corollary 3.1 gives points $(r^i, s_i) \in \operatorname{conv}_{k_i} C_i$ with $1 \leq k_i \leq m + 1$ such that

$$b + \kappa \geq b + \kappa - \sum_{j=1}^{l} \alpha_j w^j = \sum_{i=1}^{n} r^i, \quad \eta \geq \sum_{j=1}^{l} \alpha_j v(z^j) \geq \sum_{i=1}^{n} s_i$$

and $\sum_{i=1}^{n} k_i \leq m + n$. Since $(r^i, s_i) \in \operatorname{conv}_{k_i} C_i$, there exists $\tilde{x}^{ij} \in \mathbb{R}^{n_i}$, $\beta_{ij} \geq 0$ for $j = 1, \ldots, k_i$ such that $f_i(\tilde{x}^{ij}) < +\infty$, $g_i(\tilde{x}^{ij}) < +\infty$, $\sum_{j=1}^{k_i} \beta_{ij} = 1$ and

$$r^i = \sum_{j=1}^{k_i} \beta_{ij} g_i(\tilde{x}^{ij}), \quad s_i = \sum_{j=1}^{k_i} \beta_{ij} f_i(\tilde{x}^{ij}).$$

For each $i = 1, \ldots, n$, define $\hat{x}^i = \sum_{j=1}^{k_i} \beta_{ij} \tilde{x}^{ij}$. If $h^{k_i}(\hat{x}^i; f_i, g_i) = +\infty$, we also have $\rho_i^{k_i} = +\infty$ and the theorem is trivial in this case. Otherwise, observe that

$$\sum_{j=1}^{k_i} \beta_{ij} g_i(\tilde{x}^{ij}) \in G^{k_i}(\hat{x}^i; g_i),$$

because $\tilde{x}^{ij}$ and β_{ij} satisfy all the constraints given in (5.2). As a result, there will be $\hat{q}^i \in \mathbb{R}^{n_i}$ such that

$$g_i(\hat{q}^i) \leq \sum_{j=1}^{k_i} \beta_{ij} g_i(\tilde{x}^{ij}),$$

$$f_i(\hat{q}^i) \leq h^{k_i}(\hat{x}^i; f_i, g_i) + \epsilon.$$

Thus,

$$\kappa \geq \sum_{i=1}^{n} \sum_{j=1}^{k_i} \beta_{ij} g_i(\tilde{x}^{ij}) - b \geq \sum_{i=1}^{n} g_i(\hat{q}^i) - b,$$

and

$$\sum_{i=1}^{n} \rho_i^{k_i} + \eta \geq \sum_{i=1}^{n} \left(\rho_i^{k_i} + \sum_{j=1}^{k_i} \beta_{ij} f_i(\tilde{x}^{ij}) \right) \geq \sum_{i=1}^{n} h_i^{k_i}(\hat{x}^i; f_i, g_i) \geq \sum_{i=1}^{n} f_i(\hat{q}^i) - n\epsilon.$$

Now $(\hat{q}^1, \ldots, \hat{q}^n)$ is a feasible solution to the perturbed problem $v(\kappa)$. Following the original proof, we have

$$v^{**}(0) + \epsilon + \sum_{i=1}^{n} \rho_i^{k_i} \geq v(0) - (n+1)\epsilon$$

and then finish the proof by letting $\epsilon \to 0$. □

5.2 Application to Routing with Granularity Constraints

The network utility maximization problem with multipath routing is a convex optimization problem that has been well addressed, and in contrast Section 4.1 studied the nonconvex problem with single-path routing. The single-path routing and multipath routing can be regarded as two extreme cases of a spectrum of

routing schemes of different flexibility, and there are different ways to interpret this flexibility. One obvious way is based on the number of paths (W) that are allowed for each source-destination pair. Single-path routing corresponds to the case of $W = 1$, and multipath routing is the limit case of $W \to \infty$. For any integer $W > 1$, it is expected that the routing solution would perform somewhere between single-path routing and general multipath routing. Routing optimization with these path cardinality constraints will be examined in Appendix A.

Here we introduce a different angle to generalize single-path and multipath routing by considering routing problems with certain traffic splitting restrictions. More specifically, each user is only allowed to choose its traffic split ratio as a multiple of $1/p$, where the integer p is the split granularity parameter.[1] Clearly, the case of $p = 1$ is equivalent to single-path routing, while $p \to \infty$ corresponds to multipath routing. Routing optimizations with split ratio granularity constraints are studied using heuristic methods in previous works such as [29, 30]. In this section, we will use the tools developed in Section 5.1 to estimate the duality gap for the nonconvex network utility maximization problem with traffic splitting granularity constraints. The relationship among the different types of constraints mentioned here is summarized in Figure 5.1.

5.2.1 The Traffic Splitting Granularity Constraints

As a general way to deal with granularity constraints and many other practical restrictions on the sending rates, we modified the network utility maximization

[1] There are other possible split ratio granularity constraints. For example, in [29] the constraint is that the split ratio is a multiple of $1/q$, where q is bounded by some integer Q.

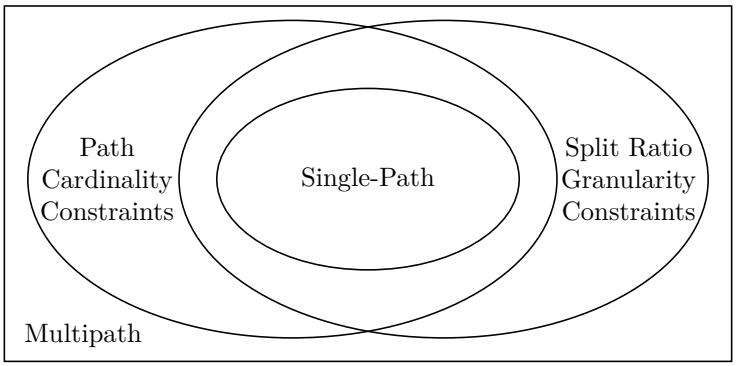


Figure 5.1: The relationship among the different types of constraints that can be imposed in the routing optimization problem. Note that the single-path constraint is as a special case of both path cardinality constraints and split ratio granularity constraints.

problem (4.3) by introducing the rate constraint set $\mathcal{S}_K$:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} f_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{N} R^i x^i \leq c, \\ & x^i \in \mathcal{S}_{K^i}, \quad \forall i = 1, \ldots, N, \end{aligned} \tag{5.4}$$

where

$$f_i(x^i) = \begin{cases} -U_i(x^i), & \text{if } 0 \leq x^i \leq \|c\|_\infty, \\ +\infty, & \text{otherwise.} \end{cases}$$

Furthermore, the problem (5.4) can be converted into

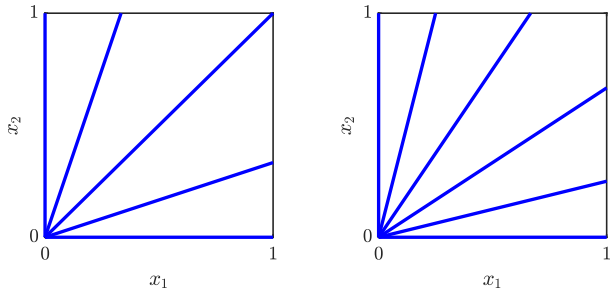
$$\begin{aligned} \min \quad & \sum_{i=1}^{N} f_i(x^i) \\ \text{s. t.} \quad & \sum_{i=1}^{N} g_i(x^i) \leq c. \end{aligned} \tag{5.5}$$

as a special case of problem (5.1) by defining the functions

$$g_i(x^i) = \begin{cases} R^i x^i, & \text{if } x \in \mathcal{S}_{K^i}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Without loss of generality, we assume that $\|c\|_\infty = 1$ and

$$\mathcal{S}_K \subseteq \mathcal{I}_K = \{x \in \mathbb{R}^K | 0 \leq x \leq 1\}.$$



(a) $p = 4$ (b) $p = 5$

Figure 5.2: The set $\mathcal{S}_2$ (shown as the bold lines above) under the cases $p = 4$ and $p = 5$.

In the case of split ratio granularity constraints, the split ratio needs to be a multiple of $1/p$, where p is a given positive integer. Hence we can choose

$$\mathcal{S}_K = \{0\} \cup \left\{ x \in \mathcal{I}_K \middle| x \neq 0, \frac{px_k}{\|x\|_1} \in \mathbb{Z}, k = 1, \ldots, K \right\}. \tag{5.6}$$

For convenience, usually we will work with an equivalent definition for the set $\mathcal{S}_K$: A rate vector $x \in \mathcal{I}_K$ satisfies the split ratio granularity constraints, i.e., $x \in \mathcal{S}_K$, if and only if there exists a vector $\alpha \in \mathbb{R}^K$ satisfying:

1. $\alpha_k \geq 0$, $\alpha_k \in \mathbb{Z}$, for $k = 1, \ldots, K$.
2. $\sum_{k=1}^{K} \alpha_k = p$.
3. There exists some real number $\lambda \geq 0$ such that $x = \lambda\alpha$.

The vector α will be known as the (scaled) split ratio of x.

Figure 5.2 illustrates the set $\mathcal{S}_K$ for $K = 2$ and different choices of p. In general, the set $\mathcal{S}_K$ is the union of rays each of which passes through a point α satisfying the first and second conditions above, and $\mathcal{S}_K$ contains the corner $(1, 1, \ldots, 1)$ of $\mathcal{I}_K$ if and only if p is a multiple of K.

5.2.2 The Duality Gap

Now we are going to estimate the duality gap for the problem (5.5) using Theorem 5.1. The key step is to compute the nonconvexity $\rho^k(f_i, g_i)$ for functions f_i and g_i, which further requires us to estimate the function $h^k(x^i; f_i, g_i)$.

First we consider the case of the linear utility function $U_i(x^i) = \|x^i\|_1$. Note that in this problem all the matrices R^i are nonnegative. Therefore, for given $x^i \in \mathcal{I}_{K^i}$, if we choose a rate vector $z \in \mathbb{R}^{K^i}$ with the property that $z \leq x^i$ and $z \in \mathcal{S}_{K^i}$, then the constraint in the infimum of (5.3) will be automatically satisfied. The optimal choice of such rate vector z should be the one with the least throughput loss. This illuminates us to consider the following optimization problem that finds the optimal rounding $z \in \mathcal{S}_K$ for a rate vector $x \in \mathcal{I}_K$ such that the throughput loss is minimized:

$$\begin{aligned} \min \quad & \|x\|_1 - \|z\|_1 \\ \text{s.t.} \quad & z \leq x, \\ & z \in \mathcal{S}_K. \end{aligned} \tag{5.7}$$

For each rate vector $x \in \mathcal{I}_K$, define the *throughput loss function* $\phi_K(x)$ to be the optimal value of problem (5.7). By the above analysis,

$$h^k(x^i; f_i, g_i) \leq f_i(x^i) + \phi_{K^i}(x^i).$$

Define

$$\rho_K = \max_{x \in \mathcal{I}_K} \phi_K(x)$$

to be the maximum throughput loss of a user with K available paths. As a con-

vention, $\rho_0 = 0$. Then

$$\begin{aligned}\rho^k(f_i, g_i) &\leq \sup\left\{ f_i\left(\sum_{j=1}^{k} \alpha_j x^{ij}\right) - \sum_{j=1}^{k} \alpha_j f_i(x^{ij}) + \phi_{K^i}\left(\sum_{j=1}^{k} \alpha_j x^{ij}\right)\right\} \\ &\leq \rho_{K^i},\end{aligned}$$

since our objective function f_i is convex. Assume the users are sorted in a way such that

$$\rho_{K^1} \geq \rho_{K^2} \geq \cdots \geq \rho_{K^N}.$$

Then the value of the optimization problem given in Theorem 5.1 and thus the duality gap is bounded by

$$\sum_{i=1}^{\min\{N,L\}} \rho_{K^i}. \tag{5.8}$$

The throughput loss function $\phi_K(x)$ will play an important role in our analysis as the bound (5.8) for the duality gap directly depends on its maximum value ρ_K. In the next, we will give an explicit formula for the function $\phi_K(x)$. Using the equivalent definition for the set $\mathcal{S}_K$ given in Section 5.2.1, i.e., write $y = \lambda\alpha$ where α is the split ratio of y, (5.7) can be rewritten as a problem over variables λ and α:

$$\begin{aligned} \min \quad & \|x\|_1 - p\lambda \\ \text{s.t.} \quad & \lambda\alpha_k \leq x_k, \quad \forall k = 1, \ldots, K, \\ & \lambda \geq 0,\ \sum_{k=1}^{K} \alpha_k = p, \\ & \alpha_k \geq 0,\ \alpha_k \in \mathbb{Z}, \quad \forall k = 1, \ldots, K. \end{aligned} \tag{5.9}$$

Fix the split ratio α. To minimize the throughput loss, the largest possible λ should be to chosen, which is given by

$$\lambda^* = \min_{k=1,\ldots,K} \frac{x_k}{\alpha_k}.$$

Here we make the convention that $x_k/\alpha_k = +\infty$ when $\alpha_k = 0$. Plugin the formula for λ^* into (5.9), the optimal value

$$\phi_K(x) = \|x\|_1 - p \max_{\alpha} \min_{k=1,\ldots,K} \frac{x_k}{\alpha_k}, \tag{5.10}$$

where the maximization is taken over all the vectors $\alpha \geq 0$ with $\sum_{k=1}^{K} \alpha_k = p$ and $\alpha_k \in \mathbb{Z}$ for $k = 1, \ldots, K$. The following are some simple properties of $\phi_K(x)$:

1. $\phi_K(x)$ is continuous on the domain $\mathcal{I}_K$.
2. If $x \in \mathcal{I}_K$ and $\lambda x \in \mathcal{I}_K$, $\phi_K(\lambda x) = \lambda \phi_K(x)$.

The throughput loss function $\phi_2(x)$ for the case $p = 4$ is illustrated in Figure 5.3(a), and the actual optimal rounding for a rate vector in this case is described in Figure 5.3(b). The local maximizers for $\phi_2(x)$ are

$$(1/4, 1), (1, 1/4), (2/3, 1), (1, 2/3).$$

The later two are global maximizers and $\rho_2 = 1/3$. The situations in higher dimensions are similar but the number of local maximizers is much larger (see Figure 5.3(c)).

The case for the logarithmic utility function $U_i(x^i) = \log\|x^i\|_1$ is similar. Here if the original rate vector of user i is x^i and the rounded rate vector is z, then

$$f_i(x^i) - f_i(z) = -\log \frac{\|x^i\|_1}{\|z\|_1} = \log\left(1 - \frac{\|x^i\|_1 - \|z\|_1}{\|x^i\|_1}\right).$$

Accordingly, consider the maximum relative throughput loss

$$\rho_K^R = \max_{x \in \mathcal{I}_K, x \neq 0} \frac{\phi_K(x)}{\|x\|_1}.$$

Now the duality gap is bounded by

$$-\sum_{i=1}^{\min\{N,L\}} \log(1 - \rho_{K^i}^R) \tag{5.11}$$

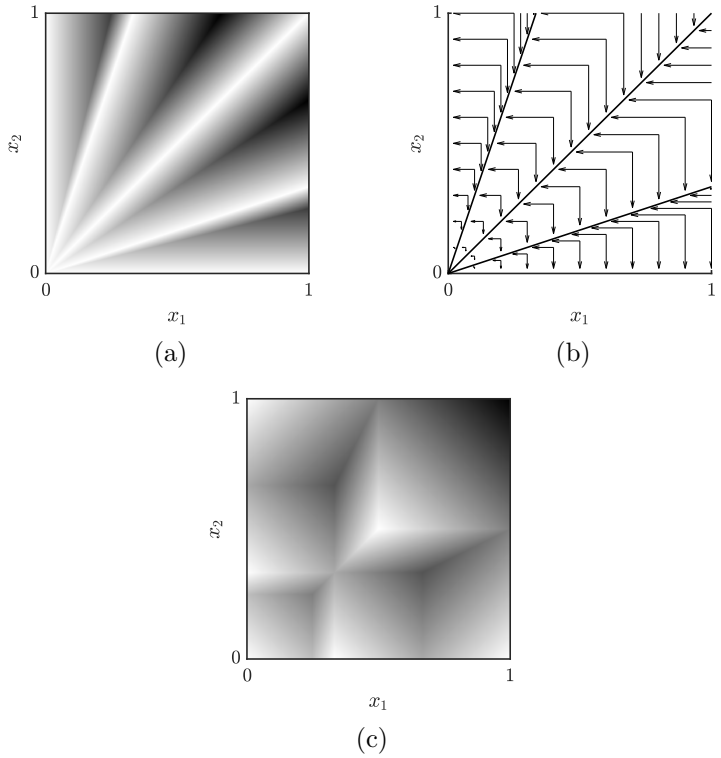


Figure 5.3: (a) The throughput loss function $\phi_2(x)$ for the case $p = 4$. The function value is close to zero (white area) on points near the set $\mathcal{S}_2$ and becomes large (dark area) on points far away from the set $\mathcal{S}_2$. (b) The optimal rounding for a rate vector in $\mathcal{I}_2$. To find the optimal rounding, start with that vector and move according to the arrow until a point in $\mathcal{S}_2$ (represented by the solid lines) is reached. (c) The value of function $\phi_3(x)$ on the plane $x_3 = 1$. Darker area indicates larger function value.

if the users are sorted by

$$\rho^R_{K^1} \geq \rho^R_{K^2} \geq \cdots \geq \rho^R_{K^N}.$$

5.2.3 Maximum Relative Throughput Loss for Logarithmic Utility

The bounds (5.8) and (5.11) for the duality gap in the linear and logarithmic utility case depend on ρ_K and ρ^R_K, respectively. The remaining task is to compute these

values. We start with the easier case of logarithmic utility, which requires us to calculate the maximum relative throughput loss ρ_K^R.

Theorem 5.2. *For every $x \in \mathcal{I}_K$ with $x \neq 0$, the relative throughput loss*

$$\frac{\phi_K(x)}{\|x\|_1} \leq \frac{K-1}{p+K-1}.$$

Proof. Since the function $\phi_K(x)/\|x\|_1$ is continuous, we only need to prove this inequality for every x in a dense subset of $\mathcal{I}_K$. Let $\mathcal{I}'_K$ be the set containing all nonzero $x \in \mathcal{I}_K$ with the following property: For any pair of paths $1 \leq k, l \leq K$, $k \neq l$ and any pair of integers $1 \leq a, b \leq p$, $x_k/a \neq x_l/b$. $\mathcal{I}'_K$ is a dense subset of $\mathcal{I}_K$ since it is a finite intersection of open dense subsets, so it is sufficient to prove the inequality on $\mathcal{I}'_K$.

For any $x \in \mathcal{I}'_K$, let α be the split ratio of the optimal rounding for x, i.e., α is the one attaining the maximum in (5.10). Without loss of generality, we also assume that

$$\min_{k=1,\ldots,K} \frac{x_k}{\alpha_k} = \frac{x_1}{\alpha_1}.$$

In this case, $\alpha_1 > 0$, which implies that $x_1 > 0$ and $\alpha_l < p$ for all other $l = 2, \ldots, K$. For these l, we are going to prove by contradiction that

$$\frac{x_1}{\alpha_1} \geq \frac{x_l}{\alpha_l + 1}. \tag{5.12}$$

If the above inequality does not hold for a particular path l, define a new split ratio β by setting

$$\beta_k = \begin{cases} \alpha_1 - 1, & \text{if } k = 1, \\ \alpha_l + 1, & \text{if } k = l, \\ \alpha_k, & \text{otherwise.} \end{cases}$$

By our assumption that x_1/α_1 is the minimum,

$$\frac{x_k}{\beta_k} = \frac{x_k}{\alpha_k} \geq \frac{x_1}{\alpha_1},\ k \neq 1, l.$$

Furthermore, the above inequality is strict because $x \in \mathcal{I}'_K$. In addition,

$$\frac{x_1}{\beta_1} = \frac{x_1}{\alpha_1 - 1} > \frac{x_1}{\alpha_1}, \quad \frac{x_l}{\beta_l} = \frac{x_l}{\alpha_l + 1} > \frac{x_1}{\alpha_1}.$$

Therefore,

$$\min_{k=1,\ldots,K} \frac{x_k}{\beta_k} > \frac{x_1}{\alpha_1} = \min_{k=1,\ldots,K} \frac{x_k}{\alpha_k},$$

which contradicts with the fact that α is the optimal split ratio for the rounding of x.

Combining all the inequalities (5.12) for $l = 2, \ldots, K$, we have

$$\frac{x_1}{\alpha_1} \geq \frac{x_1 + \sum_{l=2}^{K} x_l}{\alpha_1 + \sum_{l=2}^{K}(\alpha_l + 1)} = \frac{\|x\|_1}{p + K - 1}$$

and

$$\begin{aligned}\frac{\phi_K(x)}{\|x\|_1} &= \frac{1}{\|x\|_1}\left(\|x\|_1 - p \min_{k=1,\ldots,K} \frac{x_k}{\alpha_k}\right) = 1 - \frac{p x_1}{\|x\|_1 \alpha_1} \\ &\leq 1 - \frac{p}{p + K - 1} = \frac{K - 1}{p + K - 1},\end{aligned}$$

which is the desired inequality. □

Theorem 5.2 shows that

$$\rho_K^R \leq \frac{K - 1}{p + K - 1}.$$

In the next subsection, we will see that this bound is in fact tight as an easy corollary of Lemma 5.1 and Lemma 5.2. In other words,

$$\rho_K^R = \frac{K - 1}{p + K - 1}.$$

5.2.4 Maximum Throughput Loss for Linear Utility

Now we calculate the maximum throughput loss ρ_K for the linear utility. To find the maximum value ρ_K for the nonconcave function $\phi_K(x)$, we first establish a type of first-order condition that all local maximizers of $\phi_K(x)$ will satisfy. Despite the fact that the number of local maximizers can be exponentially large, we are able to classify the local maximizers into a few categories, and in each category we can find the one with the maximum throughput loss. Roughly speaking, the following definition provides the necessary condition for a rate vector to be a local maximizer:

Definition 5.1. A rate vector $x \in \mathcal{I}_K$ is said to be *integral* if there exists a vector $\beta \in \mathbb{R}^K$ satisfying:

1. $\beta_k \geq 0$, $\beta_k \in \mathbb{Z}$, for $k = 1, \ldots, K$.
2. $p \leq \sum_{k=1}^{K} \beta_k \leq p + \|x\|_0 - 1$. Here $\|x\|_0$ is the number of nonzero components in x.
3. There exists some real number $\lambda \geq 0$ such that $x = \lambda\beta$.

To be specific, we will also use the terminology Γ-*integral* for integral rate vectors with $\sum_{k=1}^{K} \beta_k = \Gamma$.

Clearly, a rate vector $x \in \mathcal{I}_K$ is p-integral if and only if $x \in \mathcal{S}_K$ and $x \neq 0$. In Figure 5.3(a), the local maximizer $(1/4, 1)$ of $\phi_2(x)$ is integral and the corresponding vector $\beta = (1, 4)$. The global maximizer $(2/3, 1)$ is also integral for $\beta = (2, 3)$.

Here are the major steps to compute ρ_K:

1. Prove that any maximizer of the function $\phi_K(x)$ must be integral. Hence we only need to figure out the maximum throughput loss over all integral rate vectors.
2. Find the relative throughput loss for integral rate vectors. We will see that all rate vectors that are Γ-integral have the same relative throughput loss, which is $1 - p/\Gamma$.
3. Show that the maximum throughput among all Γ-integral rate vectors is given by
$$\frac{\Gamma}{\lceil \Gamma/K \rceil}.$$

Finally, we can conclude that

$$\begin{aligned}\rho_K &= \max_{\Gamma=p,\ldots,p+K-1} \left(1 - \frac{p}{\Gamma}\right) \frac{\Gamma}{\lceil \Gamma/K \rceil} \\ &= \max_{\Gamma=p,\ldots,p+K-1} \frac{\Gamma - p}{\lceil \Gamma/K \rceil}.\end{aligned}$$

In the next, we will first address the last two steps in Lemma 5.1 and Lemma 5.2. The first step needs more elaboration and will be dealt with thereafter.

Lemma 5.1. *If a rate vector $x \in \mathcal{I}_K$ is Γ-integral, then its relative throughput loss*

$$\frac{\phi_K(x)}{\|x\|_1} = 1 - \frac{p}{\Gamma}.$$

Proof. Let $x = \lambda\beta$ as required by Definition 5.1. Since $\Gamma = \sum_{k=1}^K \beta_k \geq p$, we can find a split ratio vector α such that

$$0 \leq \alpha_k \leq \beta_k,\ \alpha_k \in \mathbb{Z},\ k = 1, \ldots, K,$$

and $\sum_{k=1}^K \alpha_k = p$. Obviously, $\lambda\alpha$ is a feasible solution to the optimal rounding problem (5.7) for x, so

$$\frac{\phi_K(x)}{\|x\|_1} \leq 1 - \frac{\sum_{k=1}^K \lambda\alpha_k}{\sum_{k=1}^K \lambda\beta_k} = 1 - \frac{p}{\Gamma}.$$

On the other hand, now assume that another split ratio α gives the optimal rounding for x. Suppose $\alpha_k < \beta_k$ for all $k = 1, \ldots, K$ with $x_k > 0$. Since α_k and β_k are integers, $\alpha_k + 1 \leq \beta_k$ for these k. Noticing that $x_k > 0$ if and only if $\beta_k > 0$, and $\alpha_k > 0$ implies $x_k > 0$, we have

$$p + \|x\|_0 = \sum_{k:x_k>0} (\alpha_k + 1) \leq \sum_{k:x_k>0} \beta_k,$$

which contradicts with the requirement on β. Thus there must exist some path l such that $\alpha_l \geq \beta_l > 0$, which implies

$$\min_{k=1,\ldots,K} \frac{x_k}{\alpha_k} \leq \frac{x_l}{\alpha_l} \leq \frac{x_l}{\beta_l} = \lambda$$

and

$$\begin{aligned} \frac{\phi_K(x)}{\|x\|_1} &= \frac{1}{\|x\|_1} \left(\|x\|_1 - p \min_{k=1,\ldots,K} \frac{x_k}{\alpha_k} \right) \\ &\geq 1 - \frac{p\lambda}{\sum_{k=1}^K \lambda\beta_k} = 1 - \frac{p}{\Gamma}. \end{aligned}$$

The desired result is followed from combining the two directions we have already shown. □

Lemma 5.2. *For integer Γ with $p \leq \Gamma \leq p + K - 1$, there exists a rate vector in $\mathcal{I}_K$ that is Γ-integral, and the maximum throughput among all Γ-integral rate vectors is*

$$\frac{\Gamma}{\lceil \Gamma/K \rceil}.$$

Proof. First, we construct a Γ-integral rate vector with the given throughput. Note that we can choose integers $\beta_k \in \{\lfloor \Gamma/K \rfloor, \lceil \Gamma/K \rceil\}$ for $k = 1, \ldots, K$ such that $\sum_{k=1}^K \beta_k = \Gamma$. Let

$$\lambda = \frac{1}{\lceil \Gamma/K \rceil}$$

and $x = \lambda\beta$. Then $x_k \leq 1$ for all k, so $x \in \mathcal{I}_K$. If $\Gamma \geq K$, all $x_k > 0$ and thus

$$\Gamma \leq p + K - 1 = p + \|x\|_0 - 1.$$

If $\Gamma < K$, then $\|x\|_0 = \Gamma$ and we also have $\Gamma \leq p + \|x\|_0 - 1$. Therefore, x is Γ-integral and its throughput

$$\|x\|_1 = \lambda\Gamma = \frac{\Gamma}{\lceil \Gamma/K \rceil}.$$

Consider an arbitrary rate vector $x \in \mathcal{I}_K$ that is Γ-integral with corresponding vector β. Because $\sum_{k=1}^{K} \beta_k = \Gamma$, there must exist some path l such that $\beta_l \geq \Gamma/K$. Since β_l is an integer, $\beta_l \geq \lceil \Gamma/K \rceil > 0$, which implies

$$\|x\|_1 = \lambda\Gamma = \frac{x_l}{\beta_l}\Gamma \leq \frac{x_l}{\lceil \Gamma/K \rceil}\Gamma \leq \frac{\Gamma}{\lceil \Gamma/K \rceil},$$

showing that the given throughput is indeed maximal. □

The remaining task is to prove that any maximizer of the function $\phi_K(x)$ is integral. However, there is a technical difficulty which can be demonstrated by the following counterexample: Assume $x \in \mathcal{I}_3$ is a rate vector and $x_1 \geq x_2 \geq x_3$. In the case of $p = 2$, since the vector $(x_2, x_2, 0)$ satisfies the split ratio granularity constraints,

$$\phi_3(x) \leq x_1 - x_2 + x_3 \leq x_1 \leq 1$$

and thus $\rho_3 \leq 1$. On the other hand, we can directly check that for all rate vectors y of the form $(1, t, t)$ where $1/2 \leq t \leq 1$, $\phi_3(y) = 1$. Thus the function $\phi_3(x)$ has infinite maximizers and almost all of them are not integral. Fortunately, our goal is not finding all the maximizers of $\phi_K(x)$ but computing the maximum of $\phi_K(x)$, and the plan presented at the beginning of this section still works as long as one of the maximizers is integral. The integrality of the maximizer $(1, 1, 1)$ dominating all other maximizers enlightens the additional requirement in Lemma 5.3.

Lemma 5.3. *If a rate vector $x \in \mathcal{I}_K$ satisfies:*

1. *x maximizes the function $\phi_K(x)$;*

2. For any $z \in \mathcal{I}_K$, if $\phi_K(z) = \phi_K(x)$ and $z \geq x$, then $z = x$,

then x is integral.

Proof. Repeating what we have done in the proof of Theorem 5.2, assume $x \neq 0$, α is the split ratio of the optimal rounding for x, and

$$\min_{k=1,\dots,K} \frac{x_k}{\alpha_k} = \frac{x_1}{\alpha_1}.$$

Define W to be the set of indices k such that there exists some integer $\beta_k > 0$ with $x_k/\beta_k = x_1/\alpha_1$. Then for any $l \notin W$ and any split ratio α' that also gives the optimal rounding for x, i.e.,

$$\min_{k=1,\dots,K} \frac{x_k}{\alpha'_k} = \min_{k=1,\dots,K} \frac{x_k}{\alpha_k} = \frac{x_1}{\alpha_1},$$

it must be true that

$$\frac{x_1}{\alpha_1} = \min_{k=1,\dots,K} \frac{x_k}{\alpha'_k} < \frac{x_l}{\alpha'_l}. \tag{5.13}$$

We want to prove that W contains all the paths. Assume the existence of a path $l \notin W$. If $x_l < 1$, consider another rate vector $z = (x_1, \dots, x_l + \delta, \dots, x_K)$ with sufficiently small $\delta > 0$. The optimal split ratio for the rounding of z must be one of the above α', and by (5.13) $\min x_k/\alpha'_k$ is not attained by path l, so

$$\min_{k=1,\dots,K} \frac{z_k}{\alpha'_k} = \min_{k=1,\dots,K} \frac{x_k}{\alpha'_k} = \frac{x_1}{\alpha_1}.$$

The throughput loss for the rate vector z is

$$\phi_K(z) = \|x\|_1 + \delta - \frac{px_1}{\alpha_1} > \phi_K(x),$$

which contradicts with the assumption that x is the maximizer.

In the case of $x_l = 1$, we have $x_k < 1$ for $k \in W$, otherwise l would belong to W. Construct a rate vector z defined by

$$z_k = \begin{cases} \delta x_k, & \text{if } k \in W, \\ x_k, & \text{if } k \notin W, \end{cases}$$

for some δ that is sufficiently close to 1. Similar to the previous case, the optimal split ratio for the rounding of z must be one of the α' giving the optimal rounding for x. By (5.13), the path minimizing x_k/α'_k is in W, which also minimizes z_k/α'_k because the difference between z and x is sufficiently small. Thus

$$\min_{k=1,\ldots,K} \frac{z_k}{\alpha'_k} = \delta \min_{k=1,\ldots,K} \frac{x_k}{\alpha'_k} = \frac{\delta x_1}{\alpha_1}.$$

The throughput loss for the rate vector z is

$$\phi_K(z) = \|x\|_1 + (\delta - 1) \sum_{k \in W} x_k - \frac{\delta x_1}{\alpha_1},$$

and

$$\phi_K(z) - \phi_K(x) = (\delta - 1) \left(\sum_{k \in W} x_k - \frac{x_1}{\alpha_1} \right).$$

By choosing appropriate δ, we can let either $\phi_K(z) > \phi_K(x)$ or $\phi_K(z) = \phi_K(x)$ but $z_k \geq x_k$ for all $k = 1, \ldots, K$ with the inequality being strict for $k = 1 \in W$, which contradicts with either of the two conditions on x.

In the above, we have proved that the set W contains all the paths. For the integers β_k we have found, $x_k/\beta_k = x_1/\alpha_1$ so $x = \lambda\beta$ for $\lambda = x_1/\alpha_1$ and $\|x\|_0 = K$. It remains to prove that $p \leq \sum_{k=1}^K \beta_k \leq p + K - 1$. First, since

$$\frac{x_k}{\beta_k} = \frac{x_1}{\alpha_1} \leq \frac{x_k}{\alpha_k},\ k = 1, \ldots, K,$$

we have $\beta_k \geq \alpha_k$ and $\sum_{k=1}^K \beta_k \geq \sum_{k=1}^K \alpha_k = p$. Moreover, the relative throughput loss of x is

$$\frac{\phi_K(x)}{\|x\|_1} = 1 - \frac{px_1/\alpha_1}{\sum_{k=1}^K \lambda\beta_k} = 1 - \frac{p}{\sum_{k=1}^K \beta_k}.$$

Theorem 5.2 tells us

$$\frac{\phi_K(x)}{\|x\|_1} \leq \frac{K-1}{p+K-1},$$

which implies that $\sum_{k=1}^{K} \beta_k \leq p+K-1$. □

Now we can prove the main result in this section:

Theorem 5.3. *The maximum throughput loss for a rate vector in $\mathcal{I}_K$ is*

$$\rho_K = \max_{\Gamma=p,\ldots,p+K-1} \frac{\Gamma-p}{\lceil \Gamma/K \rceil}. \tag{5.14}$$

Proof. Since $\phi_K(x)$ is continuous, the set

$$\{z \in \mathcal{I}_K | \phi_K(z) = \rho_K\}$$

is nonempty, closed and bounded. We can find a rate vector x maximizing the l_1 norm on the above set, which satisfies the two conditions in Lemma 5.3 and thus being integral. Now we can obtain the desired result by following the plan given at the beginning of this subsection. □

The formula (5.14) for ρ_K can be further simplified. Actually, only two possibilities of Γ have to be checked. In the range $\Gamma = p, \ldots, p+K-1$, $\lceil \Gamma/K \rceil$ can only increase by at most 1. Let $0 \leq \omega \leq K-1$ be the number that $(p+\omega)/K = \lceil p/K \rceil$, then the maximum in (5.14) is attained by either $\Gamma = p+\omega$ or $\Gamma = p+K-1$, so

$$\rho_K = \max\left\{ \frac{\omega}{\lceil p/K \rceil}, \frac{K-1}{\lceil (p+K-1)/K \rceil} \right\}. \tag{5.15}$$

Part II

Convex Relaxation with Semidefinite Programming

CHAPTER 6

PRELIMINARIES

Semidefinite program is a convex optimization problem that selects a positive semidefinite matrix to optimize a linear function subject to linear constraints. The utilization of semidefinite programming as a convex relaxation for nonconvex optimizations origins from the Goemans–Williamson algorithm for the maximum cut problem [31], which is still one of the most impressive results in this area. Even earlier, in a celebrated paper [32] of 1979, Lovász introduced the Lovász number, which is an upper bound for the Shannon capacity of graph and can also be understood as a semidefinite programming relaxation for the independence number of graph.

The technique of semidefinite programming is originally thought to be an ad hoc method for designing approximation algorithms. However, it turns out that many semidefinite programming relaxations, including the two relaxations mentioned above, can be deduced from a systematic approach called the sum-of-squares relaxation, which was originally proposed in [33]. In this framework, one usually starts from an integer programming formulation for some combinatorial problem that is hard to solve and converts it into a polynomial optimization. The obtained polynomial optimization can then be relaxed into a series of semidefinite programs called the sum-of-squares hierarchy, which provides a way of generating a series of potentially tighter relaxations with increasing size.

This chapter will give a brief introduction to the semidefinite programming and sum-of-squares hierarchy in both the primal and dual form. As surveyed in [34], the Goemans–Williamson algorithm for the maximum cut and the Lovász number for the Shannon capacity of graph are two of the most representative applications

of the semidefinite relaxation to combinatorial optimizations. It is natural to ask whether these two problems will benefit from the higher-degree sum-of-squares relaxations. In Chapter 7 and Chapter 8, we will study semidefinite-programming-based convex relaxations for the two problems and reach partially negative answer to the above question.

6.1 Semidefinite Programming

A semidefinite program is a convex optimization problem that is similar to linear programs but using symmetric matrices as decision variables. Its standard form is as following:

$$\begin{aligned} \min \quad & \operatorname{tr}(CX) \\ \text{s.t.} \quad & \operatorname{tr}(A_i X) = b_i, \quad \forall i = 1, \ldots, m, \\ & X \in \mathcal{P}_n. \end{aligned} \tag{6.1}$$

Here X is the decision variable, $\mathcal{P}_n$ is the cone of $n \times n$ positive semidefinite matrices, and the coefficients A_k and C are also $n \times n$ symmetric matrices.

As a convex optimization problem, we can write down the Lagrange dual problem of (6.1) by the approach in Chapter 2. Define the Lagrangian $L : \mathcal{P}_n \times \mathbb{R}^m \to \mathbb{R}$ by

$$L(X, y) = \operatorname{tr}(CX) - \sum_{i=1}^{m} y_i \operatorname{tr}(A_i X) + b^T y = \operatorname{tr}(DX) + b^T y,$$

where

$$D = C - \sum_{i=1}^{m} y_i A_i.$$

Then the original primal problem (6.1) can be rewritten as

$$\min_{X \in \mathcal{P}_n} \max_{y \in \mathbb{R}^m} L(x, y)$$

and the corresponding dual problem is

$$\max_{y\in\mathbb{R}^m}\min_{X\in\mathcal{P}_n} L(x,y) = \max_{y\in\mathbb{R}^m}\{b^T y + \min_{X\in\mathcal{P}_n}\operatorname{tr}(DX)\}.$$

Since a symmetric matrix P is positive semidefinite if and only if $\operatorname{tr}(PQ)\geq 0$ for any positive semidefinite matrix Q,

$$\min_{X\in\mathcal{P}_n}\operatorname{tr}(DX) = \begin{cases} 0, & \text{if } D\in\mathcal{P}_n, \\ -\infty, & \text{otherwise.} \end{cases}$$

Therefore, the dual problem can be equivalently written as

$$\begin{aligned} \max \quad & b^T y \\ \text{s.t.} \quad & C - \sum_{i=1}^{m} y_i A_i \in \mathcal{P}_n. \end{aligned}$$

Like general convex optimization problems,the weak duality always holds for semidefinite programs, but the strong duality may not. The Slater's condition that guarantees the strong duality for the semidefinite programming can be specialized as the existence of a symmetric matrix that is positive definite while satisfying all the linear constraints in (6.1).

6.2 Sum-of-squares Programming

The sum-of-squares programming is a systematic approach to relax polynomial optimization problems. In the following, we will give the minimal background for the sum-of-squares programming used in the remaining chapters. See [35] for a comprehensive introduction to the theory and application of the sum-of-squares programming.

The most basic form of a polynomial optimization problem is to decide whether a polynomial f is nonnegative or not, which is an NP-hard problem. Instead of

directly checking the nonnegativity of f, we check that whether f can be written as a sum of squares, i.e., whether there exist polynomials g_i such that

$$f(x) = \sum_i h_i^2(x).$$

This verification actually can be done by checking the feasibility of a semidefinite program:

Proposition 6.1. *A polynomial f of degree $2d$ is a sum of squares if and only if there exists a positive semidefinite matrix Q such that*

$$f(x) = z^T(x)Qz(x),$$

where $z(x)$ the vector including all monomials of degree up to d.

Proof. For one direction, if f is a sum of squares and

$$f(x) = \sum_i h_i^2(x),$$

we have

$$f(x) = \sum_i (a_i^T z(x))^2 = z^T(x) \left(\sum_i a_i a_i^T \right) z(x),$$

where a_i is the vector of coefficients of the polynomial h_i. Hence the positive semidefinite matrix Q can be chosen as $\sum_i a_i a_i^T$. For the other direction, assume $f(x) = z^T(x)Qz(x)$. Then we can decompose $Q = V^T V$, and

$$f(x) = z^T(x)V^T V z(x) = \|Vz(x)\|^2,$$

which means that f is a sum of squares. □

Now we consider polynomial optimizations maximizing a polynomial objective subject to equality polynomial constraints, i.e.,

$$\begin{aligned} \max \quad & f(x) \\ \text{s.t.} \quad & g_i(x) = 0, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{6.2}$$

where the decision variable $x \in \mathbb{R}^n$, f and g_i are all polynomials. In fact, the sum-of-squares relaxation can deal with more general polynomial optimization problems with both equality and inequality polynomial constraints, but that is not necessary for the purpose of this dissertation.

The following is a generalization of the certification through sum-of-squares decomposition for the negativity of a polynomial over the algebraic set

$$K = \{x \in \mathbb{R}^n | g_i(x) = 0,\ i = 1, \ldots, m\} \tag{6.3}$$

defined by the constraints in (6.2).

Definition 6.1. The set of polynomials q_i for $i = 1, \ldots, m$ is called a sum-of-squares proof for the nonnegativity of f over K if

$$f + \sum_{i=1}^{m} q_i g_i$$

is a sum of squares. The proof is called degree at most $2d$ if the degree of each summand in the above is at most $2d$.

The degree-$2d$ sum-of-squares relaxation for the polynomial optimization problem (6.2) is to find the minimum γ with the existence of the above degree-$2d$ sum-of-squares proof for the polynomial $\gamma - f$, which can be written as

$$\begin{aligned} \min \quad & \gamma \\ \text{s.t.} \quad & \gamma - f + \sum_{i=1}^{m} q_i g_i \text{ is a sum of squares.} \end{aligned} \tag{6.4}$$

Here $\gamma \in \mathbb{R}$ is a scalar variable and q_i is a polynomial variable whose degree satisfies

$$\deg q_i \leq 2d - \deg g_i.$$

Let opt be the optimal value of the original problem and sos_{2d} be the optimal value of degree-$2d$ sum-of-squares relaxation (6.4), then obviously

$$\text{sos}_2 \geq \text{sos}_4 \geq \cdots \geq \text{sos}_{2d} \geq \cdots \geq \text{opt}\,.$$

Using Proposition 6.1, the problem (6.4) can be rewritten as a semidefinite program with γ and the coefficients of polynomials q_i as decision variables.

6.3 Dual Formulation of the Sum-of-Squares Relaxation

Since the sum-of-squares relaxations are essentially semidefinite programs, as a special case of (6.1) we can consider their Lagrange dual problems. Skipping all the computational details, in this section we will directly give the resulting dual problem accompanied with some intuitive explanation. To simplify the presentation, the following concept of pseudo-expectation will be helpful:

Definition 6.2. A degree-$2d$ pseudo-expectation is a linear function $\tilde{\mathbb{E}}_\mu(\cdot)$ defined on the space of polynomials of degree at most $2d$ satisfying:

- $\tilde{\mathbb{E}}_\mu 1 = 1$,
- $\tilde{\mathbb{E}}_\mu h^2 \geq 0$ for all polynomials h of degree at most d.

Then the dual problem of (6.4) is

$$\begin{aligned} \max \quad & \tilde{\mathbb{E}}_\mu f \\ \text{s.t.} \quad & \tilde{\mathbb{E}}_\mu(g_i h) = 0, \quad \forall i = 1, \ldots, m, \\ & \qquad \forall \text{polynomial } h \text{ with } \deg h \leq 2d - \deg g_i, \\ & \tilde{\mathbb{E}}_\mu(\cdot) \text{ is a degree-}2d \text{ pseudo-expectation.} \end{aligned} \tag{6.5}$$

The definition of pseudo-expectation is inspired by the expectation of a polynomial of random variables. If μ is a probability distribution on $\mathbb{R}^n$ and x is a random vector with distribution μ, then

$$\mathbb{E}_\mu h^2(x) \geq 0$$

for all polynomials h. A pseudo-expectation is a linear function that behaves very similarly to a true expectation except that a degree-$2d$ pseudo-expectation only acts on polynomials of degree at most $2d$.

Assume the support of μ is in the set K defined by (6.3). Then for each constraint polynomial g_i and any polynomial h we additionally have

$$\mathbb{E}_\mu g_i(x)h(x) = 0$$

Similarly, this property is also inherited by the pseudo-expectation in the dual problem (6.5).

CHAPTER 7

MAXIMUM CUT

Consider an undirected graph $G = (V, E)$ with n vertices and weights $w_{ij} \geq 0$ for $(i, j) \in E$. The maximum cut problem is to find a cut $S \subseteq V$ maximizing the total weight of edges in the cut, which can be formulated as the following integer programming problem:

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} \frac{1}{2} w_{ij}(1 - x_i x_j) \\ \text{s.t.} \quad & x_i \in \{1, -1\}, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{7.1}$$

The above problem can be simplified by introducing the Laplacian $L = D - A$, where D is the diagonal matrix in which the ith diagonal element is the sum of weights of all edges connecting vertex i and A is the adjacency matrix of G. If we let $x = (x_1, \ldots, x_n)$, then (7.1) is equivalent to the following polynomial optimization problem:

$$\begin{aligned} \max \quad & \frac{1}{4} x^T L x \\ \text{s.t.} \quad & x_i^2 = 1, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{7.2}$$

As explained in Section 7.1, the semidefinite relaxation of Goemans–Williamson for the maximum cut problem has a 0.878 integrality gap. Whether there exists some polynomial-time solvable relaxation with strictly tighter integrality gap is a long-standing open question that is equivalent to the unique games conjecture [36]. Without the unique games conjecture, the best known result says that the integrality gap is at most 16/17 unless $P = NP$ [37].

To prove or refute the unique games conjecture, it is natural to look at the weaker question whether higher-degree relaxations from the sum-of-squares hierarchy have tighter integrality gap than the original semidefinite relaxation. After

introducing the general sum-of-squares relaxations for the maximum cut in Section 7.2 and a simplified version of the degree-4 sum-of-squares relaxation in Section 7.3, we will focus on the integrality gap for these sum-of-squares relaxations in Section 7.4. An instance of integrality gap 0.96 for the degree-4 sum-of-squares relaxation will be given first. Then we will construct instances as candidates for even looser integrality gap in Section 7.5 and Section 7.6.

7.1 Semidefinite Relaxation of Maximum Cut

In this section, we will first introduce the standard semidefinite relaxation for the maximum cut, which will be recovered in another way by the degree-2 sum-of-squares relaxation in Section 7.2. Next, we will present the Goemans-Williamson randomized rounding for the semidefinite relaxation [31], which provides a 0.878-approximation algorithm and at the same time proves that the integrality gap for the semidefinite relaxation is at least 0.878.

We start from the polynomial optimization problem (7.2). If we define an $n \times n$ matrix X by $X_{ij} = x_i x_j$, then X is positive semidefinite and the objective of problem (7.2) will equal to $\operatorname{tr}(LX)/4$. To relax the problem (7.2), we directly use the matrix X as the decision variable instead of all the x_i:

$$\begin{aligned} \max \quad & \frac{1}{4}\operatorname{tr}(LX) \\ \text{s.t.} \quad & X_{ii} = 1, \quad \forall i = 1, \ldots, n, \\ & X \in \mathcal{P}_n. \end{aligned} \tag{7.3}$$

To obtain a feasible solution to the original problem (7.2), the key step is to apply randomized rounding on the optimal solution to the relaxation (7.3). Let sdp be the optimal value of (7.3), opt be the value of the maximum cut and sol be

the value of the solution obtained from the randomized rounding. As discussed in Section 1.3, if we prove that

$$\text{sol} \geq \alpha_{GW} \, \text{sdp} \tag{7.4}$$

with $\alpha_{GW} \approx 0.878$, then we simultaneously establish the desired bound for both the approximation ratio and the integrality gap.

Now we will describe the randomized rounding procedure in details and prove the inequality (7.4). After solving the semidefinite relaxation (7.3) and obtaining the optimal solution X, we decompose $X = V^T V$ and let v_i be the ith column of the matrix V. Then we pick up a random vector $r \in \mathbb{R}^n$ subject to the uniform distribution on the unit sphere and choose the rounded solution as

$$x_i = \begin{cases} 1, & \text{if } r^T v_i \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

The expected value $\mathbb{E}[\text{sol}]$ of the obtained cut is

$$\begin{aligned} \mathbb{E}[\text{sol}] &= \sum_{(i,j)\in E} w_{ij} \Pr\{\text{sign}(r^T v_i) \neq \text{sign}(r^T v_j)\} \\ &= \sum_{(i,j)\in E} w_{ij} \frac{1}{\pi} \arccos(v_i^T v_j) \\ &= \sum_{(i,j)\in E} w_{ij} \frac{1}{\pi} \arccos X_{ij}. \end{aligned}$$

If we define

$$\alpha_{GW} = \min_{-1 \leq x \leq 1} \frac{2 \arccos x}{\pi(1-x)} \approx 0.878,$$

then

$$\mathbb{E}[\text{sol}] \geq \alpha_{GW} \sum_{(i,j)\in E} \frac{1}{2} w_{ij}(1 - X_{ij}) = \alpha_{GW} \, \text{sdp} \, .$$

Note that in the above we only prove that the expected value $\mathbb{E}[\text{sol}]$ of the rounded solution satisfies the inequality (7.4). In fact, the Goemans-Williamson

randomized rounding can be derandomized, and after that a solution satisfying (7.4) can be deterministically found in polynomial time [38].

By (7.4), we know that both the integrality gap opt / sdp and the approximation ratio sol / opt are at least α_{GW}. Now it is natural to ask whether there are any instances for which either opt / sdp or sol / opt can achieve α_{GW} or arbitrarily close to α_{GW}. It is shown in [39] that the answers to both questions are true.

7.2 Sum-of-squares Relaxation of Maximum Cut

In this section, we will see that the lowest degree of sum-of-squares relaxation is exactly the semidefinite relaxation. This is why one of the most promising directions to find a tighter relaxation for the maximum cut and consequently refute the unique games conjecture is through studying sum-of-squares relaxations of higher degree. The following is the sum-of-squares relaxation for the problem (7.2) in the primal form:

$$\begin{aligned} \min \quad & \gamma \\ \text{s.t.} \quad & \gamma - \frac{1}{4}x^T L x + \sum_{i=1}^{n} q_i(x_i^2 - 1) \text{ is a sum of squares.} \end{aligned} \tag{7.5}$$

For the degree-$2d$ sum-of-squares relaxation, the polynomial variables q_i above are required to be at most degree $2d - 2$.

To recover the semidefinite relaxation from the degree-2 sum-of-squares relax-

ation, it is easier to work with the dual form of the relaxation:

$$\begin{aligned} \max \quad & \frac{1}{4}\tilde{\mathbb{E}}_\mu(x^T L x) \\ \text{s.t.} \quad & \tilde{\mathbb{E}}_\mu(x_i^2 h) = \tilde{\mathbb{E}}_\mu h, \quad \forall i = 1, \ldots, n, \\ & \qquad \forall \text{polynomial } h \text{ of degree at most } 2d-2, \\ & \tilde{\mathbb{E}}_\mu(\cdot) \text{ is a degree-}2d \text{ pseudo-expectation.} \end{aligned} \tag{7.6}$$

In the case of degree-2 sum-of-squares relaxation, the constraint on the pseudo-expectation can be simplified as

$$\tilde{\mathbb{E}}_\mu x_i^2 = 1.$$

Because the pseudo-expectation $\tilde{\mathbb{E}}_\mu(\cdot)$ is a linear function on the vector space of polynomials up to degree 2, it is determined by its actions on all the monomials up to degree 2, i.e., by the values

$$\tilde{\mathbb{E}}_\mu x_i, \quad \tilde{\mathbb{E}}_\mu x_i x_j.$$

Assume $\tilde{\mathbb{E}}_\mu(\cdot)$ is the optimal pseudo-expectation to the problem (7.6). Then the linear function $\tilde{\mathbb{E}}_{-\mu}(\cdot)$ defined by

$$\tilde{\mathbb{E}}_{-\mu} x_i = -\tilde{\mathbb{E}}_\mu x_i, \quad \tilde{\mathbb{E}}_{-\mu} x_i x_j = \tilde{\mathbb{E}}_\mu x_i x_j$$

is also a pseudo-expectation with the same objective value. To verify the only nontrivial statement, the nonnegativity of the linear function $\tilde{\mathbb{E}}_{-\mu}(\cdot)$, we consider an arbitrary linear polynomial

$$h(x) = a^T x + b. \tag{7.7}$$

Then

$$\tilde{\mathbb{E}}_{-\mu} h^2 = \tilde{\mathbb{E}}_{-\mu}(a^T x + b)^2 = \tilde{\mathbb{E}}_\mu(-a^T x + b)^2 \geq 0.$$

As a result, the pseudo-expectation

$$\frac{1}{2}(\tilde{\mathbb{E}}_{\mu} + \tilde{\mathbb{E}}_{-\mu})$$

is also an optimal solution to the problem (7.6) satisfying the constraints that

$$\frac{1}{2}(\tilde{\mathbb{E}}_{\mu} + \tilde{\mathbb{E}}_{-\mu})(x_i) = 0.$$

Hence we can add these constraints

$$\tilde{\mathbb{E}}_{\mu} x_i = 0$$

to the problem (7.6) without changing the optimal value.

Define an $n \times n$ matrix X by $X_{ij} = \tilde{\mathbb{E}}_{\mu} x_i x_j$. With the above additional constraints, for the linear polynomial $h(x)$ in (7.7),

$$\tilde{\mathbb{E}}_{\mu} h^2(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \tilde{\mathbb{E}}_{\mu} x_i x_j = a^T X a.$$

Therefore, the nonnegativity of the pseudo-expectation $\tilde{\mathbb{E}}_{\mu}(\cdot)$ is equivalent to the positive semidefiniteness of the matrix X, and the original degree-2 sum-of-squares relaxation (7.6) becomes

$$\begin{aligned} \max \quad & \frac{1}{4} \operatorname{tr}(LX) \\ \text{s.t.} \quad & X_{ii} = 1, \quad \forall i = 1, \ldots, n, \\ & X \in \mathcal{P}_n, \end{aligned}$$

which is exactly the same semidefinite relaxation (7.3) of Goemans-Williamson.

Although higher-degree sum-of-squares relaxations provide possibly tighter relaxations to the maximum cut problem, they do not directly lead to any approximation algorithm since we do not know how to round a solution from higher-degree relaxations. But we can still ask how tight the integrality gap can be.

In Section 7.4, we will prove that the integrality gap of degree-$2d$ sum-of-squares relaxation is at most

$$1 - \frac{1}{4d^2 + 4d + 1}$$

by using complete graphs as instances, and in Section 7.6 we will present one possible direction to construct candidates of even looser integrality gap for the degree-4 sum-of-squares relaxation.

7.3 Semidefinite Relaxation with Triangle Inequalities

Despite being convex, the sum-of-squares relaxation of higher degree is still difficult to solve even for graphs of moderate size due to the large number of variables in the converted semidefinite program. Therefore, it is desirable to find a relaxation which is tighter than the semidefinite relaxation but is easier to handle.

One of the possible relaxations is the semidefinite relaxation with additional triangle constraints:

$$\begin{aligned} \max \quad & \frac{1}{4} x^T L x \\ \text{s.t.} \quad & X_{ii} = 1, \quad \forall i = 1, \dots, n, \\ & X_{ij} + X_{jk} + X_{ki} \geq -1, \quad \forall i, j, k = 1, \dots, n, \\ & X \in \mathcal{P}_n. \end{aligned} \tag{7.8}$$

The above relaxation can be obtained by weakening the constraints from the degree-4 sum-of-squares relaxation in dual form.

Consider the optimal pseudo-expectation $\tilde{\mathbb{E}}_\mu(\cdot)$ to the degree-4 sum-of-squares relaxation (7.6). If we define an $n \times n$ matrix given by

$$X_{ij} = \tilde{\mathbb{E}}_\mu(x_i x_j),$$

then by definition X is feasible to the semidefinite relaxation (7.3) and its objective value is $\mathrm{tr}(LX)/4$. Furthermore, X also satisfies all the triangle constraints. This can be seen by choosing the polynomial

$$h(x) = 1 + x_i x_j + x_j x_k + x_k x_i$$

and using the definition of the pseudo-expectation

$$\tilde{\mathbb{E}}_\mu h^2 = 4 + 4X_{ij} + 4X_{jk} + 4X_{ki} \geq 0.$$

Therefore, X is also a feasible solution to the semidefinite relaxation (7.8) with triangle inequalities. If we denote the optimal value to the problem (7.8) as $\mathrm{sdp}_\triangle$, then by the above argument

$$\mathrm{opt} \leq \mathrm{sos}_4 \leq \mathrm{sdp}_\triangle \leq \mathrm{sdp}\,.$$

However, unlike the degree-4 sum-of-squares relaxation, it is already known that the triangle inequalities will not improve the integrality gap for the semidefinite relaxation [40].

7.4 Integrality Gap of Sum-of-squares Relaxation

Numerical computation shows that higher-degree sum-of-squares relaxations are tight for many graphs, i.e., there is no integrality gap. However, as we will see in the following, for complete graph K_n of odd size, the integrality gap always exists as long as the degree $2d < n$.

For complete graph K_n, the Laplacian $L = nI_n - J_n$, where I_n and J_n are the $n \times n$ identity matrix and matrix of all ones, respectively. The polynomial in (7.5) is

$$\lambda + \frac{1}{4}\left(\sum_{i=1}^n x_i\right)^2 - \frac{n}{4}\sum_{i=1}^n x_i^2 + \sum_{i=1}^n g_i(x)(x_i^2 - 1),$$

where $g_i(x)$ are polynomials of degree at most $2d-2$. Let $x = 2y - 1$, then the above polynomial becomes

$$\lambda + \left(\sum_{i=1}^{n} y_i\right)^2 - n\sum_{i=1}^{n} y_i^2 + 4\sum_{i=1}^{n} \tilde{g}_i(y)(y_i^2 - y_i),$$

where $\tilde{g}_i(y) = g_i(2y-1)$ and its degree is also at most $2d-2$. Then the original is a degree-$2d$ sum-of-squares polynomial if and only if the new one is. Now we can apply the lower bound of the sum-of-squares relaxations for the knapsack problem:

Theorem 7.1. *Given odd integer $n > 2d$, for arbitrary polynomials δ_i of degree $2d-2$ and γ of degree $2d-1$, there does not exist sum-of-squares polynomial h such that*

$$h - \gamma\left(\sum_{i=1}^{n} y_i - \frac{n}{2}\right) - \sum_{i=1}^{n} \delta_i(y_i^2 - y_i)$$

is a negative constant.

Proof. See [41]. □

Using Theorem 7.1, we are able to show that the value sos_{2d} for graph K_n is at least $n^2/4$ if $2d < n$. If not, then there is a positive constant c and some polynomial $\tilde{g}_i(y)$ of degree $2d-2$ such that

$$h = \frac{n^2}{4} - c + \left(\sum_{i=1}^{n} y_i\right)^2 - n\sum_{i=1}^{n} y_i^2 + 4\sum_{i=1}^{n} \tilde{g}_i(y)(y_i^2 - y_i)$$

is a sum of squares. On the other hand, if we set

$$\gamma = \sum_{i=1}^{n} y_i - \frac{n}{2}, \quad \delta_i = 4\tilde{g}_i(y) - n,$$

then by Theorem 7.1, the polynomial

$$h - \left(\sum_{i=1}^{n} y_i - \frac{n}{2}\right)^2 - \sum_{i=1}^{n} (4\beta_i - n)(y_i^2 - y_i) = -c$$

cannot be a negative constant, which is a contradiction.

The maximum cut opt for K_n is $(n^2-1)/4$. Combined with the above result, we have proved that the integrality gap for K_n is at most

$$1 - \frac{1}{n^2}$$

if $2d < n$. By choosing $n = 2d + 1$, we conclude that the integrality gap of the degree-$2d$ sum-of-squares relaxation is at most

$$1 - \frac{1}{4d^2 + 4d + 1}.$$

7.5 Integrality Gap of Semidefinite Relaxation with Triangle Constraints

For the degree-4 sum-of-squares relaxation, the graph K_5 shows that its integrality gap is at most 0.96. This is not surprising since it is already known that the integrality gap is at most 16/17 unless $P = NP$. However, an instance of graph with such a loose integrality gap has not been discovered. Now we would like to investigate possible directions of constructing instances whose integrality gap is looser than K_5.

Since the degree-4 sum-of-squares relaxation is tighter than the semidefinite relaxation with triangle constraints, any instances with loose integrality gap for the degree-4 sum-of-squares relaxation must also have loose integrality gap for the semidefinite relaxation with triangle constraints. Currently, there are two known types of instances with loose integrality gap for the semidefinite relaxation with triangle constraints. The first type of instances is reduced from hard instances for unique games conjecture, whose integrality gap can be arbitrarily close to α_{GW} [40]. However, it is already proved in [42] that for these instances there is no

integrality for the degree-4 sum-of-squares relaxation. The second type of instances is constructed in [39] achieving integrality gap $\alpha_{\triangle} \approx 0.891$ which is still slightly larger than α_{GW}. They will serve as the candidates of loose integrality gap for the degree-4 sum-of-squares relaxation.

The proof of $\alpha_{\triangle}$ integrality gap in [39] for the semidefinite relaxation with triangle constraints consists of two steps. The first step is to show the existence of graphs achieving the integrality gap

$$\frac{2\beta}{\pi(1-\cos\beta)} + \epsilon$$

for the original semidefinite relaxation for any $\epsilon > 0$ and $\pi/2 < \beta < \pi$. If we choose $\beta \approx 2.33$, the integrality gap $\alpha_{GW} + \epsilon$ will be achieved, which shows that the analysis of the integrality gap for the original semidefinite relaxation in Section 7.1 is tight. On the other hand, after choosing $\beta \approx 2.07$, we will achieve the integrality gap $\alpha_{\triangle} + \epsilon$ in [39]. This result is summarized into the next theorem. Here we give a sketch of its proof because the construction in the proof will be the foundation of our construction in Section 7.6. For the complete proof, see [39].

Theorem 7.2. *For every $\epsilon > 0$ and $\pi/2 < \beta < \pi$, there exists a graph whose maximum cut is smaller than $\beta/\pi + \epsilon$ but its value of semidefinite relaxation is larger than $(1-\cos\beta)/2$.*

Sketch of proof. Instead of directly constructing a graph with n vertices, we first define an infinite graph G_c. The vertices of G_c are represented by all points of the unit sphere S^{d-1} in the d-dimensional space, and the edges of G_c are the pairs (x, y) of points whose spherical distance $\delta(x, y) \geq \beta$. Let μ be the uniform measure on the sphere S^{d-1} and μ^2 be the induced product measure on $S^{d-1} \times S^{d-1}$. Then the dimension d should be chosen large enough such that the length of almost every

edge in the graph G_c should be less than $\beta + \epsilon\pi/2$, i.e,

$$\frac{\mu^2\{(x,y) \in S^{d-1} \times S^{d-1} | \delta(x,y) \geq \beta + \epsilon\pi/2\}}{\mu^2\{(x,y) \in S^{d-1} \times S^{d-1} | \delta(x,y) \geq \beta\}} < \epsilon/2. \tag{7.9}$$

A cut of the infinite graph G_c is a measurable subset A of S^{d-1} whose cut value is defined to be

$$\mu_\beta(A) = \mu^2\{(x,y) \in S^{d-1} \times S^{d-1} | x \in A, y \notin A, \delta(x,y) \geq \beta\}.$$

By the result of [43], the maximum cut of G_c is achieved by a half sphere. Now we consider the half sphere defined by a fixed hyperplane and use an argument similar to the analysis of Goemans and Williamson. Pick up an edge (x, y) uniformly from the graph G_c. By (7.9), with probability $1 - \epsilon/2$,

$$\beta \leq \delta(x,y) < \beta + \frac{\epsilon}{2}\pi.$$

In the case where the above inequality indeed holds, the probability of the edge to be cut by the given hyperplane is at most $\beta/\pi + \epsilon/2$. Adding the remaining case with probability $\epsilon/2$, we know that the fraction of edges in the maximum cut is upper bounded by $\beta/\pi + \epsilon$. In order to get a finite graph G, we sample n points from the graph G_c and normalize the total weight in the sampled graph. If n is sufficiently large (the choice of n will be discussed later), then the maximum cut of discrete graph G will be very close to the maximum cut of G_c and thus also bounded by $\beta/\pi + \epsilon$.

Now we are going to prove that the value of semidefinite relaxation for the graph G is at least $(1 - \cos\beta)/2$ by directly constructing a feasible solution to the semidefinite relaxation. If we associate each vertex i in the sampled graph G with a vector v_i of its the location and define an $n \times n$ matrix X by letting

$$X_{ij} = v_i^T v_j.$$

Then X is a feasible solution to the semidefinite relaxation and corresponding objective value

$$\frac{1}{2}w_{ij}\sum_{(i,j)\in E}(1-X_{ij}) \geq \frac{1}{2}(1-\cos\beta),$$

which completes the proof. □

The second step is to show that the feasible solution X constructed in the above proof satisfies almost all the triangle constraints if the size of graph is appropriate, which is based on the following result:

Theorem 7.3. *For any $0 < \epsilon < 1$ and sufficiently large d, if we uniformly place $n \approx (\sqrt{27}/4 - \epsilon)^{d/2}$ points on S^{d-1}, then the expected number of pairs (x, y, z) of points that violate the triangle inequality, i.e.*

$$x^T y + y^T z + z^T x < -1,$$

is at most ϵn.

Proof. See [39]. □

The above result tells us that if the number of points sampled from the infinite graph G_c is $n \approx (\sqrt{27}/4)^{d/2}$, then for the feasible semidefinite relaxation solution X the number of violated triangle constraints in (7.8) is small. Naturally, we need to adjust the associated vector v_i for the sampled points such that all the triangle constraints will be satisfied. The adjustment procedure proposed by the original paper [39] first picks up a point p which is not in any violated triangle constraints. If the points i, j and k violates the triangle inequality, then move the corresponding vector v_i to the location of v_p. If the number of violations is small, then we will get a modified solution for the semidefinite relaxation whose objective value is close that of X but satisfying all triangle constraints, which implies that the integrality

gap of the semidefinite programming is almost the same in cases with or without the triangle constraints.

The above argument proposes an upper bound for the number of sampled points n in order for the semidefinite solution X of the graph G to have few violations of triangle inequalities. However, there should be an lower bound for n such that the graph G will have an approximately identical maximum cut of the infinite graph G_c. In order to obtain such a lower bound, first assume we sample an extremely large number of point n' to obtain a graph G' from G_c. Then G_c and G' will have nearly the same maximum cut. Furthermore, G' will be almost regular and the degree of each vertex will be $\Delta' \approx n'(\sin\beta)^{d-1}$. Using the result in [44], we can further sample vertices from G' by choosing only

$$O(n'/\Delta') = O((1/\sin\beta)^{d-1})$$

points such that the new sampled graph will also have nearly the same maximum cut of G', which suggests that we only need to sample $n \approx (1/\sin\beta)^{d-1}$ points from the original infinite graph G_c. Comparing the upper bound and lower bound for n, we observe that β has to satisfy

$$\sin^2(\beta) = 4/\sqrt{27}.$$

$\beta \approx 2.07$ is the only solution to the above equality in the range $\pi/2 < \beta < \pi$.

7.6 Candidates of Loose Integrality Gap

The paper [39] proves the existence of an instance with integrality gap 0.891 for the semidefinite relaxation with triangle constraints. However, a quick numerical calculation shows that the size of the resulted graph is extremely large if we directly

follow the construction in the proof, and the size is the far beyond what is able to be solved with the degree-4 sum-of-squares relaxation. In order to verify that whether the above construction will provide an instance for the degree-4 sum-of-squares relaxation whose integrality gap is looser than K_5, we first need to find instances of small graphs with such integrality gap for the semidefinite relaxation with triangle inequalities.

This section will present a new method to construct candidates of graphs with small size and potentially loose integrality gap. The first step is to follow the construction in the proof of Theorem 7.2 to obtain a randomly sampled graph and the corresponding feasible semidefinite solution X. Then we use an efficient way of adjusting the feasible solution X to satisfy all of the triangle inequalities. The key idea is to do the adjustment incrementally. For the point i, we are trying to move the corresponding vector v_i to eliminate the violations of triangle inequalities among points i, j and k for $j < k < i$. In order to minimize the change of the integrality gap, the movement of the vector v_i should be minimized. These observations illuminate us to consider the following convex optimization problem for the adjustment of the point i:

$$\begin{aligned} \min \quad & \|p - v_i\| \\ \text{s.t.} \quad & v_j^T p + v_k^T p + v_j^T v_k \geq -1, \quad \forall 1 \leq j < k < i, \\ & v_i^T p \geq 1. \end{aligned} \tag{7.10}$$

Here the decision variable p is the new location of the vector v_i. The last constraint in the above optimization problem only guarantees that $\|p\| \geq 1$. In order to keep the new vector still on the unit sphere, we need to normalize the vector p and set the new value of the vector v_i to be $p' = p/\|p\|$. Now we need to verify that all triangle constraints for points $j < k < i$ are satisfied. In the case when $v_j^T p + v_k^T p \geq 0$,

since $v_j^T v_k \geq -1$ and $v_j^T p' + v_k^T p' \geq 0$, we must have

$$v_j^T p' + v_k^T p' + v_j^T v_k \geq -1.$$

In the case when $v_j^T p + v_k^T p < 0$,

$$v_j^T p' + v_k^T p' + v_j^T v_k \geq v_j^T p + v_k^T p + v_j^T v_k \geq -1.$$

Therefore, after the adjustment for the point i, there will be no violations for the triangle constraints among any points before the point i. After doing all the adjustments for each point in the graph, we will obtain a feasible solution satisfying all triangle constraints. The procedure above is summarized as following:

Algorithm 7.1.

Choose n vectors $v_1, \ldots, v_n$ uniformly on the unit sphere S^{d-1}.

Set the adjacency matrix A of the graph by letting

$$A_{ij} \leftarrow \begin{cases} 1, & \text{if } \arccos(v_i^T v_j) \geq \beta, \\ 0, & \text{otherwise.} \end{cases}$$

for $i \leftarrow 1$ to n **do**

Solve the optimization problem (7.10). Let p be the optimal solution.

$v_i \leftarrow p/\|p\|$.

end for

Return the feasible semidefinite solution X by letting $X_{ij} \leftarrow v_i^T v_j$.

Using the above technique, we are able to construct a candidate of small graph with loose integrality gap. Since the number of minimum sampled points from the infinite graph G_c is in the order of $O((1/\sin\beta)^{d-1})$, β should be as small as possible. Our goal is to beat the integrality gap 0.96, so we choose $\beta = 1.78$ and the corresponding integrality gap that we can theoretically approach is

$$\frac{2\beta}{\pi(1-\cos\beta)} + \epsilon \approx 0.9383.$$

The choice of this particular β could reduce the size of constructed instance while reserving space for the random fluctuation during the construction and the increment of the integrality gap during the later adjustment process.

The graph G_c we constructed is in the dimension of 42 and we sampled 2500 points from G_c to get the desired instance. To verify the integrality gap for this instance, the actual value of the maximum cut has to be known. Since it is very hard to compute the maximum cut accurately on a graph of such size, the actual value of maximum cut has to be substituted by the approximate value from the Goemans-Williamson algorithm. We run the adjustment in Algorithm 7.1 to eliminate all the violations of triangle inequalities for the feasible semidefinite solution. The final ratio between the value of the semidefinite relaxation with triangle constraints and the value of the approximate solution from the Goemans-Williamson algorithm increases to 0.9577.

For the instance constructed above, the obtained ratio 0.9577 is only a lower bound for the integrality gap of this instance. The major difficulty here is that the size of the instance is still too large. For such a graph we are still not able to figure out both the exact values of the degree-4 sum-of-squares relaxation and the actual maximum cut. Whether the instance constructed here has an integrality gap looser than 0.96 or not for the degree-4 sum-of-squares relaxation remains open.

CHAPTER 8

THE SHANNON CAPACITY OF GRAPH

The *Shannon capacity of a graph* is a graph invariant originated from computing the maximum achievable rate to transmit information with zero possibility of error through a noisy channel [45]. To state the definition of Shannon capacity, we need the following notions in graph theory: For an undirected graph G, let $V(G)$ and $E(G)$ be its vertex set and edge set, respectively. Let $\alpha(G)$ be the independence number (aka stability number) of G, i.e., the size of the maximum independent set in G. For two vertices $i, j \in V(G)$, the notation $i \sim_G j$ means either $i = j$ or $(i, j) \in E(G)$. The *strong product* $G \boxtimes H$ of two graphs G and H is a graph such that

- its vertex set $V(G \boxtimes H)$ is the Cartesian product $V(G) \times V(H)$ and
- $(i, j) \sim_{G \boxtimes H} (k, l)$ if and only if $i \sim_G k$ and $j \sim_H l$.

The Shannon capacity $\Theta(G)$ of graph G is defined by

$$\Theta(G) = \sup_k \sqrt[k]{\alpha(G^k)},$$

where G^k is the strong product of G with itself for k times.

The Shannon capacity is unknown for most graphs, including certain simple cases such as odd cycles C_{2n+1} when $n \geq 3$. By definition, for any positive integer k, $\sqrt[k]{\alpha(G^k)}$ provides a direct lower bound for the Shannon capacity $\Theta(G)$, although it is still hard to calculate due to the NP-hardness of maximum independent set problem and the exponential growth of the size of G^k. Finding a good upper bound for $\Theta(G)$ is even more difficult. One well-known upper bound is the Lovász number $\vartheta(G)$ proposed in [32], which can be efficiently computed by solving a

semidefinite program (SDP). The most famous application of Lovász number is the establishment of the Shannon capacity for the pentagon graph C_5:

$$\sqrt{5} = \sqrt{\alpha(C_5^2)} \leq \Theta(C_5) \leq \vartheta(C_5) = \sqrt{5}.$$

However, for 7-cycle C_7, $\vartheta(C_7) \approx 3.3177$, while the best known lower bound [46] at the time of writing is

$$\Theta(C_7) \geq \sqrt[5]{\alpha(C_7^5)} \geq \sqrt[5]{367} \approx 3.2578.$$

Determining the exact value for the Shannon capacity $\Theta(C_7)$ remains an open problem.

One interesting direction is to look for a tighter upper bound for the Shannon capacity than the Lovász number. Since the definition of the Shannon capacity is closely related to the independence number, and in fact the Lovász number itself can be derived from approximating the independence number of a graph, it is tempting to find better upper bounds for the Shannon capacity by using tighter approximations for the independence number. The major challenge here is to ensure that the new approximation is still an upper bound for the Shannon capacity. In Section 8.1, we will look at general conic programming approximation for the independence number, which is a natural generalization of the SDP-based Lovász number. Next, in Section 8.2, we will propose a condition called the *product property* over the cones appeared in the above approximate optimization problem. This property guarantees that the optimal value of the approximation is an upper bound for the Shannon capacity. Surprisingly, in Section 8.3 it is shown that the semidefinite cone used by the Lovász number is the largest cone with such a property, thus ruling out the possibility of improving the estimation of the Shannon capacity along this way.

8.1 Conic Programming for the Independence Number

In this section, we will first formulate the maximum independent set problem as a copositive program. If the semidefinite cone is used as an inner approximation for the copositive cone in this program, the obtained objective value is exactly the Lovász number. As a generalization, we consider all the possible cones that are subsets of the copositive cone, and the corresponding conic programs will be the candidates to generate better upper bounds for the Shannon capacity.

Our starting point is the Motzkin-Straus theorem, which gives the exact value of the independence number of a graph:

Theorem 8.1 (Motzkin-Straus)**.** *If A is the adjacency matrix of a graph G with n vertices, then the independence number of G is given by*

$$\frac{1}{\alpha(G)} = \min_{x \in \mathbb{R}^n_+, \sum_i x_i = 1} x^T (I + A) x.$$

In [47], the optimization problem in Theorem 8.1 is converted into the following equivalent form:

$$\begin{aligned} \alpha(G) = \min \quad & \lambda \\ \text{s.t.} \quad & \lambda(I + A) - J_n \in \mathcal{C}_n. \end{aligned} \tag{8.1}$$

Here J_n is the $n \times n$ matrix of all ones. In order to make the above problem (8.1) closer to the formulation for the Lovász number, we are going to further rewrite it

as follows:

$$\begin{aligned} \min \quad & \lambda \\ \text{s.t.} \quad & Y - J_n \in \mathcal{C}_n, \\ & Y_{ii} = \lambda, \quad \forall i = 1, \ldots, n, \\ & Y_{ij} = 0, \quad \forall i \not\sim_G j, \\ & Y \in \mathcal{S}_n. \end{aligned} \tag{8.2}$$

Since problem (8.1) can be viewed as problem (8.2) with the additional constraint $Y = \lambda(I + A)$, problem (8.2) is a relaxation of the original problem (8.1). To show that these two problems are indeed equivalent, the following property of copositive matrices will be useful:

Lemma 8.1. *Assume Q is a copositive matrix whose diagonal entries are all equal to μ. R is another symmetric matrix of the same size. If for each entry of R either $R_{ij} = Q_{ij}$ or $R_{ij} = \mu$, then R is also copositive.*

Proof. We only need to consider the case in which $R = Q$ except for some off-diagonal entry $R_{st} = \mu$ (and also $R_{ts} = \mu$), since the general result can be obtained by repeating the same argument for each difference between R and Q. For any $x \in \mathbb{R}^n_+$ with $\sum_i x_i = 1$,

$$x^T R x = \mu x_s^2 + \mu x_t^2 + 2\mu x_s x_t + \sum_{\substack{(i,j) \neq (s,s), \\ (s,t),(t,s),(t,t)}} Q_{ij} x_i x_j. \tag{8.3}$$

Fix x_i, $i \neq s, t$, as constants and regard $x^T R x$ as a function of x_s by replacing

$$x_t = 1 - x_s - \sum_{i \neq s,t} x_i.$$

Then the first part of (8.3)

$$\mu x_s^2 + \mu x_t^2 + 2\mu x_s x_t = \mu(x_s + x_t)^2 = \mu \left(1 - \sum_{i \neq s,t} x_i\right)^2$$

becomes a constant. Since the remaining terms in (8.3) are all linear functions of x_s, $x^T Rx$ is also linear as a function of x_s and thus must achieve the minimum when $x_s = 0$ or $x_s = 1$. However, in both cases, $x^T Rx = x^T Qx \geq 0$, which implies that R is also copositive. □

Now we can prove that the problems (8.1) and (8.2) have the same optimal value. Consider an arbitrary feasible solution (λ, Y) to problem (8.2). Let

$$Q = Y - J_n, \quad R = \lambda(I + A) - J_n.$$

All the diagonal entries of Q are $\lambda - 1$. By Lemma 8.1, the matrix R is also copositive and thus $\lambda \geq \alpha(G)$ by (8.1). On the other hand, the solution $\lambda^* = \alpha(G)$, $Y^* = \alpha(G)(I + A)$ is feasible to (8.2), so it must be optimal.

The copositive cone constraint in (8.2) makes the problem hard to solve. If we substitute the copositive cone $\mathcal{C}_n$ in (8.2) with the semidefinite cone $\mathcal{P}_n$, the optimal value for the modified problem is the Lovász number $\vartheta(G)$. Since $\mathcal{P}_n \subseteq \mathcal{C}_n$, we immediately get $\alpha(G) \leq \vartheta(G)$. Naturally, to find a tighter bound for the Shannon capacity $\Theta(G)$, we can replace the copositive cone $\mathcal{C}_n$ in (8.2) by some cone between $\mathcal{C}_n$ and $\mathcal{P}_n$, which may lead to some problem whose optimal value is potentially between the Shannon capacity $\Theta(G)$ and the Lovász number $\vartheta(G)$.

The above discussion illuminates us to construct more general approximations for the independence number $\alpha(G)$ by introducing some arbitrary cone $\mathcal{A}_n \subseteq \mathcal{C}_n$,

i.e., the problem

$$\begin{aligned} \min \quad & \lambda \\ \text{s.t.} \quad & Y - J_n \in \mathcal{A}_n, \\ & Y_{ii} = \lambda, \quad \forall i = 1, \ldots, n, \\ & Y_{ij} = 0, \quad \forall i \not\sim_G j, \\ & Y \in \mathcal{S}_n. \end{aligned} \tag{8.4}$$

In the case when the cone $\mathcal{A}_n$ is chosen to be the semidefinite cone $\mathcal{P}_n$, the above problem (8.4) gives the Lovász number $\vartheta(G)$. To provide some other examples of $\mathcal{A}_n$, one can approximate the copositive cone $\mathcal{C}_n$ based on sum-of-squares programming [47]. Note that the copositivity of a matrix $Q \in \mathcal{S}_n$ can be equivalently written as

$$p_Q(x) = \sum_{i,j} Q_{ij} x_i^2 x_j^2 \geq 0, \quad \forall x \in \mathbb{R}^n. \tag{8.5}$$

Like determining copositivity, it is NP-hard to decide whether the polynomial $p_Q(x)$ is nonnegative or not. However, if $p_Q(x)$ can be written as a sum of squares, i.e.,

$$p_Q(x) = \sum_k g_k^2(x),$$

where $g_k(x)$ are arbitrary polynomials of $x \in \mathbb{R}^n$, then clearly $p_Q(x)$ is nonnegative. All symmetric matrices $Q \in \mathcal{S}_n$ whose corresponding polynomial $p_Q(x)$ given by (8.5) is a sum of squares constitute a cone, which will be denoted as $\mathcal{C}_n^{(0)}$ in the following. Obviously $\mathcal{C}_n^{(0)} \subseteq \mathcal{C}_n$, and furthermore it is tractable to determine whether a matrix Q is in the cone $\mathcal{C}_n^{(0)}$ through SDP. In fact, $\mathcal{C}_n^{(0)}$ has a simple characterization [48]:

$$\mathcal{C}_n^{(0)} = \mathcal{P}_n + \mathcal{N}_n.$$

In other words, the polynomial $p_Q(x)$ is a sum of squares if and only if the matrix Q can be written as a sum of a positive semidefinite matrix and a nonnegative

symmetric matrix.

For any graph G, the optimal value of problem (8.4), in which $\mathcal{A}_n = \mathcal{C}_n^{(0)}$, is called $\vartheta'(G)$, the Schrijver ϑ'-function [49]. Since

$$\mathcal{P}_n \subseteq \mathcal{C}_n^{(0)} \subseteq \mathcal{C}_n,$$

for any graph G,

$$\alpha(G) \leq \vartheta'(G) \leq \vartheta(G).$$

Moreover, there exists some graph for which the second inequality is strict (see [49]). Given these properties, $\vartheta'(G)$ looks promising for being a better upper bound for the Shannon capacity $\Theta(G)$.

More generally, we can find even better approximations for the copositive cone $\mathcal{C}_n$ by using higher order sum-of-squares polynomials. For each nonnegative integer r, define a set $\mathcal{C}_n^{(r)}$ as follows: a symmetric matrix Q belongs to $\mathcal{C}_n^{(r)}$ if and only if there exists a polynomial $h(x)$ of degree at most $2r$ such that both $h(x)$ and $h(x)p_Q(x)$ are sum of squares. Then $\mathcal{C}_n^{(r)}$ is a cone, and

$$\mathcal{P}_n \subseteq \mathcal{C}_n^{(0)} \subseteq \mathcal{C}_n^{(1)} \subseteq \cdots \subseteq \mathcal{C}_n^{(r)} \subseteq \cdots \subseteq \mathcal{C}_n.$$

Similar to the Schrijver ϑ'-function, we denote the optimal value of the corresponding problem (8.4) as $\vartheta^{(r)}(G)$.

For higher-order sum-of-squares cones $\mathcal{C}_n^{(r)}$ where $r > 0$, although $\vartheta^{(r)}(G)$ is a tighter upper bound for the independence number than $\vartheta(G)$ or $\vartheta'(G)$, it is too tight for being an upper bound for the Shannon capacity $\Theta(G)$. For instance, for the pentagon graph C_5, if $r > 0$,

$$\alpha(C_5) = \vartheta^{(r)}(C_5) = 2 < \vartheta(C_5) = \vartheta'(C_5) = \Theta(C_5) = \sqrt{5}.$$

To obtain a correct upper bound for the Shannon capacity from cones $\mathcal{C}_n^{(r)}$, we have to add extra constraints in the problem (8.4) to restrict these cones. Whatever the exact form of constraints is, we can still analyze the restricted problem as a special case of (8.4) as long as these constraints define a cone.

In the following, we will assume $\mathcal{A}_n$ in the above problem (8.4) to be an arbitrary cone satisfying $\mathcal{A}_n \subseteq \mathcal{C}_n$, and the optimal value will be called $f(G)$. To ensure that $f(G)$ is still an upper bound for the Shannon capacity $\Theta(G)$, in the next section we will look at the key property of the semidefinite cone $\mathcal{P}_n$ used by the Lovász number $\vartheta(G)$ that guarantees $\Theta(G) \leq \vartheta(G)$ and then try to enforce the same property on the cone $\mathcal{A}_n$ in (8.4).

8.2 Product Property and Upper Bounds for the Shannon Capacity

One fundamental property[1] of the Lovász number is

$$\vartheta(G \boxtimes H) \leq \vartheta(G)\vartheta(H) \tag{8.6}$$

for any graphs G and H, which immediately implies that

$$\sqrt[k]{\alpha(G^k)} \leq \sqrt[k]{\vartheta(G^k)} \leq \vartheta(G)$$

for all positive integers k, and thus

$$\Theta(G) = \sup_k \sqrt[k]{\alpha(G^k)} \leq \vartheta(G).$$

The above argument can also be applied to the graph function $f(G)$ defined as the optimal value of (8.4). Since $\alpha(G) \leq f(G)$, as long as $f(G)$ satisfies the

[1] In fact, the equality holds in (8.6), but the reverse direction is not relevant for our purpose.

similar inequality

$$f(G \boxtimes H) \leq f(G)f(H) \tag{8.7}$$

for any graphs G and H, $f(G)$ will also be an upper bound for the Shannon capacity $\Theta(G)$. To find the condition that leads to the inequality (8.7), we need to generalize the proof for the property (8.6) of the Lovász number, which itself is a special case of the general product rules in semidefinite programming [50].

Consider two graphs G of n vertices and H of m vertices. Assume (λ', Y') and (λ'', Y'') are the optimal solutions to the problem (8.4) for graph G and H, respectively. Let $Y = Y' \otimes Y''$, i.e., the Kronecker product of Y' and Y'', which is an $nm \times nm$ matrix given by

$$Y = \begin{pmatrix} Y'_{11}Y'' & \cdots & Y'_{1n}Y'' \\ \vdots & \ddots & \vdots \\ Y'_{n1}Y'' & \cdots & Y'_{nn}Y'' \end{pmatrix}.$$

We index the rows of Y by pairs (i, j) and the columns by pairs (k, l), then the above definition can be rewritten as

$$Y_{(i,j)(k,l)} = Y'_{ik}Y''_{jl}.$$

If $(i, j) \not\sim_{G\boxtimes H} (k, l)$, then either $i \not\sim_G k$ or $j \not\sim_H l$, which implies either $Y'_{ik} = 0$ or $Y''_{jl} = 0$ and thus $Y_{(i,j)(k,l)} = 0$. Since all the diagonal entries of Y equal to $\lambda'\lambda''$, if we can show that $Y - J_{nm} \in \mathcal{A}_{nm}$, then $(\lambda'\lambda'', Y)$ will be a feasible solution to the problem (8.4) for the product graph $G \boxtimes H$. In this case, we have

$$f(G \boxtimes H) \leq \lambda'\lambda'' = f(G)f(H).$$

Let

$$Q = Y' - J_n, \quad R = Y'' - J_m.$$

Then $Q \in \mathcal{A}_n$, $R \in \mathcal{A}_m$. The only missing part that remains to show is

$$Y - J_{nm} = Y' \otimes Y'' - J_{nm} = (Q + J_n) \otimes (R + J_m) - J_{nm} \in \mathcal{A}_{nm},$$

which will be encapsulated into the following definition:

Definition 8.1. Given two symmetric matrices $Q \in \mathcal{S}_n, R \in \mathcal{S}_m$, define

$$Q \odot R = (Q + J_n) \otimes (R + J_m) - J_{nm}.$$

A sequence of cones $\mathcal{A}_n \subseteq \mathcal{S}_n$ is said to have the *product property* if for any matrices $Q \in \mathcal{A}_n, R \in \mathcal{A}_m$, we have $Q \odot R \in \mathcal{A}_{nm}$.

Based on this definition, the above argument can summarized as follows:

Theorem 8.2. *If the cones $\mathcal{A}_n$ in problem* (8.4) *satisfy $\mathcal{A}_n \subseteq \mathcal{C}_n$ and the product property, then $\Theta(G) \leq f(G)$ for any graph G.*

As an example, we check that the product property holds for the sequence of cones $\mathcal{P}_n$ in the Lovász number. Assume matrices $Q \in \mathcal{P}_n$, $R \in \mathcal{P}_m$. Then the matrix

$$Q \odot R = (Q + J_n) \otimes (R + J_m) - J_{nm} = Q \otimes R + Q \otimes J_m + J_n \otimes R$$

is positive semidefinite, because the Kronecker product of two positive semidefinite matrices is still positive semidefinite. Therefore, Theorem 8.2 implies that the Lovász number $\vartheta(G) \geq \Theta(G)$.

The product property is a sufficient condition for the functional inequality (8.7) and further for being an upper bound for the Shannon capacity. However, neither the product property nor the inequality (8.7) is necessary for being the upper bound. In any case, from the proof of Theorem 8.2, one can see that the product property is the most natural condition to guarantee $\Theta(G) \leq f(G)$. Next, we will study the product property holds for what choice of cones $\mathcal{A}_n$.

8.3 Optimality of the Lovász Number

In the last section, we stated the product property, the condition for our new bound $f(G)$ defined in (8.4) to be an upper bound for the Shannon capacity. At the same time, we do not want $f(G)$ to be much larger than the Lovász number for the same graph G. Note that the Lovász number satisfies the following sandwich theorem:

$$\alpha(G) \leq \vartheta(G) \leq \chi(\bar{G}),$$

where $\chi(\bar{G})$ is the chromatic number for the complement graph of G. If we choose $G = \bar{K}_2$, the edgeless graph of two vertices, then

$$2 = \alpha(\bar{K}_2) \leq \vartheta(\bar{K}_2) \leq \chi(K_2) = 2.$$

If the new bound $f(G)$ satisfies the similar sandwich theorem, then $f(\bar{K}_2) = 2$, which means that the matrix

$$\Lambda = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \in \mathcal{A}_2.$$

We want to find a sequence of cones $\mathcal{A}_n$ satisfying all the above desired condition. However, it turns out that the only possible cones $\mathcal{A}_n$ must be subsets of the corresponding semidefinite cones $\mathcal{P}_n$, and the obtained upper bound $f(G)$ would be at least the Lovász number.

Theorem 8.3. *If a sequence of cones $\mathcal{A}_n$ satisfies the following properties:*

1. *$\mathcal{A}_n \subseteq \mathcal{C}_n$ for all n.*
2. *The matrix $\Lambda \in \mathcal{A}_2$.*
3. *The sequence $\mathcal{A}_n$ has the product property.*

Then we must have $\mathcal{A}_n \subseteq \mathcal{P}_n$ for all n.

Proof. Prove by contradiction. Assume there is a matrix $A \in \mathcal{A}_n$ which is not positive semidefinite and v is an eigenvector of A such that $v^T A v < 0$.

We first construct a matrix $B \in \mathcal{A}_m$ with $m = 2n$ which has a eigenvector w such that

$$w^T B w < 0, \quad \sum_i w_i = 0.$$

For any $k > 0$, let $B = A \odot (k\Gamma)$, then by the cone property and the product property,

$$B = \begin{pmatrix} 2kA + J_n & -J_n \\ -J_n & 2kA + J_n \end{pmatrix} \in \mathcal{A}_m.$$

If we let

$$w = \begin{pmatrix} v \\ -v \end{pmatrix},$$

then

$$w^T B w = 4k v^T A v + 4 v^T J_n v.$$

In the above argument, we can choose k with

$$k > \frac{v^T J_n v}{v^T A v},$$

and now the matrix B and its eigenvector w will have all the desired properties.

Next, we are going to construct another matrix $C \in \mathcal{A}_{2m}$ which is not copositive, which will lead to a contradiction. Define

$$x = \max(w, 0), \quad y = \max(-w, 0).$$

Then $x, y \geq 0$ and $w = x - y$.

For any $k > 1$, by the product property again, the matrix

$$C = (k\Gamma) \odot B = \begin{pmatrix} (k+1)B + kJ_m & -(k-1)B - kJ_m \\ -(k-1)B - kJ_m & (k+1)B + kJ_m \end{pmatrix} \in \mathcal{A}_{2m}.$$

Consider

$$\begin{pmatrix} x^T & y^T \end{pmatrix} C \begin{pmatrix} x \\ y \end{pmatrix} = (k+1)(x^T Bx + y^T By) - 2(k-1)x^T By$$
$$+k(x^T J_m x + y^T J_m y) - 2kx^T J_m y,$$

where the second part,

$$k(x^T J_m x + y^T J_m y) - 2kx^T J_m y = k \sum_i \sum_j (x_i x_j + y_i y_j - x_i y_j - y_i x_j)$$
$$= k \sum_i \sum_j (x_i - y_i)(x_j - y_j) = k \left(\sum_i w_i \right)^2 = 0.$$

Since

$$w^T Bw = (x-y)^T B(x-y) = x^T Bx + y^T By - 2x^T By < 0,$$

for sufficiently large k, we have

$$x^T Bx + y^T By < 2\frac{k-1}{k+1} x^T By,$$

which will imply

$$\begin{pmatrix} x^T & y^T \end{pmatrix} C \begin{pmatrix} x \\ y \end{pmatrix} < 0.$$

Now we obtain a matrix $C \in \mathcal{A}_{2m}$ and C is not copositive, which is a contradiction.

□

Theorem 8.3 tells us that either the cones do not have the product property or the resulting function $f(G) \geq \vartheta(G)$. As a result, it is impossible to derive an upper bound for the Shannon capacity that is better than the Lovász number by enforcing the product property on cones $\mathcal{A}_n$.

For the Schrijver ϑ'-function, the corresponding cones $\mathcal{C}_n^{(0)}$ satisfy the first and second condition of Theorem 8.3 but not the conclusion $\mathcal{C}_n^{(0)} \subseteq \mathcal{P}_n$. Therefore, by Theorem 8.3, the cones $\mathcal{C}_n^{(0)}$ do not have the product property. Although not having the product property for $\mathcal{C}_n^{(0)}$ does not directly imply that the Schrijver ϑ'-function is not an upper bound for the Shannon capacity, it strongly suggests such negative result. In fact, it is very hard to disprove that the Schrijver ϑ'-function is an upper bound, because at least for graphs G of moderate size the two values $\vartheta'(G)$ and $\vartheta(G)$ are very close to each other. In order to prove that $\vartheta'(G)$ is not an upper bound, we have to find some sufficiently large k and show that

$$\vartheta'(G) < \sqrt[k]{\alpha(G^k)} \leq \vartheta(G),$$

which is extremely difficult even if G contains only a few vertices. We believe that the Schrijver ϑ'-function is not an upper bound for the Shannon capacity due to its lack of the product property, but whether this is actually true or not remains open.

CHAPTER 9

CONCLUSION

In Part I of this dissertation, we propose a novel bound for the duality gap of separable problems. The improvement over the existing results is attributed to two sources. First, instead of using a single number measurement, a series of numbers are introduced to characterize the nonconvexity of a function in a potentially much finer manner. Such a fine measure of nonconvexity is based on the concept of kth convex hull of a set, which allows us to differentiate different levels of nonconvexity for nonconvex sets. Second, for a separable nonconvex problem, we do not approximate each subproblem individually as people had done before. Instead, by considering all subproblems jointly and noticing that the total deviation of every subproblem to a convex problem is bounded, we reach a much tighter estimation.

In Part II of this dissertation, we first look at the integrality gap for the sum-of-squares relaxations of the maximum cut. We give an instance of integrality gap 0.96 for the degree-4 sum-of-squares relaxation. However, it is already known that there should exist instances achieving integrality gap $16/17 \approx 0.941$ unless $P = NP$ although such instances have never been explicitly found. Next, we give a direction of searching for candidates of integrality gap looser than 0.96 by constructing instances with loose integrality gap for the semidefinite relaxation with triangle constraints. Although the obtained instance by our method is already greatly smaller than the one constructed based on the previous proof, we need to further optimize the construction such that the size of the instance is within the range in which we are able to verify its integrality gap computationally.

Finally, we consider the possibility of using generalized conic programming to

upper bound the Shannon capacity of graph. The product property is a natural condition that guarantees the relaxation being an upper bound of the Shannon capacity, and we prove that it is impossible for the value of certain conic programs, including the sum-of-squares relaxations for the independence number, to satisfy the product property, except the original Lovász number. However, although unlikely, it is possible that there exist other upper bounds without satisfying the product property, and it is of our future work to investigate such possibilities.

To reach an approximation algorithm for nonconvex optimizations, we need to have not only a relaxation but also the rounding algorithm for the optimal solution obtained from the relaxation. In this dissertation, we focus on estimating the integrality gap or the duality gap for the convex relaxations. A natural future direction is to study the approaches of designing rounding algorithms and analyzing the performance guarantees of the obtained solution based on our deeper understanding of the convex relaxations achieved in this dissertation.

APPENDIX A

MULTIPATH RELAXATION FOR NETWORK UTILITY MAXIMIZATION

The two convex relaxations based on Lagrangian duality and semidefinite programming are general approaches that can be applied to varieties of nonconvex optimization problems. However, for particular problems, there may be other relaxations for which we are able to prove tighter integrality gap. In this appendix, we will reconsider the network utility maximization problem with rate constraints (5.4) in which the rate constraint set $\mathcal{S}_K$ are chosen to model the following two types of constraints:

- *Path cardinality constraints* where each user is allowed to send positive rates on at most W paths. This case is the generalization of the single-path routing problem discussed in Section 4.1, and $\mathcal{S}_K$ is chosen to be

$$\mathcal{S}_K = \{x \in \mathbb{R}^K | \|x\|_0 \leq W\}.$$

  Here $\|x\|_0$ is the number of nonzero components in the vector x.

- *Split ratio granularity constraints* where the split ratio is each user must be a multiple of $1/p$. This case is already studied in Section 5.2 using the general tool of convex relaxation based on Lagrangian duality, and $\mathcal{S}_K$ is chosen to be (5.6).

For simplicity, in this appendix we assume the linear utility $U^i(x^i) = \|x^i\|_1$ for each user and $\|c\|_\infty = 1$. In addition, in the case of path cardinality constraints, we assume $W \geq K^i$ for every user i.

Now we consider the multipath relaxation of the original network utility maximization problem (5.4):

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} f_i(x^i) \\ \text{s.t.} \quad & \sum_{i=1}^{N} R^i x^i \le c, \\ & x^i \in \mathcal{T}_{K^i}, \quad \forall i = 1, \ldots, N, \end{aligned} \tag{A.1}$$

where

$$\mathcal{T}_K = \{x \in \mathcal{I}_K | \|x\|_1 \le C_K\}$$

and

$$C_K = \max_{x \in \mathcal{S}_K} \|x\|_1.$$

For the case of path cardinality constraints, it is easy to see that $C_K = W$. For the case of split ratio granularity constraints, because a nonzero rate vector is in $\mathcal{S}_K$ if and only if it is p-integral, by Lemma 5.2 we immediately know that

$$C_K = \frac{p}{\lceil p/K \rceil}.$$

Let opt be the optimal value of the original problem (5.4) and opt_C be the optimal value of the relaxed problem (A.1), then $\text{opt} \le \text{opt}_C$. In the remaining, we are going to upper bound the integrality gap $\text{opt}_C - \text{opt}$ for both the path cardinality constraints case and the split ratio granularity constraints.

A.1 Path Cardinality Constraints

In this section, we will bound the integrality gap for the problem (A.1) with path cardinality constraints. Since the feasible region of (A.1) is bounded, the optimal solution to (A.1) can be attained at one vertex of the feasible region, which will

be called an optimal vertex solution. Our starting point is to show that optimal vertex solutions have the following sparsity property:

Lemma A.1. *Assume x is an optimal vertex solution to the problem* (A.1). *Let N' be the number of users with at least $W+1$ positive flows in rate allocation x, then*

$$N' \leq \frac{L}{W}.$$

At the same time, x has at most $N'+L$ positive flows.

Proof. Recall that a vertex must have $K = \sum_{i=1}^{N} K^i$ independent active constraints (constraints hold with equality). If in vertex x, a user i has less than W positive flows, its corresponding constraint

$$\|x^i\|_1 \leq W \tag{A.2}$$

must be inactive.

Now assume user i has exact W positive flows and its corresponding constraint (A.2) holds with equality. Without loss of generality, we can assume that its first W paths are used. Then the only possible case is

$$\begin{aligned} x_k^i &= 1, && k = 1, \ldots, W, \\ x_k^i &= 0, && k = W+1, \ldots, K^i. \end{aligned}$$

For $k = 1, \ldots, W$, if l is an arbitrary link on path k, then $c_l = 1$. The capacity constraint for link l must have the form $x_k^i = c_l$, i.e. the link l is fully occupied by user i. Therefore, the constraint (A.2) of user i is not an independent constraint because it can be written as a linear combination of the active constraints in $Rx \leq c$ and $x \geq 0$.

Hence, there are at most N' independent active constraints from (A.2). Since at most L active constraints can be obtained from $Rx \leq c$, there are at least $K - N' - L$ active constraints among $x \geq 0$, i.e. x must have at most $N' + L$ positive flows.

However, N' is the number of users who have at least $W + 1$ positive flows. By the previous result, we have

$$(W + 1)N' \leq N' + L,$$

which implies $N' \leq L/W$. □

For any optimal solution x to the relaxation (A.1), we can round it into a feasible solution to the original problem (5.4) by sending rates only on the paths that have W largest rates for each user and setting the rates on the other paths to be zero. Based on Lemma A.1, we can bound the loss of the total rates after the rounding of an optimal vertex solution, which leads to the following bound for the integrality gap:

Theorem A.1. *The integrality gap*

$$\text{opt}_C - \text{opt} \leq \Psi(L, W),$$

where

$$\Psi(L, W) = \max_{n=1,\ldots,\lfloor L/W \rfloor} \left(n - \frac{Wn^2}{n + L} \right) W. \tag{A.3}$$

Proof. Assume x is an optimal vertex solution to the problem (A.1) and y is the rounded solution from x. Let G^i be the number of positive flows of user i in rate allocation x. Let $\mathcal{S}$ denote the set of users with at least $W + 1$ positive flows, i.e.

$$\mathcal{S} = \{i | G^i \geq W + 1, i = 1, \ldots, N\}.$$

Note that set $\mathcal{S}$ contains the users who will be affected by the rounding, and N' is just the number of users in $\mathcal{S}$.

If $i \in \mathcal{S}$, then the average of positive rates of user i in x must be less than or equal to that in y, i.e.

$$\frac{\|x^i\|_1}{G^i} \leq \frac{\|y^i\|_1}{W},$$

because y^i only contains the flows of user i with W largest rates. Now

$$\begin{aligned} \|x^i\|_1 - \|y^i\|_1 &\leq \left(1 - \frac{W}{G^i}\right) \|x^i\|_1 \\ &\leq \left(1 - \frac{W}{G^i}\right) W, \end{aligned} \tag{A.4}$$

where (A.4) holds because x is feasible for (A.1). If $i \notin \mathcal{S}$, then

$$\|x^i\|_1 - \|y^i\|_1 = 0. \tag{A.5}$$

Adding up (A.4) and (A.5) for all users, we have

$$\begin{aligned} \text{opt}_C - \text{opt} &\leq \sum_{i=1}^{N} \|x^i\|_1 - \sum_{i=1}^{N} \|y^i\|_1 \\ &\leq \sum_{i \in \mathcal{S}} \left(1 - \frac{W}{G^i}\right) W \\ &= \left(N' - W \sum_{i \in \mathcal{S}} \frac{1}{G^i}\right) W \\ &\leq \left(N' - W \frac{N'^2}{\sum_{i \in \mathcal{S}} G^i}\right) W \\ &\leq \left(N' - \frac{W N'^2}{N' + L}\right) W, \end{aligned}$$

in which the last second inequality holds because

$$\frac{N'}{\sum_{i \in \mathcal{S}} 1/G^i} \leq \frac{1}{N'} \sum_{i \in \mathcal{S}} G^i,$$

and the last inequality is from Lemma A.1.

Since N' is an integer between 0 and L/W, by enumerating all possibilities for N', we establish the desired result. □

For single-path routing ($W = 1$),

$$\left(n - \frac{Wn^2}{n+L}\right) W = \frac{L}{1 + L/n}.$$

The maximizer inside (A.3) is always attained by $n = L$, so $\Psi(L, 1) = L/2$, which is the same as the result obtained in Section 4.1. The general case is complex due to the floor function in (A.3), but we have the following result:

Proposition A.1. *If $W \geq 2$,*

$$\Psi(L, W) \leq (\sqrt{W} - \sqrt{W-1})^2 WL.$$

Proof. For any integer $n = 1, \ldots, \lfloor L/W \rfloor$,

$$\begin{aligned}
&\left(n - \frac{Wn^2}{n+L}\right) W \\
&= \left((2W-1)L - (n+L)(W-1) - \frac{WL^2}{n+L}\right) W \\
&\leq \left((2W-1)L - 2\sqrt{(W-1)WL^2}\right) W \\
&= (\sqrt{W} - \sqrt{W-1})^2 WL.
\end{aligned}$$

□

A.2 Split Ratio Granularity Constraints

For the case of split ratio granularity constraints, we also have the following sparsity property for an optimal vertex solution to (A.1). We omit the proof here since it is similar to the proof of Lemma A.1.

Lemma A.2. *Assume x is an optimal vertex solution to the problem* (A.1)*. Then x has at most $N + L$ positive flows.*

Define

$$\rho_K^C = \max_{x \in \mathcal{T}_K} \phi_K(x),$$

where $\phi_K(x)$ is the throughput loss function given by (5.10). Assume

$$(\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^N)$$

is an optimal vertex solution to the problem (A.1) using at most $N + L$ paths in total. Let $\tilde{K}^i$ be the number of nonzero components in the vector $\bar{x}^i$. Then

$$0 \leq \tilde{K}^i \leq K^i, \quad \sum_{i=1}^{N} \tilde{K}^i \leq N + L,$$

and the throughput loss for user i during the rounding step is bounded by $\rho_{\tilde{K}^i}$ as all the zero components can be ignored during the rounding. Since $\tilde{K}^i$ are unknown, we have to enumerate all the possible values of $\tilde{K}^i$ and find the worst case in order to bound the total throughput loss, which leads to the following upper found for the integrality gap:

$$\begin{aligned} \text{opt}_C - \text{opt} \leq \max \quad & \sum_{i=1}^{N} \rho_{\tilde{K}^i}^C \\ \text{s. t.} \quad & \sum_{i=1}^{N} \tilde{K}^i \leq N + L, \\ & 0 \leq \tilde{K}^i \leq K^i,\ \tilde{K}^i \in \mathbb{Z}, \quad \forall i = 1, \ldots, N. \end{aligned}$$

To obtain the exact bound, we need to calculate the numbers ρ_K^C. There are two simple cases in which ρ_K^C can be calculated easily. First, if there exists a rate vector $x \in \mathcal{T}_K$ having the maximum throughput and maximum relative throughput loss simultaneously, that is $\|x\|_1 = C_K$ and the relative performance loss

$$\frac{\phi_K(x)}{\|x\|_1} = \frac{K-1}{p+K-1},$$

then we directly know that

$$\rho_K^C = \phi_K(x) = \frac{K-1}{p+K-1} C_K$$

Second, if there is a rate vector in $\mathcal{T}_K$ whose throughput loss attains the maximum throughput loss ρ_K in the set $\mathcal{S}_K$, then obviously $\rho_K^C = \rho_K$. In fact, the following theorem shows that either of the above two case will occur.

Theorem A.2. *There exists a rate vector $x \in \mathcal{I}_K$ whose relative performance loss is*

$$\frac{\phi_K(x)}{\|x\|_1} = \frac{K-1}{p+K-1}$$

making one of the following two statements hold: (i) $\|x\|_1 \geq C_K$; (ii) $\phi_K(x) = \rho_K$.

Proof. The result is obvious for $K = 1$. If $K > 1$, by Lemma 5.2, there exists an integral rate vector x such that

$$\|x\|_1 = \frac{p+K-1}{\lceil (p+K-1)/K \rceil}$$

and its relative throughput loss is

$$\frac{\phi_K(x)}{\|x\|_1} = \frac{K-1}{p+K-1}.$$

For this particular x, if $\|x\|_1 \geq C_K$, statement (i) holds. Otherwise,

$$\frac{p+K-1}{\lceil (p+K-1)/K \rceil} = \|x\|_1 < C_K = \frac{p}{\lceil p/K \rceil}.$$

Let $\lceil p/K \rceil = \Theta$. The only possibility that the above inequality holds is $\lceil (p+K-1)/K \rceil = \Theta + 1$. Now this inequality can be rewritten as

$$\frac{p+K-1}{\Theta+1} < \frac{p}{\Theta},$$

which implies that

$$(K-1)\Theta < p. \tag{A.6}$$

If statement (ii) does not hold, i.e.,

$$\phi_K(x) = \frac{K-1}{\lceil (p+K-1)/K \rceil} < \rho_K,$$

we will end up with a contradiction. In (5.15), the maximization is not attained by $\Gamma = p + K - 1$ and thus

$$\rho_K = \frac{\omega}{\Theta} > \frac{K-1}{\Theta+1}, \tag{A.7}$$

where ω is the integer such that $(p+\omega)/K = \Theta$. Note that $\omega < K-1$, so

$$\frac{(K-1)\Theta}{\Theta+1} < \omega \leq K-2,$$

which implies

$$\frac{K-2}{K-1} > \frac{\Theta}{\Theta+1}$$

and thus $K - 1 > \Theta + 1$.

On the other hand, we can substitute $\omega = K\Theta - p$ into inequality (A.7), which gives

$$\frac{K-1}{\Theta+1} < \frac{K\Theta - p}{\Theta} = K - \frac{p}{\Theta} < 1,$$

where the last inequality is from (A.6). Now we have $K - 1 < \Theta + 1$, which is a contradiction. □

Theorem A.2 offers the following approach to calculate ρ_K^C: First, we find the integral rate vector x with the maximum relative throughput loss as described in the proof of Lemma 5.2. If $\|x\|_1 \geq C_K$, then scale down x to make $\|x\|_1 = C_K$ without changing the relative throughput loss. In this case,

$$\rho_K^C = \frac{K-1}{p+K-1} C_K.$$

If $\|x\|_1 < C_K$, then $x \in \mathcal{T}_K$ and Theorem A.2 guarantees that $\phi_K(x) = \rho_K = \rho_K^C$.

BIBLIOGRAPHY

[1] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[2] S. H. Low, "Convex relaxation of optimal power flow—part i: Formulations and equivalence," *IEEE Control Netw. Syst.*, vol. 1, no. 1, pp. 15–27, Mar. 2014.

[3] V. V. Vazirani, *Approximation Algorithms*. Springer, 2003.

[4] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

[5] R. Horst and P. M. Pardalos, *Handbook of Global Optimization*. Springer, 1995.

[6] Z.-Q. Luo and S. Zhang, "Duality gap estimation and polynomial time approximation for optimal spectrum management," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2675–2689, Jul. 2009.

[7] H. Zhang, J. Shao, and R. Salakhutdinov, "Deep neural networks with multi-branch architectures are intrinsically less non-convex," in *Proc. Mach. Learn. Res.*, vol. 89, Apr. 2019, pp. 1099–1109.

[8] A. Askari, A. d'Aspremont, and L. El Ghaoui, "Naive feature selection: Sparsity in naive Bayes," Jul. 2019, arXiv:1905.09884.

[9] M. Udell and S. Boyd, "Bounding duality gap for separable problems with linear constraints," *Comput. Optim. Appl.*, vol. 64, no. 2, pp. 355–378, Jun. 2016.

[10] S. H. Low, "A duality model of TCP and queue management algorithms," *IEEE/ACM Trans. Netw.*, vol. 11, no. 4, pp. 525–536, Aug. 2003.

[11] K. J. Arrow, H. D. Block, and L. Hurwicz, "On the stability of the competitive equilibrium, ii," *Econometrica*, vol. 27, no. 1, pp. 82–109, Jan. 1959.

[12] H. Scarf, "Some examples of global instability of the competitive equilibrium," *Int. Econ. Rev.*, vol. 1, no. 3, pp. 157–172, Sep. 1960.

[13] O. H. Ibarra and C. E. Kim, "Fast approximation algorithms for the knapsack and sum of subset problems," *J. ACM*, vol. 22, no. 4, pp. 463–468, Oct. 1975.

[14] S. Arora, "Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems," *J. ACM*, vol. 45, no. 5, pp. 753–782, Sep. 1998.

[15] J. Håstad, "Clique is hard to approximate within $n^{1-\epsilon}$," *Acta Math.*, vol. 182, no. 1, pp. 105–142, Mar. 1999.

[16] S. Khot, "On the power of unique 2-prover 1-round games," in *Proc. ACM STOC*, May 2002, pp. 767–775.

[17] S. Arora and C. Lund, "Hardness of approximations," in *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Company, 1997, pp. 399–446.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[19] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II*. Springer, 1993.

[21] J. P. Aubin and I. Ekeland, "Estimates of the duality gap in nonconvex optimization," *Math. Oper. Res.*, vol. 1, no. 3, pp. 225–245, Aug. 1976.

[22] R. M. Starr, "Quasi-equilibria in markets with non-convex preferences," *Econometrica*, vol. 37, no. 1, pp. 25–38, Jan. 1969.

[23] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[24] D. P. Bertsekas, *Convex Optimization Theory*. Athena Scientific, 2009.

[25] J. Lawrence and V. Soltan, "Carathéodory-type results for the sums and unions of convex sets," *Rocky Mt. J. Math.*, vol. 43, no. 5, pp. 1675–1688, Oct. 2013.

[26] E. Asplund, "A k-extreme point is the limit of k-exposed points," *Isr. J. Math.*, vol. 1, no. 3, pp. 161–162, Sep. 1963.

[27] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[28] D. P. Bertsekas and N. R. Sandell Jr., "Estimates of the duality gap for large-scale separable nonconvex optimization problems," in *Proc. IEEE CDC*, Dec. 1982, pp. 782–785.

[29] K. Németh, A. Kőrösi, and G. Rétvári, "Optimal OSPF traffic engineering using legacy equal cost multipath load balancing," in *Proc. IFIP Networking*, May 2013.

[30] Y. Lee, Y. Seok, Y. Choi, and C. Kim, "A constrained multipath traffic engineering scheme for MPLS networks," in *Proc. IEEE ICC*, Apr. 2002, pp. 2431–2436.

[31] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. ACM*, vol. 42, no. 6, pp. 1115–1145, Nov. 1995.

[32] L. Lovász, "On the Shannon capacity of a graph," *IEEE Trans. Inf. Theory*, vol. 25, no. 1, pp. 1–7, Jan. 1979.

[33] J. B. Lasserre, "A sum of squares approximation of nonnegative polynomials," *SIAM Rev.*, vol. 49, no. 4, pp. 651–669, Nov. 2007.

[34] M. J. Todd, "Semidefinite optimization," *Acta Numer.*, vol. 10, pp. 515–560, May 2001.

[35] B. Barak and D. Steurer. (2016) Proofs, beliefs, and algorithms through the lens of sum-of-squares. [Online]. Available: https://www.sumofsquares.org/

[36] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell, "Optimal inapproximability results for MAX-CUT and other 2-variable CSPs?" *SIAM J. Comput.*, vol. 37, no. 1, pp. 319–357, May 2007.

[37] J. Håstad, "Some optimal inapproximability results," *J. ACM*, vol. 48, no. 4, pp. 798–859, Jul. 2001.

[38] S. Mahajan and H. Ramesh, "Derandomizing approximation algorithms based on semidefinite programming," *SIAM J. Comput.*, vol. 28, no. 5, pp. 1641–1663, May 1999.

[39] U. Feige and G. Schechtman, "On the optimality of the random hyperplane rounding technique for MAX CUT," *Random Struct. Algor.*, vol. 20, no. 3, pp. 403–440, May 2002.

[40] S. A. Khot and N. K. Vishnoi, "The unique games conjecture, integrality gap for cut problems and embeddability of negative-type metrics into ℓ_1," *J. ACM*, vol. 62, no. 1, pp. 8:1–8:39, Feb. 2015.

[41] D. Grigoriev, "Complexity of Positivstellensatz proofs for the knapsack," *Comput. Complex.*, vol. 10, no. 2, pp. 139–154, Dec. 2001.

[42] B. Barak, F. G. Brandão, A. W. Harrow, J. Kelner, D. Steurer, and Y. Zhou, "Hypercontractivity, sum-of-squares proofs, and their applications," in *Proc. ACM STOC*, May 2012, pp. 307–326.

[43] A. Baernstein II and B. A. Taylor, "Spherical rearrangements, subharmonic functions, and $*$-functions in n-space," *Duke Math. J.*, vol. 43, no. 2, pp. 245–268, Jun. 1976.

[44] A. Bhaskara, S. Daruki, and S. Venkatasubramanian, "Sublinear algorithms for MAXCUT and correlation clustering," Feb. 2018, arXiv:1802.06992.

[45] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[46] S. C. Polak and A. Schrijver, "New lower bound on the Shannon capacity of C_7 from circular graphs," *Inf. Process. Lett.*, vol. 143, pp. 37–40, Mar. 2019.

[47] E. de Klerk and D. V. Pasechnik, "Approximation of the stability number of a graph via copositive programming," *SIAM J. Optim.*, vol. 12, no. 4, pp. 875–892, Mar. 2002.

[48] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, 2000.

[49] A. Schrijver, "A comparison of the Delsarte and Lovász bounds," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 425–429, Jul. 1979.

[50] R. Mittal and M. Szegedy, "Product rules in semidefinite programming," in *Proc. FCT*, Aug. 2007, pp. 435–445.