Fractional Programming for Communication System Design

by

Kaiming Shen

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Electrical and Computer Engineering
University of Toronto

Abstract

Fractional Programming for Communication System Design

Kaiming Shen
Doctor of Philosophy
Graduate Department of Electrical and Computer Engineering
University of Toronto
2020

Fractional programming (FP) refers to the class of optimization involving ratio terms. The traditional techniques for FP, *i.e.*, the Charnes-Cooper method and Dinkelbach's method, however can only be applied to the single-ratio problem (or the max-min-ratio problem). This thesis aims to develop new techniques for FP to address a broader range of optimization problems consisting of multiple ratio terms, and to investigate their applications in communication system design.

Toward this end, we propose a new technique named the quadratic transform to tackle multiple ratios—which can even be nested in some nonlinear functions. We also propose a Lagrangian dual transform in order to tackle the logarithmic ratio problems that are frequently encountered in communication system design. As a further theoretical contribution, the classic scalar FP is generalized to a multidimensional space wherein the ratios are of a matrix form. Remarkably, both of the quadratic transform and the Lagrangian dual transform can be extended to higher dimensions. Moreover, we justify the proposed FP methods by connecting them to the minorization-maximization (MM) theory.

On the application side of FP in communication system design, this thesis begins with the continuous optimization of transmit powers and beamforming vectors in the signal-to-interference-plus-noise ratio (SINR) terms. An interesting result about power control is that FP leads to a fixed-point iteration with provable convergence whereas the existing algorithms of the same type cannot guarantee convergence in general. For the discrete optimization, we advocate a novel way of using FP, thereby reformulating the complicated integer programming problem as a weighted bipartite matching problem. A critical aspect of this approach is that it encompasses the weighted minimum mean square error (WMMSE) algorithm as a special case. Finally, we demonstrate the role of matrix FP in optimizing the multi-data-stream transmission for device-to-device (D2D) systems as well as in designing the nonorthogonal pilot sequences for massive multiple-input multiple-out (MIMO).

To my parents

Acknowledgements

According to the ancient Chinese poet *Han Yu* (768—824), an excellent mentor ought to propagate "Tao" (loosely translated as doctrine), impart professional knowledge, and resolve doubts. Presumably, Han Yu himself had met such an ideal advisor, and I have the same fortune as he does. I owe a great debt of gratitude to Professor Wei Yu for his insightful criticism and advice over the past eight years, without which this thesis would not have been possible, and most importantly, for training me in the skills required for a first-rate scientist and engineer.

I wish to thank my committee members Professor Ashish Khisti and Professor Ben Liang as well as my external appraiser Professor Dongning Guo for their invaluable suggestions and insights. I am especially indebted to Professor Yonina C. Eldar for hours of stimulating discussions on pilot design—which led to the second half of Chapter 5 of the present thesis. For first seeing the connection between fractional programming and MM algorithm, I would like to thank Professor Daniel P. Palomar, who was an indispensable guide in establishing the work of Section 2.6 of the thesis. It is a particular pleasure to pay tribute to Professor Ya-Feng Liu for many constructive inputs and for being an inspiring friend. I am grateful to Professor Frank R. Kschischang for opening to me the wonderful world of error control codes and for his tremendous support in my career. I also wish to thank Professor Stark C. Draper for organizing the weekly seminar from which I have benefited a lot.

My deepest thanks also to my colleagues. I have had the good fortune to work with Reza K. Farsani on information theory, and with Wei Cui on machine learning. I wish to thank Hei Victor Cheng for assisting with the massive MIMO study of mine. It is always fun to chat with Zhilin Chen about a variety of research frontiers in our leisure time. I also sincerely thank the other group members including Liang Liu, Jinming Wen, Binbin Dai, Foad Sohrabi, Tao Jiang, Pratik Patil, and Arvin Ayoughi for their friendship. My thanks also go to the visitors Xihan Chen, Saeed R. Khosravirad, Xiaoyang Li, and Alaa Alameer Ahmad, with whom I have collaborated on various subjects.

Moreover, I want to acknowledge the University of Toronto for providing funding for this endeavor.

Finally, I wish to thank my soulmate Liyao Xiang for her encouragement and support. And I would like to thank my parents Longchi Shen and Shengfeng Ni for believing in me from the start. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

List of Acronyms

AWGN	Additive White Gaussian Noise
BCD	Block Coordinate Descent
BS	Base Station
CSI	Channel State Information
CDF	Cumulative Distribution Function
D2D	Device-to-Device
FP	Fractional Programming
GDoF	Generalized Degrees-of-Freedom
GP	Geometric Programming
i.i.d.	Independent and Identically Distributed
LP	Linear Programming
MIMO	Multiple-Input Multiple-Output
MM	Minorization-Maximization
MMSE	Minimum Mean Square Error
MRC	Maximum-Ratio Combining
MSE	Mean Square Error
PSD	Power Spectral Density
QAM	Quadrature Amplitude Modulation
SINR	Signal-to-Interference-plus-Noise Ratio
SISO	Single-Input Single-Output
TIN	Treating Interference as Noise
WMMSE	Weighted Minimum Mean Square Error

Chapter 1

Introduction

Fractional program is a family of optimization problems containing one or more ratio terms, *e.g.,* the single-ratio problem

$$\underset{x}{\text{maximize}} \quad \frac{A(x)}{B(x)} \tag{1.1a}$$

$$\text{subject to} \quad x \in \mathcal{X} \tag{1.1b}$$

and the sum-of-ratios problem

$$\underset{x}{\text{maximize}} \quad \sum_{i=1}^{n} \frac{A_i(x)}{B_i(x)} \tag{1.2a}$$

$$\text{subject to} \quad x \in \mathcal{X}, \tag{1.2b}$$

where the numerator functions $A(x)$ and $A_i(x)$ are assumed to be nonnegative while the denominator functions $B(x)$ and $B_i(x)$ are assumed to be strictly positive, along with a nonempty and compact constraint set $\mathcal{X}$ composed of a finite number of inequalities.

The study of fractional programming (FP) was arguably initiated by *John von Neumann* in his celebrated paper "A Model of General Economic Equilibrium" [1] first published in German in 1937. It has since been considered extensively in broad areas in economics, management science, information theory, optics, graph theory, and computer science [2–4]. The early works of FP concentrate on the single-ratio problem in (1.1), typically under the *concave-convex* condition (see Definition 1 in Section 2.1); the Charnes-Cooper method [5, 6] and Dinkelbach's method [7] are the standard tools in this area. Others in the existing literature seek the global optimum of the sum-of-ratios problem in (1.2) by means of branch-and-bound search [8–10]. Nevertheless, as pointed out in [9, 11], the solution to a sum-of-ratios problem with more than twenty ratios is already beyond the reach of any known algorithm within reasonable time.

This thesis comprises two main parts. In the first part we examine the theoretical basis of FP, aiming to improve upon the classic approach with two respects. *First*, we propose a new optimization technique named the quadratic transform to address the multi-ratio fractional pro-

gram (including the sum-of-ratios problem as a special case). This is in contrast to the classic techniques like the Charnes-Cooper method and Dinkelbach's method that only can be applied to the single-ratio or the max-min-ratio case. In addition, a novel Lagrangian dual transform is proposed for the logarithmic ratio problem. We then demonstrate a deep connection between the proposed methods and the minorization-maxization (MM) algorithm. *Second*, as a further theoretical contribution to FP, we propose a multidimensional generalization whereby the numerators, the denominators, and even the ratios between them can all be matrices. Remarkably, it is shown that the quadratic transform and the Lagrangian dual transform can be extended to higher dimensions.

Of equal importance are a variety of application cases of FP as shown in the second part of the thesis, ranging from power control to beamforming, energy efficiency maximization, link scheduling, spatial multiplexing, multi-data-stream device-to-device (D2D) transmission, and pilot sequence design for massive multiple-input multiple-output (MIMO). These selected examples are meant to exhibit the diversity in using the FP technique for continuous optimization, discrete optimization, and matrix optimization. It is worth highlighting the key role played by the quadratic transform in these applications. Owing to its capability to deal with a broad range of problems with multiple ratios, the quadratic transform stimulates new directions on extensive issues in communication system design, most of which have never been viewed from an FP perspective in the past research.

1.1 Motivation

What is the physical motivation of fractional term in communication systems? The present thesis is committed to answering this question. Dinkelbach's method has recently been applied in [12–15] to solve the energy efficiency maximization problem for wireless communication systems. FP is ideally suited for this problem scenario, because the objective function is already in a ratio form. In contrast, the aim of this thesis is to extend the use of FP to address a broader range of optimization problems in communication system design, particularly the ones not expressed in a single-ratio form.

We focus on communication systems in which the data rate is computed as $\log(1+\mathrm{SINR})$[1], where SINR represents the signal-to-interference-plus-noise ratio, *i.e.*,

$$\mathrm{SINR} = \frac{\text{Desired Signal Strength}}{\text{Interfering Signal Strength} + \text{Background Noise Level}}. \tag{1.3}$$

The prominent role played by SINR in communication systems makes FP an invaluable tool for network design and optimization.

Although a vast array of works already exist for FP, them mostly specialize in the *single-ratio* problem. For example, prior works on communication system design [12–15] that rely on classic

[1]For ease of notation, we use the natural logarithm in $\log(1+\mathrm{SINR})$ throughout the thesis.

FP techniques have had to confine their system models to the scenario involving only one single ratio. Although multi-ratio problems are dealt with in [16], they are restricted to some specific forms (*e.g.*, the max-min problem). System-level communication network design, however, often has to deal with multiple ratios, because the overall system performance is typically a function of multiple fractional parameters (*e.g.*, SINRs) from multiple interfering links. Solving *multiple-ratio* FP is however NP-complete [17]. The state-of-the-art methods for finding the globally optimal solution all require exponential running time (*e.g.*, using branch-and-bound search [8–10]). As to finding a stationary-point solution of the multiple-ratio problem, only general-purpose techniques such as successive convex approximation are known.

This thesis addresses the multiple-ratio FP problem from a new viewpoint. Our core theoretical contribution is a novel technique called the *quadratic transform* that introduces some suitable auxiliary variables, then recasts the original problem into a form amenable to iterative optimization. Specifically, this new technique decouples the numerator and the denominator of each ratio term, similar to the conventional *Dinkelbach's transform* (but works with multiple ratio as opposed to single ratio or max-min-ratio for the classic method). This ratio-decoupling feature of the proposed quadratic transform is particularly suited for the coordinated resource optimization across multiple cells in a wireless cellular network. For instance, the multicell power spectrum optimization is a challenging nonconvex problem, because the transmit power levels of the different links strongly impact each other through the interference terms in SINR. Our proposed FP approach decouples the signal and the interference terms of the multiple links through a set of auxiliary variables, thereby converting the original nonconvex problem into a sequence of convex problems.

In addition to SINR, the thesis further shows that the minimum mean square error (MMSE), a fundamental measure (*e.g.*, for signal inference and for channel estimation) in digital communications, is closely related to the matrix FP we have developed. Three insights are provided in this regard. First, it turns out that the well-known weighted mean square error (WMMSE) algorithm amounts to a particular way of ratio decoupling by the quadratic transform. Second, the MMSE of channel estimation for massive MIMO can be recognized as a matrix FP problem. Third, we connect the MMSE of channel estimation to the weighted sum rate maximization objective by using the Lagrangian dual transform.

1.2 Outline

This thesis offers a unifying FP framework that approaches diverse aspects of communication system design. The succeeding chapters can be divided into two parts: Chapter 2 discusses the theoretical basis of FP while Chapters 3 to 5 emphasize the application side.

Chapter 2 starts with the Charnes-Cooper method and Dinkelbach's method. Although the two classic techniques of FP can only deal with the single-ratio problem (or the max-min-ratio problem), looking back at them is worthwhile in that it reveals a crucial idea behind these

preexisting methods—*ratio decoupling*, which guides the construction of our new method. As the main result of Chapter 2, a new technique named the quadratic transform is shown to work for a much broader range of FP problems than the conventional methods, especially in the presence of multiple ratio terms. While the prior studies of multi-ratio FP typically focus on the sum-of-ratios problem, our work introduces a novel generalizations wherein the multiple ratio terms can be nested in some particular functions, and shows that the quadratic transform still works in this scenario.

In order to facilitate the optimization involving the rate expression $\log(1 + \text{SINR})$, we devise another new technique named the *Lagrangian dual transform* that is capable of moving the fractional terms to the outside of logarithm, thus reformulating the logarithmic problem as a sum-of-ratios problem. This technique is extensively used across Chapters 3 to 5 when the logarithmic rate function is involved. Taken together, the quadratic transform and the Lagrangian dual transform have a powerful synergism, especially when solving discrete optimization problems.

At this stage of the thesis we further propose an extension of FP to a multidimensional space, assuming that the numerators, the denominators, and even the ratios between them can be of a matrix form. It is a remarkable result that the quadratic transform and the Lagrangian dual transform both extends to this scenario. Moreover, we connect this matrix version of FP to the MM theory—a popular framework for nonconvex optimization, and in return justify the performance of our new techniques by taking insight from MM.

The above theoretical studies provide a new means of optimizing communication systems. Chapter 3 concentrates on a series of continuous problems: Power control, beamforming, and energy efficiency maximization. These examples illustrate that the quadratic transform enables an efficient iterative optimization algorithm with provable convergence to a stationary point. There are two results worth highlighting. First, the proposed closed-form method for power control can be interpreted as a particular way of fixed-point iteration. As compared to the existing fixed-point iteration methods in [18–20], a critical advantage of this FP approach lies in that it guarantees convergence. Second, we propose a novel idea of treating the numerator itself as an inner multiple-ratio problem nested in the outer single-ratio energy efficiency problem; this double transformation enables the energy efficiency maximization across multiple wireless links at a system level. Although using FP for energy efficiency problem is already considered in the prior works [12–15], they rely on Dinkelbach's method and thus have to limit the optimization to the link-level.

In Chapter 4 we will start a new line of research that uses the quadratic transform and the Lagrangian dual transform jointly to optimize the discrete variables in the SINR terms. Unlike the continuous problems, discrete or mixed discrete-continuous problems normally cannot be recast as convex problems. As opposed to the common heuristic of relaxing the discrete variables [21], we propose an FP-based reformulation in a weighted bipartite matching form that can be readily addressed by the standard combinatorial algorithms. We illustrate this approach by

solving the important and challenging problem of uplink coordinated multi-cell user scheduling in wireless cellular systems. Uplink scheduling is more challenging than downlink scheduling, because uplink user scheduling decisions significantly affect the interference pattern in nearby cells. A crucial insight is that the well-known WMMSE algorithm can also be recognized as a particular way of ratio decoupling. But we show that our proposed way of ratio decoupling is more suited for discrete optimization.

In Chapter 5, we shift away from the conventional scalar FP and enter the realm of matrix FP. Note that some foregoing examples already involve multidimensional variables, *e.g.*, joint scheduling and beamforming, but they still assume that the ratio terms are all scalar-valued (even though the numerators and the denominators are not necessarily scalars). We will consider joint scheduling and beamforming for a D2D system with multiple data streams transmitted on the same link, whereas the previous example in Chapter 4 assumes at most one data stream per link. As a result, the SINR term of each link is now a matrix, so the matrix FP is necessary. Furthermore, we recognize the MMSE channel estimation for massive MIMO as a matrix fractional problem. The matrix quadratic transform can be applied, for example, to the pilot sequence design. We also use the matrix Lagrangian dual transform to build a connection between the MMSE channel estimation objective and the sum rate maximization objective.

1.3 Notation

Throughout the thesis, we use bold lower-case (or upper-case) letters to denote vectors (or matrices), $\|\cdot\|$ the Euclidean norm, $\bar{(\cdot)}$ the entry-wise conjugate of vector or matrix, $(\cdot)^\top$ the transpose, $(\cdot)^H$ the conjugate transpose, $\mathrm{vec}(\cdot)$ the vectorization, $\mathrm{tr}(\cdot)$ the trace, $\otimes$ the Kronecker product, and $(\cdot)^{\frac{1}{2}}$ (or sometimes $\sqrt{\cdot}$ for ease of notation) the square root of a matrix. Let $\mathbb{E}[\cdot]$ be the expectation, $\mathbb{R}$ the set of real numbers, $\mathbb{R}_+$ the set of nonnegative numbers, $\mathbb{R}_{++}$ the set of strictly positive numbers, $\mathbb{C}^{m\times n}$ the set of $m \times n$ complex matrices, $\mathbb{H}_+^{n\times n}$ the set of $n \times n$ positive semidefinite Hermitian matrices, $\mathbb{H}_{++}^{n\times n}$ the set of $n \times n$ positive definite Hermitian matrices, $\mathrm{diag}(\cdot)$ the diagonal matrix, j the imaginary unit, $\Re$ (or $\Im$) the real (or imaginary) part of a complex number, $\boldsymbol{I}_n$ the $n\times n$ identity matrix, and $\mathcal{N}$ (or $\mathcal{CN}$) a (complex) Gaussian distribution. In addition, we use underline to denote a collection of variables, *e.g.*, $\underline{\mathbf{X}} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$. For two random variables X and Y, we use $X \perp\!\!\!\perp Y$ to denote that they are independent.

Chapter 2

Fractional Programming Theory

This chapter defines different types of fractional program and introduces new mathematical tools that form the building block of this thesis. We begin with the basic case—the single-ratio problem. Two classic techniques, the Charnes-Cooper method and Dinkelbach's method, are reviewed in brief. Although these traditional approaches are typically restricted to the single-ratio problem, they provide valuable inspiration in the development of the new technique. We thereafter propose a novel method named the *quadratic transform*. It is shown that the quadratic transform is much more capable than the Charnes-Cooper method and Dinkelbach's method in dealing with the multi-ratio problems, *e.g.*, the sum-of-ratios problem, the sum-of-functions-of-ratio problem, and the function-of-multi-ratio problem. We also devise a *Lagrangian dual transform* to address the logarithmic ratio problems that are frequently encountered in communication system design. Subsequently, we extend the above results to higher dimensions where the fractional term is of a matrix form. These concepts and ideas are for the most part new in the FP theory. The final achievement of this chapter is to demonstrate the connection between FP and the MM algorithm.

2.1 Single-Ratio Problem

Considering a nonempty and compact constraint set $\mathcal{X} \subseteq \mathbb{C}^d$ (thus composed of a finite number of inequalities), a nonnegative function A: $\mathbb{C}^d \mapsto \mathbb{R}_+$, and a strictly positive function B: $\mathbb{C}^d \mapsto \mathbb{R}_{++}$, where $d \in \mathbb{N}$, the *single-ratio* (maximization) problem is

$$\underset{\mathbf{x}}{\text{maximize}} \quad \frac{A(\mathbf{x})}{B(\mathbf{x})} \tag{2.1a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}. \tag{2.1b}$$

We then present two simple examples of the single-ratio problem as follows.

Example 1 (*Linear Single-Ratio Problem*)**.** The problem in (2.1) is said to be a linear single-ratio problem if it meets these two conditions: (i) $A(\mathbf{x})$ and $B(\mathbf{x})$ are both affine functions;

(ii) $\mathcal{X}$ consists of a set of linear constraints. This nonconvex problem can be converted to an equivalent linear programming (LP) problem [22] by using the Charnes-Cooper method [5, 6].

Example 2 (*Rayleigh Quotient*)**.** The single-ratio problem (2.1) is said to be a Rayleigh quotient problem if it can be written as

$$\underset{\mathbf{x}}{\text{maximize}} \quad \frac{\mathbf{x}^H \mathbf{M} \mathbf{x}}{\mathbf{x}^H \mathbf{x}} \tag{2.2a}$$

$$\text{subject to} \quad \mathbf{x} \neq \mathbf{0}, \tag{2.2b}$$

where $\mathbf{M} \succeq \mathbf{0}$ is a positive semidefinite complex Hermitian matrix. The solution to (2.2) can be optimally determined as $\mathbf{x}^\star = \mathbf{v}$ despite the nonconvexity, where $\mathbf{v}$ is the eigenvector of $\mathbf{M}$ corresponding to the maximum eigenvalue. The well-known principal component analysis (PCA) algorithm for data analysis builds on this result.

Furthermore, the following type of the single-ratio problem has been studied extensively in the classic literature of FP.

Definition 1 (*Concave-Convex Single-Ratio Problem*)**.** The single-ratio problem (2.1) is said to be concave-convex if $A(\mathbf{x})$ is a concave function while $B(\mathbf{x})$ is a convex function. Note that the concave-convex single-ratio problem is nonconvex in general.

2.1.1 Classic Methods

We state the Charnes-Cooper method (in 1962) and Dinkelbach's method (in 1967) in the following two theorems without proofs. They both aim to decouple the numerator function $A(\mathbf{x})$ and the denominator function $B(\mathbf{x})$ for the single-ratio problem (2.1); the merit of doing so is discussed at the end of this subsection.

Theorem 1 (*Charnes-Cooper Method* [5, 6])**.** *The single-ratio problem (2.1) is equivalent to*

$$\underset{z,\mathbf{y}}{\textit{maximize}} \quad zA\left(\frac{\mathbf{y}}{z}\right) \tag{2.3a}$$

$$\textit{subject to} \quad zB\left(\frac{\mathbf{y}}{z}\right) = 1 \tag{2.3b}$$

$$z \in \mathcal{Z} \tag{2.3c}$$

$$\mathbf{y} \in \mathcal{Y} \tag{2.3d}$$

in the sense that the optimal solution $\mathbf{x}^\star$ of (2.1) can be recovered by either

$$z = \frac{1}{B(\mathbf{x})} \tag{2.4}$$

or

$$\mathbf{y} = \frac{\mathbf{x}}{B(\mathbf{x})} \tag{2.5}$$

given the optimal solution $(z^\star, \mathbf{y}^\star)$ *of (2.3), where* $\mathcal{Z}$ *and* $\mathcal{Y}$ *are the constraint sets of* z *and* $\mathbf{y}$ *according to (2.4) and (2.5), respectively, as* $\mathbf{x} \in \mathcal{X}$.

Remark 1. The above method is first proposed by Charnes and Cooper [5] for the linear single-ratio problem then extended by Schaible [6] to the concave-convex single-ratio problem. In particular, the linear single-ratio FP in Example 1 can be recast to LP under the transformation by the Charnes-Cooper method.

Theorem 2 (*Dinkelbach's Method* [7])**.** *Consider a sequence of optimization problems*

$$\underset{\mathbf{x}}{\text{maximize}} \quad A(\mathbf{x}) - yB(\mathbf{x}) \tag{2.6a}$$
$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.6b}$$

with the auxiliary variable y *iteratively updated as*

$$y^{(t+1)} = \frac{A(\mathbf{x}^{(t)})}{B(\mathbf{x}^{(t)})}, \tag{2.7}$$

where the superscript t *is the iteration index. If* $\mathbf{x}^\star$ *is the optimal solution of (2.6) at convergence, then it is the optimal solution of (2.1) as well. Note that convergence is guaranteed by alternatively updating* y *according to (2.7) and solving for* $\mathbf{x}$ *in (2.6), because* y *is nondecreasing after each iteration. Actually,* y *converges to the optimum value of* $A(\mathbf{x})/B(\mathbf{x})$.

Observe that the Charnes-Cooper method decouples the denominator and the numerator by moving the $B(\mathbf{x})$ to the constraint (2.3b) while retaining $A(\mathbf{x})$ in the objective function, whereas Dinkelbach's method decouples $A(\mathbf{x})$ and $B(\mathbf{x})$ directly in the objective function. In comparison, the Charnes-Cooper method is more complicated in that: (i) additional constraints are introduced; (ii) $\mathcal{Z}$ and $\mathcal{Y}$ need to be characterized, which may be numerically difficult.

But what is the benefit of ratio decoupling? Consider a concave-convex single-ratio problem for example. Although the problem itself is nonconvex, the reformulated problem in (2.3) by the Charnes-Cooper method or (2.6) by Dinkelbach's method turns out to be convex under the concave-convex condition, so the optimal solution can be efficiently determined after proper transformation. Actually, the two classic techniques were devised exactly for the concave-convex single-ratio problem. In addition to the above simple showcase, the thesis will demonstrate the merits of ratio decoupling in many other problem scenarios that are neither single-ratio nor concave-convex.

2.1.2 Quadratic Transform

Classic techniques for FP work well for single-ratio problems, but they cannot be easily generalized to multiple-ratio FP. This is because although these classic transforms have the property that the original FP and the transformed problem have the same optimal solution, the optimal value of the objective function of the transformed problem is not necessarily the same as the

original FP objective function value. Thus, when multiple ratios are involved, one cannot apply the transform to each ratio individually.

The thesis proposes a new transform, which is motivated by Dinkelbach's transform, but with an added constraint that the value of the objective function must stay the same. It is named the *quadratic transform* because it involves quadratic terms.

First, we formally state the properties that the desired transformed objective function is expected to have, when reformulating the original FP objective function in (2.1):

- C1: *(Decoupling)* The new objective has the form $g(\mathbf{x}, y) = f(A(\mathbf{x}))q_1(y) + h(B(\mathbf{x}))q_2(y)$, where y is an auxiliary variable.
- C2: *(Equivalent Solution)* Variable $\mathbf{x}^\star$ maximizes $A(\mathbf{x})/B(\mathbf{x})$ if and only if $\mathbf{x}^\star$ together with some $y^\star$ maximizes $g(\mathbf{x}, y)$.
- C3: *(Equivalent Objective)* If $y^\star = \arg\max_y g(\mathbf{x}, y)$, then $g(\mathbf{x}, y^\star) = A(\mathbf{x})/B(\mathbf{x})$.
- C4: *(Concavity)* Function $g(\mathbf{x}, y)$ is concave over y for fixed $\mathbf{x}$, *i.e.*, $\partial^2 g/\partial y^2 \leq 0$.

The above four conditions are all naturally motivated. C1 and C2 follow from the idea of the classic FP transforms in order to decouple the optimization of $A(\mathbf{x})$ and $B(\mathbf{x})$ through y; C3 makes a stronger equivalence with the original problem as motivated by the desired application for multiple-ratio problems; C4 allows for convex optimization over y for fixed $\mathbf{x}$. Note that C3 implies C2 but not vice versa. In fact, Dinkelbach's transform satisfies C1, C2 and C4, but does not satisfy C3. (Specifically, at the optimum, Dinkelbach's transform has $y^\star = A(\mathbf{x})/B(\mathbf{x})$ according to (2.7), therefore its $g(\mathbf{x}, y^\star) = 0$.)

The new technique named the *quadratic transform* meets all these conditions C1-C4, as stated in the following theorem.

Theorem 3 (*Quadratic Transform*)**.** *The quadratic transform*

$$g(\mathbf{x}, y) = 2y\sqrt{A(\mathbf{x})} - y^2 B(\mathbf{x}) \tag{2.8}$$

satisfies the conditions C1-C4. Further, if C4 is strengthened to require that $\partial^2 g/\partial y^2$ is independent of y, then any $g(\mathbf{x}, y)$ that satisfies C1-C4 must be of the form

$$g(\mathbf{x}, y) = 2(t_1 y + t_2)\sqrt{A(\mathbf{x})} - (t_1 y + t_2)^2 B(\mathbf{x}) \tag{2.9}$$

for some $t_1 \neq 0$ and some $t_2 \in \mathbb{R}$. Thus, the proposed quadratic transform is without loss of generality up to an affine transformation in y.

Proof. The equivalence between (2.1) and (2.8) can be readily verified. When $\mathbf{x}$ is fixed, observe that $g(\mathbf{x}, y)$ is a concave function of y, so y can be optimally determined by setting $\partial g/\partial y$ to zero, *i.e.*, $y^\star = \sqrt{A(\mathbf{x})}/B(\mathbf{x})$. Substituting this $y^\star$ expression into $g(\mathbf{x}, y)$ recovers the original objective function $A(\mathbf{x})/B(\mathbf{x})$ and thus establishes the equivalence. The proof of the uniqueness is relegated to Appendix A. □

The quadratic form of g is advocated here in order to let $\partial^2 g/\partial y^2$ be a constant given $\mathbf{x}$, thereby making the condition C4 easier to verify. But we remark that there may exist other ways of constructing g.

2.2 Multi-Ratio Problems

We now proceed to the multi-ratio FP by considering n pairs of the nonnegative numerator function $A_i(\mathbf{x}) \geq 0$ and the strictly positive denominator function $B_i(\mathbf{x}) > 0$, $i = 1, \ldots, n$. The concave-convex condition is then extended accordingly.

Definition 2 (*Concave-Convex Condition*)**.** A multi-ratio problem is said to be concave-convex if each $A_i(\mathbf{x})$ is a concave function while each $B_i(\mathbf{x})$ is a convex function.

Four types of multi-ratio problem are discussed in what follows: Max-min-ratio problem, sum-of-ratios problem, the sum-of-functions-of-ratio problem, and the function-of-multi-ratio problem. The latter two types have not yet been studied in the prior works of FP. The goal of this section is to show that the quadratic transform in Theorem 3 can be readily extended to all the above multi-ratio problems, whereas the classic Dinkelbach's method is limited to the max-min-ratio scenario.

2.2.1 Max-Min-Ratio Problem

With a nonempty and compact constraint set $\mathcal{X}$ imposed on the variable $\mathbf{x}$, the max-min-ratio problem is

$$\underset{\mathbf{x}}{\text{maximize}} \qquad \min_i \left\{ \frac{A_i(\mathbf{x})}{B_i(\mathbf{x})} \right\} \tag{2.10a}$$

$$\text{subject to} \qquad \mathbf{x} \in \mathcal{X}. \tag{2.10b}$$

We remark that the above problem was introduced by von Neumann to model the *economic growth* over time [1, 22].

Actually, the max-min-ratio problem is a rare case of the multi-ratio FP for which the classic Dinkelbach's method can be easily generalized, as stated in the following theorem.

Theorem 4 (*Generalized Dinkelbach's Method* [16])**.** *Consider a sequence of problems*

$$\underset{\mathbf{x}}{\textit{maximize}} \qquad \min_i \left\{ A_i(\mathbf{x}) - y B_i(\mathbf{x}) \right\} \tag{2.11a}$$

$$\textit{subject to} \qquad \mathbf{x} \in \mathcal{X} \tag{2.11b}$$

with the auxiliary variable y iteratively updated as

$$y^{(t+1)} = \min_i \left\{ \frac{A_i(\mathbf{x}^{(t)})}{B_i(\mathbf{x}^{(t)})} \right\}, \tag{2.12}$$

where the superscript t is the iteration index. If $\mathbf{x}^\star$ is the optimal solution of (2.11) at convergence, then it must be the optimal solution of the max-min-ratio problem in (2.10) as well. Note that convergence is guaranteed by alternatively updating $\underline{y}$ according to (2.12) and solving for $\mathbf{x}$ in (2.11), because $\underline{y}$ is nondecreasing after each iteration.

We further show that the new technique, the quadratic transform, is capable of decoupling multiple ratios simultaneously for the max-min-ratio problem as well.

Corollary 1. *The max-min-ratio problem (2.10) is equivalent to*

$$\underset{\mathbf{x},\,\underline{y}}{\text{maximize}} \quad \min_i \left\{ 2y_i\sqrt{A_i(\mathbf{x})} - y_i^2 B_i(\mathbf{x}) \right\} \tag{2.13a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.13b}$$

$$y_i \in \mathbb{R},\ \forall i \tag{2.13c}$$

in the sense that $\mathbf{x}^\star$ is the optimal solution of (2.10) if and only if it is the optimal solution of (2.13) with some $\underline{y}^\star$.

Observe that Dinkelbach's method requires only one auxiliary variable but the quadratic transform introduces an auxiliary variable y_i for each ratio term, so Dinkelbach's method could be preferable when dealing with the max-min-ratio problem.

Because the max-min-ratio problem is just to maximize the pointwise minimum of multiple ratio terms, the solution of the single-ratio problem can be easily extended to this particular structure. Nevertheless, when facing the subsequent multi-ratio problems, the classic Dinkelbach's method no longer works and yet the quadratic transform is still applicable.

2.2.2 Sum-of-Ratios Problem

Again, assume a sequence of the numerator functions $A_i(\mathbf{x})$ and the denominator functions $B_i(\mathbf{x})$, $i = 1, \ldots, n$, along with a nonempty and compact constraint set $\mathcal{X}$. We define the *sum-of-ratios* problem as

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{n} \frac{A_i(\mathbf{x})}{B_i(\mathbf{x})} \tag{2.14a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}. \tag{2.14b}$$

The state-of-the-art methods for finding the globally optimal solution all require exponential running time (*e.g.*, using branch-and-bound search [8–10]). In fact, as pointed out in [9] and [11], the solution to a general FP problem consisting of more than twenty ratio terms is already beyond the reach of known approaches within reasonable time. As to finding stationary-point solution of the multiple-ratio problem, only general-purpose techniques such as successive convex approximation are known. It is worth mentioning that combining multiple ratios into

a single ratio (by finding a common denominator) does not make problem easier in general because it typically cannot preserve the concave-convex condition.

The quadratic transform in Theorem 3 can be readily extended for the sum-of-ratios problem due to C3 as shown below.

Corollary 2. *The sum-of-ratios problem (2.14) is equivalent to*

$$\underset{\mathbf{x},\underline{y}}{\text{maximize}} \quad \sum_{i=1}^{n}\Big(2y_i\sqrt{A_i(\mathbf{x})}-y_i^2B_i(\mathbf{x})\Big) \tag{2.15a}$$

$$\text{subject to} \quad \mathbf{x}\in\mathcal{X} \tag{2.15b}$$

$$y_i\in\mathbb{R},\ \forall i \tag{2.15c}$$

in the sense that $\mathbf{x}^\star$ *is the optimal solution of (2.14) if and only if it is the optimal solution of (2.15) with some* $\underline{y}^\star$.

Condition C3 is critical for extending the idea of decoupled optimization of numerators and denominators to the sum-of-ratios problem. As mentioned before, Dinkelbach's transform does not satisfy C3. Without the equivalence in the optimal objective function value, it is normally difficult to extend Dinkelbach's transform to the multiple-ratio case (except in special cases such as the max-min problem [16]). A straightforward extension of Dinkelbach's transform, such as

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{n}\big(A_i(\mathbf{x})-y_iB_i(\mathbf{x})\big) \tag{2.16a}$$

$$\text{subject to} \quad \mathbf{x}\in\mathcal{X} \tag{2.16b}$$

with the auxiliary variable y_i iteratively updated as the previous A_i/B_i, does not guarantee the equivalence to (2.14).

Although not immediately recognized at the time the work of the quadratic transform was first published [23], the above quadratic transform is akin to the earlier work of Benson [8, 24], as restated below.

Proposition 1 (*Benson's Transform* [8, 24])**.** *The sum-of-ratios problem (2.14) is equivalent to*

$$\underset{\mathbf{x},\underline{u},\underline{v}}{\text{maximize}} \quad \sum_{i=1}^{n}\Big(2u_i\sqrt{A_i(\mathbf{x})}-v_iB_i(\mathbf{x})\Big) \tag{2.17a}$$

$$\text{subject to} \quad \mathbf{x}\in\mathcal{X} \tag{2.17b}$$

$$u_i^2-v_i\le 0,\ \forall i \tag{2.17c}$$

in the sense that $\mathbf{x}^\star$ *is the optimal solution of (2.14) if and only if it is the optimal solution of (2.17) with some* $(\underline{u}^\star,\underline{v}^\star)$.

The transform (2.17) is proposed by Benson [8,24] in order to facilitate a branch-and-bound search for the global optimum of the sum-of-ratios problem. It can be shown that at the optimum, we must have $u_m^2 = v_m$, thus if we had made them equal *a priori*, this reduces to the quadratic transform. Finally, we remark that the Benson's transform does not apply to the sum-of-ratios minimization problem.

2.2.3 Sum-of-Functions-of-Ratio Problem

We now assume that the ratio terms can be wrapped in some functions. With a sequence of nondecreasing functions $f_i(\cdot)$, the sum-of-functions-of-ratio problem is

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{n} f_i\left(\frac{A_i(\mathbf{x})}{B_i(\mathbf{x})}\right) \tag{2.18a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}. \tag{2.18b}$$

The above type of problem has not yet been considered in the existing works of FP. We are motivated to study it by the capacity function $\log(1 + \text{SINR})$, which is common in the communication system but has never been viewed from an FP perspective. It turns out that the quadratic transform still works for the sum-of-functions-of-ratio problem, as stated below.

Corollary 3. *The sum-of-functions-of-ratio problem (2.18) is equivalent to*

$$\underset{\mathbf{x},\, \underline{y}}{\textit{maximize}} \quad \sum_{i=1}^{n} f_i\left(2y_i\sqrt{A_i(\mathbf{x})} - y_i^2 B_i(\mathbf{x})\right) \tag{2.19a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.19b}$$

$$y_i \in \mathbb{R},\ \forall i \tag{2.19c}$$

in the sense that $\mathbf{x}^\star$ is the optimal solution of (2.18) if and only if it is the optimal solution of (2.19) with some $\underline{y}^\star$.

Proof. We first rewrite the problem (2.18) as $\max_{\mathbf{x},\, \underline{r}} \sum_{i=1}^{n} f_i(r_i)$ subject to $\mathbf{x} \in \mathcal{X}$ and $r_i = A_i(\mathbf{x})/B_i(\mathbf{x})$; because of the condition C3, variable r_i can be replaced with $\max_y \left(2y_i\sqrt{A_i(\mathbf{x})} - y_i^2 B_i(\mathbf{x})\right)$; further, since f_i is nondecreasing, $\max_{\mathbf{x}} \sum_{i=1}^{n} f_i\big(\max_{\underline{y}}(2y_i\sqrt{A_i(\mathbf{x})} - y_i^2 B_i(\mathbf{x}))\big)$ can be rewritten as in (2.19a) by merging $\max_{\mathbf{x}}$ and $\max_{\underline{y}}$. □

Finally, we remark that the proposed method does not work for the sum-of-ratios minimization problem in general.

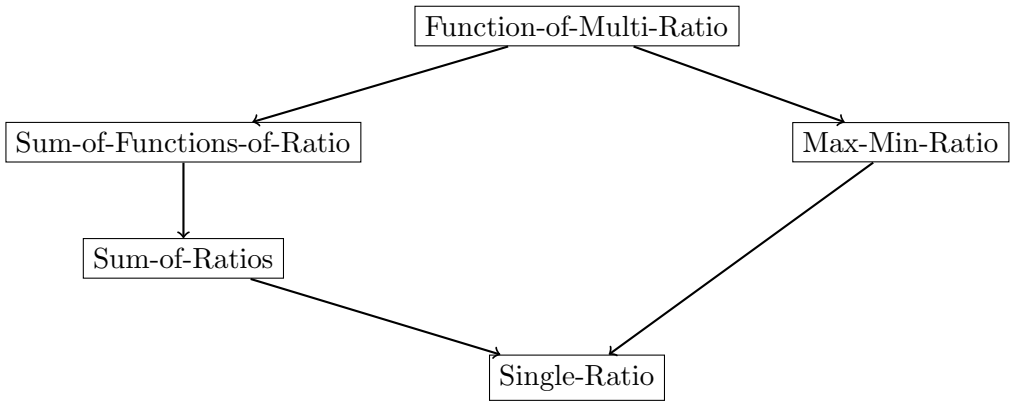


Figure 2.1: Relations between different types of FP problems. We use $\clubsuit \to \diamondsuit$ to indicate that $\diamondsuit$ is a special case of $\clubsuit$.

2.2.4 Function-of-Multi-Ratio Problem

The sum-of-functions-of-ratio problem can be generalized further as

$$\underset{\mathbf{x}}{\text{maximize}} \quad f\left(\frac{A_1(\mathbf{x})}{B_1(\mathbf{x})}, \ldots, \frac{A_n(\mathbf{x})}{B_n(\mathbf{x})}\right) \tag{2.20a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{2.20b}$$

where $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a decreasing function in the sense that $f(r_1, \ldots, r_n) \geq f(r'_1, \ldots, r'_n)$ given any $(r_1, \ldots, r_n) \succeq (r'_1, \ldots, r'_n)$. Observe that the *function-of-multi-ratio* problem encompasses the sum-of-functions-ratio problem and the max-min-ratio problem as special cases. The relations between the different FP problems are summarized in Fig. 2.1 displayed on the next page.

The following corollary shows that the quadratic transform can be further extended to the function-of-multi-ratio scenario.

Corollary 4. *The function-of-multi-ratio problem (2.20) is equivalent to*

$$\underset{\mathbf{x}, \underline{y}}{\textit{maximize}} \quad f\left(2y_1\sqrt{A_1(\mathbf{x})} - y_1^2 B_1(\mathbf{x}), \ldots, 2y_n\sqrt{A_n(\mathbf{x})} - y_n^2 B_n(\mathbf{x})\right) \tag{2.21a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.21b}$$

$$y_i \in \mathbb{R}, \ \forall i \tag{2.21c}$$

in the sense that $\mathbf{x}^\star$ is the optimal solution of (2.20) if and only if it is the optimal solution of (2.21) with some $\underline{y}^\star$. The proof is omitted since it resembles that of Corollary 3.

Observe that the above different forms of the quadratic transform are in essence of the same structure. This observation leads us to a unifying algorithmic framework for multi-ratio FP as stated in the next section.

Algorithm 1: Iterative Optimization

1 Initialize the variable $\mathbf{x}$ to a feasible value;
2 **repeat**
3 Update $\underline{y}$ by (2.22);
4 Update $\mathbf{x}$ by solving the new problem obtained by the quadratic transform, with the formerly updated $\underline{y}$ substituted in;
5 **until** *the value of the new objective function converges*;

2.3 Iterative Optimization via Quadratic Transform

Equipped with the above different forms of quadratic transform, we now consider optimizing the original variable $\mathbf{x}$ and the auxiliary variable $\underline{y}$ alternatively in terms of the new objective function. In particular, when $\mathbf{x}$ is held fixed, $\underline{y}$ can be optimally determined by the condition C3 in Section 2.1.2, that is

$$y^\star = \frac{\sqrt{A(\mathbf{x})}}{B(\mathbf{x})} \quad \text{or} \quad y_i^\star = \frac{\sqrt{A_i(\mathbf{x})}}{B_i(\mathbf{x})}. \tag{2.22}$$

This iterative optimization is summarized in Algorithm 1 as displayed on the next page.

2.3.1 Optimality Analysis

We examine the performance of Algorithm 1 under the different problem settings.

Theorem 5 (*Monotonic Increment*)**.** *For the function-of-multi-ratio problem (hence including the special cases like the sum-of-functions-of-ratio problem, the sum-of-ratios problem, the max-min-ratio problem, and the single-ratio problem), the value of the original objective function is monotonically nondecreasing after each iteration of Algorithm 1.*

Proof. The above result can be obtained directly by using the condition C3. And we will see this verified alternatively from an MM algorithm perspective in Section 2.6. □

Theorem 6 (*Stationary Point*)**.** *If the FP problem further meets the concave-convex condition, i.e., when every $A_i(\mathbf{x})$ is concave and every $B_i(\mathbf{x})$ is convex along with a convex constraint set $\mathcal{X}$, then Algorithm 1 consists of a sequence of convex optimizations in step 4 that lead to a stationary point of (2.18).*

Proof. The algorithm is basically a block coordinate ascent algorithm for the reformulated problem (2.19), which is a convex optimization problem due to the concave-convex form of (2.18), so it converges to a stationary point $(\mathbf{x}^\star, \underline{y}^\star)$ of (2.19). Due to the equivalence in the solution (namely the condition C2) and the equivalence in the objective value (namely the condition C3), the first-order condition on $\mathbf{x}$ for (2.19) under the optimal $\underline{y}^\star$ is the same as for

the original problem (2.18), hence the algorithm also converges to a stationary point of (2.18). We remark that the above result can be also proved by using an MM interpretation as shown in Section 2.6. □

In comparison to the gradient-based approach, the proposed iterative algorithm provides efficient closed-form optimization and thus gets rid of the step size tuning. Regarding how far the stationary point is from the global optimum, it is theoretically difficult to provide a bound in general, but we mention that the existing analysis based on the MM theory [25] applies to our method because of an MM interpretation as shown later in Section 2.6.

Theorem 7 (*Global Optimality*)**.** *For the single-ratio problem (2.1) and the max-min-ratio (2.10) problem with differentiable $A(\mathbf{x})$ and $B(\mathbf{x})$, Algorithm 1 converges to the globally optimal solution of the respective problems so long as the concave-convex condition is satisfied.*

Proof. For the singel-ratio problem, the key is to verify that any stationary point must be the local optimum in the special cases of single-ratio or max-min problems. This can be established by showing that the concave-convex single-ratio FP problem is *pseudo-convex* (see Appendix B). Further, since the problem has only one local optimum, it must be global optimum. This fact has been proved in [26] for the case where $A(\mathbf{x})$ and $B(\mathbf{x})$ are differentiable and $A(\mathbf{x})$ is concave and $B(\mathbf{x})$ is convex. Thus for single-ratio FP, Algorithm 1 converges to a global optimum. Furthermore, by the result in [27] that any local optimum solution is also the global optimum solution for the problem max $\min_i\{f_i\}$ given that each f_i is a pseudoconcave function, the global optimality is established in the max-min-ratio case. □

Fig. 2.2 shows an example of a single-ratio concave-convex FP problem whose unique local optimum is the global optimum. We note that this property of converging to the globally optimal solution holds also for the Charnes-Cooper method and Dinkelbach's method. This is true despite that the original problem is not necessarily convex.

2.3.2 Rate of Convergence

We analyze the convergence speed of Algorithm 1 as compared to the classic transforms. Note that if the single-ratio problem is concave-convex, solving the problem by Dinkelbach's transform amounts to a sequence of convex optimizations (2.6) over $\mathbf{x}$ with the auxiliary variable y iteratively updated by (2.7). It is shown in [26] that the iteration by Dinkelbach's transform converges at a superlinear rate, *i.e.*,

$$\lim_{t\to\infty} \frac{|y^\star - y_{t+1}|}{|y^\star - y_t|} = 0 \tag{2.23}$$

where subscript t is the index of iteration, and $y^\star$ is the auxiliary variable value at the convergence. For ease of comparison, we evaluate the convergence of Algorithm 1 for the single-ratio

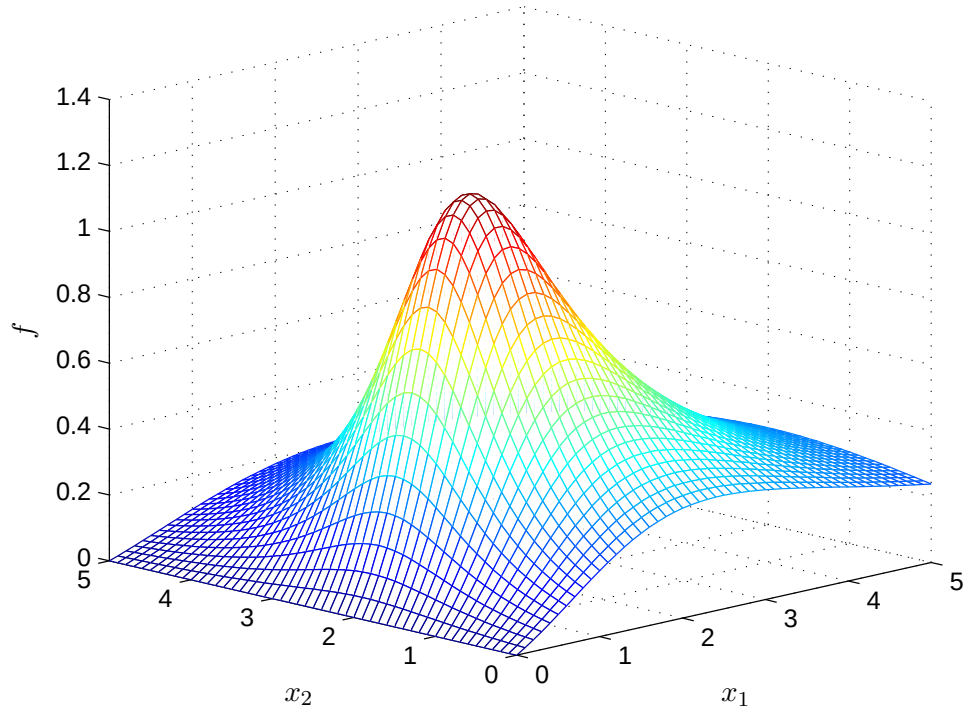


Figure 2.2: Maximizing $f(x_1, x_2) = x_1/((x_1-1)^2 + (x_2-2)^2 + 1)$ over $x_1 \geq 0$ and $x_2 \geq 0$ is a single-ratio concave-convex FP problem. Although $f(x_1, x_2)$ is not concave, its local optimum is also the global optimum.

problem as well. As compared to Dinkelbach's transform, the quadratic transform (*i.e.*, Algorithm 1) can be considerably slower. The following example shows that the convergence rate of Algorithm 1 can be strictly slower than superlinear.

Example 3 (*Quadratic Transform is Slower than Dinkelbach's Method*)**.** Consider an example of the single-ratio concave-convex problem

$$\underset{x}{\text{maximize}} \quad \frac{x}{x^2+1} \tag{2.24a}$$

$$\text{subject to} \quad x \geq 0. \tag{2.24b}$$

The quadratic transform reformulates its objective function as

$$g(x, y) = 2y\sqrt{x} - y^2(x^2+1). \tag{2.25}$$

Introduce a subscript t to denote the iteration index. When x is set to some x_t, the auxiliary variable y is optimally updated according to (2.22), *i.e.*,

$$y_{t+1} = \frac{\sqrt{x_t}}{x_t^2+1}. \tag{2.26}$$

After y is updated to y_{t+1}, the optimal x is

$$x_{t+1} = (2y_{t+1})^{-\frac{2}{3}}. \tag{2.27}$$

Combined together, the above two equations yield an iterative update of y:

$$y_{t+1} = \frac{(2y_t)^{-\frac{1}{3}}}{(2y_t)^{-\frac{4}{3}} + 1}. \tag{2.28}$$

With y initialized to 0.1 (so $y_0 = 0.1$), it can be shown that y_{t+1} in (2.28) converges to $\frac{1}{2}$ in a nondecreasing fashion. We then have

$$\begin{aligned}
&\lim_{t\to\infty} \frac{|y^\star - y_{t+1}|}{|y^\star - y_t|} \\
&= \lim_{t\to\infty} \frac{y^\star - y_{t+1}}{y^\star - y_t} && \text{(2.29a)} \\
&= \lim_{y\to\frac{1}{2}} \frac{1}{\frac{1}{2} - y} \left(\frac{1}{2} - \frac{(2y)^{-\frac{1}{3}}}{(2y)^{-\frac{4}{3}} + 1} \right) && \text{(2.29b)} \\
&= \frac{1}{3}. && \text{(2.29c)}
\end{aligned}$$

Thus, Algorithm 1 in this example converges more slowly than the iterative optimization based on Dinkelbach's transform. The convergence of these two methods is illustrated in Fig. 2.3.

Moreover, we will see an MM interpretation of our FP methods in Section 2.5, so it is worth mentioning that the *spectral radius*

$$\rho = 1 - \min_{\mathbf{u}\neq 0} \frac{\mathbf{u}^H \cdot \nabla^2 f(\mathbf{x}) \cdot \mathbf{u}}{\mathbf{u}^H \cdot \nabla^2 g(\mathbf{x}|\hat{\mathbf{x}}_\infty) \cdot \mathbf{u}} \tag{2.30}$$

has been proposed as a metric reflecting the rate of convergence for the MM algorithm [28]. In principle, ρ reflects how well the surrogate function $g(\mathbf{x}|\hat{\mathbf{x}})$ approximates the original objective function $f(\mathbf{x})$ in terms of the second moment—smaller ρ indicates tighter approximation and thus faster convergence. However, this type of analysis has limited value in our problem case because: (i) it requires the updating function of $\mathbf{x}$ to be differentiable whereas our problem involves discrete variables; (ii) computing ρ entails solving a difficult nonconvex problem; (iii) it only characterizes the local convergence in the proximity of $\mathbf{x}_\infty$.

It is stressed that although the conventional Dinkelbach's transform can result in a faster convergence rate than the proposed quadratic transform, the use of the former technique is restricted to the single-ratio problem whereas the latter is capable of dealing with multiple ratios. Further, for multiple-ratio FP problems where global convergence is not guaranteed, slower convergence can sometime be advantageous as it allows the algorithm to more fully explore the solution space.

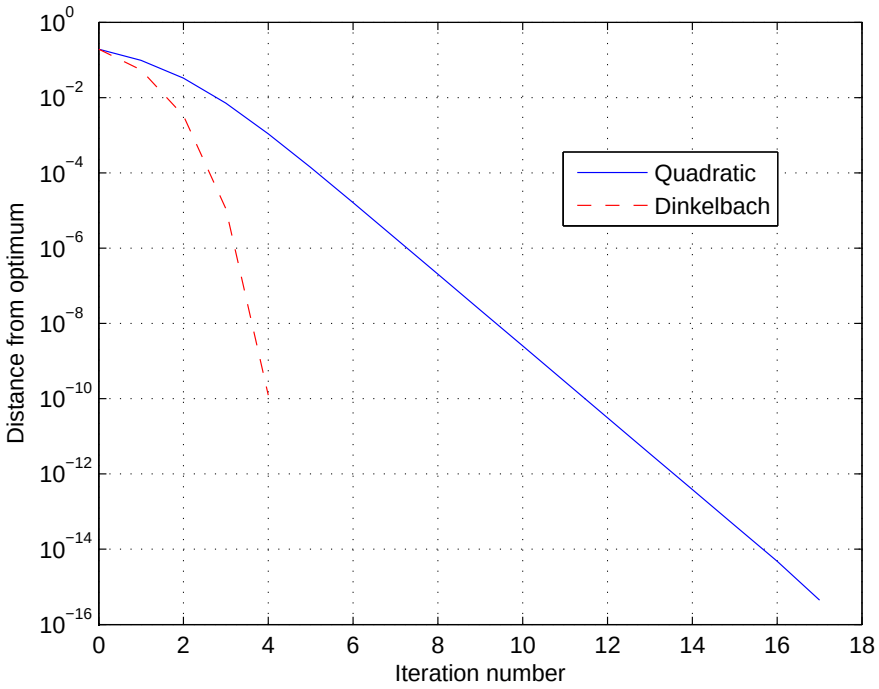


Figure 2.3: When applied to the single-ratio problem (2.24), Dinkelbach's method converges faster than the quadratic transform.

2.4 Lagrangian Dual Transform

The quadratic transform as stated in the former section is the core FP technique used for treating the continuous problems. When it comes to the discrete problems of user scheduling, we need to introduce a new FP technique named the *Lagrangian dual transform*. Its main role is to reformulate the $\log(1 + \text{SINR})$ maximization problem as a weighted bipartite matching problem in conjunction with the quadratic transform, typically applied to the scenario where the numerator and denominator of SINR are both a linear combination of the optimizing variable $\mathbf{x}$.

Optimization problem for communication system design often involves data rates expressed as logarithmic functions of SINR, *i.e.*, $\log(1 + \text{SINR})$. We propose two different approaches for applying FP to such problems. In the direct FP, the quadratic transform is immediately applied to the log-function of the ratio to decouple the numerator and denominator, while in the closed-form FP, a Lagrangian dual transform is first applied to take the ratio out of the logarithm. For continuous optimization problems, the two approaches give comparable performance. However, for discrete scheduling problems involving $\log(1 + \text{SINR})$, the second approach of using Lagrangian dual transform becomes indispensable.

We develop the Lagrangian dual transform technique that accomplishes the task of "moving" SINR to the outside of logarithm. This technique plays a crucial role in addressing the

discrete scheduling problems, because it allows a subsequent quadratic transform to express all optimization variables in linear terms. This section gives a detailed derivation of the Lagrangian dual transform technique with a constructive proof of the main result.

The target problem is a weighted sum-of-logarithms maximization:

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{n} w_i \log\left(1 + \frac{A_i(\mathbf{x})}{B_i(\mathbf{x})}\right) \tag{2.31a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{2.31b}$$

where w_i's are nonnegative weights, $A_i(\mathbf{x})$'s are nonnegative functions and $B_i(\mathbf{x})$'s are positive functions for all i, and $\mathcal{X}$ is a nonempty constraint set. The above formulation is often used to model the weighted sum rate maximization problem of a communication network. The ratio $A_i(\mathbf{x})/B_i(\mathbf{x})$ can be physically interpreted as the SINR term. The problem (2.31) has no known convex reformulation. Further, the constraint represented by $\mathcal{X}$ is not necessarily compact, *i.e.*, the variable $\mathbf{x}$ may be discrete or mixed discrete-continuous.

The main result is the following Lagrangian dual transform capable of converting (2.31) to a sum-of-ratios form.

Theorem 8 (*Lagrangian Dual Transform*)**.** *The weighted sum-of-logarithms problem (2.31) is equivalent to*

$$\underset{\mathbf{x},\,\underline{\gamma}}{\textit{maximize}} \quad f_r(\mathbf{x}, \underline{\gamma}) \tag{2.32a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{2.32b}$$

where the new objective function f_r is defined as

$$f_r(\mathbf{x}, \underline{\gamma}) = \sum_{i=1}^{n} w_i \log(1+\gamma_i) - \sum_{i=1}^{n} w_i \gamma_i + \underbrace{\sum_{i=1}^{n} \frac{w_i(1+\gamma_i)A_i(\mathbf{x})}{A_i(\mathbf{x}) + B_i(\mathbf{x})}}_{\textit{Sum-of-ratio term}}. \tag{2.33}$$

The two problems are equivalent in the sense that $\mathbf{x}$ is the solution to (2.31) if and only if it is the solution to (2.32) with some $\underline{\gamma}^\star$, and the optimal objective values of these two problems are also equal.

Proof. Observe that f_r is a concave differentiable function over $\underline{\gamma}$ when $\mathbf{x}$ is held fixed, so $\underline{\gamma}$ can be optimally determined by setting each $\partial f_r/\partial \gamma_i$ to zero, *i.e.*, $\gamma_i^\star = A_i(\mathbf{x})/B_i(\mathbf{x})$. Substituting this $\gamma_i^\star$ back in f_r recovers the weighted sum-of-logarithms objective function in (2.31a) exactly. The equivalence is therefore established. □

To provide insight on how the above transform is obtained, we revisit the weighted sum-of-logarithms problem (2.31) from a Lagrangian dual perspective, and provide an alternative constructive proof of Theorem 8.

First, by introducing a new variable γ_i to replace each ratio term inside the logarithm, (2.31) can be rewritten as

$$\begin{aligned} \underset{\mathbf{x},\underline{\gamma}}{\text{maximize}} \quad & \sum_{i=1}^{n} w_i \log\left(1+\gamma_i\right) && \text{(2.34a)} \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X} && \text{(2.34b)} \\ & \gamma_i \leq \frac{A_i(\mathbf{x})}{B_i(\mathbf{x})},\ \forall i. && \text{(2.34c)} \end{aligned}$$

The above optimization can be thought of as an outer optimization over $\mathbf{x}$ and an inner optimization over γ_i with fixed $\mathbf{x}$. The inner optimization given $\mathbf{x}$ is formulated as

$$\begin{aligned} \underset{\underline{\gamma}}{\text{maximize}} \quad & \sum_{i=1}^{n} w_i \log(1+\gamma_i) && \text{(2.35a)} \\ \text{subject to} \quad & \gamma_i \leq \frac{A_i(\mathbf{x})}{B_i(\mathbf{x})},\ \forall i. && \text{(2.35b)} \end{aligned}$$

The solution to this inner optimization is obviously that γ_i should satisfy (2.35b) with equality. But, let's express the problem in a different way. Note that (2.35) is a convex optimization in $\underline{\gamma}$, so the *strong duality* [22] holds. Introduce the dual variable λ_i for each inequality constraint in (2.35b) and form the Lagrangian function

$$L(\underline{\gamma},\underline{\lambda}) = \sum_{i=1}^{n} w_i \log(1+\gamma_i) - \sum_{i=1}^{n} \lambda_i \left(\gamma_i - \frac{A_i(\mathbf{x})}{B_i(\mathbf{x})}\right). \tag{2.36}$$

Due to strong duality, the optimization (2.35) is equivalent to the dual problem

$$\underset{\underline{\lambda}\succeq\mathbf{0}}{\text{minimize}} \quad \underset{\underline{\gamma}}{\text{maximize}} \quad L(\underline{\gamma},\underline{\lambda}), \tag{2.37}$$

where $\succeq$ is the pointwise greater-than-or-equal-to symbol.

Let $(\underline{\gamma}^\star,\underline{\lambda}^\star)$ be the saddle point of the above. It must satisfy the first-order condition $\partial L/\partial\gamma_i = 0$:

$$\lambda_i^\star = \frac{w_i}{1+\gamma_i^\star}. \tag{2.38}$$

But from the trivial solution to the optimization problem (2.35), we already know that $\gamma_i^\star = A_i(\mathbf{x})/B_i(\mathbf{x})$, so

$$\lambda_i^\star = \frac{w_i B_i(\mathbf{x})}{A_i(\mathbf{x})+B_i(\mathbf{x})}. \tag{2.39}$$

Note that $\lambda_i^\star \geq 0$ is automatically satisfied here. Using (5.33) in (2.37), problem (2.35) can then be reformulated as

$$\underset{\underline{\gamma}}{\text{maximize}} \quad L(\underline{\gamma},\underline{\lambda}^\star). \tag{2.40}$$

Furthermore, combining with the outer maximization over $\mathbf{x} \in \mathcal{X}$ and after some algebra, we

find (2.40) to be exactly the same as the maximization of (2.33) in Theorem 8.

We remark that the Lagrangian dual transform can also be interpreted as constructing a surrogate function from an MM perspective as specified in Section 2.6. Furthermore, if ratio is nested in some other functions, then we need to change the surrogate function accordingly.

2.5 Matrix Fractional Programming

Generalization of the conventional scalar FP to the multidimensional space constitutes another major contribution of this chapter. Two cases are examined: (i) the numerator is vector and the denominator is matrix, so the ratio is still scalar; (ii) the numerator and the denominator are both matrices, so their ratio is also a matrix.

2.5.1 Vector Numerators & Matrix Denominators

We first consider FP assuming that the numerators are vectors and the denominators are matrices. This class of FP arises in dealing with multi-antenna communication systems. Given a sequence of function $\mathbf{a}_i(\mathbf{x}) \in \mathbb{C}^d$ and function $\mathbf{B}_i(\mathbf{x}) \in \mathbb{H}_{++}^{d\times d}$, for $i = 1, \ldots, n$, along with a nonempty and compact constraint set $\mathcal{X}$, where $d \in \mathbb{N}$, a sum-of-ratios problem with vector numerators and matrix denominators is is

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{n} \mathbf{a}_i^H(\mathbf{x})\mathbf{B}_i^{-1}(\mathbf{x})\mathbf{a}_i(\mathbf{x}) \tag{2.41a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}. \tag{2.41b}$$

The extension of the quadratic transform for this multidimensional case is stated below.

Theorem 9 (*Vector Quadratic Transform*)**.** *The vector-numerator-and-matrix-denominator single-ratio problem (2.41) is equivalent to*

$$\underset{\mathbf{x},\,\underline{\mathbf{y}}}{\textit{maximize}} \quad \sum_{i=1}^{n} \left(2\Re\left\{\mathbf{y}_i^H\mathbf{a}_i(\mathbf{x})\right\} - \mathbf{y}_i^H\mathbf{B}_i(\mathbf{x})\mathbf{y}_i\right) \tag{2.42a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.42b}$$

$$\mathbf{y}_i \in \mathbb{C}^d,\ \forall i. \tag{2.42c}$$

Proof. Recognize each term in the summation of (2.42a) as $\mathbf{y}_i^H\mathbf{a}_i + \mathbf{a}_i^H\mathbf{y}_i - \mathbf{y}_i^H\mathbf{B}_i\mathbf{y}_i$ and then further rewrite it as $\mathbf{a}_i^H\mathbf{B}_i^{-1}\mathbf{a}_i - (\mathbf{y}_i - \mathbf{B}_i^{-1}\mathbf{a}_i)^H\mathbf{B}_i(\mathbf{y}_i - \mathbf{B}_i^{-1}\mathbf{a}_i)$ by completing the square. It is easy to see that the optimal solution of (2.42) is $\mathbf{y}_i^\star = \mathbf{B}_i^{-1}(\mathbf{x})\mathbf{a}_i(\mathbf{x})$ and the optimal value of (2.42a) equals to $\mathbf{a}_i^H\mathbf{B}_i^{-1}\mathbf{a}_i$ exactly. The equivalence to (2.41) is therefore established. □

The above transformation can be extended to the sum-of-functions-of-ratio problem and the function-of-multi-ratio problem in Section 2.2. Theorem 9 focuses on the sum-of-ratios

case because that is the main type of matrix FP problem we will face when dealing with spatial multiplexing in wireless networks. Moreover, the Lagrangian dual transform can also be extended for the vector-numerator-and-matrix-denominator scenario, as stated in the following theorem.

Theorem 10 (*Vector Lagrangian Dual Transform*)**.** *Given a sequence of nonnegative weights* $w_i \geq 0$*, vector-valued functions* $\mathbf{a}_i(\mathbf{x}) \in \mathbb{C}^d$*, and matrix-valued function* $\mathbf{B}_i(\mathbf{x}) \in \mathbb{H}_{++}^{d\times d}$*, for* $i = 1, \ldots, n$*, along with a nonempty and compact constraint set* $\mathcal{X}$*, where* $d \in \mathbb{N}$*, a weighted sum logarithm FP problem*

$$\underset{\mathbf{x}}{\textit{maximize}} \quad \sum_{i=1}^{n} w_i \log\left(1 + \mathbf{a}_i^H(\mathbf{x})\mathbf{B}_i^{-1}(\mathbf{x})\mathbf{a}_i(\mathbf{x})\right) \tag{2.43a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.43b}$$

is equivalent to

$$\underset{\mathbf{x},\,\underline{\gamma}}{\textit{maximize}} \quad f_r(\mathbf{x}, \underline{\gamma}) \tag{2.44a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{2.44b}$$

where the new objective function f_r *is*

$$f_r(\mathbf{x}, \underline{\gamma}) = \sum_{i=1}^{n} w_i \log(1+\gamma_i) - \sum_{i=1}^{n} w_i\gamma_i + \sum_{i=1}^{n} w_i(1+\gamma_i)\mathbf{a}_i^H(\mathbf{x})\big(\mathbf{a}_i(\mathbf{x})\mathbf{a}_i^H(\mathbf{x}) + \mathbf{B}_i(\mathbf{x})\big)^{-1}\mathbf{a}_i(\mathbf{x}). \tag{2.45}$$

Proof. Since f_r is analytic in the complex plane and also f_r is concave over $\boldsymbol{\gamma}$ for fixed $\mathbf{x}$, we take its complex derivative and solve each $\partial f_r/\partial\gamma_i = 0$. The optimal $\gamma_i^\star$ is easily seen as $\mathbf{a}_i^H(\mathbf{x})\mathbf{B}_i^{-1}(\mathbf{x})\mathbf{a}_i(\mathbf{x})$. Substituting this $\gamma_i^\star$ back in f_r recovers the weighted sum-of-logarithms objective function in (2.43a) exactly, thus establishing the equivalence. □

We have assumed that the numerator and the denominator can be matrices but the ratio between them is still scalar-valued. What follows is a further extension to account for the matrix-form fractional terms.

2.5.2 Matrix Numerators & Matrix Denominators

The definition of ratio can be naturally generalized to the matrix case. Recall that $\sqrt{\mathbf{A}} \in \mathbb{C}^{d\times d}$ is a square root of matrix $\mathbf{A} \in \mathbb{H}_{+}^{d\times d}$ if $\sqrt{\mathbf{A}}\sqrt{\mathbf{A}}^H = \mathbf{A}$. (Note that the square root of matrix may not be unique.) For any pair of $\mathbf{A} \in \mathbb{H}_{+}^{d\times d}$ and $\mathbf{B} \in \mathbb{H}_{++}^{d\times d}$, let $\sqrt{\mathbf{A}}$ be a square root of $\mathbf{A}$, then $\sqrt{\mathbf{A}}^H\mathbf{B}^{-1}\sqrt{\mathbf{A}}$ is said to be a matrix ratio between $\mathbf{A}$ and $\mathbf{B}$. The FP transforms of Theorem 3 and Theorem 8 can now be generalized accordingly. We state these new results in the following.

Theorem 11 (*Matrix Quadratic Transform*)**.** *Given a sequence of numerator functions* $\mathbf{A}_i(\mathbf{x}) \in \mathbb{H}_+^{d\times d}$*, denominator functions* $\mathbf{B}_i(\mathbf{x}) \in \mathbb{H}_{++}^{d\times d}$*, and nondecreasing matrix functions* $f_i(\mathbf{Z}) \in \mathbb{R}$ *in the sense that* $f_i(\mathbf{Z}') \geq f_i(\mathbf{Z})$ *if* $\mathbf{Z}' \succeq \mathbf{Z}$*, for* $i = 1, \ldots, n$*, along with a nonempty and compact constraint set* $\mathcal{X}$*, where* $d \in \mathbb{N}$*, the matrix-numerator-and-matrix-denominator sum-of-ratios problem*

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{n} f_i\big(\sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{B}_i^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_i(\mathbf{x})\big) \tag{2.46a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.46b}$$

is equivalent to

$$\underset{\mathbf{x},\,\underline{\mathbf{Y}}}{\text{maximize}} \quad \tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}) \tag{2.47a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.47b}$$

$$\mathbf{Y}_i \in \mathbb{C}^{d\times d},\ \forall i, \tag{2.47c}$$

where the new objective function is

$$\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}) = \sum_{i=1}^{n} f_i\Big(2\Re\{\sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{Y}_i\} - \mathbf{Y}_i^H\mathbf{B}_i(\mathbf{x})\mathbf{Y}_i\Big). \tag{2.48}$$

Note that the above transformation has implicitly required that the argument of $f_i(\cdot)$ *in (2.48) is a positive semidefinite matrix.*

Proof. To show that (2.46) is equivalent to (2.47), we first optimize over $\mathbf{Y}_i$ for fixed $\mathbf{x}$ in (2.47). This can be done for each term in the summation in $\tilde{f}_q$ separately. Since $f_i(\cdot)$ is assumed to be monotonic, we only need to optimize its argument, which is a quadratic function of $\mathbf{Y}_i$. This optimization has a closed-form solution by completing the square, *i.e.*,

$$\begin{aligned} 2\Re\{\sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{Y}_i\} - \mathbf{Y}_i^H\mathbf{B}_i(\mathbf{x})\mathbf{Y}_i &= \sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{Y}_i + \mathbf{Y}_i^H\sqrt{\mathbf{A}}_i(\mathbf{x}) - \mathbf{Y}_i^H\mathbf{B}_i(\mathbf{x})\mathbf{Y}_i \\ &= \sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{B}_i^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_i(\mathbf{x}) - \boldsymbol{\Delta}_i^H\mathbf{B}_i(\mathbf{x})\boldsymbol{\Delta}_i, \end{aligned} \tag{2.49}$$

where $\boldsymbol{\Delta}_i = \mathbf{Y}_i - \mathbf{B}_i^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_i(\mathbf{x})$. We then obtain the optimal $\mathbf{Y}_i^\star = \mathbf{B}_i^{-1}(\mathbf{x})\sqrt{\mathbf{A}}_i(\mathbf{x})$. Substituting this $\mathbf{Y}_i^\star$ in $\tilde{f}_q$ recovers the original problem. □

We also give the matrix version of the Lagrangian dual transform in the following theorem.

Theorem 12 (*Matrix Lagrangian Dual Transform*)**.** *Given a nonempty constraint set* $\mathcal{X}$ *as well as a sequence of the numerator functions* $\mathbf{A}_i(\mathbf{x}) \in \mathbb{H}_+^{d\times d}$*, the denominator functions* $\mathbf{B}_i(\mathbf{x}) \in \mathbb{H}_{++}^{d\times d}$*, and the nonnegative weights* $w_i \geq 0$*, for* $i = 1, \ldots, n$*, where* $d \in \mathbb{N}$*, the sum-of-weighted-*

logarithmic-matrix-ratios problem

$$\underset{\mathbf{x}}{\textit{maximize}} \quad \sum_{i=1}^{n} w_i \log \left| \boldsymbol{I}_d + \sqrt{\mathbf{A}}_i^H(\mathbf{x}) \mathbf{B}_i^{-1}(\mathbf{x}) \sqrt{\mathbf{A}}_i(\mathbf{x}) \right| \tag{2.50a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.50b}$$

is equivalent to

$$\underset{\mathbf{x}, \underline{\boldsymbol{\Gamma}}}{\textit{maximize}} \quad f_r(\mathbf{x}, \underline{\boldsymbol{\Gamma}}) \tag{2.51a}$$

$$\textit{subject to} \quad \mathbf{x} \in \mathcal{X} \tag{2.51b}$$

$$\boldsymbol{\Gamma}_i \in \mathbb{H}_+^{d \times d}, \ \forall i, \tag{2.51c}$$

where the new objective function is

$$f_r(\mathbf{x}, \underline{\boldsymbol{\Gamma}}) = \sum_{i=1}^{n} w_i \Big(\log |\boldsymbol{I}_d + \boldsymbol{\Gamma}_i| - \mathrm{tr}(\boldsymbol{\Gamma}_i) + \mathrm{tr}\Big((\boldsymbol{I}_d + \boldsymbol{\Gamma}_i) \sqrt{\mathbf{A}}_i^H(\mathbf{x}) \big(\mathbf{A}_i(\mathbf{x}) + \mathbf{B}_i(\mathbf{x}) \big)^{-1} \sqrt{\mathbf{A}}_i(\mathbf{x}) \Big) \Big). \tag{2.52}$$

Proof. Using the Woodbury matrix identity

$$\big(\mathbf{D} + \mathbf{UCV} \big)^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{U} \big(\mathbf{C}^{-1} + \mathbf{VD}^{-1}\mathbf{U} \big)^{-1} \mathbf{VD}^{-1}, \tag{2.53}$$

we can rewrite (2.52) as

$$f_r(\mathbf{x}, \underline{\boldsymbol{\Gamma}}) = \sum_{i=1}^{n} w_i \Big(\log |\boldsymbol{I}_d + \boldsymbol{\Gamma}_i| + n - \mathrm{tr}\Big((\boldsymbol{I}_d + \boldsymbol{\Gamma}_i) \big(\boldsymbol{I}_d + \sqrt{\mathbf{A}}_i^H(\mathbf{x}) \mathbf{B}_i^{-1}(\mathbf{x}) \sqrt{\mathbf{A}}_i(\mathbf{x}) \big)^{-1} \Big) \Big). \tag{2.54}$$

We then consider the optimization of the above new form of f_r. Note that the optimization over $\boldsymbol{\Gamma}_i$ can be done separately for each term of the summation. Since each of the terms is concave over $\boldsymbol{\Gamma}_i$ when $\mathbf{x}$ is fixed, the optimal $\boldsymbol{\Gamma}_i$ can be determined by setting $\partial f_r / \partial \boldsymbol{\Gamma}_i$ to zero, *i.e.*,

$$(\boldsymbol{I}_d + \boldsymbol{\Gamma}_i)^{-1} - \Big(\boldsymbol{I}_d + \sqrt{\mathbf{A}}_i^H(\mathbf{x}) \mathbf{B}_i^{-1}(\mathbf{x}) \sqrt{\mathbf{A}}_i(\mathbf{x}) \Big)^{-1} = \mathbf{0}. \tag{2.55}$$

Note that the derivative $\partial f_r / \partial \boldsymbol{\Gamma}_i$ exists in this case because f_r is a spectral function [29]. Thus, we obtain the optimal $\boldsymbol{\Gamma}_i^\star = \sqrt{\mathbf{A}}_i^H(\mathbf{x}) \mathbf{B}_i^{-1}(\mathbf{x}) \sqrt{\mathbf{A}}_i(\mathbf{x})$. Substituting this $\boldsymbol{\Gamma}_i^\star$ in (2.54) recovers the original problem, thereby establishing the theorem. □

Observe that the proposed matrix quadratic transform of Theorem 11 can be applied to decouple the ratio terms of f_r in (2.52) to further transform the matrix FP, as stated in the corollary below.

Corollary 5. *The sum-of-weighted-logarithmic-matrix-ratios problem (2.50) is equivalent to*

$$\underset{\mathbf{x},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}}{\text{maximize}} \quad f_q(\mathbf{x},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}) \tag{2.56a}$$

$$\text{subject to} \quad \mathbf{x}\in\mathcal{X} \tag{2.56b}$$

$$\mathbf{\Gamma}_i \in \mathbb{H}_+^{d\times d},\ \forall i \tag{2.56c}$$

$$\mathbf{Y}_i \in \mathbb{C}^{d\times d},\ \forall i, \tag{2.56d}$$

where the new objective function is

$$f_q(\mathbf{x},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}) = \sum_{i=1}^{n}\Big(w_i\log|\boldsymbol{I}_d+\mathbf{\Gamma}_i| - w_i\mathrm{tr}\big(\mathbf{\Gamma}_i\big) + \mathrm{tr}\Big((\boldsymbol{I}_d+\mathbf{\Gamma}_i)\cdot \Big(2\sqrt{w_i}\sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{Y}_i - \mathbf{Y}_i^H\big(\mathbf{A}_i(\mathbf{x})+\mathbf{B}_i(\mathbf{x})\big)\mathbf{Y}_i\Big)\Big)\Big). \tag{2.57}$$

Note that $\Re\{\cdot\}$ can be dropped for the term $\sqrt{\mathbf{A}}_i^H(\mathbf{x})\mathbf{Y}_i$ because of trace.

Proof. Treating $f_i(\mathbf{Z}) = \mathrm{tr}\big((\boldsymbol{I}_d+\mathbf{\Gamma}_i)\mathbf{Z}\big)$ as the nondecreasing function, $\sqrt{w_i}\sqrt{\mathbf{A}}_i(\mathbf{x})$ as the square root of the numerator, and $\mathbf{A}_i(\mathbf{x})+\mathbf{B}_i(\mathbf{x})$ as the denominator, we apply the matrix quadratic transform of Theorem 11 to the last term of f_r in (2.52) to obtain the above reformulation. □

Note that the new objective function f_q is an affine function of each of the square-root terms of the numerator $\sqrt{w_i}$ and $\sqrt{\mathbf{A}}_i(\mathbf{x})$ and the denominator term $\mathbf{B}_i(\mathbf{x})$, while keeping all other terms fixed. This facilitates algorithm design for solving the matrix FP problem. We also remark that there are also other ways of applying the matrix quadratic transform to f_r in (2.52) by choosing different matrix ratios and functions $f_i(\cdot)$. The different ways of decomposition are discussed in detail in Section 4.4.

The former vector FP can be used to deal with the MIMO communication where each link has at most one data stream. To reap the full benefit of MIMO, each link needs to carry multiple data streams. In this case, the matrix FP of Theorem 11 is indispensable. These multidimensional FP techniques are typically applied to the $\log(1+\mathrm{SINR})$-type rate maximization problem. In particular, if the numerator and denominator of SINR are both affine functions of the target variables (e.g., beamforming vectors), then the variables can be optimized in closed form in the new problem by completing the square.

2.6 Connection to MM Algorithm

A critical theoretical insight is that the FP methods proposed above can be recast in the MM framework. We focus on the matrix FP since it is a generalization of the scalar version.

First, we give a brief introduction to MM. Consider a general optimization problem:

$$\underset{\mathbf{x}}{\text{maximize}} \quad f(\mathbf{x}) \tag{2.58a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{2.58b}$$

where $f(\mathbf{x})$ is not assumed to be concave. Because of the nonconvexity, it is not always easy to solve the problem directly. The core idea behind the MM algorithm is to successively solve a sequence of *well-chosen* approximations of the original problem [30, 31]. Specifically, at point $\hat{\mathbf{x}} \in \mathcal{X}$, the MM algorithm approximates problem (2.58) as

$$\underset{\mathbf{x}}{\text{maximize}} \quad g(\mathbf{x}|\hat{\mathbf{x}}) \tag{2.59a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}, \tag{2.59b}$$

where $g(\mathbf{x}|\hat{\mathbf{x}})$ is referred to as the *surrogate function* and is defined by these two conditions:

- M1: $g(\mathbf{x}|\hat{\mathbf{x}}) \leq f(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$;
- M2: $g(\hat{\mathbf{x}}|\hat{\mathbf{x}}) = f(\hat{\mathbf{x}})$.

The surrogate function $\hat{g}(\mathbf{x}|\hat{\mathbf{x}})$ takes a form such that $\mathbf{x}$ can be optimized easily, conventionally constructed as a concave function of $\mathbf{x}$ in the continuous optimization case. However, the choice of $\hat{g}(\mathbf{x}|\hat{\mathbf{x}})$ as proposed in this thesis is not always a concave function, especially when dealing with the discrete optimization.

The MM algorithm updates $\hat{\mathbf{x}}$ iteratively as follows:

$$\hat{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}|\hat{\mathbf{x}}_t), \tag{2.60}$$

where subscript t is the iteration index. Note that the function value of $f(\hat{\mathbf{x}})$ is nondecreasing after each iteration because

$$f(\hat{\mathbf{x}}_{t+1}) \overset{(a)}{\geq} g(\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t) \overset{(b)}{\geq} g(\hat{\mathbf{x}}_t|\hat{\mathbf{x}}_t) \overset{(c)}{=} f(\hat{\mathbf{x}}_t), \tag{2.61}$$

where (a) follows by M1, (b) follows by the optimality of $\hat{\mathbf{x}}_{t+1}$ in (2.60), and (c) follows by M2. This is illustrated in Fig. 2.4 on the next page.

The following proposition gives a convergence analysis of the MM algorithm.

Proposition 2. *Let $\hat{\mathbf{x}}_t$ be the solution produced by the MM update (2.60) after t iterations. The function value $f(\hat{\mathbf{x}}_t)$ converges in a nondecreasing fashion in t. Further, the variable $\hat{\mathbf{x}}_t$ converges to a stationary point solution to the original optimization problem (2.58) if the following three conditions are satisfied: (i) $f(\mathbf{x})$ is continuous over a convex closed set $\mathcal{X}$; (ii) $g(\mathbf{x}|\hat{\mathbf{x}})$ is continuous in $(\mathbf{x}, \hat{\mathbf{x}})$; (iii) $f(\mathbf{x})$ and $g(\mathbf{x}|\hat{\mathbf{x}})$ are differentiable with respect to $\mathbf{x}$ given $\hat{\mathbf{x}}$.*

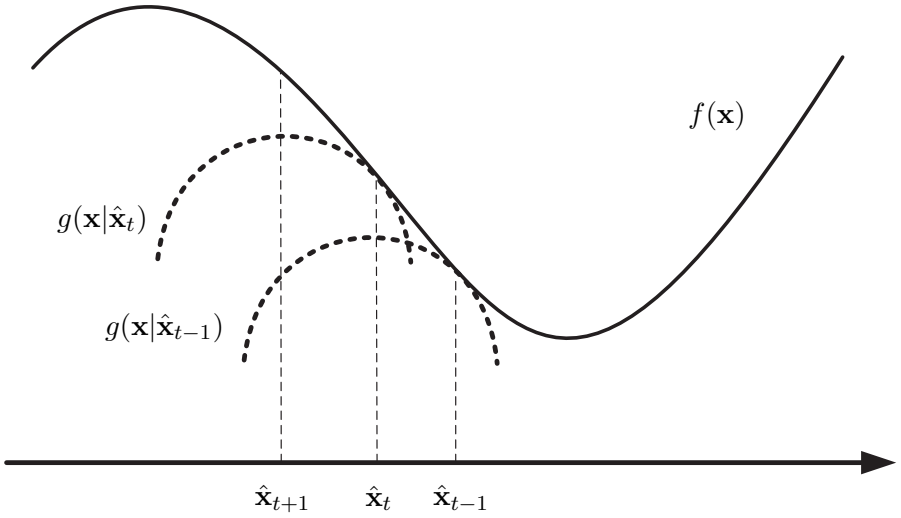


Figure 2.4: The iterative optimization by the MM algorithm. Observe that $f(\hat{\mathbf{x}})$ is monotonically nondecreasing after each iteration.

Proof. The non-decreasing convergence of $f(\hat{\mathbf{x}})$ is already verified in (2.61). Further, combining the above condition (iii) with the conditions M1 and M2, we obtain that $f(\mathbf{x})$ and $g(\mathbf{x}|\hat{\mathbf{x}})$ have the same gradient with respect to $\mathbf{x}$ at $\mathbf{x} = \hat{\mathbf{x}}$. This result, along with the above conditions (i) and (ii), guarantees that $\hat{\mathbf{x}}_t$ converges to a stationary point solution to the original optimization problem (2.58) according to [30]. We remark that the proof can be adapted to the case where $\mathbf{x}$ is a complex variable; the argument is similar to that of [32]. □

The MM algorithm is a framework rather than an algorithmic prescription, because the algorithm depends on the specific choice of the surrogate function. If $f(\cdot)$ is twice differentiable, its second order Taylor expansion is often the first candidate to check to see whether it is suitable as a surrogate function. For more general functions, many of the ingenious ways of constructing a surrogate function have been documented in [31].

The main point of this section is that the proposed matrix FP transforms can be interpreted in the MM framework as a way of constructing surrogate functions of the original problems, as stated below.

Theorem 13. *Consider the matrix quadratic transform in Theorem 11, if we consider the optimal $\mathbf{Y}_i^\star$ as a function of $\hat{\mathbf{x}}$ and substitute it into $\tilde{f}_q$ in (2.48), then the new objective function $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$, where*

$$\mathbf{Y}_i(\hat{\mathbf{x}}) = \mathbf{B}_i^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_i(\hat{\mathbf{x}}) \tag{2.62}$$

is a surrogate function of the objective function of the optimization problem (2.46).

Proof. Use $f_{\mathrm{I}}(\mathbf{x})$ to denote the objective function in (2.46a). Substitute $\mathbf{Y}_i(\hat{\mathbf{x}}) = \mathbf{B}_i^{-1}(\hat{\mathbf{x}})\sqrt{\mathbf{A}}_i(\hat{\mathbf{x}})$ back in $\tilde{f}_q$. We aim to show that $g(\mathbf{x}|\hat{\mathbf{x}}) = \tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$ is a surrogate function of $f_{\mathrm{I}}(\mathbf{x})$.

As already shown in the proof of Theorem 11, $\underline{\mathbf{Y}}(\mathbf{x})$ is the optimum solution for the maximization of $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}})$ over $\underline{\mathbf{Y}}$ when $\mathbf{x}$ is fixed. So, $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}})) \leq \tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\mathbf{x})), \forall \hat{\mathbf{x}}, \mathbf{x}$. Further, it can be seen that $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\mathbf{x})) = f_{\mathrm{I}}(\mathbf{x})$ for any $\mathbf{x}$.

Thus, for each fixed $\hat{\mathbf{x}}$, we have $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}})) \leq f_{\mathrm{I}}(\mathbf{x})$, $\forall \mathbf{x}$, and $\tilde{f}_q(\hat{\mathbf{x}}, \underline{\mathbf{Y}}(\hat{\mathbf{x}})) = f_{\mathrm{I}}(\hat{\mathbf{x}})$, thus verifying the conditions M1 and M2 for $\tilde{f}_q(\mathbf{x}, \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$ to be a surrogate function of $f_{\mathrm{I}}(\mathbf{x})$. □

Theorem 14. *Consider the matrix Lagrangian dual transform in Theorem 12, if we consider the optimal $\mathbf{\Gamma}_i^\star$ as a function of $\hat{\mathbf{x}}$ and substitute it into f_r in (2.52), then the new objective function $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$, where*

$$\mathbf{\Gamma}_i(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}_i^H(\hat{\mathbf{x}}) \mathbf{B}_i^{-1}(\hat{\mathbf{x}}) \sqrt{\mathbf{A}}_i(\hat{\mathbf{x}}) \tag{2.63}$$

is a surrogate function of the objective function of the optimization problem (2.50).

Proof. We use $f_{\mathrm{II}}(\mathbf{x})$ to denote the objective function in (2.50a). We further substitute $\mathbf{\Gamma}_i(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}_i^H(\hat{\mathbf{x}}) \mathbf{B}_i^{-1}(\hat{\mathbf{x}}) \sqrt{\mathbf{A}}_i(\hat{\mathbf{x}})$ back in f_r, aiming to show that $g(\mathbf{x}|\hat{\mathbf{x}}) = f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$ is a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$.

As shown in the proof of Theorem 12, $\underline{\mathbf{\Gamma}}(\mathbf{x})$ is the optimal solution to maximizing $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}})$ over $\underline{\mathbf{\Gamma}}$ when $\mathbf{x}$ is fixed, so $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) \leq f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\mathbf{x})), \forall \mathbf{x}, \hat{\mathbf{x}}$. Also, it holds true that $f_r(\hat{\mathbf{x}}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) = f_{\mathrm{II}}(\hat{\mathbf{x}})$, $\forall \hat{\mathbf{x}}$. Combining the above results, we see that the conditions M1 and M2 are satisfied, thus $f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$ is a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$. □

Corollary 6. *Consider the transform in Corollary 5, if we consider the optimal $\mathbf{\Gamma}_i^\star$ and the optimal $\mathbf{Y}_i^\star$ as two functions of $\hat{\mathbf{x}}$, and substitute them into into f_q, then the new objective function $f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}), \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$, where*

$$\mathbf{\Gamma}_i(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}_i^H(\hat{\mathbf{x}}) \mathbf{B}_i^{-1}(\hat{\mathbf{x}}) \sqrt{\mathbf{A}}_i(\hat{\mathbf{x}}) \tag{2.64}$$

and

$$\mathbf{Y}_i(\hat{\mathbf{x}}) = \big(\mathbf{A}_i(\hat{\mathbf{x}}) + \mathbf{B}_i(\hat{\mathbf{x}})\big)^{-1} \big(\sqrt{w_i} \sqrt{\mathbf{A}}_i(\hat{\mathbf{x}})\big), \tag{2.65}$$

is a surrogate function of the objective function of the optimization problem (2.50).

Proof. Again, let $f_{\mathrm{II}}(\mathbf{x})$ be the objective function in (2.50a). Introduce two new variables $\hat{\mathbf{x}}$ and $\hat{\hat{\mathbf{x}}}$, and substitute $\mathbf{\Gamma}_i(\hat{\mathbf{x}}) = \sqrt{\mathbf{A}}_i^H(\hat{\mathbf{x}}) \mathbf{B}_i^{-1}(\hat{\mathbf{x}}) \sqrt{\mathbf{A}}_i(\hat{\mathbf{x}})$ and $\mathbf{Y}_i(\hat{\hat{\mathbf{x}}}) = (\mathbf{A}_i(\hat{\hat{\mathbf{x}}}) + \mathbf{B}_i(\hat{\hat{\mathbf{x}}}))^{-1} (\sqrt{w_i} \sqrt{\mathbf{A}}_i(\hat{\hat{\mathbf{x}}}))$ back in f_q and f_r. Let $g_1(\mathbf{x}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}}) = f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}), \underline{\mathbf{Y}}(\hat{\hat{\mathbf{x}}}))$, and $g_2(\mathbf{x}|\hat{\mathbf{x}}) = f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$.

According to Theorem 14, $g_2(\mathbf{x}|\hat{\mathbf{x}})$ is a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$ in the sense that $g_2(\mathbf{x}|\hat{\mathbf{x}}) \leq f_{\mathrm{II}}(\mathbf{x})$ and $g_2(\hat{\mathbf{x}}|\hat{\mathbf{x}}) = f_{\mathrm{II}}(\hat{\mathbf{x}}), \forall \mathbf{x}, \hat{\mathbf{x}}$. According to Theorem 13, $g_1(\mathbf{x}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}})$ is a surrogate function with respect to f_r in the sense that $g_1(\mathbf{x}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}}) \leq f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}))$ and $g_1(\hat{\hat{\mathbf{x}}}|\hat{\hat{\mathbf{x}}}, \hat{\mathbf{x}}) = f_r(\hat{\hat{\mathbf{x}}}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})), \forall \mathbf{x}, \hat{\mathbf{x}}, \hat{\hat{\mathbf{x}}}$.

Combining these results and fixing $\hat{\hat{\mathbf{x}}} = \hat{\mathbf{x}}$, we obtain $g_1(\mathbf{x}|\hat{\mathbf{x}}, \hat{\mathbf{x}}) \leq f_r(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) = g_2(\mathbf{x}|\hat{\mathbf{x}}) \leq f_{\mathrm{II}}(\mathbf{x}), \forall \mathbf{x}$ and $g_1(\hat{\mathbf{x}}|\hat{\mathbf{x}}, \hat{\mathbf{x}}) = f_r(\hat{\mathbf{x}}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}})) = g_2(\hat{\mathbf{x}}|\hat{\mathbf{x}}) = f_{\mathrm{II}}(\hat{\mathbf{x}})$, thereby verifying the conditions M1 and M2 for $f_q(\mathbf{x}, \underline{\mathbf{\Gamma}}(\hat{\mathbf{x}}), \underline{\mathbf{Y}}(\hat{\mathbf{x}}))$ to be a surrogate function of $f_{\mathrm{II}}(\mathbf{x})$. □

The above connection between FP and MM provides a theoretical basis to a sequence of new algorithms based on the quadratic transform and the Lagrangian dual transform as proposed in the rest of the thesis. In principle, the basic properties that have been discovered about MM can be carried over to our FP approach automatically. A direct benefit we will see is to simplify the convergence proof of Proposition 5 in Section 3.1.3 by using the MM interpretation, which would have been much more cumbersome [33] otherwise.

2.7 Summary

This chapter focuses on the theoretical aspect of FP. The main contributions are four fold. First, we introduce the sum-of-functions-of-ratio problem and the function-of-multi-ratio problem as two new types of multi-ratio problems. Second, we propose a new ratio-decoupling technique named the quadratic transform that works for a broader range of FP problems than the classic methods, especially when the problem consists of multiple ratio terms; we also propose a Lagrangian dual transform to address the logarithmic FP problem. Third, we put forward a new idea of generalizing FP to a multidimensional space, namely matrix FP. Last, the proposed techniques are shown to be closely connected to the MM algorithm.

Chapter 3

Continuous Optimization Problems

This chapter characterizes the role of FP in dealing with the continuous problems of communication system design, including power control, beamforming, and energy efficiency maximization. We aim to show that the proposed quadratic transform can greatly facilitate the optimization involving multiple ratios by recasting the original nonconvex problem into a sequence of convex problems. This ratio-decoupling transformation gives rise to an efficient iterative optimization algorithm that guarantees convergence to a stationary point. We further show that the proposed algorithm for power control can be recognized as a fixed-point iteration. Differing from the existing power control algorithms of the same type, our fixed-point iteration guarantees convergence to a stationary point regardless of channels.

3.1 Power Control

The proposed approach is applied to the optimization of transmit powers pursuing a weighted sum-rate maximization across a single-input single-output (SISO) wireless cellular network, which is a notorious challenging problem. We propose two methods: (i) A direct approach that applies the quadratic transform directly to SINR, then updates the power variables iteratively via a sequence of convex optimizations; (ii) a more sophisticated approach that uses the quadratic transform in conjunction with the Lagrangian dual transform to derive closed-form iterative updates. We further interpret the second approach as a fixed-point iteration and compare it with the existing fixed-point iteration methods for power control.

3.1.1 Problem Formulation

The first example is the classic power control problem for a downlink SISO cellular network with a set of single-antenna base stations (BSs) $\mathcal{B}$, each serving a single-antenna user. Let $h_{i,j} \in \mathbb{C}$ be the downlink channel from BS j to user i; let σ^2 be the power level of additive white Gaussian noise (AWGN). Introduce variable p_i for each BS i as its transmit power level,

constrained by a power budget of $P_{\max}$. The downlink data rate of user i is computed as

$$R_i = \log\left(1 + \frac{|h_{i,i}|^2 p_i}{\sum_{j\neq i} |h_{i,j}|^2 p_j + \sigma^2}\right). \tag{3.1}$$

We consider the maximization of a weighted sum rate objective function

$$f_o(\underline{p}) = \sum_{i\in\mathcal{B}} w_i R_i, \tag{3.2}$$

where w_i accounts for the priority of the ith BS-user downlink. The power control problem is formulated as

$$\underset{\underline{p}}{\text{maximize}} \quad f_o(\underline{p}) \tag{3.3a}$$

$$\text{subject to} \quad 0 \leq p_i \leq P_{\max},\ \forall i \in \mathcal{B}. \tag{3.3b}$$

This problem is numerically difficult due to its nonconvexity. Indeed, the problem can be solved globally by using a polyblock approximation approach [34], but not in polynomial time. Moreover, for the case where all the SINRs are sufficiently high so that $\log(1 + \text{SINR})$ can be approximated as $\log(\text{SINR})$, the problem can be globally solved via *geometric programming* (GP) [35]. Moreover, the structure of the interference functions is investigated in [36, 37] for solving the power control problem. The goal of this section is to find at least a stationary point of the power control problem in an efficient manner.

3.1.2 Direct Approach

Although the power control problem is not in a direct ratio form, the main components of its objective function, the SINR terms, are fractional. Because each SINR term resides inside the logarithm function, which is nondecreasing and concave, the condition of Theorem 6 is satisfied in this problem. Specifically, after applying the quadratic transform to each SINR term, we arrive at the following reformulation:

$$\underset{\underline{p},\underline{y}}{\text{maximize}} \quad f_q^{\text{DIR}}(\underline{p},\underline{y}) \tag{3.4a}$$

$$\text{subject to} \quad 0 \leq p_i \leq P_{\max},\ \forall i \in \mathcal{B} \tag{3.4b}$$

$$y_i \in \mathbb{R},\ \forall i \in \mathcal{B}, \tag{3.4c}$$

where $\underline{y}$ refers to the collection $\{y_i\}_{i\in\mathcal{B}}$. The new objective function f_q^{DIR} is

$$f_q^{\text{DIR}}(\underline{p},\underline{y}) = \sum_{i\in\mathcal{B}} w_i \log\left(1 + 2y_i\sqrt{|h_{i,i}|^2 p_i} - y_i^2\left(\sum_{j\neq i} |h_{i,j}|^2 p_j + \sigma^2\right)\right) \tag{3.5}$$

with an auxiliary variable y_i introduced by the quadratic transform for each downlink i.

Algorithm 2: Direct FP for Power Control

1 Initialize $\underline{p}$ to a feasible value;
2 **repeat**
3 Update $\underline{y}$ by (3.6);
4 Update $\underline{p}$ by solving the convex optimization problem (3.4) over $\underline{p}$ for fixed $\underline{y}$;
5 **until** *the value of function* f_q^{DIR} *in (3.5) converges*;

Following Algorithm 1, we optimize $\underline{y}$ and $\underline{p}$ in an iterative fashion. The optimal $\underline{y}$ for fixed $\underline{p}$ is

$$y_i^\star = \frac{\sqrt{|h_{i,i}|^2 p_i}}{\sum_{j\neq i} |h_{i,j}|^2 p_j + \sigma^2}. \tag{3.6}$$

Then, finding the optimal $\underline{p}$ for fixed $\underline{y}$ is a convex problem. This power control method is summarized in Algorithm 2.

By Theorem 6 in Section 2.3.1, Algorithm 2 guarantees a convergence to a stationary point of problem (3.3). We remark that Algorithm 2 can be easily extended to the multiple-band system, where the frequency band is partitioned into T sub-bands, and the user rate is computed as

$$R_i = \sum_{t=1}^{T} \frac{1}{T} \log\left(1 + \frac{|h_{i,i}^t|^2 p_i^t}{\sum_{j\neq i} |h_{i,j}^t|^2 p_j^t + \sigma^2}\right). \tag{3.7}$$

Here, $h_{i,j}^t$ and p_j^t represent the channel and the transmit power level in the tth sub-band, respectively. The power constraint (3.3b) now becomes

$$\sum_{t=1}^{T} p_i^t \leq P_{\max},\ \forall i \in \mathcal{B} \tag{3.8}$$

and

$$p_i^t \geq 0,\ \forall i \in \mathcal{B},\ t = 1, \ldots, T. \tag{3.9}$$

To adapt Algorithm 2 to this multiple-band scenario, we just need to introduce an auxiliary variable y_i^t for each (i,t) pair and include a sum-power constraint across the sub-bands for each transmitter, so step 3 remains closed-form update and step 4 remains convex optimization. We then show that the direct FP method can be extended for a general utility function.

Proposition 3 (*Power Control for General Utility Maximization*)**.** *Given a nondecreasing concave utility function U_i of rate R_i for each user i, e.g., the log-utility function, the sum utility maximization problem*

$$\underset{\underline{p}}{\textit{maximize}} \quad \sum_{i\in\mathcal{B}} U_i(R_i) \tag{3.10a}$$

$$\textit{subject to} \quad 0 \leq p_i \leq P_{\max},\ \forall i \in \mathcal{B} \tag{3.10b}$$

is equivalent to

$$\underset{\underline{p},\underline{y}}{\textit{maximize}} \quad \sum_{i\in\mathcal{B}} U_i(Q_i) \tag{3.11a}$$

$$\textit{subject to} \quad 0 \le p_i \le P_{\max},\ \forall i \in \mathcal{B} \tag{3.11b}$$

$$y_i \in \mathbb{R},\ \forall i \in \mathcal{B}, \tag{3.11c}$$

where

$$Q_i = \log\left(1 + 2y_i|h_{i,i}|\sqrt{p_i} - y_i^2 \sum_{j\neq i} |h_{i,j}|^2 p_j - y_i^2\sigma^2\right). \tag{3.12}$$

The new problem as stated above can be solved (to a stationary point) as follows. When $\underline{p}$ is fixed, variable $\underline{y}$ is optimally determined by (3.6); when $\underline{y}$ is fixed, optimizing $\underline{p}$ in (3.11) is a convex problem.

Furthermore, we may encounter a SINR cap in practice, *e.g.,* at 60 dB, so the data rate can no longer increase when SINR exceeds 60 dB. In this case, we simply impose a ceiling constraint $\mathrm{SINR}_{\mathrm{cap}}$ on the ratio term, and the new problem becomes

$$\underset{\underline{p},\underline{y}}{\text{maximize}} \quad \sum_{i\in\mathcal{B}} U_i\big(\min\{Q_i, \mathrm{SINR}_{\mathrm{cap}}\}\big) \tag{3.13a}$$

$$\text{subject to} \quad 0 \le p_i \le P_{\max},\ \forall i \in \mathcal{B} \tag{3.13b}$$

$$y_i \in \mathbb{R},\ \forall i \in \mathcal{B}. \tag{3.13c}$$

Note that the convexity of the above new problem is preserved because the ceiling function is concave.

3.1.3 Closed-Form Approach

This subsection shows a different use of FP for solving the power control problem. This new approach is based on a Lagrangian dual reformulation of the power control problem as stated below. This leads to an algorithm in which each iteration is performed in closed form, rather than having to solve a convex optimization problem numerically, which is often more desirable than the direct FP approach introduced in the previous subsection.

Proposition 4. *By virtue of the Lagrangian dual transform in Theorem 8, the original power control problem (3.3) is reformulated as*

$$\underset{\underline{p},\underline{\gamma}}{\textit{maximize}} \quad f_r^{CF}(\underline{p},\underline{\gamma}) \tag{3.14a}$$

$$\textit{subject to} \quad 0 \le p_i \le P_{\max},\ \forall i \in \mathcal{B}, \tag{3.14b}$$

Algorithm 3: Closed-Form FP for Power Control

1 Initialize $\underline{p}$ and $\underline{\gamma}$ feasible values;
2 **repeat**
3 Update $\underline{y}$ by (3.19);
4 Update $\underline{\gamma}$ by (3.16);
5 Update $\underline{p}$ by (3.18);
6 **until** *the value of function* f_q^{CF} *in (3.17) converges*;

where the new objective function is

$$f_r^{CF}(\underline{p},\underline{\gamma}) = \sum_{i\in\mathcal{B}} w_i \log(1+\gamma_i) - \sum_{i\in\mathcal{B}} w_i\gamma_i + \sum_{i\in\mathcal{B}} \frac{w_i(1+\gamma_i)|h_{i,i}|^2 p_i}{\sum_{j\in\mathcal{B}}|h_{i,j}|^2 p_j + \sigma^2}. \tag{3.15}$$

We propose an iterative algorithm based on the above reformulation. When p_i is held fixed, the optimal γ_i is obtained by setting $\partial f_r^{\mathrm{CF}}/\partial\gamma_i$ to zero, *i.e.*,

$$\gamma_i^\star = \frac{|h_{i,i}|^2 p_i}{\sum_{j\neq i}|h_{i,j}|^2 p_j + \sigma^2}. \tag{3.16}$$

Note that the optimal γ_i is equal to the downlink SINR of BS i. When γ_i is held fixed, only the last term of f_r^{CF}, which has a sum-of-ratio form, is involved in the optimization of p_i. By the quadratic transform, we further recast f_r^{CF} to

$$f_q^{\mathrm{CF}}(\underline{p},\underline{\gamma},\underline{y}) = \sum_{i\in\mathcal{B}} 2y_i\sqrt{w_i(1+\gamma_i)|h_{i,i}|^2 p_i} - \sum_{i\in\mathcal{B}} y_i^2\left(\sum_{j\in\mathcal{B}}|h_{i,j}|^2 p_j + \sigma^2\right) + \mathrm{const}(\underline{\gamma}), \tag{3.17}$$

where $\mathrm{const}(\underline{\gamma})$ refers to a constant term when $\underline{\gamma}$ is fixed. For maximizing f_q^{CF} iteratively over p_i and y_i, we find closed-form update equations as

$$p_i^\star = \min\left\{P_{\max}, \frac{y_i^2 w_i(1+\gamma_i)|h_{i,i}|^2}{\left(\sum_{j\in\mathcal{B}} y_j^2|h_{j,i}|^2\right)^2}\right\} \tag{3.18}$$

and

$$y_i^\star = \frac{\sqrt{w_i(1+\gamma_i)|h_{i,i}|^2 p_i}}{\sum_{j\in\mathcal{B}}|h_{i,j}|^2 p_j + \sigma^2}. \tag{3.19}$$

These updating steps amount to an iterative optimization as summarized in Algorithm 3.

Unlike the direct FP approach, the above algorithm is not a conventional block coordinate descent (BCD), because the optimizing objective is not fixed, *i.e.*, γ_i is optimally updated for f_r^{CF} while y_i and p_i are optimally updated for f_q^{CF}. Nonetheless, its convergence to the stationary point can still be established by Theorem 6 in Section 2.3.1.

Proposition 5. *The weighted sum rate is nondecreasing after each iteration of Algorithm 3,*

so the objective function of the optimization problem is guaranteed to converge. Furthermore, at convergence, the solution $\underline{p}$ is a stationary point of the original problem.

Proof. We prove convergence based on the MM interpretation of the FP transforms as shown in Section 2.6. The Step 3 and Step 4 of the algorithm construct the surrogate functions as defined in Theorem 14 and Theorem 13 in Section 2.6. Step 5 of Algorithm 3 performs the maximization step of MM, so the weighted sum rate must be nondecreasing after Step 5, by (2.61). Since the optimization objective is nondecreasing and is bounded above, Algorithm 3 must converge in objective value.

The weighted sum rate is a differentiable function over $\underline{p}$. Further, the conditions of Theorem 6 in Section 2.3.1 are satisfied. So, at convergence, the solution of $\underline{p}$ given by Algorithm 3 must be a stationary point according to the proof of Theorem 6. □

We remark that proving the convergence of Algorithm 3 without the MM interpretation would have been much more cumbersome. We remark also that Algorithms 2 and 3 can be initialized with some simple but reasonable heuristic, *e.g.*, setting the initial powers to the half of the maximum powers. In our simulations, however, in order to guarantee fair comparisons, we use random starting points then average out the results. Moreover, we set some small constant $\delta > 0$ and use the convergence criterion $|f_q^{(t)} - f_q^{(t-1)}| < \delta$ where t is the iteration index.

We remark also that user i may get stuck at the off state if p_i is close to zero, namely premature turning-off. To resolve this issue, a heuristic method is to set a positive lower bound on p_i at the early stage of iterations. In Chapter 4 we will revisit this premature turning-off issue in the discrete optimization case.

3.1.4 Connection to Fixed-Point Iteration

This subsection illustrates that Algorithm 3 can be interpreted as a fixed-point iteration on the first-order condition of the power optimization problem. Attaining a stationary-point solution of the power control problem is equivalent to finding a solution to the first-order condition for (3.3), *i.e.*,

$$\frac{\partial f_o(\underline{p})}{\partial p_i} = 0, \tag{3.20}$$

which can be written as

$$\frac{1}{p_i} \cdot \underbrace{\frac{w_i \gamma_i(\underline{p})}{1+\gamma_i(\underline{p})}}_{T_{1i}(\underline{p})} - \underbrace{\sum\nolimits_{j \neq i} \frac{w_j \gamma_i^2(\underline{p}) |h_{j,i}|^2}{(1+\gamma_i(\underline{p}))|h_{j,j}|^2 p_j}}_{T_{2i}(\underline{p})} = 0 \tag{3.21}$$

where $\gamma_i(\underline{p})$ represents the SINR function of $\underline{p}$ in cell i as defined in (3.16). To find a set of powers that satisfy the above condition, one strategy [18–20] is to isolate p_i at one side of the equation—this automatically results in an update equation for power, which, if converging, would achieve at least a stationary point of the power control problem.

However, it is in general not easy to decide which part of the left-hand side of (3.20) should be fixed in order to ensure the convergence of fixed-point iteration. For instance, [19] proposes to fix T_{1i} and T_{2i} as shown in (3.21) and arrives at the following fixed-point method for power control

$$p_i^{(t+1)} = \min\left\{P_{\max}, \frac{T_{1i}(\underline{p}^{(t)})}{T_{2i}(\underline{p}^{(t)})}\right\}, \tag{3.22}$$

where the superscript t is the iteration index. However, this fixed-point iteration does not necessarily converge. (In fact, [19] proves that this iteration is guaranteed to converge when the resulting SINR values are all sufficiently high.)

With $\underline{\gamma}^\star$ and $\underline{y}^\star$ substituted in (3.18), the update equation (3.18) can also be thought of as a fixed-point iteration of the first-order condition for power control, exactly like (3.21) except that different components $\widetilde{T}_{1i}$ and $\widetilde{T}_{2i}$, shown below, are fixed

$$\frac{1}{\sqrt{p_i}} \cdot \underbrace{\frac{w_i\gamma_i(\underline{p})}{\sqrt{p_i}}}_{\widetilde{T}_{1i}(\underline{p})} - \underbrace{\sum\nolimits_j \frac{w_j\gamma_i^2(\underline{p})|h_{j,i}|^2}{(1+\gamma_i(\underline{p}))|h_{j,j}|^2 p_j}}_{\widetilde{T}_{2i}(\underline{p})} = 0. \tag{3.23}$$

In this case, the transmit power variable p_i update becomes

$$p_i^{(t+1)} = \min\left\{P_{\max}, \left(\frac{\widetilde{T}_{1i}(\underline{p}^{(t)})}{\widetilde{T}_{2i}(\underline{p}^{(t)})}\right)^2\right\}, \tag{3.24}$$

which, along with an additional projection step onto the constraint set, can be seen to be (3.18) after some algebra. Thus, the power control part of Algorithm 3 is just a fixed-point iteration, but with a crucial advantage that convergence is guaranteed, in contrast to the updates proposed in [18–20].

Finally, it is worth highlighting that the role of the auxiliary variables $(\underline{\gamma}, \underline{y})$ is to fix a set of $\underline{p}$ in the fixed-point iteration with provable convergence.

3.1.5 Numerical Results

We now evaluate the performance of FP for power control on a downlink cellular network consisting of seven wrapped-around hexagonal cells. Within each cell, the BS is located at the center and the downlink users are randomly placed. The BS-to-BS distance is set to be 0.8 km. The maximum transmit power level at the BS side is set to be -27 dBm/Hz, and the AWGN power level is set to be -170 dBm/Hz. A 10 MHz frequency band is fully reused across all the cells. The downlink distance-dependent path-loss is simulated by $128.1 + 37.6\log_{10}(d) + \tau$ in dB, where d represents the BS-to-user distance in km, and τ is a zero-mean Gaussian random variable with 8 dB standard deviation for the shadowing effect. We consider sum rate maximization by setting all the weights to 1.

The proposed FP approaches are compared to several benchmarks: first, direct optimiza-

tion based on a modified Newton's method [38], which deals with the power constraints via the nearest-point projection (the full Newton's method is too computationally complex), and second, an approach based on a modified version of GP called SCALE [39]. The version of SCALE implemented here involves solving a GP in every iteration.

Fig. 3.1 shows the performance of various power control algorithms in flat-fading channels. The closed-form FP takes the largest number of iterations to converge, but its computation per iteration is the lowest because of the closed-form updates in every iteration. In contrast, SCALE and direct FP both require solving a convex problem in each iteration. The closed-form FP also has lower complexity than Newton's method on per-iteration basis. In our simulation experience, the closed-form FP is the fastest.

Fig. 3.2 simulates a frequency selective fading scenario, in which the bandwidth is divided into 4 subbands; one downlink user is scheduled per tone. The resulting power control differs from the flat-fading case because of the sum power constraint across the subbands, *i.e.*, $\sum_n p_i^n \leq P_{\max}$ where p_i^n denotes the power level in tone n at BS i. In this case, Newton's method has to apply a heuristic nearest-point projection in order to satisfy the sum power constraint, but this no longer guarantees a stationary-point solution. As can be seen in the simulation, Newton's method now has much worse performance.

To conclude, the FP based approaches are competitive with the state-of-the-art algorithms in power control, with the closed-form FP having lower overall complexity due to its lower per-iteration cost. Note that the converged values of different algorithms may differ depending on the starting point, as only stationary-point convergence is guaranteed in all cases.

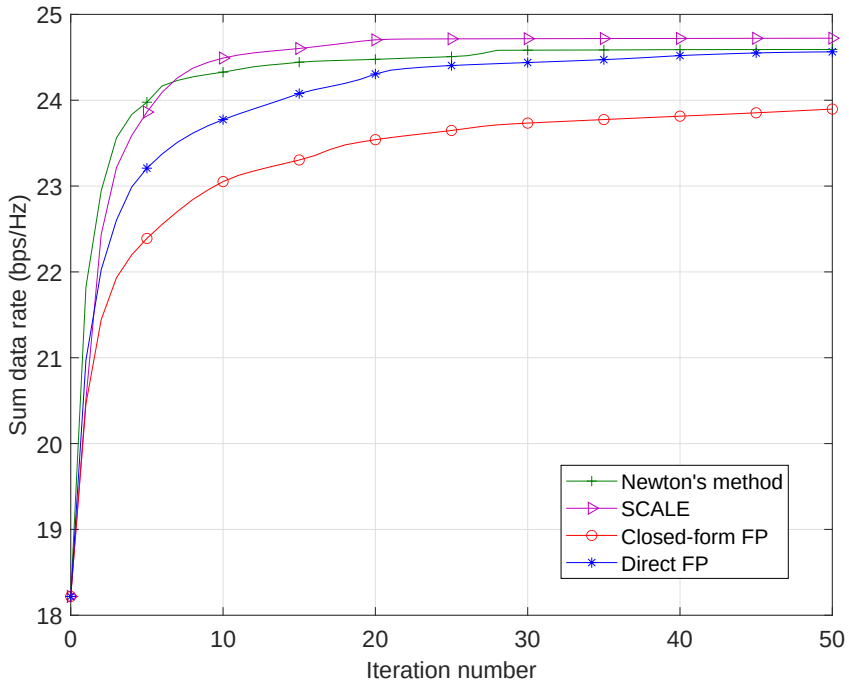


Figure 3.1: Power control in flat-fading channels.

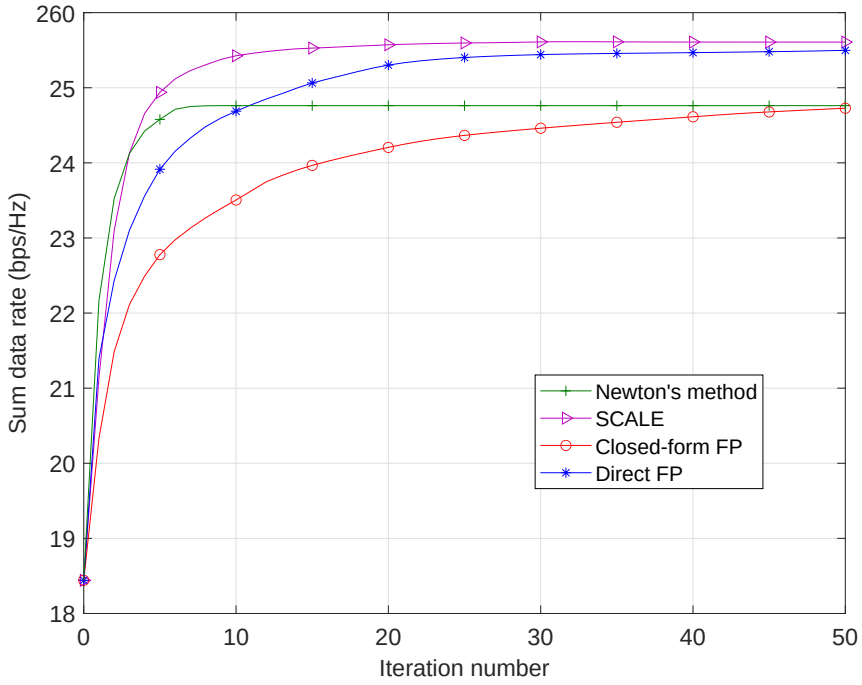


Figure 3.2: Power control in frequency-selective fading channels.

3.2 Beamforming

Chapter 2 has introduced a multidimensional version of FP with vector numerators and matrix denominators (albeit the ratio terms remain scalar-valued). This section shows that this new technique is ideally suited for the beamforming design in the spatial multiplexing wireless networks.

3.2.1 Problem Formulation

The second example is an application of multidimensional FP to the beamforming optimization problem. Consider a downlink MIMO cellular network with a set of BSs $\mathcal{B}$. Assume that each BS has M antennas and each user terminal has N antennas; then at most M downlink data streams are supported per cell via spatial multiplexing. Let $\mathbf{H}_{im,j} \in \mathbb{C}^{N\times M}$ be the downlink channel from BS j to the user who is scheduled in the mth data stream at BS i. Let σ^2 be the AWGN power level. Introduce variable $\mathbf{v}_{im} \in \mathbb{C}^M$ as the downlink transmit beamformer at BS i for its m-th data stream. The data rate of stream (i,m), R_{im}, is computed as

$$R_{im}(\underline{\mathbf{v}}) = \log\left(1 + \mathbf{v}_{im}^H \mathbf{H}_{im,i}^H \left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n)\neq(i,m)} \mathbf{H}_{im,j}\mathbf{v}_{jn}\mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H\right)^{-1} \mathbf{H}_{im,i}\mathbf{v}_{im}\right). \quad (3.25)$$

Let weight w_{im} be the priority of user scheduled in the m-th data stream at BS i. We seek to maximize the weighted sum rate over the beamforming vectors:

$$\underset{\underline{\mathbf{v}}}{\text{maximize}} \quad \sum_{i,m} w_{im} R_{im}(\underline{\mathbf{v}}) \quad (3.26a)$$

$$\text{subject to} \quad \sum_{m=1}^{M} \|\mathbf{v}_{im}\|_2^2 \leq P_{\max},\ \forall i \in \mathcal{B}, \quad (3.26b)$$

where we use $P_{\max}$ to denote the transmit power constraint at the BS side. This is a challenging nonconvex problem with vector variables.

3.2.2 Direct Approach

Similar to the power control case, the direct FP approach applies the vector quadratic transform of Theorem 9 in Section 9 to each SINR term. This leads to a new objective function f_q^{DIR} as

$$f_q^{\text{DIR}}(\underline{\mathbf{v}}, \underline{\mathbf{y}}) = \sum_{(i,m)} w_{im} \log\Bigg(1 + 2\Re\left\{\mathbf{y}_{im}^H \mathbf{H}_{im,i}\mathbf{v}_{im}\right\} - \mathbf{y}_{im}^H\left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n)\neq(i,m)} \mathbf{H}_{im,j}\mathbf{v}_{jn}\mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H\right)\mathbf{y}_{im}\Bigg), \quad (3.27)$$

Algorithm 4: Direct FP for Beamforming

1 Initialize $\underline{\mathbf{v}}$ to a feasible value;
2 **repeat**
3 Update $\underline{\mathbf{y}}$ by (3.29);
4 Update $\underline{\mathbf{v}}$ by solving the convex problem (3.28) over $\underline{\mathbf{v}}$ for fixed $\underline{\mathbf{y}}$;
5 **until** *the value of function* f_q^{DIR} *in (3.27) converges*;

where an auxiliary variable $\mathbf{y}_{im} \in \mathbb{C}^N$ is introduced for each data stream (i, m). The optimization problem (3.26) can now be recast to

$$\underset{\underline{\mathbf{v}},\, \underline{\mathbf{y}}}{\text{maximize}} \quad f_q^{\text{DIR}}(\underline{\mathbf{v}}, \underline{\mathbf{y}}) \tag{3.28a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} \|\mathbf{v}_{im}\|_2^2 \leq P_{\max},\ \forall i \in \mathcal{B} \tag{3.28b}$$

$$\mathbf{y}_{im} \in \mathbb{C}^N,\ \forall (i, m). \tag{3.28c}$$

Decoupled by the multidimensional quadratic transform, the SINR term is converted to a concave function of $\underline{\mathbf{v}}$. Since the outer logarithmic function is nondecreasing and concave, the optimization problem (3.28) is a convex problem of $\underline{\mathbf{v}}$ when the auxiliary variable $\underline{\mathbf{y}}$ is held fixed.

We follow Algorithm 1 in Section 2.3 to maximize f_q^{DIR} over $\underline{\mathbf{v}}$ and $\underline{\mathbf{y}}$ iteratively. Each $\mathbf{y}_{im}$ for fixed $\underline{\mathbf{v}}$ is optimally determined as

$$\mathbf{y}_{im}^{\star} = \left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n) \neq (i,m)} \mathbf{H}_{im,j} \mathbf{v}_{jn} \mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H \right)^{-1} \mathbf{H}_{im,i} \mathbf{v}_{im}. \tag{3.29}$$

For fixed $\underline{\mathbf{y}}$, the optimal $\mathbf{v}_{im}$ can be obtained by convex optimization. The resulting algorithm, stated as Algorithm 4, has a provable convergence to a stationary point due to Theorem 6 in Section 2.3.1. This algorithm requires solving a convex problem numerically in every iteration. In the next section, we illustrate another use of FP that yields a closed-form optimization in every iteration.

3.2.3 Closed-Form Approach

As for power control, a closed-form FP approach can also be developed for the beamforming problem. The main idea is the same as in power control, but in a multidimensional vector space. The sum logarithm problem is first reformulated in a sum-of-ratios form using a Lagrangian dual transform; the quadratic transform is subsequently applied to the ratios. After applying the vector Lagrangian dual transform in Theorem 10 in Section 2.5.1 to (3.26), we arrive at a

sum-of-ratios reformulation with $f_r^{\mathrm{CF}}(\underline{\mathbf{v}}, \underline{\gamma})$ as

$$f_r^{\mathrm{CF}}(\underline{\mathbf{v}}, \underline{\gamma}) = \sum_{(i,m)} w_{im} \Bigg(\log(1+\gamma_{im}) - \gamma_{im} + (1+\gamma_{im}) \mathbf{v}_{im}^H \mathbf{H}_{im,i}^H \cdot \left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n)} \mathbf{H}_{im,j} \mathbf{v}_{jn} \mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H \right)^{-1} \mathbf{H}_{im,i} \mathbf{v}_{im} \Bigg). \quad (3.30)$$

When $\underline{\mathbf{v}}$ is fixed, the optimal γ_{im} can be found by setting $\partial f_r^{\mathrm{CF}} / \partial \gamma_{im}$ to zero with respect to each (i,m) pair, that is

$$\gamma_{im}^\star = \mathbf{v}_{im}^H \mathbf{H}_{im,i}^H \left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n) \neq (i,m)} \mathbf{H}_{im,j} \mathbf{v}_{jn} \mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H \right)^{-1} \mathbf{H}_{im,i} \mathbf{v}_{im}. \quad (3.31)$$

The multidimensional quadratic transform in Theorem 9 in Section 2.5.1 can then be readily applied to further recast f_r^{CF} to

$$f_q^{\mathrm{CF}}(\underline{\mathbf{v}}, \underline{\gamma}, \underline{\mathbf{y}}) = \sum_{(i,m)} \Bigg(2\sqrt{w_{im}(1+\gamma_{im})}\, \Re\{ \mathbf{v}_{im}^H \mathbf{H}_{im,i}^H \mathbf{y}_{im} \} - \mathbf{y}_{im}^H \left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n)} \mathbf{H}_{im,j} \mathbf{v}_{jn} \mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H \right) \mathbf{y}_{im} \Bigg) + \mathrm{const}(\underline{\gamma}), \quad (3.32)$$

where $\mathrm{const}(\underline{\gamma})$ is a constant term when $\underline{\gamma}$ is fixed.

The above f_q^{CF} reformulation is obtained by treating $\sqrt{w_{im}(1+\gamma_{im})} \mathbf{H}_{im,i} \mathbf{v}_{im}$ as the numerator vector and also treating $\left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n)} \mathbf{H}_{im,j} \mathbf{v}_{jn} \mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H \right)$ as the denominator matrix in Theorem 9. The weighted sum-rate maximization problem (3.26) is then reformulated as

$$\begin{aligned}
\underset{\underline{\mathbf{v}}, \underline{\gamma}, \underline{\mathbf{y}}}{\text{maximize}} \quad & f_q^{\mathrm{CF}}(\underline{\mathbf{v}}, \underline{\gamma}, \underline{\mathbf{y}}) && (3.33a) \\
\text{subject to} \quad & \sum_{m=1}^{M} \|\mathbf{v}_{im}\|_2^2 \leq P_{\max}, \ \forall i \in \mathcal{B} && (3.33b) \\
& \gamma_{im} \in \mathbb{R}, \ \forall (i,m) && (3.33c) \\
& \mathbf{y}_{im} \in \mathbb{C}^N, \ \forall (i,m). && (3.33d)
\end{aligned}$$

The merit of reformulating f_r^{CF} as f_q^{CF} is to facilitate iterative optimization over $\mathbf{v}_{im}$. With the other variables fixed, the optimal $\mathbf{y}_{im}$ can be found by solving $\partial f_q^{\mathrm{CF}} / \partial \mathbf{y}_{im} = \mathbf{0}$, *i.e.*,

$$\mathbf{y}_{im}^\star = \left(\sigma^2 \boldsymbol{I}_N + \sum_{(j,n)} \mathbf{H}_{im,j} \mathbf{v}_{jn} \mathbf{v}_{jn}^H \mathbf{H}_{im,j}^H \right)^{-1} \sqrt{w_{im}(1+\gamma_{im})} \mathbf{H}_{im,i} \mathbf{v}_{im}. \quad (3.34)$$

Algorithm 5: Closed-Form FP for Beamforming

1 Initialize $\underline{\mathbf{v}}$ and $\underline{\gamma}$ to feasible values;
2 **repeat**
3 Update $\underline{\mathbf{y}}$ by (3.34);
4 Update $\underline{\gamma}$ by (3.31);
5 Update $\underline{\mathbf{v}}$ by (3.35);
6 **until** *the value of function* f_q^{CF} *in (3.32) converges*;

Likewise, when the other variables are fixed, the optimal $\underline{\mathbf{v}}$ is

$$\mathbf{v}_{im}^{\star} = \left(\eta_i \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{jn,i}^H \mathbf{y}_{jn} \mathbf{y}_{jn}^H \mathbf{H}_{jn,i}\right)^{-1} \sqrt{w_{im}(1+\gamma_{im})} \mathbf{H}_{im,i}^H \mathbf{y}_{im}, \tag{3.35}$$

where η_i is a dual variable introduced for the power constraint, optimally determined by (due to complementary slackness)

$$\eta_i^{\star} = \inf\left\{\eta_i \geq 0 : \sum_{m=1}^{M} \|\mathbf{v}_{im}(\eta_i)\|_2^2 \leq P_{\max}\right\}. \tag{3.36}$$

Note that the optimal η_i in (3.36) can be determined efficiently by bisection search. Algorithm 5 summarizes the above steps.

We remark that the proposed FP framework in this particular beamforming case, *i.e.*, Algorithm 5, is equivalent to the well-known WMMSE algorithm [40, 41]. This can be verified by substituting $\underline{\gamma}$ and $\underline{\mathbf{y}}$ in the updating formula of $\underline{\mathbf{v}}$. We will explore this connection further in Section 4.4. Like Algorithm 3, Algorithm 5 is not a BCD but its convergence can be established, *e.g.*, by the MM interpretation from Section 2.6.

3.2.4 Numerical Results

The simulation model assumes the same setting as in Section 3.1.5 for network topology, AWGN, distance-dependent pathloss, max transmit power, except that two users are randomly located within each cell and that the BSs and the users are now equipped with 2 antennas each. Consider Rayleigh fading for the channel coefficients. We pursue a maximization of sum rate in the network by setting all the weights $w_{im} = 1$.

Fig. 3.3 compares the different FP approaches. It shows that direct FP converges in fewer iterations than the closed-form FP (which is equivalent to the WMMSE algorithm [41]), e.g., the former achieves a sum rate of 470 Mbps within 10 iterations but the latter needs 25 iterations. However, counting just the number of iterations is misleading. The closed-form FP is in fact much more efficient than direct FP on a per-iteration basis, because closed-form FP updates all variables in closed form, while direct FP requires solving a convex optimization in each

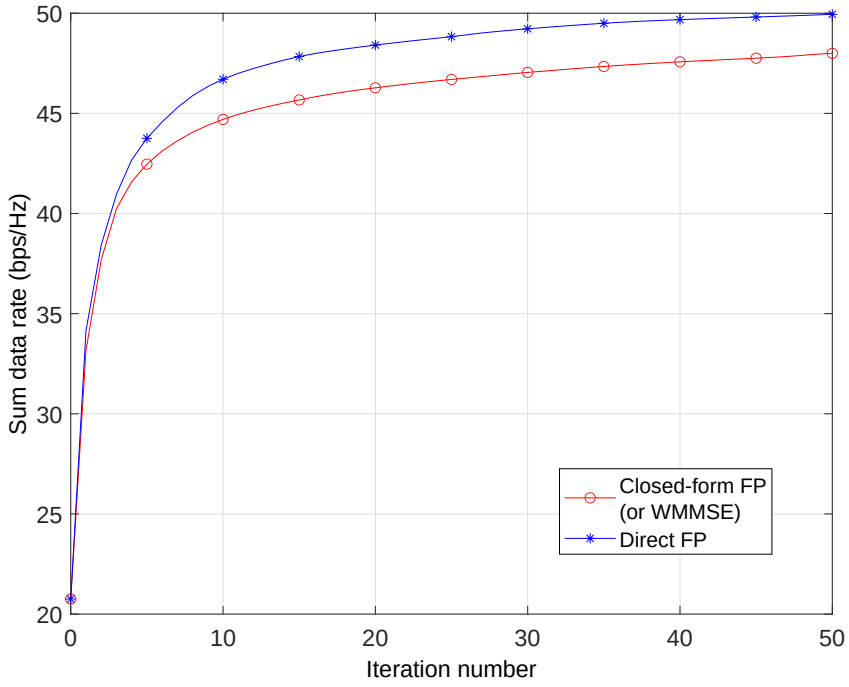


Figure 3.3: Beamforming for sum data rate maximization.

iteration. Therefore, the closed-form FP algorithm is much preferred as compared to the direct approach.

3.3 Energy Efficiency Maximization

As a final application example in this chapter, we illustrate the use of FP for solving energy efficiency maximization problems, both for the single-link case which has been treated in prior FP literature, and for the multiple-link case which requires the new techniques developed in this thesis.

3.3.1 Link-Level Problem Formulation

Consider an isolated end-to-end wireless link; the sender and the receiver are equipped with one antenna each. Let $h \in \mathbb{C}$ be the link channel, and let σ^2 be the AWGN power level. The total power consumption consists of two parts: the transmit power p which is constrained by a power budget $P_{\max}$, and a constant link ON-power P_{on}. The objective is to maximize the ratio of data rate to the total power consumption, namely the energy efficiency, by optimizing p, *i.e.*,

$$\underset{p}{\text{maximize}} \quad \frac{\log\left(1+|h|^2p/\sigma^2\right)}{p+P_{\text{on}}} \tag{3.37a}$$

$$\text{subject to} \quad 0 \le p \le P_{\max}. \tag{3.37b}$$

This problem is nonconvex in general.

For this single-link case, (3.37) is a single-ratio concave-convex FP problem and thus its globally optimal solution can be found using the conventional FP technique (*e.g.*, Dinkelbach's method), as already shown in the past literature [12–15]. An alternative is to apply our proposed quadratic transform. The problem is then reformulated as

$$\underset{p,\,y}{\text{maximize}} \quad 2y\sqrt{\log\left(1+\frac{|h|^2p}{\sigma^2}\right)} - y^2\left(p+P_{\text{on}}\right) \tag{3.38a}$$

$$\text{subject to} \quad 0 \le p \le P_{\max}. \tag{3.38b}$$

Clearly, the optimal y for fixed p is

$$y^\star = \frac{\sqrt{\log\left(1+|h|^2p/\sigma^2\right)}}{p+P_{\text{on}}}. \tag{3.39}$$

Then solving p for fixed y is a convex problem. This iteration converges to the global optimum according to Theorem 7.

Furthermore, [12] suggests a simple extension to include multiple links. Consider a total of n links. Let p_i be the transmit power of the ith link, and let h_i be its channel. The power constraint is imposed on the sum transmit power across the links. A crucial assumption here is that these links use separate spectrum bands and thus do not interfere with each other. A max-min energy efficiency problem is formulated as

$$\underset{\underline{p}}{\text{maximize}} \quad \min_i \left\{\frac{\log\left(1+|h_i|^2p_i/\sigma^2\right)}{p_i+P_{\text{on}}}\right\} \tag{3.40a}$$

$$\text{subject to} \quad p_i \ge 0,\ \forall i \tag{3.40b}$$

$$\sum_{i=1}^{n} p_i = P_{\max}. \tag{3.40c}$$

Observe that the above problem is a concave-convex max-min-ratio problem, so either the generalized Dinkelbach's method of Theorem 4 or the quadratic transform can be used to find the globally optimum solution.

Although the above energy efficiency maximization problem involves more than one wireless links, its objective function the pointwise minimum across the multiple links and thus can be easily tackled. Actually, the power variables of the different links are still optimized by the generalized Dinkelbach's method in Section 2.2.1 on a per-link basis to a large extent, so we

still deem (3.40) as a link-level problem.

A key component missing in the above link-level setup is "interference". It is far more challenging and yet far more worthwhile to consider energy efficiency at a system level with cross-link interference taken into consideration.

3.3.2 System-Level Problem Formulation

Energy efficient maximization across multiple interfering links is a more challenging problem. Consider a spatial multiplexing multiple-antenna broadcast channel model with one sender equipped with M antennas to send individual data to its M receivers. Assume that every receiver has N antennas and supports one data stream. Let $\mathbf{H}_m \in \mathbb{C}^{N\times M}$ be the channel between the sender and the mth receiver; let $\mathbf{v}_m \in \mathbb{C}^M$ be the beamformer for the transmission to the mth receiver. The energy efficiency maximization problem in this case is formulated as

$$\underset{\underline{\mathbf{v}}}{\text{maximize}} \quad \frac{\sum_{m=1}^{M} R_m(\underline{\mathbf{v}})}{\sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 + P_{\text{on}}} \tag{3.41a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 \le P_{\max}, \tag{3.41b}$$

where the rate function $R_m(\underline{\mathbf{v}})$ of receiver m is

$$R_m(\underline{\mathbf{v}}) = \log\left(1 + \mathbf{v}_m^H \mathbf{H}_m^H \left(\sigma^2 \boldsymbol{I}_N + \sum_{n\neq m} \mathbf{H}_m \mathbf{v}_n \mathbf{v}_n^H \mathbf{H}_m^H\right)^{-1} \mathbf{H}_m \mathbf{v}_m\right). \tag{3.42}$$

We first describe the approach in [12–15]. Dinkelbach's method recasts the objective function to

$$f_d(\underline{\mathbf{v}}, y) = \sum_{m=1}^{M} R_m(\underline{\mathbf{v}}) - y\left(\sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 + P_{\text{on}}\right). \tag{3.43}$$

However, unlike the single-link case, the reformulation f_d is no longer a concave function of $\underline{\mathbf{v}}$, so optimizing $\underline{\mathbf{v}}$ for fixed y is numerically difficult. Hence, the iterative algorithm based on Dinkelbach's method cannot be easily extended to the multiple-link scenario. In fact, [15] considers multiple links only under the assumption that the resulting SINRs are all sufficiently high; [14] globally solves the f_d maximization problem using a monotonic optimization approach (which has an exponential-time complexity), and also proposes a polynomial-time algorithm to attain a stationary point when the transmitter has a single antenna (*i.e.*, when $\mathbf{v}_m$ reduces to a scalar).

Moreover, [42] proposes a gradient method to maximize the nonconcave function f_d in (3.43), and [43] advocates successive convex approximation. But none of them can find in polynomial time the globally optimal $\underline{\mathbf{v}}$ that maximizes f_d. We remark that the optimality of $\underline{\mathbf{v}}$ in maximizing f_d is critical to the convergence of the Dinkelbach's method [7], so these existing polynomial-time algorithms are not guaranteed to converge in general. By contrast,

our approach does not rely on the Dinkelbach's method, and has provable convergence. As a further remark, if the sum rate objective function is changed to the superposition coding inner bound, the new problem after the Dinkelbach's method would have been convex and can be optimally solved by a water-filling scheme [44].

3.3.3 Iterative Optimization by Nested Fractional Programming

We advocate a novel use of the quadratic transform to address the problem. First, apply the single-ratio quadratic transform (*i.e.*, Theorem 3) to decouple the energy efficiency ratio as

$$f_q(\underline{\mathbf{v}}, y) = 2y\left(\sum_{m=1}^{M} R_m(\underline{\mathbf{v}})\right)^{\frac{1}{2}} - y^2\left(\sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 + P_{\text{on}}\right). \tag{3.44}$$

The same issue now arises as with Dinkelbach's method—the reformulated objective function is not concave in $\underline{\mathbf{v}}$. It is crucial to note that the function $x^{\frac{1}{2}}$ is nondecreasing and concave, and also that the second term in (3.44) is concave. Thus, the concavity of f_q over $\underline{\mathbf{v}}$ can be restored if the term inside the square root $\sum_{m=1}^{M} R_m$ is recast as a concave function.

Following this idea, we apply the (multidimensional) quadratic transform to each SINR term inside the R_m expression (3.42) in f_q, and further recast f_q to f_{qq}:

$$f_{qq}(\underline{\mathbf{v}}, y, \underline{\mathbf{z}}) = 2y\left(\sum_{m=1}^{M} \log\left(1 + 2\Re\{\mathbf{z}_m^H \mathbf{H}_m \mathbf{v}_m\} - \mathbf{z}_m^H\left(\sigma^2 \boldsymbol{I}_N + \sum_{n\neq m} \mathbf{H}_m \mathbf{v}_n \mathbf{v}_n^H \mathbf{H}_m^H\right)\mathbf{z}_m\right)\right)^{\frac{1}{2}} - y^2\left(\sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 + P_{\text{on}}\right). \tag{3.45}$$

The ultimate reformulation of (3.41) after the two uses of the quadratic transform now becomes

$$\begin{aligned} \underset{\underline{\mathbf{v}}, y, \underline{\mathbf{z}}}{\text{maximize}} \quad & f_{qq}(\underline{\mathbf{v}}, y, \underline{\mathbf{z}}) && \text{(3.46a)} \\ \text{subject to} \quad & \sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 \leq P_{\max} && \text{(3.46b)} \\ & \mathbf{z}_m \in \mathbb{C}^N, \ \forall m. && \text{(3.46c)} \end{aligned}$$

We emphasize that y is introduced by the first use of FP for decoupling the energy efficiency ratio while $\mathbf{z}_m$ is introduced by the second use of FP for decoupling the SINR terms.

We propose an iterative optimization. When all the other variables are held fixed, the optimal $\mathbf{z}_m$ is

$$\mathbf{z}_m^\star = \left(\sigma^2 \boldsymbol{I}_N + \sum_{n\neq m} \mathbf{H}_m \mathbf{v}_n \mathbf{v}_n^H \mathbf{H}_m^H\right)^{-1} \mathbf{H}_m \mathbf{v}_m, \ \forall m. \tag{3.47}$$

Algorithm 6: Nested FP for Energy Efficiency Maximization

```
Initialize v to a feasible value;
repeat
    Update z by (3.47);
    Update y by (3.48);
    Update v by solving the convex optimization problem
    (3.46) for fixed z and y;
until the value of function f_qq in (3.45) converges;
```

After the update of $\underline{\mathbf{z}}$, the optimal y is

$$y^\star = \frac{\sqrt{\sum_{m=1}^{M} R_m(\underline{\mathbf{v}})}}{\sum_{m=1}^{M} \|\mathbf{v}_m\|_2^2 + P_{\text{on}}}. \tag{3.48}$$

Most importantly, when $\underline{z}$ and y are both fixed, (3.46) is a convex problem of $\mathbf{v}_m$, and therefore the optimal $\mathbf{v}_m$ can be efficiently found using the standard numerical method.

This iterative optimization is summarized in Algorithm 6. We refer to it as the nested FP approach, because the reformulating procedure involves an outer FP for the energy efficiency ratio as well as an inner FP for the nesting SINR terms. Based on the equivalence of objective function property C3 stated in Section 2.1.2, it is easy to verify the convergence of Algorithm 6 to a stationary point of the original problem (3.41) with the energy efficiency value nondecreasing after each iteration.

3.3.4 Numerical Results

The simulation model assumes flat-fading channel(s) over a 1 MHz-wide frequency band. The maximum transmit power level is set to be -39 dBm/Hz; the ON-power level is set to be 5 dBm; the background noise level is set to be -160 dBm/Hz. We test the proposed algorithm for two network scenarios:

- Single-link case: Consider one pair of sender and receiver, equipped with one antenna each; the channel coefficient between them is modeled with -120 dB pathloss.

- Multiple-link case: Consider one sender and three receivers; the sender has 3 antennas and the receivers have 2 antennas each. The channel coefficients between the transmit and receive antennas are modeled with independent and identically distributed (i.i.d.) Rayleigh fading component plus -120 dB pathloss.

Fig. 3.4 compares the Dinkelbach's transform approach [12–15] and the proposed quadratic transform in maximizing energy efficient for the single-link case. It can be observed that Dinkelbach's transform gives a faster convergence. To attain the optimal energy efficiency,

Dinkelbach's transform needs 4 iterations while the quadratic transform needs 8 iterations. This result agrees with the convergence rate analysis in Section 2.3.2.

Fig. 3.5 evaluates the performance of Algorithm 6 in maximizing the multiple-link energy efficiency. We reiterate that Dinkelbach's transform [12–15] is not applicable in this case. As can be seen from the figure, Algorithm 6 raises the energy efficiency significantly to more than four-fold after just 8 iterations.

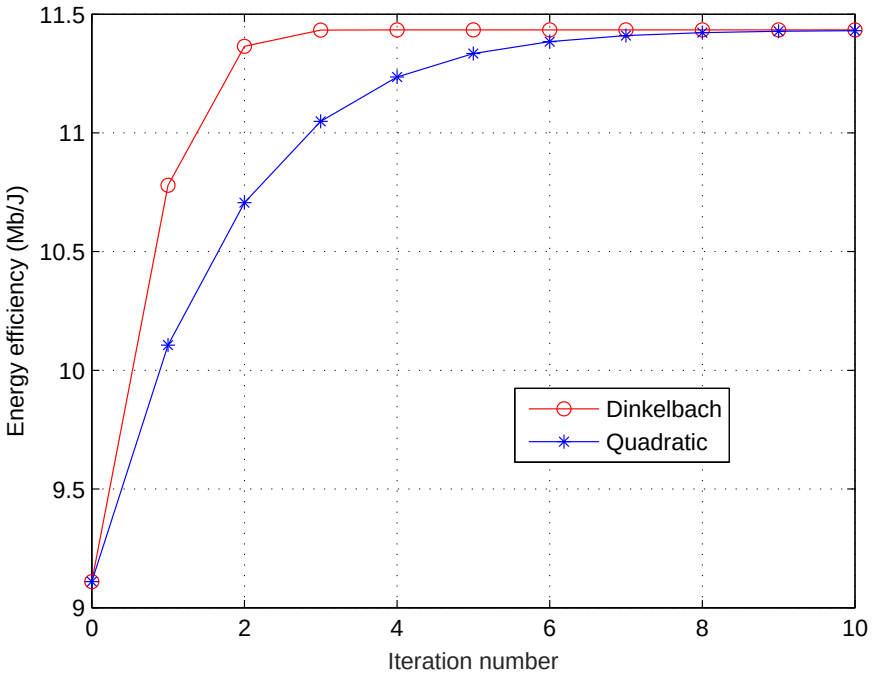


Figure 3.4: Energy efficiency maximization for a single link.

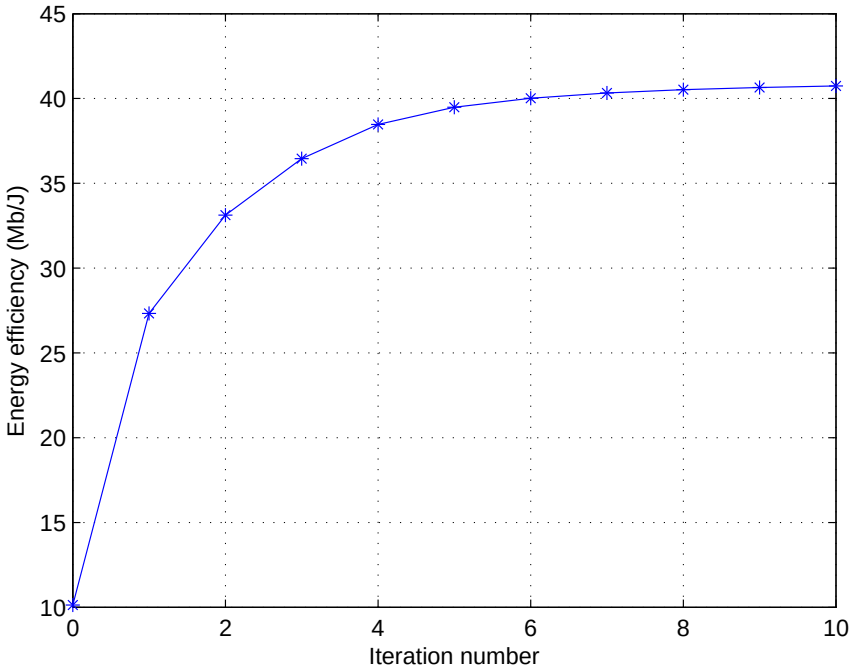


Figure 3.5: Energy efficiency maximization for a broadcast network.

3.4 Summary

We use the new FP technique, the quadratic transform, to tackle a broad range of continuous FP problems with multiple ratios in contrast to the conventional techniques which can only handle single ratio or the max-min case. Based on the quadratic transform, a variety of FP approaches are devised for solving the continuous problems in communication systems, *i.e.*, power control, beamforming, and energy efficiency maximization. The proposed FP approaches recast the original nonconvex problem to a sequence of convex problems, thereby allowing efficient iterative optimization with provable convergence to a stationary point solution.

Chapter 4

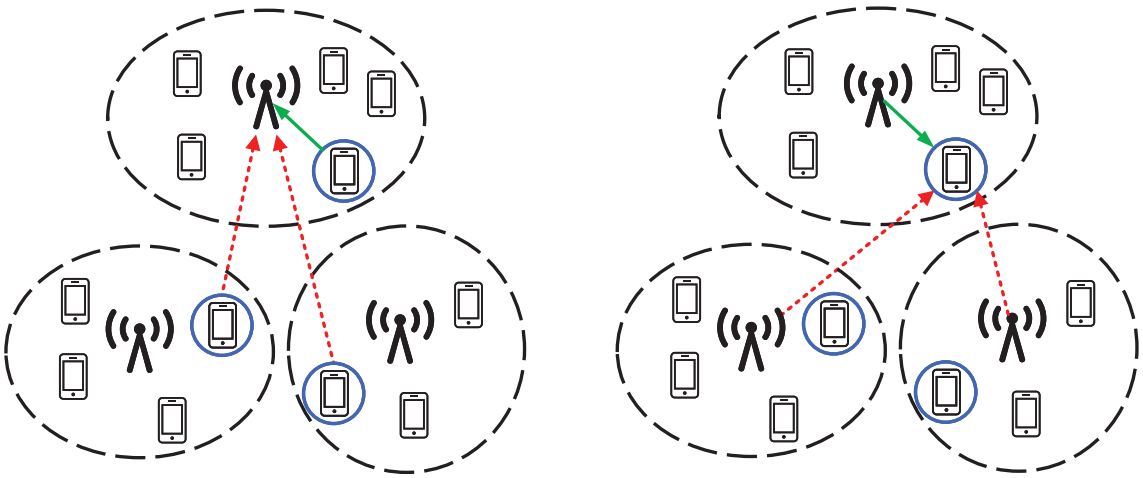
Discrete Optimization Problems

This chapter tackles the discrete problems, such as those involving user scheduling, which are considerably more difficult to solve. Unlike the continuous problems, discrete or mixed discrete-continuous problems normally cannot be recast as convex problems. In contrast to the common heuristic of relaxing the discrete variables [21], this work reformulates the original problem in an FP form amenable to distributed combinatorial optimization. We illustrate this approach by considering an important and challenging problem of uplink coordinated multi-cell user scheduling in wireless cellular systems. Uplink scheduling is more challenging than downlink scheduling, because uplink user scheduling decisions significantly affect the interference pattern in nearby cells. Further, the discrete scheduling variable needs to be optimized jointly with continuous variables such as transmit power levels and beamformers. The central idea of the proposed FP approach is to decouple the interaction among the interfering links, thereby permitting a distributed and joint optimization of the discrete and continuous variables with provable convergence. Importantly, it is shown that the well-known WMMSE algorithm is equivalent to a particular way of FP, but the proposed way is far more suited for optimizing discrete variables like the scheduling decision.

4.1 Single-Antenna Uplink User Scheduling

The goal of this section is to optimally schedule uplink users and to set their transmit power levels jointly across multiple cells so as to maximize the network utility in a SISO network. The problem involves mixed continuous variables (power) and discrete variables (uplink scheduling); it is quite challenging, because scheduling and power decisions in each cell significantly affect the interference patterns in neighboring cells. This section proposes an FP-based reformulation that allows power control and uplink scheduling to be determined jointly and in a distributed fashion with the assistance of some auxiliary variables. We remark that this approach can be further extended to apply to the full-duplex [45] scenario where uplink and downlink coexist.

We explore the use of FP for optimization problems that involve discrete variables within the $\log(1+\mathrm{SINR})$ rate expressions—in particular the problem of coordinated multi-cell uplink user



(a) Uplink multicell network (b) Downlink multicell network

Figure 4.1: Interference pattern depends on the user scheduling in the neighboring cells in the uplink, but not so in the downlink. Here, the solid lines represent the desired signal; the dashed lines represent the interfering signal; the scheduled user terminal in each cell is circled.

scheduling in wireless cellular networks, where the optimization parameters are the selection of which users to schedule in each cell, along with their power and beamforming vectors.

The user scheduling problem in the uplink is more challenging than in the downlink, because the uplink interference pattern depends strongly on the scheduling decisions of the neighboring cells, whereas in the downlink, the interference pattern does not depend on scheduling decisions if the powers are fixed at the BSs, as illustrated in Fig. 4.1. Nonetheless, if the powers are not fixed, then the interference in downlink network can be affected by scheduling decisions as well—a BS without any user scheduled would be deactivated.

4.1.1 Problem Formulation

Consider the uplink of a wireless cellular network. Let $\mathcal{B}$ be the set of BSs deployed in the network, and let $\mathcal{K}_i$ be the set of users who are associated with BS i. Each BS i together with its associated users in $\mathcal{K}_i$ forms a cell. In every time-slot, users are scheduled for uplink transmission on a cell basis. In this section, the BSs and the users are assumed to be equipped with a single antenna each; extension to the multiple-antenna case involving beamforming optimization is considered in the next section. For the user scheduling and power control purpose, introduce variable $s_i \in \mathcal{K}_i$ to denote the user to be scheduled at BS i, and introduce variable p_k to denote the transmit power level of user k if it gets scheduled for uplink transmission. Let $h_{i,k} \in \mathbb{C}$ be the uplink channel coefficient from user k to BS i; let σ^2 be the power level of AWGN. Given a set of weights w_k that reflect the user priorities in each time-slot, we have the following

weighted sum rate maximization objective:

$$f_o(\underline{s},\underline{p}) = \sum_{i\in\mathcal{B}} w_{s_i} \log\left(1+\frac{|h_{i,s_i}|^2 p_{s_i}}{\sum_{j\neq i}|h_{i,s_j}|^2 p_{s_j}+\sigma^2}\right). \tag{4.1}$$

The joint scheduling and power control problem in an uplink SISO network can be written as

$$\begin{aligned} \underset{\underline{s},\underline{p}}{\text{maximize}} \quad & f_o(\underline{s},\underline{p}) && \text{(4.2a)}\\ \text{subject to} \quad & 0\leq p_k \leq P_{\max},\ \forall k && \text{(4.2b)}\\ & s_i \in \mathcal{K}_i \cup \{0\},\ \forall i, && \text{(4.2c)} \end{aligned}$$

where $P_{\max}$ is the maximum transmit power level of the user and 0 refers to the decision of not scheduling any user. Because of the SISO setting, at most one user can be scheduled in each cell i; we set $s_i = k$ if some user k is scheduled in the cell, and set $s_i = 0$ otherwise. In particular, w_{s_i} and p_{s_i} are both implicitly set to zero if $s_i = 0$.

The above problem is difficult to tackle directly due to the fact that the uplink scheduling decisions have significant impact on the interference pattern. A particular scheduling decision s_i in cell i strongly influences the scheduling decisions s_j in its neighboring cells. In addition, even when the discrete variable $\underline{s}$ is held fixed, solving for the power variable $\underline{p}$ in (4.2) is still nontrivial, because the objective function is nonconvex.

4.1.2 Implicit Scheduling by Power Control

Before proceeding to the proposed FP approach, we discuss an alternative perspective of treating the uplink scheduling problem as a power control problem, and explain why the corresponding optimization method would not produce good results numerically.

As opposed to formulating the joint uplink scheduling and power control as a mixed discrete-continuous problem as in (4.2), we could replace the scheduling variable $\underline{s}$ with the power variable $\underline{p}$, based on the observation that a user k is scheduled if and only if its power level p_k is positive. Then, the problem can be converted to a continuous power optimization over all users. To formalize this idea, we rewrite the objective function as

$$f_o(\underline{p}) = \sum_{i\in\mathcal{B}}\sum_{k\in\mathcal{K}_i} w_k \log\left(1+\frac{|h_{i,k}|^2 p_k}{\sum_{k'\neq k}|h_{i,k'}|^2 p_{k'}+\sigma^2}\right), \tag{4.3}$$

where k' refers to any other user in the network, including those who are in the same cell as user k, *i.e.*, $k' \in \bigcup_{i\in\mathcal{B}} \mathcal{K}_i$. The uplink scheduling problem can then be rewritten as a power

optimization problem involving only the power variable $\underline{p}$:

$$\underset{\underline{p}}{\text{maximize}} \quad f_o(\underline{p}) \tag{4.4a}$$

$$\text{subject to} \quad 0 \leq p_k \leq P_{\max},\ \forall k. \tag{4.4b}$$

Although strictly speaking, the above optimization problem does not have the constraint that at most only one user can be active, the optimal solution of (4.4) does take such a form in most practical regime of interest. In this case, the two problems (4.4) and (4.2) are equivalent, *i.e.*, the optimal solution $(\underline{s}^\star, \underline{p}^\star)$ of (4.2) can recover the optimal $\underline{p}^\star$ for (4.4), and vice versa.

Problem (4.4) is nonconvex, but it can be solved by using the gradient method to attain a stationary point, or by using the FP methods from the previous chapter, either the direct approach or the closed-form approach. After solving (4.4), we simply schedule those users with positive p_k.

However, as a subtle point we wish to highlight, using a power control algorithm to solve the scheduling problem has a serious deficiency. The main problem is that due to the highly nonconvex nature of the objective function, the stationary point of a power control algorithm is highly sensitive to the initial condition. As a result, this class of methods suffers from a serious *premature turning-off* issue. If some link is deactivated in the early stage of the iterative optimization, it can never be reactivated in the later iterations, because its local gradient would strongly discourage it from doing so. Past efforts to convexify this power control problem, e.g, by approximating the problem as a geometric program [35], essentially smooths out the local optima; but it works only at high SINR. For the scheduling problem, most of the links have low SINRs—in fact, due to intra-cell interference, at most one link in each cell can have its SINR higher than 1.

The main contribution of this part of work is to show that a novel use of the Lagrangian dual transform, coupled with the quadratic transform, can avoid the premature turning-off issue through weighted bipartite matching.

4.1.3 Pricing Method by Fractional Programming

The scheduling decision and the transmit power level of the scheduled user in each cell interact with its neighboring cells through the interference term in the denominator of rate expression in the objective function. A naive way for tackling the problem would be to make scheduling and power allocation decisions on an individual per-cell basis, assuming that the interference is fixed, then update the interference terms, and iterate between the cells. But such an approach does not work well, because the interference pattern can drastically change when a different user is scheduled; there is no guarantee that the iteration would even converge.

The main idea of our proposed method is to devise a way of using FP to enable the individual update of scheduling and power on a per-cell basis, while ensuring convergence. Toward this end, the quadratic transform and the Lagrangian dual transform are used together to recast the

problem in a sequence of equivalent forms. We remark that applying the quadratic transform alone cannot achieve this desired decoupling.

First, apply the Lagrangian dual transform to reformulate the original objective function $f_o(\underline{s}, \underline{p})$ as

$$f_r(\underline{s}, \underline{p}, \underline{\gamma}) = \sum_{i \in \mathcal{B}} w_{s_i} \log(1 + \gamma_i) - \sum_{i \in \mathcal{B}} w_{s_i} \gamma_i + \sum_{i \in \mathcal{B}} \frac{w_{s_i}(\gamma_i + 1)|h_{i,s_i}|^2 p_{s_i}}{\sum_j |h_{i,s_j}|^2 p_{s_j} + \sigma^2}. \tag{4.5}$$

The original problem (4.2) is now reformulated as

$$\underset{\underline{s}, \underline{p}, \underline{\gamma}}{\text{maximize}} \quad f_r(\underline{s}, \underline{p}, \underline{\gamma}) \tag{4.6a}$$

$$\text{subject to} \quad \text{(4.2b), (4.2c).} \tag{4.6b}$$

We propose to optimize all the variables iteratively. When $(\underline{s}, \underline{p})$ are held fixed, the optimal $\underline{\gamma}$ can be explicitly determined by setting $\partial f_r / \partial \gamma_i$ to zero, *i.e.*,

$$\gamma_i^\star = \frac{|h_{i,s_i}|^2 p_{s_i}}{\sum_{j \neq i} |h_{i,s_j}|^2 p_{s_j} + \sigma^2}. \tag{4.7}$$

Next, we apply the quadratic transform on the fractional term in (4.5) in order to to optimize $(\underline{s}, \underline{p})$ in f_r for fixed $\underline{\gamma}$. Introduce an auxiliary variable y_i for each ratio $\frac{w_{s_i}(\gamma_i+1)|h_{i,s_i}|^2 p_{s_i}}{\sum_j |h_{i,s_j}|^2 p_{s_j} + \sigma^2}$ in the last term of $f_r(\underline{s}, \underline{p}, \underline{\gamma})$. We use the vector Lagrangian dual transform in Theorem 10 to further reformulate $f_r(\underline{s}, \underline{p}, \underline{\gamma})$ as $f_q(\underline{s}, \underline{p}, \underline{\gamma}, \underline{y})$ in (4.8):

$$\begin{aligned} f_q(\underline{s}, \underline{p}, \underline{\gamma}, \underline{y}) = \sum_{i \in \mathcal{B}} w_{s_i} \log(1 + \gamma_i) - \sum_{i \in \mathcal{B}} w_{s_i} \gamma_i + \sum_{i \in \mathcal{B}} \Bigg(& 2 y_i \sqrt{w_{s_i}(\gamma_i + 1) \left|h_{i,s_i}\right|^2 p_{s_i}} - \\ & y_i^2 \left(\sum_{j \in \mathcal{B}} \left|h_{i,s_j}\right|^2 p_{s_j} + \sigma^2 \right) \Bigg) \end{aligned} \tag{4.8}$$

$$\begin{aligned} = \sum_{i \in \mathcal{B}} \Bigg(& w_{s_i} \log(1 + \gamma_i) - w_{s_i} \gamma_i - y_i^2 \sigma^2 + 2 y_i \sqrt{w_{s_i}(\gamma_i + 1) \left|h_{i,s_i}\right|^2 p_{s_i}} - \\ & \sum_{j \in \mathcal{B}} y_j^2 \left|h_{j,s_i}\right|^2 p_{s_i} \Bigg), \end{aligned} \tag{4.9}$$

Hence, in order to solve problem (4.6) over $(\underline{s}, \underline{p})$, we can equivalently consider the following problem over $(\underline{s}, \underline{p}, \underline{y})$:

$$\underset{\underline{s}, \underline{p}, \underline{y}}{\text{maximize}} \quad f_q(\underline{s}, \underline{p}, \underline{\gamma}, \underline{y}) \tag{4.10a}$$

$$\text{subject to} \quad \text{(4.2b), (4.2c).} \tag{4.10b}$$

The overall strategy is then to iteratively optimize $\underline{\gamma}$ according to (4.7) and optimize $(\underline{s}, \underline{p}, \underline{y})$ as in (4.10).

The newly introduced objective function f_q groups the terms related to the same s_i together. The key observation is that the scheduling and power variables $(\underline{s}, \underline{p})$ are now *decoupled* in this new formulation (4.10). Specifically, the scheduling and power optimization in each cell, *i.e.*, (s_i, p_i), can be done independently in each cell, as long as $\underline{\gamma}$ and $\underline{y}$ are fixed. This motivates an iterative approach for solving (4.10).

We propose to maximize f_q over variables $\underline{\gamma}$, $\underline{y}$, $\underline{s}$ and $\underline{p}$ in an iterative manner as follows. The update of $\underline{\gamma}$ is already shown as in (4.7). When all the other variables are fixed, the optimal $\underline{y}$ can be obtained by setting $\partial f_q / \partial y_i$ to zero, *i.e.*,

$$y_i^\star = \frac{\sqrt{w_{s_i}(1+\gamma_i)|h_{i,s_i}|^2 p_{s_i}}}{\sum_{j\in\mathcal{B}} |h_{i,s_j}|^2 p_{s_j} + \sigma^2}. \tag{4.11}$$

Fixing $\underline{y}$ and $\underline{\gamma}$, if user k is to be scheduled by its associated BS j, we can derive its optimal transmit power level p_k by setting $\partial f_q / \partial p_k$ to zero. Subject to a maximum power constraints, the optimal p_k can be explicitly determined by

$$p_k = \min \left\{ P_{\max}, \frac{w_k(1+\gamma_i)\left|h_{i,k}\right|^2 y_i^2}{\left(\sum_{j\in\mathcal{B}} \left|h_{j,k}\right|^2 y_j^2\right)^2} \right\}, \ \forall k \in \mathcal{K}_i. \tag{4.12}$$

The most important part of the algorithm is the optimization of the scheduling variable $\underline{s}$. As stated previously, the objective function f_q has the desirable property that the optimization of $\underline{s}$ is decoupled on a per-cell basis, *i.e.*, the optimization of s_i does not depend on the other s_j variables for $j \neq i$, when $\underline{\gamma}$ and $\underline{y}$ are fixed. Now, since the optimal transmit power level p_k is already determined by (4.12) if user k is scheduled, we can substitute the optimized power p_k into f_q and make optimal scheduling decision through a simple search to find the user that maximizes f_q in each cell. Moreover, we can rewrite f_q in the form of difference between two positive functions, and formally state the scheduling decision as follows:

$$s_i^\star = \begin{cases} 0, \text{ if } \max\limits_{k\in\mathcal{K}_i} \left\{ G_i(k) - \sum\limits_{j\neq i} D_j(k) \right\} \leq 0; \\ \arg\max\limits_{k\in\mathcal{K}_i} \left\{ G_i(k) - \sum\limits_{j\neq i} D_j(k) \right\}, \text{ otherwise,} \end{cases} \tag{4.13}$$

where the functions $G_i(k)$ and $D_j(k)$ are defined as

$$G_i(k) = w_k \log(1+\gamma_i) - w_k\gamma_i - p_k y_i^2 \left|h_{j,k}\right|^2 + 2y_i\sqrt{w_k(1+\gamma_i)\left|h_{i,k}\right|^2 p_k}, \ \forall k \in \mathcal{K}_i \tag{4.14}$$

Algorithm 7: Joint Uplink Scheduling and Power Control

```
Initialize s, p, and γ to feasible values;
repeat
    Update y by (4.11);
    Update γ by (4.7);
    Update (s, p) jointly by (4.13) and (4.12);
until the value of function f_q in (4.8) converges;
```

and

$$D_j(k) = y_j^2 \left|h_{j,k}\right|^2 p_k, \ \forall k \notin \mathcal{K}_j. \tag{4.15}$$

In the above equation (4.13), we interpret $G_i(k)$ and $D_j(k)$ as the utility and penalty functions, respectively, so that the scheduling decision has an intuitive utility-minus-price structure. More precisely, $G_i(k)$ is the utility gain of scheduling user k at BS i and $D_j(k)$ is the penalty for interfering a neighboring cell j by scheduling user k. The best user to schedule is the one that balances these two effects. Note that the scheduling and power control are done on a per-cell basis. This enables distributed implementation.

Furthermore, when the max value of $G_i(k) - \sum_{j \neq i} D_j(k)$ at BS i is less than zero, it implies that no user should be scheduled at this BS i in the time slot in order to reduce the intercell interference suffered by the neighboring BSs. This situation possibly occurs in an ultra-dense uplink network scenario.

We summarize the proposed joint scheduling and power control strategy in Algorithm 7. Note that the algorithm is not a conventional block coordinate ascent method, because the optimizing objective function is not fixed, *i.e.*, $\underline{s}$, $\underline{p}$ and $\underline{y}$ are optimally updated for f_q while $\underline{\gamma}$ is optimally updated for f_r. Nevertheless, its convergence can be established by using the MM interpretation in Section 2.6.

Proposition 6. *Algorithm 7 is guaranteed to converge, with the weighted sum rate f_o monotonically nondecreasing after each iteration. The converged solution is a stationary point of f_o with respect to $\underline{p}$ if $\underline{s}$ is assumed to be fixed.*

We note that due to the nonconvex nature of the problem with respect to $\underline{p}$, finding a stationary point in $\underline{p}$ is likely to be the best that one can do. Moreover, since $\underline{s}$ is a discrete variable, it is difficult to assert any optimality with respect to $\underline{s}$. In fact, we can show that even with $\underline{p}$ fixed, finding the optimal $\underline{s}$ is NP-hard.

To see the NP-hardness, we can use an argument inspired by [46] in which the NP-hardness of the power control problem is established. Construct a simplified example, in which each BS receives interference from a subset of neighboring users only, and the interference level is large so whenever interference is present the rate is effectively zero, and otherwise the rate is one. Selecting one user in each cell to maximize the overall sum rate now amounts to solving a maximum independent set problem on a graph, which is NP-hard. Further, unless P = NP,

Table 4.1: Sum log-utilities of FP-based coordinated uplink scheduling and power control as compared to the baselines.

Algorithm	Total Log-Utility
Power Control by WMMSE	27.17
Fixed Interference	52.16
Proposed FP Method	60.15

it is impossible even to solve the problem within a constant approximation ratio in polynomial time [47].

Observe here that Algorithm 7 avoids premature turning-off. Even if a user k is not activated in the tth iterate, the related auxiliary variable y_i is still nonzero according to (4.11), so long as at least some other user is scheduled in its cell. Thus, user k still stands a chance to be reactivated in future iterations when the interference pattern becomes favorable, as indicated by (4.12). Furthermore, if all the users in cell i have been turned off, then the corresponding BS will be turned off as well, and consequently no user in the cell will be rescheduled in the further iterations. To resolve this issue, one heuristic method is to put a positive lower bound on y_i in order to avoid shutting BS i off.

As a final remark, throughout this chapter we have assumed that the full channel state information (CSI) is available. In practical implementations, the cost of channel estimation for all users however could be prohibitive. Further, including all users in the scheduling step can incur large computational complexity. The complexity in implementing Algorithm 7 can be lowered in practice using a two-stage scheduling strategy. We first roughly choose a subset of potential users according to their weights, then apply Algorithm 7 to refine the scheduling decision. This can greatly reduce the run-time complexity and the cost of acquiring CSI.

4.1.4 Numerical Results

To evaluate the performance of the proposed joint uplink scheduling and power control algorithm, numerical simulation is performed in a 7-cell wrapped-around topology with a total of 84 users uniformly placed in the network. The BS-to-BS distance is 0.8 km. Each user is associated with the strongest BS. The maximum transmit power spectral density (PSD) of the users is -47 dBm/Hz; the background noise PSD is set to be -169 dBm/Hz over 10 MHz bandwidth. The wireless channel model includes a distance-dependent pathloss component at $128.1 + 37.6 \log_{10}(d)$ dB (where the distance d is in km) and a log-normal shadowing component with 8dB standard deviation.

In the simulation, the joint user scheduling and power control problem is solved across the multiple cells in each time-slot with the user priority weights updated as the reciprocals of long-term average user rates over the time, in order to ensure proportional fairness across the users. Over time, this setting of the weights maximizes the log-utility, $\sum_k \log(\bar{R}_k)$, over all users in the network, where $\bar{R}_k$ is the long-term average rate of user k, expressed in Mbps in

the numerical results below.

The following two baseline uplink scheduling strategies are also simulated for comparison purpose:

- *Power Control*: The uplink scheduling and power control problem can also be thought of as a global power control problem, in which users not being scheduled are assigned zero power. Thus, we can run power control for all the users in the network at the same time. Most users will be assigned zero power; users assigned positive transmit power levels (typically at most one per cell) are the ones scheduled. This global power control problem is highly nonconvex. In the simulation, we use the WMMSE algorithm [40, 41] for power control to arrive at a local optimum.

- *Fixed Interference Method*: In this method, uplink scheduling and power control are performed iteratively. Each user is initialized with some powers. In the scheduling stage, the user that maximizes the weighted rate in each cell is chosen, assuming fixed interference pattern from the previous iteration. In the power control stage, the powers of the scheduled users are updated by solving a weighted sum rate maximization problem. We iterate between the two steps until convergence or a fixed number of iterations is reached.

Fig. 4.3 shows the cumulative distribution function (CDF) of the user data rates in the network and Table 4.1 lists the log-utility[1] achieved by the different methods for uplink user scheduling and power control. We see that the baseline of power control provides poor performance, mainly because the power control algorithm tends to stuck in a locally optimal solution of the nonconvex problem. The fixed-interference method is also not capable of arriving at a desirable solution. In contrast, the proposed algorithm performs much better in terms of utility, as shown in Table 4.1. Fig. 4.3 shows that the 10th-percentile user rate of the proposed algorithm is at least 50% more than that of the fixed-interference method. Since these low-rate users are typically located at the cell-edge where cross-cell interference is the strongest, this shows that the proposed FP-based algorithm is effective in alleviating interference by coordinating uplink scheduling and power control. We remark that this gain is achieved despite the low overall complexity of the FP method.

Moreover, in order to demonstrate that the proposed FP method indeed outperforms the other methods in maximizing the weighted sum rate, we now use the same rate weight sequence for all the algorithms, as displayed in Fig. 4.2; the weight sequence is generated by the FP method. For the sake of display clarity, we plot the normalized weighted sum rate, namely weighted sum rate divided by sum weight, versus the time slot. The figure shows that the proposed method outperforms the benchmarks consistently in maximizing the weighted sum rate.

[1]The utility is computed for data rate in Mbps throughout the thesis.

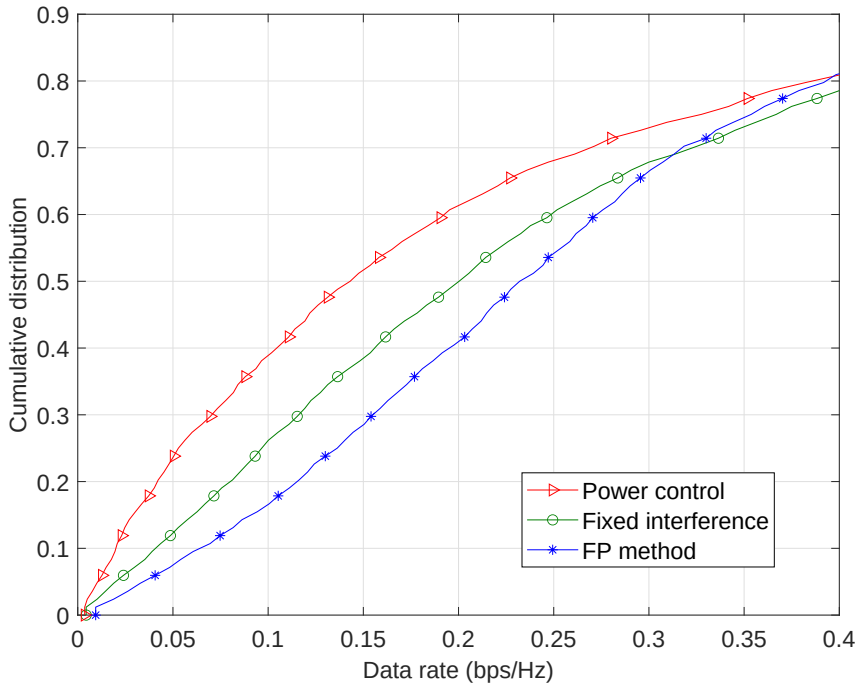


Figure 4.2: Comparison of the proposed FP-based coordinated uplink user scheduling and power control method with two baseline methods in terms of CDF of user rates.

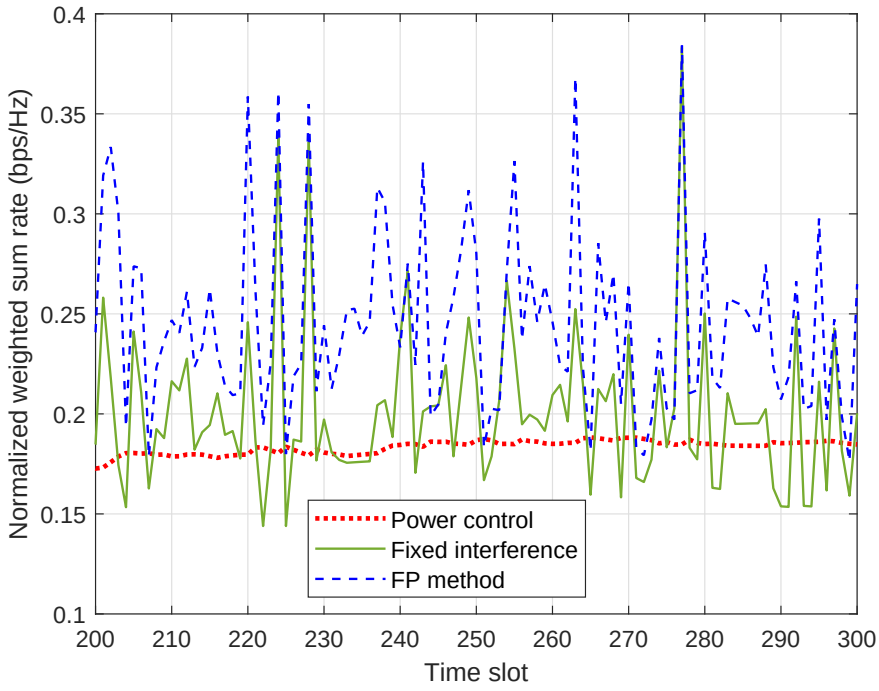


Figure 4.3: Normalized weighted sum rate vs. time slot when the weight sequence by the proposed FP method is used for all the methods.

4.2 Multi-Antenna Uplink User Scheduling

The objective here is to schedule uplink users and to set their transmit beamformers jointly across multiple cells so as to maximize the network utility in a MIMO network. The key step is to incorporate a further FP reformulation involving vector variables. The resulting reformulation is structured as a weighted bipartite matching problem and allows the optimization of discrete and continuous variables in a joint and distributed fashion by using the standard techniques, *e.g.*, the auction algorithm [48] and Hungarian algorithm [49] for solving the weighted bipartite matching problem globally.

4.2.1 Problem Formulation

Following the notation in Section 4.1.1, we use $\mathcal{B}$ to denote the set of BSs in the network, $\mathcal{K}_i$ the set of users who are associated with BS i, σ^2 the power level of AWGN, w_k the weight of user k, and $P_{\max}$ the maximum transmit power level at the user side.

We now assume that each user is equipped with N antennas and each BS is equipped with M antennas. Spatial multiplexing can therefore support up to M data streams per cell (but some data streams may have zero throughput). Let s_{im} be the index of the user who is scheduled in the mth stream at BS i. Let $\mathbf{v}_k \in \mathbb{C}^N$ be the transmit beamformer of user k if it gets scheduled. Let $\mathbf{H}_{i,k} \in \mathbb{C}^{M\times N}$ be the uplink channel from user k to BS i. The joint uplink user scheduling and beamforming problem with a weighted sum-rate maximizing objective can be formulated as

$$\begin{aligned} \underset{\underline{s},\,\underline{\mathbf{v}}}{\text{maximize}} \quad & f_o(\underline{s},\underline{\mathbf{v}}) && \text{(4.16a)}\\ \text{subject to} \quad & \|\mathbf{v}_{im}\|_2^2 \le P_{\max},\ \forall (i,m) && \text{(4.16b)}\\ & s_{im} \in \mathcal{K}_i \cup \{0\},\ \forall (i,m), && \text{(4.16c)} \end{aligned}$$

where the objective function f_o is

$$f_o(\underline{s},\underline{\mathbf{v}}) = \sum_{(i,m)} w_{s_{im}} \log\left(1+\mathbf{v}_{s_{im}}^H \mathbf{H}_{i,s_{im}}^H \left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)\neq(i,m)} \mathbf{H}_{i,s_{jn}}\mathbf{v}_{s_{jn}}\mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H\right)^{-1} \mathbf{H}_{i,s_{im}}\mathbf{v}_{s_{im}}\right). \quad (4.17)$$

As before, $w_{s_{im}}$ and $\mathbf{v}_{s_{im}}$ are both implicitly set to zero if $s_{im}=0$. Note that under this MIMO setting we allow scheduling up to M users per cell. Using the number of antennas, namely the degrees of freedom, to limit the number of scheduled users is reasonable in most of the practical scenarios. Theoretically speaking, however, this scheme can be suboptimal in some special cases like the low-SNR regime.

The above problem is more challenging than the uplink user scheduling and power control problem (4.2) of the SISO case. In addition to the intercell interference, we also need to take

into account the interference coming from the same cell because multiple users can be scheduled at each BS simultaneously.

4.2.2 Joint Fractional Programming and Matching

Recall that in Section 4.1.1 we make use of the quadratic transform and the Lagrangian dual transform to derive a reformulation for the joint uplink scheduling and power control problem, whereby the power and scheduling variables can be grouped on a per-cell basis. This reformulating procedure can be adapted to the multidimensional case for problem (4.16). First, apply the multidimensional Lagrangian dual transform in Theorem 10 to reformulate the original objective function $f_o(\underline{s}, \underline{\mathbf{v}})$ as $f_r(\underline{s}, \underline{\mathbf{v}}, \underline{\gamma})$:

$$f_r(\underline{s}, \underline{\mathbf{v}}, \underline{\gamma}) = \sum_{(i,m)} w_{s_{im}} \Bigg(\log(1+\gamma_{im}) - \gamma_{im} + (1+\gamma_{im}) \mathbf{v}_{s_{im}}^H \mathbf{H}_{i,s_{im}}^H \cdot \left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H \right)^{-1} \mathbf{H}_{i,s_{im}} \mathbf{v}_{s_{im}} \Bigg) \tag{4.18}$$

with an auxiliary variable γ_{im} introduced for each data stream (i,m). Thus, the original problem (4.16) is equivalent to

$$\underset{\underline{s}, \underline{\mathbf{v}}, \underline{\gamma}}{\text{maximize}} \quad f_r(\underline{s}, \underline{\mathbf{v}}, \underline{\gamma}) \tag{4.19a}$$

$$\text{subject to} \quad \text{(4.16b), (4.16c).} \tag{4.19b}$$

Following Algorithm 7, we propose to optimize the variables in (4.19) in an iterative fashion. When the primal variables $\underline{s}$ and $\underline{\mathbf{v}}$ are both held fixed, maximizing f_r over $\underline{\gamma}$ is a convex problem which can be efficiently solved by setting $\partial f_r / \partial \gamma_{im}$ to zero, that is

$$\gamma_{im}^\star = \mathbf{v}_{s_{im}}^H \mathbf{H}_{i,s_{im}}^H \left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n) \neq (i,m)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H \right)^{-1} \mathbf{H}_{i,s_{im}} \mathbf{v}_{s_{im}}. \tag{4.20}$$

Note that the optimal γ_{im} is equal to the resulting uplink SINR in data stream (i,m) exactly.

We then consider optimizing $\underline{s}$ and $\underline{\mathbf{v}}$ for fixed $\underline{\gamma}$. This subproblem only involves the last term of f_q which has a multidimensional sum-of-ratios form. By treating $\sqrt{w_{s_{im}}(1+\gamma_{im})}\mathbf{H}_{i,s_{im}}\mathbf{v}_{s_{im}}$ as the numerator vector $\mathbf{a}_i$ and $\left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H\right)$ as the denominator matrix

$\mathbf{B}_i$ in Theorem 9 from Section 2.5.1, we arrive at a new objective function

$$f_q(\underline{s},\underline{\mathbf{v}},\underline{\gamma},\underline{\mathbf{y}}) = \sum_{(i,m)} w_{s_{im}} \log(1+\gamma_{im}) - \sum_{(i,m)} w_{s_{im}}\gamma_{im} + \sum_{(i,m)} \Bigg(2\sqrt{w_{s_{im}}(1+\gamma_{im})}\cdot \Re\left\{\mathbf{v}_{s_{im}}^H \mathbf{H}_{i,s_{im}}^H \mathbf{y}_{im}\right\} - \mathbf{y}_{im}^H \Bigg(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H \Bigg) \mathbf{y}_{im} \Bigg), \quad (4.21)$$

where an auxiliary variable $\mathbf{y}_{im} \in \mathbb{C}^M$ is introduced with respect to each data stream (i,m). Thus, the optimization of f_r in (4.19) is further recast to

$$\underset{\underline{s},\underline{\mathbf{v}},\underline{\gamma},\underline{\mathbf{y}}}{\text{maximize}} \qquad f_q(\underline{s},\underline{\mathbf{v}},\underline{\gamma},\underline{\mathbf{y}}) \quad (4.22a)$$

$$\text{subject to} \qquad (4.16b),\ (4.16c). \quad (4.22b)$$

With the update of $\underline{\gamma}$ already shown in (4.20), we now consider the optimization of $\underline{s}$, $\underline{\mathbf{v}}$ and $\underline{\mathbf{y}}$ in f_q. First, when all the other variables are fixed, the optimal $\underline{\mathbf{y}}$ can be explicitly determined by setting $\partial f_r / \partial \mathbf{y}_{im}$ to zero, that is

$$\mathbf{y}_{im}^\star = \left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H \right)^{-1} \sqrt{w_{s_{im}}(1+\gamma_{im})} \mathbf{H}_{i,s_{im}} \mathbf{v}_{s_{im}}. \quad (4.23)$$

Observe that the optimal $\mathbf{y}_{im}$ is exactly an MMSE receiver scaled by a factor of $\sqrt{w_{s_{im}}(1+\gamma_{im})}$, with respect to each data stream (i,m).

It remains to optimize the variables $\underline{s}$ and $\underline{\mathbf{v}}$ in f_q. We gain incorporate the idea of *weighted bipartite matching* for the joint optimization of these two variables. The key observation is that the scheduling of user s_{im} and its transmit beamformer $\mathbf{v}_{im}$ in a particular data stream (i,m) contribute to the objective function (4.21) in a way that is *independent* of the scheduling and beamformer choices in other streams. More specifically, if some user k is scheduled in the data stream (i,m), *i.e.*, $s_{im} = k$, then the optimal transmit beamformer of user k with respect to (i,m), denoted as $\boldsymbol{\tau}_{k,im}$, can be determined by solving $\partial f_q / \partial \mathbf{v}_{s_{im}} = \mathbf{0}$, that is

$$\boldsymbol{\tau}_{k,im} = \left(\sum_{(j,n)} \mathbf{H}_{j,k}^H \mathbf{y}_{jn} \mathbf{y}_{jn}^H \mathbf{H}_{j,k} + \eta_{k,im}^\star \boldsymbol{I}_N \right)^{-1} \sqrt{w_k(1+\gamma_{im})} \mathbf{H}_{i,k}^H \mathbf{y}_{im}, \quad (4.24)$$

where the dual variable $\eta_{k,im}^\star$ accounts for power constraint (4.16b) and is optimally determined by the complementary slackness condition

$$\eta_{k,im}^\star = \inf\left\{ \eta_{k,im} \geq 0 : \|\boldsymbol{\tau}_{k,im}(\eta_{k,im})\|_2^2 \leq P_{\max} \right\}. \quad (4.25)$$

This $\eta_{k,im}^\star$ can be efficiently evaluated via bisection search.

Therefore, the utility value (in terms of f_q) of scheduling user k in one particular data

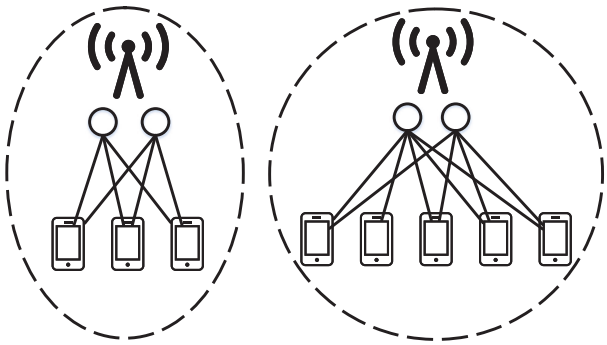


Figure 4.4: The scheduling variables s_{im}'s are decoupled on a per-cell basis after the FP-based reformulation. Optimizing the scheduling variable s in (4.22) can be characterized as a weighted bipartite matching between the users and the data streams in each cell, with the matching weights defined by (4.26).

stream (i, m) can be determined analytically. This allows solving $\underline{s}$ and $\underline{\mathbf{v}}$ jointly by weighted bipartite matching. To formalize the idea, we define the utility value of assigning user k to data stream (i, m) as $\xi_{k,im}$:

$$\xi_{k,im} = w_k \log(1+\gamma_{im}) - w_k\gamma_{im} + 2\sqrt{w_k(1+\gamma_{im})}\,\Re\left\{\boldsymbol{\tau}_{k,im}^H \mathbf{H}_{i,k}^H \mathbf{y}_{im}\right\} - \sigma^2\|\mathbf{y}_{im}\|_2^2 - \sum_{(j,n)} \mathbf{y}_{jn}^H \mathbf{H}_{j,k}\boldsymbol{\tau}_{k,im}\boldsymbol{\tau}_{k,im}^H \mathbf{H}_{j,k}^H \mathbf{y}_{jn}. \quad (4.26)$$

As a result, the f_q maximizing problem (4.22) reduces to the following weighted bipartite matching problem:

$$\begin{aligned}
\underset{\underline{x}}{\text{maximize}} \quad & \sum_{k\in\mathcal{K}_i}\sum_{m=1}^{N} \xi_{k,im}x_{k,im} && (4.27a)\\
\text{subject to} \quad & \sum_{k\in\mathcal{K}_i} x_{k,im} \le 1,\ \forall(i,m) && (4.27b)\\
& \sum_{m=1}^{N} x_{k,im} \le 1,\ \forall(k,i) && (4.27c)\\
& x_{k,im} \in \{0,1\},\ \forall(k,i,m), && (4.27d)
\end{aligned}$$

where the binary variable $x_{k,im}$ indicates whether or not user k is scheduled in the mth data stream at its associated BS i. We remark that the above matching problem is considered at each BS i *individually*, as illustrated in Fig. 4.4.

Weighted bipartite matching is a well-studied problem in the field of combinatorics [50]. It can be efficiently solved by the existing algorithms with polynomial-time computational complexity using, e.g., the Hungarian algorithm [49] and the auction algorithm [48], with a computational complexity of $O((K+M)^3)$. Further, because in practice the matching weights $\xi_{k,im}$ are always evaluated with finite precision, in this finite-precise case, the complexity of

Algorithm 8: Joint Uplink Scheduling, Power Control, and Beamforming

```
Initialize s, v and γ to feasible values;
repeat
    Update y by (4.23);
    Update γ by (4.20);
    Update (s, v) jointly by (4.24), (4.28) and (4.29);
until the value of function f_q in (4.21) converges;
```

matching can be reduced to $O((K+M)^2)$ using the algorithm in [51].

After solving for $\underline{x}$ in problem (4.27), we recover the optimal scheduling variable $\underline{s}^\star$ by

$$s^\star_{im} = \begin{cases} k, \text{ if } x^\star_{k,im} = 1 \text{ for some } k \in \mathcal{K}_i; \\ 0, \text{ otherwise,} \end{cases} \tag{4.28}$$

where the decision 0 is made in data stream (i,m) if any user scheduled in the stream would have contributed to f_q negatively. Note that $x^\star_{k,im}$ must be zero if $\xi_{k,im} < 0$. In practice, we can further facilitate weighted matching by removing the edges corresponding to negative $\xi_{k,im}$ from the bipartite graph. The transmit beamformers of the scheduled users are then set to the optimal values in (4.24) accordingly:

$$\mathbf{v}^\star_k = \boldsymbol{\tau}_{k,im}, \quad \text{if } x^\star_{k,im} = 1 \text{ for some } (i,m). \tag{4.29}$$

We summarize the proposed iterative distributed optimization in Algorithm 8. Like Algorithm 7, this algorithm guarantees convergence although it is not a block coordinate ascent method, as stated in the following proposition.

Proposition 7. *Algorithm 8 is guaranteed to converge, with the weighted sum rate f_o monotonically nondecreasing after each iteration. The converged solution is a stationary point of f_o with respect to $\underline{\mathbf{v}}$ if $\underline{s}$ is assumed to be fixed.*

We note that the SISO algorithm in Section 4.1.3 is a special case of the weighted bipartite matching approach for the MIMO problem. Further, we can use the same argument to show that computing the optimal $\underline{s}$ for fixed $\underline{\mathbf{v}}$ is already NP-hard, so the above convergence result is likely the best one can hope for.

As a final remark, Algorithms 7 and 8 can be initialized with simple but reasonable heuristic. For example, in a 2×2 MIMO network, the two users with the highest weights in each cell can be scheduled at the beginning, and their beamformers can be set to maximize the signal strength. Moreover, we set some small constant $\delta > 0$ and use the convergence criterion $|f_q^{(t)} - f_q^{(t-1)}| < \delta$ where t is the iteration index.

4.2.3 Complexity Analysis

We now compare the complexities of Algorithm 8 and the WMMSE method [40, 41] (which is modified to include scheduling as stated in Section 4.1.2). For ease of analysis, assume that each cell has the same number of users. Let K be the number of users per cell; let B be the total number of BSs deployed throughout the network. Following [41], we evaluate the algorithm complexity with respect to each round of iteration.

First consider the *communication complexity*. In Algorithm 8, every BS needs to collect $(\underline{s}, \underline{\mathbf{v}}, \underline{\mathbf{y}})$ except $\underline{\gamma}$ with respect to each (j, n) pair, so the overall communication complexity of Algorithm 8 is $O(M^2B^2 + MNB^2)$, which is independent of K. In the WMMSE method, each BS needs to collect $(\underline{\mathbf{v}}, \underline{\gamma}, \underline{\mathbf{y}})$ with respect to every user in the network, thus the overall communication complexity of WMMSE is $O(MKB^2 + NKB^2)$. WMMSE in general has a much higher communication complexity, because normally $K \gg M$ (*i.e.*, only a small portion of users in the cell are scheduled in each time-slot).

We further analyze the *computational complexity*. Assuming that the classic Hungarian algorithm is used for weighted bipartite matching, the overall computational complexity of Algorithm 8 per iteration can be shown to be $O(c_{\text{FP}})$, where

$$c_{\text{FP}} = M^4B^2 + MN^3KB + (M^3N + MN^2)KB^2 + (K + M)^3B. \tag{4.30}$$

Each iteration of the WMMSE algorithm involves a matrix multiplication with respect to every user-BS pair in the network. Consequently, it requires a computational complexity of $O(c_{\text{WMMSE}})$, where

$$c_{\text{WMMSE}} = \left(M^3 + N^3\right) KB + M^2KB^2 + \left(MN + N^2\right) K^2B^2. \tag{4.31}$$

We remark that *matrix chain ordering* needs to be optimized for both of the algorithms to find the most efficient way of multiplying matrices. For simplicity, we further assume that M and N are fixed and also that K is much greater than both M and N. Then, the above computational complexities become

$$c_{\text{FP}} = KB^2 + K^3B \;\text{ and }\; c_{\text{WMMSE}} = K^2B^2, \tag{4.32}$$

so Algorithm 8 is more complex if the number of users K is large. However, as already mentioned in Section 4.2.2, because the matching weights are in practice expressed with finite precision, the efficiency of bipartite matching can be improved from $O(K^3)$ to $O(K^2)$ by using the algorithm of [51]. Then, we have

$$c_{\text{FP}} = KB^2 + K^2B < K^2B^2 = c_{\text{WMMSE}}. \tag{4.33}$$

In this case, Algorithm 8 is overall more computationally efficient than WMMSE.

Finally, we mention the recent work [52] that uses a *deep neural network* to learn the mapping from the geometric locations of wireless devices directly to an interference-aware link activation pattern, thus bypassing the channel estimation stage. In this work, FP is applied

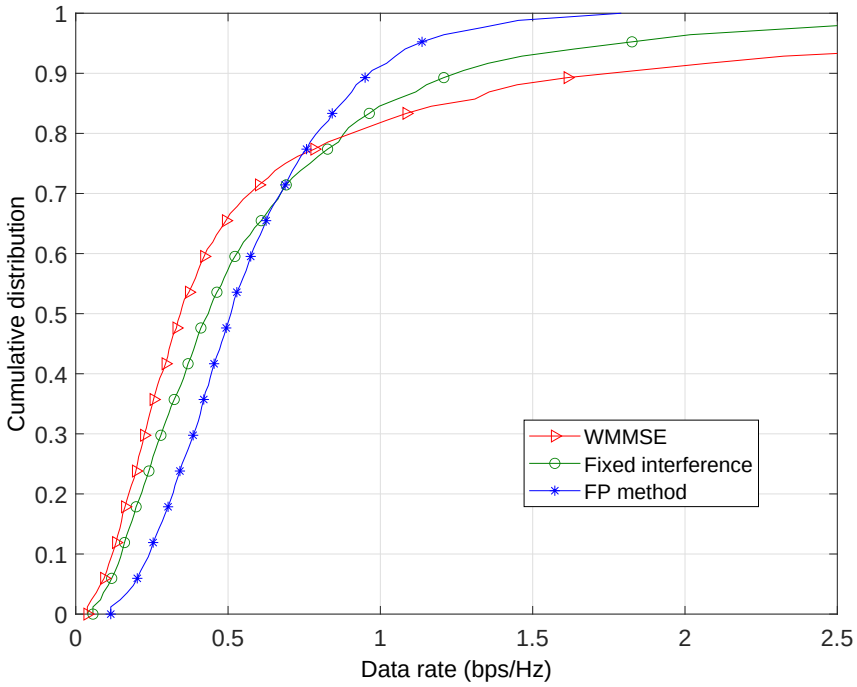


Figure 4.5: Comparison of the proposed FP-based coordinated uplink user scheduling and beamforming method with two baseline methods in terms of CDF of user rates.

to generate the training data set. This supervised learning method is adopted as a benchmark in [53, 54] to compare with some other learning-based methods.

4.2.4 Numerical Results

We validate the proposed FP-based approach by simulating a network consisting of 7 cells in a wrapped around topology. A total of 84 users randomly distributed in the network are associated with the BS to which the channel is the strongest. Each user is equipped with 2 antennas and each BS is equipped with 4 antennas. The uplink MIMO channels consist of two components: a large-scale fading component (including pathloss and shadowing), which follows the model discussed in Section 4.1.4, and a Rayleigh fading component. The user weights in every time-slot are updated as the reciprocal of the long-term average rates in order to maximize a proportional fairness utility. All other parameters, *i.e.*, the channel pathloss model, AWGN, maximum transmit power, and spectrum bandwidth, follow the settings in Section 4.1.4.

The following methods are introduced as benchmarks:

- *WMMSE*: The WMMSE algorithm is introduced in [40, 41]. To use WMMSE for user scheduling, we initialize all the users in the network with some random beamformers, then run the WMMSE algorithm to optimize weighted sum rate. At convergence, most users would be assigned zero beamformer; those assigned nonzero beamformers are scheduled.

Table 4.2: Sum log-utilities of the proposed coordinated uplink scheduling and beamforming method as compared to the two baseline schemes.

Algorithm	Total Log-Utility
WMMSE	175.87
Fixed Interference	183.45
Proposed FP Method	193.79

User scheduling is therefore determined *implicitly* by beamforming. In the SISO case, the beamforming step reduces to power control.

- *Fixed interference method*: This heuristic method extends the fixed interference method in Section 4.1.4. Iteratively, apply a beamforming method (*e.g.,* WMMSE) for fixed user scheduling variable $\underline{s}$, and then optimize $\underline{s}$ for fixed beamformers. This works well in the downlink because the optimal scheduling can be explicitly determined [38]. For the uplink, the heuristic is to emulate the downlink by assuming fixed interference from the neighboring cells.

The proposed algorithm is compared with the aforementioned two baselines. As shown in Fig. 4.5, the proposed FP-based method has a significant advantage over the baselines particularly for low-rate users. For example, the rates of the 10th-percentile users is improved by at least 50% the proposed algorithm. These low-rate users are mostly located close to the cell edges, highlighting the important role of coordinated uplink scheduling and beamforming in interference mitigation. Table 4.2 shows that the proposed FP method substantially improves the sum log-utility in the network as compared to the benchmarks, verifying that interference management by coordinating user schedules and beamformers is crucial to the network performance.

4.3 Discrete Beamforming

Thus far it is assumed that each beamformer $\mathbf{v}_{im}$ can be set to an arbitrary vector as long as the power constraint $\|\mathbf{v}_{im}\|_2^2 \leq P_{\max}$ is satisfied. We now consider a discrete scenario for beamforming where the choice for $\mathbf{v}_{im}$ is restricted to a codebook

$$\mathcal{V} = \left\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \cdots, \boldsymbol{\phi}_{|\mathcal{V}|}\right\}. \tag{4.34}$$

In the above, each $\boldsymbol{\phi}_n \in \mathbb{C}^N$ (for $n = 1, \ldots, |\mathcal{V}|$) represents a possible beamforming vector.

In this case, if some user k is scheduled in the data stream (i, m), then its optimal transmit beamformer $\boldsymbol{\tau}_{k,im}$ in terms of f_q can be obtained by searching through the codebook, that is

$$\boldsymbol{\tau}_{k,im} = \arg\max_{\boldsymbol{\phi}\in\mathcal{V}} \left\{ 2\sqrt{w_k(1+\gamma_{im})}\, \Re\left\{\boldsymbol{\phi}^H \mathbf{H}_{i,k}^H \mathbf{y}_{im}\right\} - \sum_{(j,n)} \mathbf{y}_{jn}^H \mathbf{H}_{j,k} \boldsymbol{\phi}\boldsymbol{\phi}^H \mathbf{H}_{j,k}^H \mathbf{y}_{jn} \right\}. \tag{4.35}$$

The bipartite matching process (4.27) can then be performed with $\xi_{k,im}$ set according to (4.26) but using the above $\boldsymbol{\tau}_{k,im}$. After matching, the optimal $\underline{\mathbf{v}}$ is recovered by (4.29).

To find the optimal $\mathbf{v}_{im}$ in the discrete search (4.35) requires at most a computational complexity of $O(|\mathcal{V}|)$. This complexity can be further reduced to $O(\log|\mathcal{V}|)$ by taking advantage of the functional structure of (4.35). The idea is to first maximize f_q over $\underline{\mathbf{v}}$ without considering the discrete constraint (4.34) and then find the discrete solution $\boldsymbol{\phi} \in \mathcal{V}$ that is closest to the relaxed solution $\tilde{\mathbf{v}}_{im}$, for every (i,m) pair, *i.e.*,

$$\boldsymbol{\tau}_{k,im} = \arg\min_{\boldsymbol{\phi}\in\mathcal{V}} \|\boldsymbol{\phi} - \tilde{\mathbf{v}}_{im}\|_2, \tag{4.36}$$

where the relaxed solution $\tilde{\mathbf{v}}_{im}$ is the $\boldsymbol{\tau}_{k,im}$ in (4.24) without the discrete codebook constraint. Observe that the right-hand side of (4.35) is a concave quadratic function of variable $\boldsymbol{\phi}$, and then after completing the square, it can be shown that updating $\boldsymbol{\tau}_{k,im}$ by (4.36) yields exactly the same solution as in (4.35). Therefore, although the above relax-and-then-round approach in (4.36) is a common heuristic for discrete beamforming, our FP framework gives a theoretical justification by showing that this approach actually maximizes the reformulated objective f_q.

An efficient way to perform the optimization (4.35) can now be devised based on (4.36), as stated in the following proposition.

Proposition 8 (*Nearest Point Projection for Discrete Beamforming*)**.** *The optimal update (4.35) for discrete beamforming can be realized by the nearest point projection as in (4.36) with a computational complexity of* $O(\log|\mathcal{V}|)$.

Proof. Construct a k-d tree [55] for all the elements of $\mathcal{V}$ in advance. The following three steps produce the nearest-point projection (4.36): Insert $\tilde{\mathbf{v}}_{im}$ in the k-d tree; then search for the nearest neighbor of $\tilde{\mathbf{v}}_{im}$ in the tree and output it as the projection result; finally delete $\tilde{\mathbf{v}}_{im}$ from the tree. The insertion, search, and deletion operations all have an average complexity $O(\log|\mathcal{V}|)$. □

We remark that a similar result can be derived for discrete power control in the SISO case, in which case the search through the k-d tree reduces to a one-dimensional bisection search.

4.4 Connection to WMMSE Algorithm

The WMMSE algorithm originally derives from the signal inference [40, 41]. In what follows, we give another derivation for WMMSE based on the proposed quadratic transform. Recall that after the use of Lagrangian dual transform, the original objective function $f_o(\underline{s},\underline{\mathbf{v}})$ is recast to $f_r(\underline{s},\underline{\mathbf{v}},\underline{\gamma})$, in which the primal variables $\underline{s}$ and $\underline{\mathbf{v}}$ only appear in the last sum-of-ratio term. Specifically, each ratio contained in the sum-of-ratio term of f_r can be written as

$$d_{im}\mathbf{v}_{s_{im}}^H\mathbf{H}_{i,s_{im}}^H\mathbf{B}_{im}^{-1}\mathbf{H}_{i,s_{im}}\mathbf{v}_{s_{im}}, \tag{4.37}$$

where two new notations d_{im} and $\mathbf{B}_{im}$ are introduced to simplify notation:

$$d_{im} = w_{s_{im}}(1+\gamma_{s_{im}}) \tag{4.38}$$

and

$$\mathbf{B}_{im} = \sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H. \tag{4.39}$$

Recall that in deriving the further reformulation of f_q, we propose in Section 4.2.2 to apply the multidimensional quadratic transform in Theorem 9 by identifying the ratio pattern of (4.37) as

$$\mathbf{a}_{im}^H \mathbf{B}_{im}^{-1} \mathbf{a}_{im}, \tag{4.40}$$

where the vector numerator $\mathbf{a}_{im}$ is recognized as

$$\mathbf{a}_{im} = \sqrt{d_{im}} \mathbf{H}_{i,s_{im}} \mathbf{v}_{s_{im}}. \tag{4.41}$$

However, this is not the only way to implement the FP technique. In fact, we could have applied the multidimensional quadratic transform to the ratios in a different way:

$$d_{im} \cdot \left(\check{\mathbf{a}}_{im}^H \mathbf{B}_{im}^{-1} \check{\mathbf{a}}_{im} \right), \tag{4.42}$$

where the numerator vector is newly recognized as

$$\check{\mathbf{a}}_{im} = \mathbf{H}_{i,s_{im}} \mathbf{v}_{s_{im}}. \tag{4.43}$$

In this case, we would have arrived at a different reformulation $\check{f}_q$ as

$$\begin{aligned} \check{f}_q(\underline{s}, \underline{\mathbf{v}}, \underline{\gamma}, \underline{\mathbf{y}}) = \sum_{(i,m)} w_{s_{im}} \Bigg(\log(1+\gamma_{im}) - \gamma_{im} + (1+\gamma_{im}) \Bigg(2\Re\left\{ \mathbf{v}_{s_{im}}^H \mathbf{H}_{i,s_{im}}^H \mathbf{y}_{im} \right\} \\ - \mathbf{y}_{im}^H \left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H \right) \mathbf{y}_{im} \Bigg) \Bigg). \end{aligned} \tag{4.44}$$

This reformulation gives the following iterative algorithm for optimizing beamformers. Finding the optimal $\underline{\mathbf{y}}$ by solving $\partial \check{f}_q / \partial \mathbf{y}_{im} = \mathbf{0}$ with respect to each (i,m) pair amounts to

$$\check{\mathbf{y}}_{im} = \left(\sigma^2 \boldsymbol{I}_M + \sum_{(j,n)} \mathbf{H}_{i,s_{jn}} \mathbf{v}_{s_{jn}} \mathbf{v}_{s_{jn}}^H \mathbf{H}_{i,s_{jn}}^H \right)^{-1} \mathbf{H}_{i,s_{im}} \mathbf{v}_{s_{im}}. \tag{4.45}$$

Note that the above $\check{\mathbf{y}}_{im}$ solution is exactly an MMSE receiver. Likewise, the optimal transmit

beamformer is

$$\check{\mathbf{v}}_{s_{im}} = \left(\sum_{(j,n)} d_{jn} \mathbf{H}_{j,s_{im}}^H \mathbf{y}_{jn} \mathbf{y}_{jn}^H \mathbf{H}_{j,s_{im}} + \eta_{im}^\star \boldsymbol{I}_N \right)^{-1} d_{im} \mathbf{H}_{i,s_{im}}^H \mathbf{y}_{im}, \tag{4.46}$$

where

$$\eta_{im}^\star = \inf \left\{ \eta_{im} \geq 0 : \|\check{\mathbf{v}}_{s_{im}}(\eta_{im})\|_2^2 \leq P_{\max} \right\} \tag{4.47}$$

is the optimal dual variable for the power constraint (4.16b) by complementary slackness. Finally, the update of $\underline{\gamma}$ remains the same as in (4.20). When iteratively applying the above updates of $\underline{\gamma}$, $\underline{\mathbf{v}}$ and $\underline{\mathbf{y}}$ for the fixed scheduling variable $\underline{s}$, we arrive at exactly the WMMSE algorithm for beamforming. Therefore, WMMSE can be interpreted as a specific way of using FP to solve the optimal beamforming problem.

However, unlike our proposed reformulation f_q in (4.21), this $\check{f}_q$ does not allow an explicit distributed solution for $\underline{s}$, because the discrete variables s_i's are *not decoupled* in the last term of $\check{f}_q$ as shown in (4.44). While the FP-based method formerly proposed in this chapter is able to use weighted bipartite matching to find the optimal $\underline{s}$, the WMMSE algorithm can only optimize the scheduling variable implicitly by optimizing beamformers for all the users in the network. This implicit scheduling of WMMSE is not only more computationally complex, but also has inferior performance as shown in the previous section.

4.5 Summary

This chapter explores the application of FP for the discrete (or mixed discrete-continuous) problems. The central idea is to decouple the complicated interfering interactions among the different links by a novel quadratic transform and a Lagrangian dual transform, thereby allowing efficient and distributed optimization. We illustrate the proposed FP approach by considering the uplink user scheduling, power control, and beamforming problem for wireless cellular networks. By incorporating weighted bipartite matching, we devise a novel use of FP whereby the discrete scheduling variables can be jointly optimized with the continuous variables such as power and beamformers. As compared to the existing methods, the proposed FP approach treats discrete optimization rigorously without relaxation. It is further shown that the WMMSE algorithm is a particular form of FP, but in contrast to the proposed approach, WMMSE is not well equipped to deal with discrete user scheduling variables.

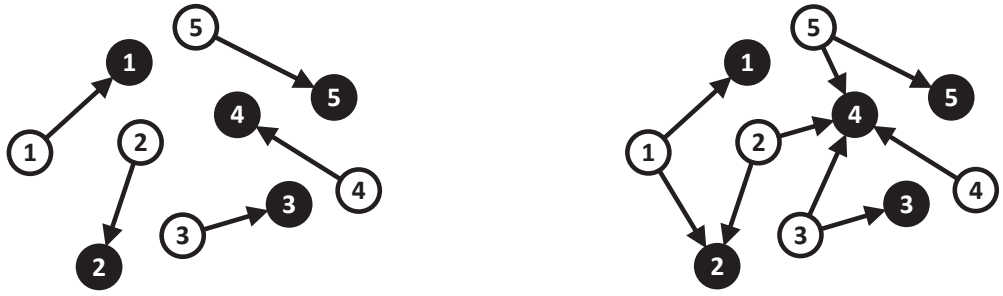
Chapter 5

Matrix Optimization Problems

We have seen some applications of FP for optimizing the multidimensional variables in the foregoing chapters, *e.g.*, the multi-link energy efficiency maximization in a broadcast network in Section 3.3 as well as the joint uplink scheduling and beamforming in Section 4.2, but they all assume that the ratio term is a scalar-valued function of multidimensional variables. This chapter goes further to account for the matrix-valued fractional function. We will start with a multi-data-stream transmission scenario in which the SINR term has a matrix form; this is a natural generalization of the aforementioned applications based on the scalar SINR. In the second example, we propose to treat the channel estimation problem with a weighted sum mean square error (MSE) minimizing objective as a matrix fractional program, then apply the matrix version of the quadratic transform in Section 2.5.2. Moreover, making use of the matrix Lagrangian dual transform, we relate the weighted MMSE of channel estimation to a weighted sum-rate maximization problem.

5.1 Multi-Data-Stream Transmission in Flexibly Associated D2D Networks

Interference management is a fundamental issue in D2D communications whenever the transmitter-and-receiver pairs are located in close proximity and frequencies are fully reused, so active links may severely interfere with each other. We put forward an optimization strategy named FPLinQ to coordinate the link scheduling decisions among the interfering links, along with power control and beamforming. The key enabler is a novel optimization method called matrix FP that generalizes previous scalar and vector forms of FP in allowing multiple data streams per link. From an application perspective, it is shown that as compared to the existing methods for coordinating scheduling in the D2D network, such as FlashLinQ, ITLinQ, and ITLinQ+, the proposed FPLinQ approach is more general in allowing multiple antennas at both the transmitters and the receivers, and further in allowing arbitrary and multiple possible associations between the devices via matching.



(a) Fixed Single Association (b) Flexible Association

Figure 5.1: D2D network with white circles denoting the transmitters and black circles denoting the receivers. In the fixed single association model (a), the transmitters have a fixed one-to-one mapping to the receivers. This section considers a more general setting (b) in which each transmitter has the flexibility of associating with one of multiple receivers, and each receiver has the flexibility of associating with one of multiple transmitters.

5.1.1 Problem Formulation

Consider a wireless D2D network with a set of transmitters $\mathcal{I}$ and a set of receivers $\mathcal{J}$. We assume that each transmitter may have data to transmit to one or more receivers, and likewise each receiver may wish to receive data from one or more transmitters. Thus, the communication scenario considered here is a generalization of traditional D2D network with fixed single association between each pair of transmitter and receiver to a scenario with multiple possible associations between the transmitters and the receivers as shown in Fig. 5.1. We assume that in each scheduling time slot, each transmitter (or receiver) can only communicate with at most one of its associated receivers (or transmitters)[1], respectively, so that the mapping between the transmitters and the receivers is one-to-one. The task of scheduling is to identify which set of links over the network to activate in each slot. Further, we assume that the transmitters and the receivers are each equipped with N antennas and permit multiple data streams to be carried via MIMO transmission. The task of beamforming and power control is to design the transmit beamformers for each of these multiple data streams in each active link in the scheduling slot.

Mathematically, let $\mathcal{K}_j \subseteq \mathcal{I}$ be the set of transmitters associated with each particular receiver $j \in \mathcal{J}$. Likewise, let $\mathcal{L}_i \subseteq \mathcal{J}$ be the set of receivers associated with each transmitter $i \in \mathcal{I}$. Let $\mathbf{H}_{ji} \in \mathbb{C}^{N\times N}$ be the channel from transmitter i to receiver j in the scheduling slot. The joint scheduling, beamforming, and power control problem can be written down as that of optimizing two sets of variables: s_j, the index of the transmitter scheduled at receiver j, and $\mathbf{V}_i \in \mathbb{C}^{N\times N}$, the collection of beamforming vectors at transmitter i in each scheduling slot so as to maximize some network wide objective function. Throughout this section, we assume that the channel state information is completely known and network optimization is performed in a centralized

[1]Note that the D2D model considered here is more general than the traditional wireless cellular network model of [33] in effectively allowing multiple and arbitrary associations between the base-stations and the mobile terminals, but on the other hand, is also more restrictive in that it does not allow spatial multiplex at either the receiver or the transmitter.

manner. It is proved in [46,56] that this network optimization problem is NP-hard, even under such idealized assumptions.

We use the weighted sum rate as the optimization objective in each scheduling slot, where the weights are adjusted from slot to slot in an outer loop in order to maximize a network utility of long-term average rates. We assume that interference is treated as noise, so that the achievable data rate in each scheduling slot can be computed from the receiver's perspective, *i.e.*, for each receiver j, as [57]

$$R_j = \log \left| \boldsymbol{I}_N + \mathbf{V}_{s_j}^H \mathbf{H}_{js_j}^H \mathbf{F}_j^{-1} \mathbf{H}_{js_j} \mathbf{V}_{s_j} \right| \tag{5.1}$$

with the interference-plus-noise term

$$\mathbf{F}_j = \sigma^2 \boldsymbol{I}_N + \sum_{j' \in \mathcal{J} \setminus \{j\}} \mathbf{H}_{js_{j'}} \mathbf{V}_{s_{j'}} \mathbf{V}_{s_{j'}}^H \mathbf{H}_{js_{j'}}^H, \tag{5.2}$$

where σ^2 is the power of AWGN. Given a set of nonnegative weights $w_{ji} \geq 0$, the optimization problem is therefore

$$\begin{aligned}
\underset{\underline{\mathbf{V}}, \underline{s}}{\text{maximize}} \quad & \sum_{j \in \mathcal{J}} w_{js_j} R_j && \text{(5.3a)} \\
\text{subject to} \quad & \operatorname{tr}\big(\mathbf{V}_i^H \mathbf{V}_i\big) \leq P_{\max}, \ \forall i \in \mathcal{I} && \text{(5.3b)} \\
& s_j \in \mathcal{K}_j \cup \{\varnothing\}, \ \forall j \in \mathcal{J} && \text{(5.3c)} \\
& s_j \neq s_{j'} \text{ or } s_j = \varnothing, \ \forall j \neq j', && \text{(5.3d)}
\end{aligned}$$

where we have assumed a per-scheduling-slot and per-node power constraint $P_{\max}$ and $\varnothing$ denotes the decision of not scheduling any transmitter at receiver j. We remark that $\mathbf{H}_{js_j}$, $\mathbf{V}_{s_j}$, and w_{js_j} are set to zero if $s_j = \varnothing$. Constraint (5.3d) states that the same transmitter cannot be scheduled for more than one receiver at a time, as required by the assumption that the association between the transmitters and the receivers in the D2D network must be one-to-one. Problem (5.3) involves a discrete optimization over $\underline{s}$ and a nonconvex continuous optimization over $\underline{\mathbf{V}}$, which make it a challenging optimization problem. Below, we first review several conventional approaches including the BCD algorithm and the greedy algorithms.

5.1.2 Existing Algorithms: FlashLinQ, ITLinQ, and ITLinQ+

We further examine the current state-of-the-art methods for D2D link scheduling in the literature: FlashLinQ [58], ITLinQ [59], and ITLinQ+ [60]. These works assume that each terminal has a single antenna, and further that each transmitter (or receiver) is only associated with one receiver (or transmitter) respectively, namely the fixed single association case shown in Fig. 5.1(a).

Because deciding the ON-OFF state for all the links at the same time is difficult, all three

Algorithm 9: Sequential Link Selection

```
1 Initialize the set of activated links A to ∅;
2 for each link (i, j) do
3 |   if (i, j) does not "conflict" with any link in A then
4 |   |   Schedule link (i, j) and add it to A;
5 |   end
6 end
```

algorithms propose to schedule the links in a greedy fashion sequentially, as stated in Algorithm 9. The main difference between FlashLinQ [58], ITLinQ [59], and ITLinQ+ [60] lies in the criterion for deciding whether the new link conflicts with already scheduled ones in Step 3 of Algorithm 9.

The FlashLinQ scheme [58] applies a threshold θ to SINR, assuming that adding link i to $\mathcal{A}$ does not cause conflict if and only if all the activated links have their SINRs higher than θ. The performance of FlashLinQ is highly sensitive to the value of θ, but choosing θ properly can be difficult in practice. Further, using the same θ for all the links is usually suboptimal when the weight varies from link to link.

From an information theory perspective, a seminal study [61] on the multi-user Gaussian interference channel provides a sufficient (albeit not necessary) condition for the optimality of treating interference as noise (TIN) for maximizing the generalized degrees-of-freedom (GDoF)[2] as follows:

$$\log|h_{ji}| \geq \max_{j' \neq j} \left\{\log|h_{j'i}|\right\} + \max_{i' \neq i} \left\{\log|h_{ji'}|\right\}, \tag{5.4}$$

where $h_{ji} \in \mathbb{C}$ is the channel of the single-antenna case. We refer to this result as the TIN condition.

The central idea of ITLinQ and ITLinQ+ is to schedule a subset of links that meet this TIN condition. Because the TIN condition in (5.4) is often too stringent to activate sufficient number of links, ITLinQ and ITLinQ+ both introduce relaxation based on design parameters. Like FlashLinQ, the performance of ITLinQ and ITLinQ+ is heavily dependent on the choice of design parameters, but they are difficult to choose optimally in practice. For example, [60] adopts two different sets of parameters for ITLinQ+ for two different network models. It is often not clear how to adapt ITLinQ and ITLinQ+ to the particular network environment of interest. It is important to point out that the theoretical basis of ITLinQ and ITLinQ+, *i.e.*, the TIN condition, only helps decide whether for some particular schedule, treating interference as noise is the optimal coding strategy from a GDoF perspective. *It does not, however, guarantee that if some schedule satisfies the TIN condition, then it must be the GDoF optimal schedule.* Thus, for a particular network, a schedule that does not satisfy the TIN condition can outperform one that does. This subtle point is illustrated in the following example.

[2]GDoF is defined as $\lim_{P\to\infty} R/\log(P)$, where R is the data rate and P is the transmit power level.

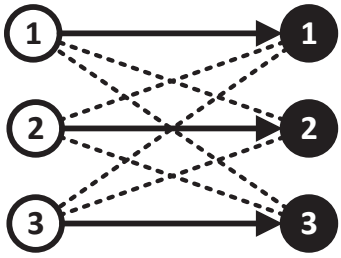


Figure 5.2: Power strength is P for each solid signal and is $P^{0.6}$ for each dashed signal. Thus, the sum GDoF equals to 1 if only one link is on, and equals to 1.2 if all links are on.

Example 4 (*Suboptimalilty of the TIN Condition for Scheduling*)**.** Consider three links as shown in Fig. 5.2. Let the desired signal strength be P and interfering signal strength be $P^{0.6}$. At most one link can be activated according to (5.4), so under the TIN condition, the total GDoF ≤ 1. But, a higher sum GDoF of 1.2 can be achieved by simply activating all the links.

Therefore, the TIN condition does not guarantee even the GDoF optimality of a given schedule. Considering further the significant gap between GDoF and the actual achievable rate, ITLinQ and ITLinQ+ can often produce quite suboptimal solutions.

5.1.3 Proposed Algorithm FPLinQ

We propose to solve the joint scheduling and beamforming problem (5.3) iteratively by first reformulating it by Corollary 5. Specifically, after specializing the variable $\mathbf{x}$ in (2.50) to be the $(\underline{\mathbf{V}}, \underline{s})$ in (5.3), we obtain the following reformulation:

Proposition 9. *The joint beamforming and link scheduling problem (5.3) is equivalent to*

$$\begin{aligned}
\underset{\underline{s},\underline{\mathbf{V}},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}}{\textit{maximize}} \quad & f_q(\underline{s},\underline{\mathbf{V}},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}) && \text{(5.5a)}\\
\textit{subject to} \quad & \textit{(5.3b), (5.3c), (5.3d)} \\
& \mathbf{\Gamma}_j \in \mathbb{H}_+^{N\times N},\ \forall j && \text{(5.5b)}\\
& \mathbf{Y}_j \in \mathbb{C}^{N\times N},\ \forall j, && \text{(5.5c)}
\end{aligned}$$

where the new objective function f_q is

$$\begin{aligned}
f_q(\underline{s},\underline{\mathbf{V}},\underline{\mathbf{\Gamma}},\underline{\mathbf{Y}}) &= \sum_{j\in\mathcal{J}} \Big(w_{js_j} \log|\boldsymbol{I}_N + \mathbf{\Gamma}_j| - w_{js_j}\mathrm{tr}\big(\mathbf{\Gamma}_j\big) + \mathrm{tr}\Big((\boldsymbol{I}_N + \mathbf{\Gamma}_j)\Big(2\sqrt{w_{js_j}}\,\mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{Y}_j^H - \\
&\qquad \mathbf{Y}_j^H\big(\mathbf{F}_j + \mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{V}_{s_j}^H\mathbf{H}_{js_j}^H\big)\mathbf{Y}_j\Big)\Big)\Big) && \text{(5.6a)}\\
&= \sum_{j\in\mathcal{J}} \Bigg[w_{js_j}\log|\boldsymbol{I}_N + \mathbf{\Gamma}_j| - w_{js_j}\mathrm{tr}\Big(\mathbf{\Gamma}_j\Big) + \mathrm{tr}\Big(2\sqrt{w_{js_j}}\,(\boldsymbol{I}_N + \mathbf{\Gamma}_j)\mathbf{H}_{js_j}\mathbf{V}_{s_j}\mathbf{Y}_j^H - \\
&\quad \sum_{j'\in\mathcal{J}} (\boldsymbol{I}_N + \mathbf{\Gamma}_{j'})\mathbf{Y}_{j'}^H\mathbf{H}_{j's_j}\mathbf{V}_{s_j}\mathbf{V}_{s_j}^H\mathbf{H}_{j's_j}^H\mathbf{Y}_{j'}\Big)\Bigg] + \sum_{j\in\mathcal{J}} \sigma^2\mathbf{Y}_j^H(\boldsymbol{I}_N + \mathbf{\Gamma}_j)\mathbf{Y}_j. && \text{(5.6b)}
\end{aligned}$$

Proof. The reformulating steps directly follow Corollary 5 in Section 2.5.2. We remark that f_q can be rewritten as in (5.6b), which enables an efficient optimization by matching. □

We now address the new problem (5.5) in an iterative manner. First, when $\underline{s}$ and $\underline{\mathbf{V}}$ are both held fixed, the auxiliary variables $\underline{\mathbf{\Gamma}}$ and $\underline{\mathbf{Y}}$ can be optimally determined as

$$\mathbf{\Gamma}_j^\star = \mathbf{V}_{s_j}^H \mathbf{H}_{js_j}^H \mathbf{F}_j^{-1} \mathbf{H}_{js_j} \mathbf{V}_{s_j} \tag{5.7}$$

and

$$\mathbf{Y}_j^\star = \left(\mathbf{F}_j + \mathbf{H}_{js_j} \mathbf{V}_{s_j} \mathbf{V}_{s_j}^H \mathbf{H}_{js_j}^H \right)^{-1} \sqrt{w_{js_j}}\, \mathbf{H}_{js_j} \mathbf{V}_{s_j}. \tag{5.8}$$

We remark that the implicit constraints as stated in Theorem 11 are automatically satisfied by the above optimal solution of the auxiliary variable $\mathbf{Y}_j^\star$.

It remains to optimize the beamforming variable $\underline{\mathbf{V}}$ and the scheduling variable $\underline{s}$. The key idea is to formulate the problem as a bipartite weighted matching problem. We consider the objective function f_q in (5.6b). The key observation is that the beamformer of each link (if it is scheduled) can be optimally determined from f_q, even without knowing the scheduling decisions for the nearby links. To formalize this idea, let $\tilde{\mathbf{V}}_{ji}$ be the tentative value of $\mathbf{V}_i^\star$ if link (i, j) is scheduled. By completing the square in f_q, we can compute $\tilde{\mathbf{V}}_{ji}$ as

$$\tilde{\mathbf{V}}_{ji} = \left(\mu_{ji} \boldsymbol{I}_N + \sum_{j' \in \mathcal{J}} \mathbf{H}_{j'i}^H \mathbf{Y}_{j'} (\boldsymbol{I}_N + \mathbf{\Gamma}_{j'}) \mathbf{Y}_{j'}^H \mathbf{H}_{j'i} \right)^{-1} \sqrt{w_{ji}}\, \mathbf{H}_{ji}^H \mathbf{Y}_j (\boldsymbol{I}_N + \mathbf{\Gamma}_j), \tag{5.9}$$

where μ_{ji} is a Lagrangian multiplier for the power constraint (5.3b), optimally determined as

$$\mu_{ji}^\star = \inf \left\{ \mu_{ji} \geq 0 : \mathrm{tr}(\tilde{\mathbf{V}}_{ji}^H \tilde{\mathbf{V}}_{ji}) \leq P_{\max} \right\} \tag{5.10}$$

which can be computed efficiently by bisection search since $\tilde{\mathbf{V}}_{ji}$ is monotonically decreasing with μ_{ji}. The solution $\tilde{\mathbf{V}}_{ji}$ in (5.9) has the same structure as an MMSE beamformer.

We now turn to the question of which $\tilde{\mathbf{V}}_{ji}$ should be chosen to be $\mathbf{V}_i$ so as to maximize f_q. This is akin to a scheduling step of choosing the best transmitter i for each receiver j. The key is to recognize this question as a weighted bipartite matching problem:

$$\begin{aligned}
\underset{\underline{q}}{\text{maximize}} \quad & \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{K}_j} \lambda_{ji} q_{ji} && \text{(5.11a)} \\
\text{subject to} \quad & \sum_{i \in \mathcal{K}_j} q_{ji} \leq 1, \ \forall j && \text{(5.11b)} \\
& \sum_{j \in \mathcal{L}_i} q_{ji} \leq 1, \ \forall i && \text{(5.11c)} \\
& q_{ji} \in \{0, 1\}, \ \forall (i, j) && \text{(5.11d)} \\
& q_{ji} = 0 \ \text{ if } i \notin \mathcal{K}_j \text{ or } j \notin \mathcal{L}_i, \ \forall (i, j), && \text{(5.11e)}
\end{aligned}$$

where the weight λ_{ji} is evaluated as

$$\lambda_{ji} = w_{ji} \log |\boldsymbol{I}_N + \boldsymbol{\Gamma}_j| - w_{ji} \mathrm{tr}(\boldsymbol{\Gamma}_j) + \mathrm{tr}\Bigg(2\sqrt{w_{ji}} (\boldsymbol{I}_N + \boldsymbol{\Gamma}_j) \mathbf{Y}_j^H \mathbf{H}_{ji} \tilde{\mathbf{V}}_{ji} - \sum_{j' \in \mathcal{J}} (\boldsymbol{I}_N + \boldsymbol{\Gamma}_{j'}) \mathbf{Y}_{j'}^H \mathbf{H}_{j'i} \tilde{\mathbf{V}}_{ji} \tilde{\mathbf{V}}_{ji}^H \mathbf{H}_{j'i}^H \mathbf{Y}_{j'} \Bigg), \quad (5.12)$$

and q_{ji} is the matching variable between the associated transmitters and receivers. This weighted bipartite matching problem can be solved optimally in polynomial time by using well-known approaches such as the Hungarian algorithm [49] or the auction algorithm [48].

Note that (5.11) is typically a *sparse* matching problem, since most pairs of $(i, j) \in \mathcal{I} \times \mathcal{J}$ are not associated, so the auction algorithm is likely to be more efficient than the Hungarian algorithm. The matching variable q_{ji} indicates $\mathbf{V}_i$ should be set to which of the $\tilde{\mathbf{V}}_{ji}$. Mathematically, $\underline{\mathbf{V}}$ is recovered as

$$\mathbf{V}_i^\star = \begin{cases} \tilde{\mathbf{V}}_{ji}, \text{ if } q_{ji} = 1 \text{ for some } j; \\ \mathbf{0}, \quad \text{otherwise.} \end{cases} \quad (5.13)$$

After updating $\underline{\mathbf{V}}$, the final step is to update the scheduling variable $\underline{s}$ for the fixed $\underline{\mathbf{V}}$. This is again a weighted bipartite matching problem, but now since $\mathbf{V}_i$ is fixed, this amounts to choosing the best receiver j for each transmitter i:

$$\begin{aligned} \underset{\underline{q}}{\text{maximize}} \quad & \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{L}_i} w_{ji} r_{ji} q_{ji} && (5.14a) \\ \text{subject to} \quad & \sum_{i \in \mathcal{K}_j} q_{ji} \leq 1, \ \forall j && (5.14b) \\ & \sum_{j \in \mathcal{L}_i} q_{ji} \leq 1, \ \forall i && (5.14c) \\ & q_{ji} \in \{0, 1\}, \ \forall (i, j) && (5.14d) \\ & q_{ji} = 0 \ \text{ if } i \notin \mathcal{K}_j \text{ or } j \notin \mathcal{L}_i, \ \forall (i, j), && (5.14e) \end{aligned}$$

where $w_{ji} r_{ji}$ is the weighted achievable rate if the receiver j is scheduled for transmitter i under fixed $\mathbf{V}_i$. Note that since $\underline{\mathbf{V}}$ is fixed, r_{ij} can be computed independently of the schedule, using an expression similar to (5.1). This problem can again be solved in polynomial time. The optimal schedule is then determined from the optimal q_{ij} as

$$s_j^\star = \begin{cases} i, \quad \text{if } q_{ji} = 1 \text{ for some } i; \\ \varnothing, \quad \text{otherwise.} \end{cases} \quad (5.15)$$

We note that the reason for having two sets of matching is because we allow a general network model in which each transmitter may associate with multiple receivers and each receiver may

Algorithm 10: Proposed FPLinQ Strategy for D2D Link Scheduling with Power Control and Beamforming

1 Initialize all the variables to feasible values;
2 **repeat**
3 Update $\underline{\mathbf{\Gamma}}$ according to (5.7);
4 Update $\underline{\mathbf{Y}}$ according to (5.27);
5 Update $\underline{\mathbf{V}}$ according to (5.13);
6 Update $\underline{s}$ by weighted bipartite matching (5.15);
7 **until** *the weighted sum rate converges*;

associate with multiple transmitters. For simpler D2D model such as the one in Fig. 5.1(a), these two matching steps would not have been necessary, as in [62].

Combining all the above steps together yields the FPLinQ strategy. Algorithm 10 summarizes the overall approach.

A desirable trait of FPLinQ as compared to FlashLinQ, ITLinQ and ITLinQ+ is that no tuning of design parameters is needed. But, FPLinQ is also somewhat more difficult to implement in a distributed fashion than FlashLinQ, ITLinQ, and ITLinQ+, because it additionally requires the update of the auxiliary variables $\underline{\mathbf{\Gamma}}$ and $\underline{\mathbf{Y}}$ per iteration.

The convergence proof of Algorithm 10 follows that of Algorithm 3 in Section 3.1.3.

Proposition 10. *The weighted sum rate across all the D2D links is nondecreasing after each iteration of Algorithm 10, so the objective function of the optimization problem is guaranteed to converge. Furthermore, at convergence, for fixed $\underline{s}$, the solution $\underline{\mathbf{V}}$ is a stationary point of the problem (5.3).*

Proof. The MM interpretation is used. Step 3 and step 4 of the algorithm is to construct the surrogate functions; step 5 carries out the maximization stage of MM with respect to $\underline{\mathbf{V}}$, and step 6 with respect to $\underline{s}$. Since the optimization objective is nondecreasing and is bounded above, Algorithm 10 must converge in objective value. When the discrete variable $\underline{s}$ is fixed, the problem reduces to a differentiable continuous optimization and meets Theorem 6 in Section 2.3.1, convergence to stationary point thus verified. □

5.1.4 Complexity Analysis

We now analyze the complexity of FPLinQ (*i.e.*, Algorithm 10). We assume that there are a total of L D2D links in the network; each transmitter/receiver is associated with a small number (*i.e.*, constant number) of neighboring devices, so that $|\mathcal{I}| = O(L)$ and $|\mathcal{J}| = O(L)$. To ease the analysis, we assume that FPLinQ runs for a fixed number of iterations.

Communication Complexity: In each iteration of FPLinQ, each transmitter i requires the tuple $(\underline{\mathbf{\Gamma}}, \underline{\mathbf{Y}}, \underline{s})$ to update $\mathbf{V}_i$, while every receiver j requires $\underline{\mathbf{V}}$ to update $\mathbf{\Gamma}_j$ and $\mathbf{Y}_j$. Each of $\mathbf{V}_i, \mathbf{\Gamma}_j, \mathbf{Y}_j$ is an $N \times N$ matrix. Further, the channel coefficients from $O(L^2)$ direct and

interfering channels are needed, with each channel being an $N \times N$ matrix. Thus, the total communication complexity of these updates is $O(N^2L^2)$. The two matchings in Step 5 and Step 6 require the matching weights of all the links, thus introducing a communication complexity of $O(L)$. The overall communication complexity of FPLinQ is then $O(N^2L^2)$. In the single-antenna single-association case, the communication complexity of FPLinQ in each iteration is $O(L^2)$; in comparison, the communication complexity of each step of FlashLinQ, ITLinQ, and ITLinQ+ is also $O(L^2)$, as they all require the $O(L^2)$ channel coefficients.

Computational Complexity: We first consider the update steps of FPLinQ prior to matching, which as analyzed in [33] has a per-iteration computational complexity of $O(N^4L^2)$. The matching step can be performed using the auction algorithm [48], which has a computational complexity of $O(L|\mathcal{I}|\log|\mathcal{I}| + L|\mathcal{J}|\log|\mathcal{J}|) = O(L^2\log(L))$. Thus, the overall per-iteration computational complexity of FPLinQ is $O(N^4L^2 + L^2\log(L))$. In the single-antenna single-association case, the per-iteration computational complexity of FPLinQ reduces to $O(L^2\log L)$, while the total computational complexities of FlashLinQ, ITLinQ, and ITLinQ+ are all equal to $O(L^2)$.

We observe that the computational complexity of FPLinQ is sensitive to the number of antennas N (mainly due to the matrix inverse). Overall, asymptotically, FPLinQ has the same communication complexity, but higher computational complexity than the greedy based approaches—FlashLinQ, ITLinQ, and ITLinQ+. Note that although the joint scheduling and power control problem is NP-hard in general [46, 56], recent results nevertheless show that scalable implementation is feasible for a metropolitan-scale network with thousands of terminals [63, 64]. In particular, [64] uses the scalar FP method of [23, 33].

Table 5.1 displayed on the next page summarizes the comparison between the proposed FPLinQ algorithm and the main benchmarks. The main advantage of FPLinQ is that it allows for flexible association, guarantees convergence without needing tuning parameters, while alleviating the potential pre-mature turn-off problem.

Table 5.1: Comparison of Link Scheduling Algorithms for D2D Networks

	FPLinQ	FlashLinQ [58]	ITLinQ [59]	ITLinQ+ [60]	BCD [38]
Scheduling & Association	Flexible	Single	Single	Single	Flexible
Power Control	✓	✗	✗	✓	✓
Beamforming	✓	✗	✗	✗	✓
Tuning Parameters	Not Needed	Required	Required	Required	Not Needed
Convergence with Fixed Schedule	Stationary Point	–	–	–	Stationary Point
Computational Complexity	$O(L^2(N^4 + \log L))$	$O(L^2)$	$O(L^2)$	$O(L^2)$	$O(L^2(N^4 + \log L))$
Communication Complexity	$O(N^2L^2)$	$O(L^2)$	$O(L^2)$	$O(L^2)$	$O(N^2L^2)$
Link Reactivation	✓	✗	✗	✗	✗

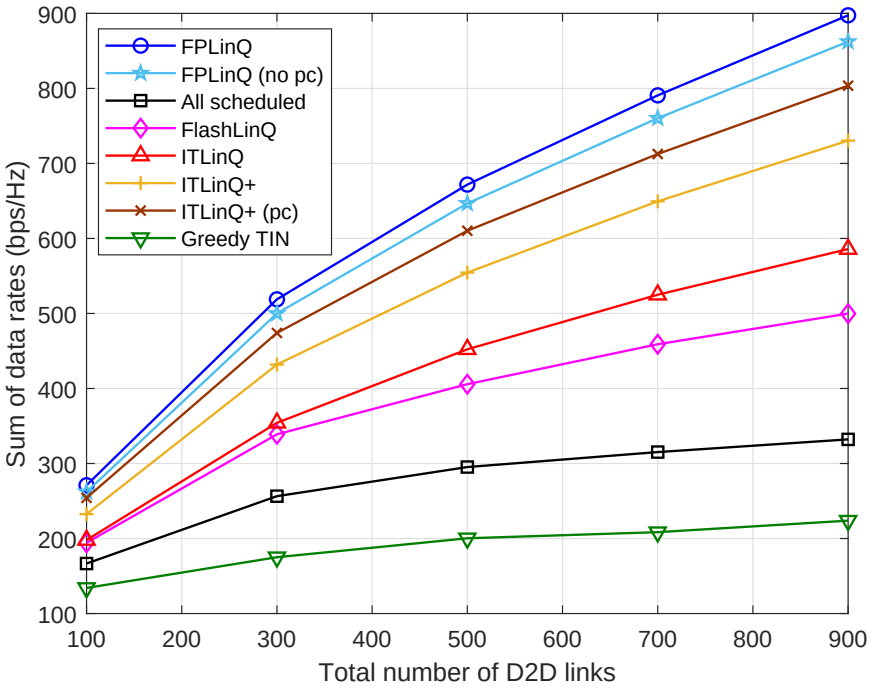


Figure 5.3: Sum-rate maximization for the single-association D2D network.

5.1.5 Numerical Results

We validate the performance of FPLinQ through comparison with the benchmark methods for a D2D network in a 1 km×1 km square area where the D2D links are randomly located. Following [58–60], we adopt the short-range outdoor channel model ITU-1411 and use a 5 MHz-wide frequency band centered at 2.4 GHz. Moreover, the antenna height of each device is 1.5 m; the antenna gain is 2.5 dBi; the PSD of AWGN is -169 dBm/Hz; the noise figure is 7 dB; the maximum transmit power is 20 dBm; the shadowing is modeled as a Gaussian random variable in decibel with the standard deviation of 10; the distance between the transmitter and receiver of each link is uniformly distributed between 2m and 65m.

The first simulation setting follows [58–60]: Given a set of links with single-antenna transmitters/receivers and fixed single association (as shown in Fig. 5.1), the aim is to maximize the sum rate across the links. We use FlashLinQ [58], ITLinQ [59], and ITLinQ [60] as benchmarks. The BCD method is equivalent to FPLinQ in this single-association case. Because the benchmark methods do not have power control, for fair comparison, we modify FPLinQ slightly to restrict the power to be either zero or the maximum, *i.e.*, round each $\mathbf{V}_i$ to $\{0, \sqrt{P_{\max}}\}$. This new version of FPLinQ without power control is referred to as "FPLinQ (no pc)". Further, we introduce two baselines: one is to activate all the links and the other is to activate the links greedily to meet the TIN condition.

Fig. 5.3 shows the sum rate versus the total number of D2D links. Observe that ITLinQ+

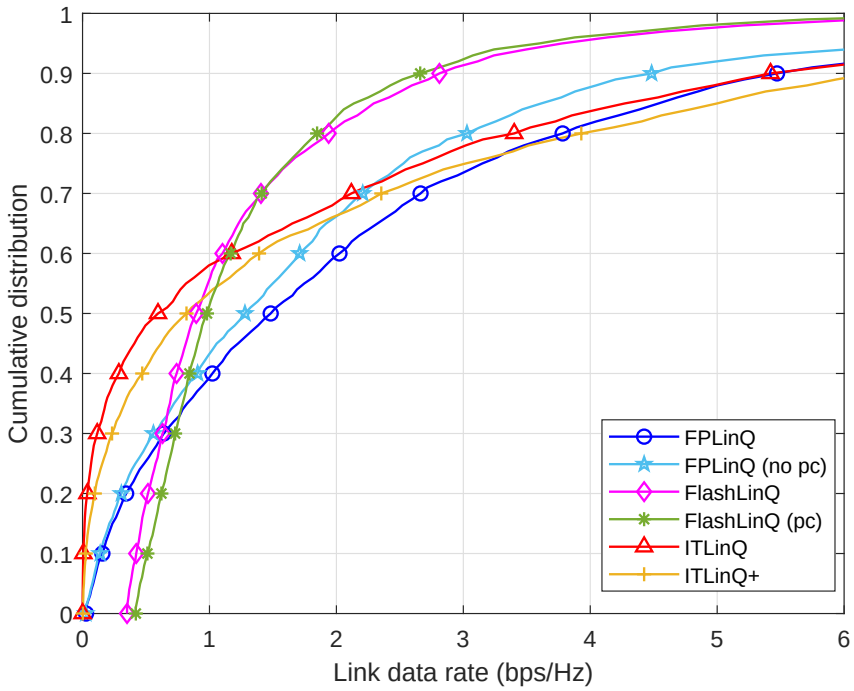


Figure 5.4: Log-utility maximization for the single-association D2D network.

outperforms ITLinQ, and ITLinQ outperforms FlashLinQ, as expected from the previous literature [59, 60]. Without power control, FPLinQ (no pc) significantly outperforms FlashLinQ, ITLinQ, and ITLinQ+, especially when the D2D links are densely located in the area. In particular, observe that Greedy TIN is even worse than simply scheduling all the links because it is too conservative about the effect of interference. Further, as suggested in [60], we run ITLinQ+ and the power control algorithm (*e.g.,* the WMMSE method) alternatively in order to account for joint scheduling and power control; this method is referred to as "ITLinQ (pc)". However, the performance of ITLinQ+ with power control is still inferior to that of FPLinQ and even that of FPLinQ (no pc).

The above simulation setting is only concerned with sum rate, as the weights are all set to one. We now consider a more demanding setting that takes priority weights into account. In this simulation, the weights are updated using the proportional fairness criterion, which is equivalent to maximizing the log-utility of the average link rates in the long run [65]. The network setting follows the previous simulation; the total number of links is fixed at 100. Fig. 5.4 compares the CDF of the link rates; the upper part of Table 5.2 compares the log-utility values. As we can see in Fig. 5.4, FPLinQ (no pc) strikes a better balance between the high-rate regime and the low-rate regime than ITLinQ and ITLinQ+. Surprisingly, FlashLinQ performs much better than ITLinQ and ITLinQ+ in this simulation; its performance is even slightly better than FPLinQ (no pc) according to Table 5.2. In particular, observe in Fig. 5.4 that the low-rate links benefit the most from FlashLinQ, so FlashLinQ is fairly effective in protecting the low-rate links

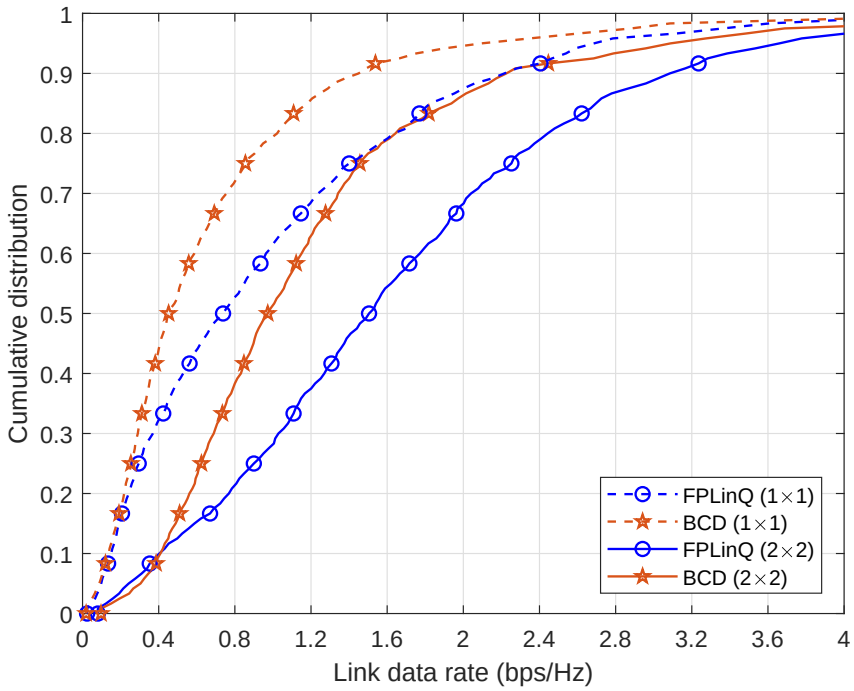


Figure 5.5: Log-utility maximization for the flexible-association D2D network.

from strong interference, but its threshold value must be chosen carefully. Further, the benefit from the low-rate links comes at a cost for high-rate links. Overall, when we include power control and compare FPLinQ with a new benchmark method that combines FlashLinQ and power control in an alternative fashion, referred to as "FlashLinQ (pc)", FPLinQ outperforms FlashLinQ (pc) in network utility, when scheduling is optimized along with transmit powers, as shown in Table 5.2 on the next page.

Finally, we consider the flexible association case. We first generate 100 disjoint D2D links as before, but also generate two extra transmitters randomly for each receiver, and further let one third of the transmitters connect with one additional geographically closest receiver (excluding the already connected one). In this setup, we frequently encounter the situation that multiple transmitters contend for the same receiver, so the premature turning-off problem is very likely to occur. We again optimize the log-utility by updating the link weights according to the proportional fairness criterion. FPLinQ is compared with the BCD method for both the single-antenna case and the 2×2 MIMO case (*i.e.*, when each device terminal has 2 antennas). Note that FlashLinQ, ITLinQ, and ITLinQ+ are not applicable here, because they do not handle MIMO. Fig. 5.5 shows the CDF of link rates, and the lower part of Table 5.2 summarizes the log-utility results. It can be seen that FPLinQ significantly outperforms BCD. In fact, as shown in Fig. 5.5, FPLinQ improves upon the BCD method by more than 50% for the 50th percentile link rate, in both the single-antenna case and the MIMO case. The corresponding log-utility of FPLinQ is also much higher. These results show that the premature turning-off can be fairly

Table 5.2: Sum Log-Utility over D2D Networks

Fixed Single Association	Log Utility
FPLinQ	177.6
FPLinQ (no pc)	162.3
FlashLinQ	163.0
FlashLinQ (pc)	170.6
ITLinQ	57.0
ITLInQ+	109.5
Flexible Association	**Log Utility**
BCD (1×1)	99.6
BCD (2×2)	186.4
FPLinQ (1×1)	139.3
FPLinQ (2×2)	224.4
FPLinQ (4×4)	298.9
FPLinQ (8×8)	369.0
Vector FP (2×2)	223.3
Vector FP (4×4)	279.0
Vector FP (8×8)	321.5

detrimental to the performance of D2D system in the flexible association case, thus making the proposed FPLinQ strategy a preferred strategy.

One of the key advantages of the proposed matrix FP strategy is its ability to accommodate multiple data streams in each MIMO link. In the next simulation, we evaluate the gain of multiple data-stream transmission over the single data-stream transmission. Toward this end, we compare FPLinQ (with matrix FP) against the vector FP method (also called multidimensional FP in [33]). The vector FP algorithm is the same as Algorithm 10 except that each transmit beamformer $\mathbf{V}_i \in \mathbb{C}^N$ is a complex vector instead of a matrix, so at most one data stream can be transmitted on each link. Fig. 5.6 shows the CDF of link rates under different MIMO settings. It can be seen that while the gain of FPLinQ as compared to the vector FP is marginal in the 2×2 MIMO case, as more antennas are deployed at each terminal, the multiple data-stream transmission by FPLinQ starts to significantly outperform. The above observations is also evident from the lower part of Table 5.2. Therefore, if the number of antenna N is small (e.g., 2), then using the vector FP in Algorithm 10 is more suited because of its lower complexity; on the other hand, if N is large (e.g., 8), then using FPLinQ with multiple data-stream transmission can boost the overall network throughput significantly.

Finally, Fig. 5.7 shows the convergence speed of FPLinQ when maximizing the sum rate for the flexible-association D2D network with 400 links. FPLinQ has fast convergence under all the three MIMO settings. Taking the 2×2 case for example, we observe from Fig. 5.7 that the majority of sum rate increment is obtained after the first 10 iterations. Observe also that the convergence of FPLinQ is slower when more antennas are deployed at each terminal. But, for the 8×8 case in Fig. 5.7, we can already reap most of the rate gain after about 40-60 iterations.

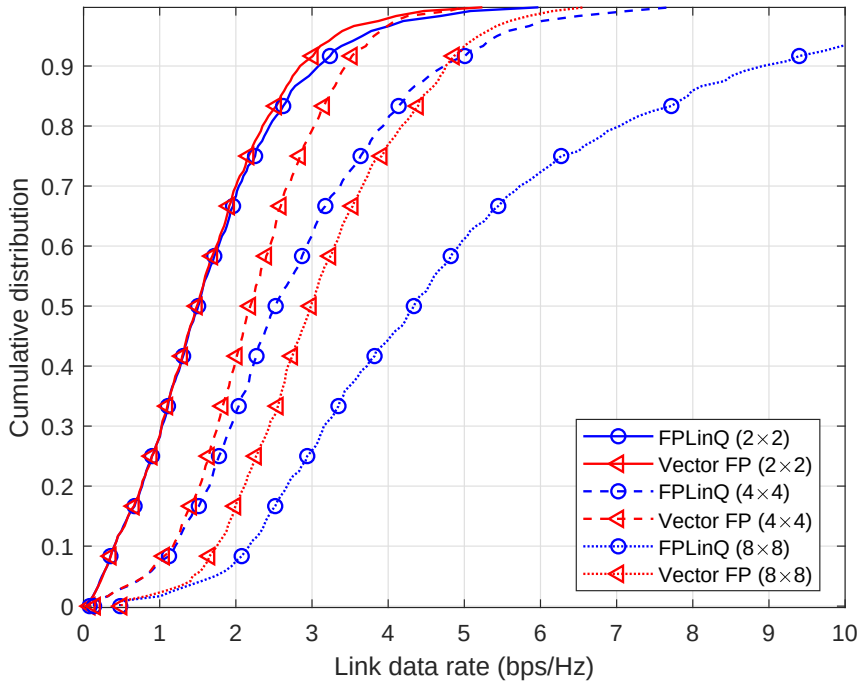


Figure 5.6: Log-utility maximization for the flexible-association D2D network: FPLinQ vs. Vector FP.

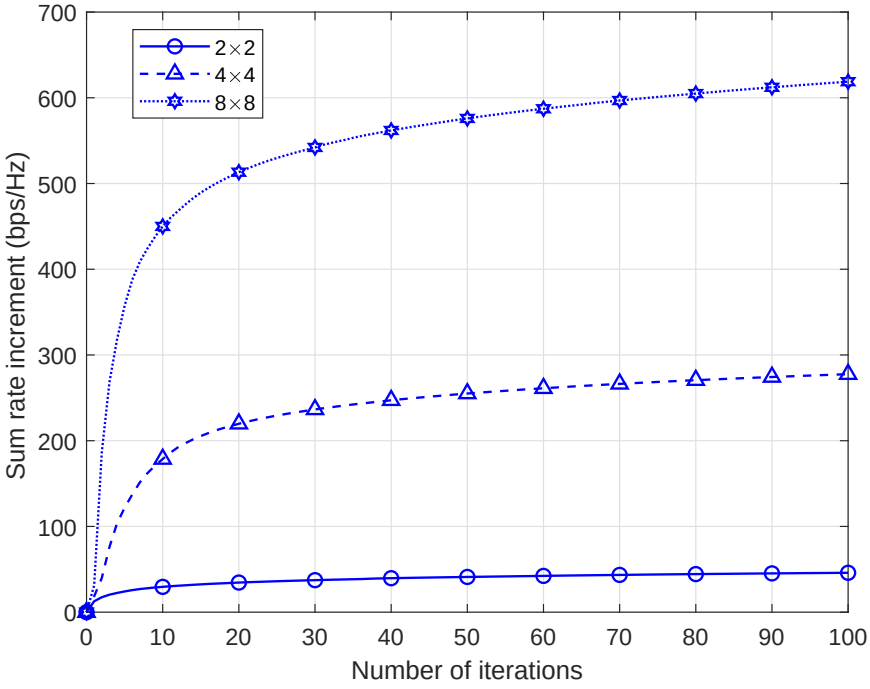


Figure 5.7: Convergence of FPLinQ in maximizing the sum rate for the flexible-association D2D network.

5.2 Nonorthogonal Pilot Design for Massive MIMO

Pilot contamination caused by the nonorthogonality of pilots is a main limiting factor in multi-cell massive MIMO systems because it can significantly impair channel estimation. Following the recent works of [66–68], our system model allows arbitrary sequences (under the power and length constraints) to be used as pilots, while the prior works mostly assume orthogonal pilots within the cell in order to eliminate the intra-cell interference in channel estimation. As illustrated in Fig. 5.8, the orthogonal scheme[3] precludes the interfering pilots from the home cell, but it results in pilot contamination from the neighboring cells (when the same set of orthogonal pilots are reused in each cell); in comparison, the nonorthogonal scheme provides more flexibility to pilot design, thereby avoiding the high correlation with the desired pilot. This section further investigates this approach using a new optimization framework. Specifically, we reformulate the problem of minimizing the weighted MSE of channel estimation as a matrix-ratio program that can be efficiently approximated as a sequence of convex optimizations via a matrix fractional programming approach. The proposed algorithm, named coordinated nonorthogonal pilot design (CNPD), provides fast convergence to a stationary point of the weighted MSE objective. We further reveal a relation between rate maximization and MSE minimization, which provides insights into the appropriate setting of weights in weighted MSE minimization.

5.2.1 Problem Formulation

Consider a total of L BSs each associated with K user terminals. We refer to the area occupied by each BS and its user terminals as a *cell*. The full spectrum band is reused in every cell. The BSs estimate the uplink channels based on the uplink pilots transmitted from the user terminals. We seek a coordinated pilot design that minimizes the channel estimation error throughout the network.

We use i or j to denote the index of each cell and its BS, and (i,k) the index of the kth user in cell i. Assume that every BS has M antennas and every user terminal has a single antenna. Let $\mathbf{h}_{j,ik} \in \mathbb{C}^M$ be the channel from user (i,k) to BS j, and let $\mathbf{H}_{ji} \in \mathbb{C}^{M\times K}$ be the channel matrix:

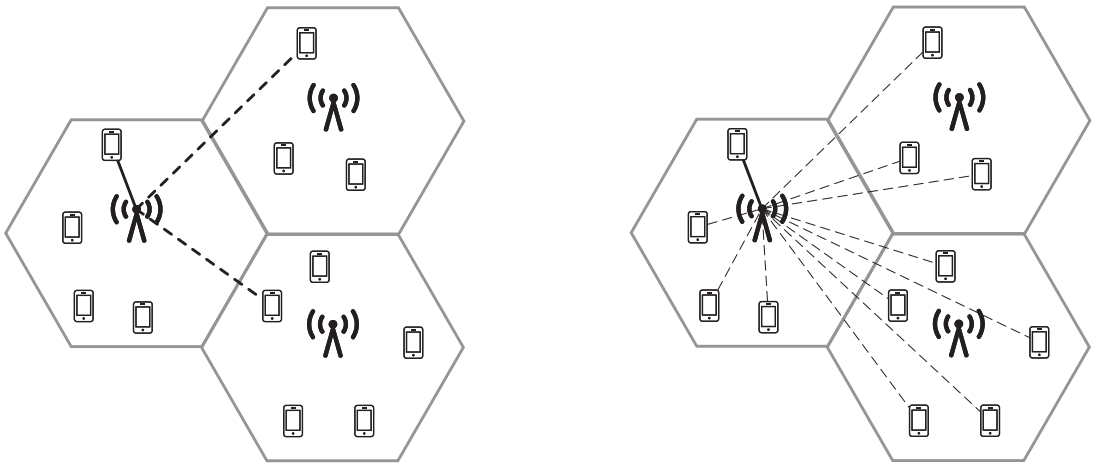
$$\mathbf{H}_{ji} = \begin{bmatrix} | & | & & | \\ \mathbf{h}_{j,i1} & \mathbf{h}_{j,i2} & \cdots & \mathbf{h}_{j,iK} \\ | & | & & | \end{bmatrix}. \tag{5.16}$$

Following the previous works [66, 67], we model each channel $\mathbf{H}_{ji}$ by using the Kronecker structure with a *partially separable correlation*, *i.e.*,

$$\mathbf{H}_{ji} = \mathbf{Q}_j^{\frac{1}{2}} \mathbf{G}_{ji} \mathbf{P}_{ji}^{\frac{1}{2}}, \tag{5.17}$$

where the receiver-side channel $\mathbf{Q}_j \in \mathbb{C}^{M\times M}$ is a deterministic positive semidefinite matrix, the

[3]Here and throughout, the orthogonal scheme refers to using orthogonal pilots in the cell, but the pilot orthogonality across cells is not guaranteed.



(a) Orthogonal (b) Nonorthogonal

Figure 5.8: Orthogonal scheme vs. nonorthogonal scheme. Solid line is desired pilot and dashed lines are interfering pilots; the width of dashed lines reflects the correlation with the desired pilot.

large-scale channel strength $\mathbf{P}_{ji} \in \mathbb{C}^{K\times K}$ is a deterministic diagonal PSD matrix

$$\mathbf{P}_{ji} = \operatorname{diag}\big[\beta_{j,i1}, \beta_{j,i2}, \ldots, \beta_{j,iK}\big] \tag{5.18}$$

with $0 \le \beta_{j,ik} \le 1$ between any pair of BS j and user (i,k), and the small-scale fading $\mathbf{G}_{ji} \in \mathbb{C}^{M\times K}$ is a random matrix with i.i.d. entries distributed as $\mathcal{CN}(0,1)$. We remark that a common channel model as considered in [69] with $\mathbf{Q}_j = \boldsymbol{I}_M$ for each $j = 1, 2, \ldots, L$ is a special case of (5.17).

Assume that each pilot sequence consists of τ symbols. Let $\mathbf{s}_{ik} \in \mathbb{C}^\tau$ be the pilot sequence transmitted from user (i,k), and further denote

$$\mathbf{S}_i = \begin{bmatrix} | & | & & | \\ \mathbf{s}_{i1} & \mathbf{s}_{i2} & \cdots & \mathbf{s}_{iK} \\ | & | & & | \end{bmatrix}. \tag{5.19}$$

Let $s_{ik}[t] \in \mathbb{C}$ be the tth symbol of the pilot sequence $\mathbf{s}_{ik}$, *i.e.*, $\mathbf{s}_{ik} = (s_{ik}[0], s_{ik}[1], \ldots, s_{ik}[\tau-1])$. The pilot signal received at BS i is

$$\mathbf{V}_i = \mathbf{H}_{ii}\mathbf{S}_i^\top + \sum_{j=1, j\neq i}^{L} \mathbf{H}_{ij}\mathbf{S}_j^\top + \mathbf{Z}_i, \tag{5.20}$$

where $\mathbf{Z}_i \in \mathbb{C}^{M\times\tau}$ is additive noise with i.i.d. entries distributed as $\mathcal{CN}(0,\sigma^2)$ for the fixed noise power level σ^2.

Upon receiving the uplink pilot signals from the users, each BS i estimates its $\mathbf{H}_{ii}$ using the

MMSE criterion from [66, 67]. Let $\hat{\mathbf{H}}_{ii}$ be the MMSE estimate of $\mathbf{H}_{ii}$; it is determined as

$$\mathrm{vec}\big(\hat{\mathbf{H}}_{ii}\big) = \big(\mathbf{P}_{ii}\mathbf{S}_i^H \otimes \boldsymbol{I}_M\big)\big(\mathbf{D}_i \otimes \boldsymbol{I}_M\big)^{-1}\mathrm{vec}\big(\mathbf{V}_i\big), \tag{5.21}$$

where

$$\mathbf{D}_i = \sigma^2 \boldsymbol{I}_\tau + \sum_{j=1}^{L} \mathbf{S}_j \mathbf{P}_{ij} \mathbf{S}_j^H. \tag{5.22}$$

The resulting MSE of user (i,k) is

$$\mathsf{MSE}_{ik} = \mathbb{E}\big[\|\hat{\mathbf{h}}_{i,ik} - \mathbf{h}_{i,ik}\|^2\big], \tag{5.23}$$

where $\hat{\mathbf{h}}_{iik}$ is the kth column of $\hat{\mathbf{H}}_{ii}$. We aim to choose the pilot sequences to minimize the sum of weighted MSEs, *i.e.*,

$$\text{minimize} \quad \sum_{i=1}^{L}\sum_{k=1}^{K} w_{ik}\mathsf{MSE}_{ik} \tag{5.24}$$

for a set of fixed nonnegative weights $w_{ik} \geq 0$. For instance, we may set $w_{ik} = 1$ to minimize the sum of MSEs as in [67], or $w_{ik} = 1/\beta_{iik}$ to minimize the sum of normalized MSEs as in [70]. For now, we focus on optimizing $\underline{\mathbf{S}}$ for the fixed weights w_{ik}. Section 5.2.5 will discuss the choice of w_{ik} for maximizing the achievable rate.

Following the steps in [67], we can formalize problem (5.24) as

$$\underset{\underline{\mathbf{S}}}{\text{maximize}} \qquad \sum_{i=1}^{L} \alpha_i \mathrm{tr}\big(\mathbf{W}_i \mathbf{P}_{ii} \mathbf{S}_i^H \mathbf{D}_i^{-1} \mathbf{S}_i \mathbf{P}_{ii}\big) \tag{5.25a}$$

$$\text{subject to} \qquad \|\mathbf{s}_{ik}\|^2 \leq \rho_{ik}, \ \forall (i,k), \tag{5.25b}$$

where $\alpha_i = \mathrm{tr}(\mathbf{Q}_i)$, $\mathbf{W}_i = \mathrm{diag}[w_{i1}, w_{i2}, \ldots, w_{iK}]$, and ρ_{ik} is the power constraint of user (i,k).

Problem (5.25) is a difficult optimization problem, because the choice of pilot sequences $\mathbf{S}_i$ appears in both the numerator and the denominator of a matrix fraction in (5.25a). The authors in [67] propose a greedy sum of ratio traces maximization (GSRTM) algorithm to optimize each row of $\mathbf{S}_i$ sequentially. Here we suggest a matrix-FP approach to optimize the entire matrices $\mathbf{S}_i$ jointly. In addition, we illustrate via simulations that it leads to more accurate channel estimation.

As studied in [66,67], the above problem formulation can be extended to the case of reduced radio-frequency (RF) chains, *i.e.*, when each BS i uses an RF chain combiner $\mathbf{U}_i \in \mathbb{C}^{N\times M}$ (for $N < M$) to reduce the dimensionality of the received signal $\mathbf{V}_i$. As already shown in [66, 67], the weight α_i in problem (5.25) then becomes $\mathrm{tr}(\mathbf{Q}_i\mathbf{U}_i^H(\mathbf{U}_i\mathbf{Q}_i\mathbf{U}_i^H)^{-1}\mathbf{U}_i\mathbf{Q}_i)$. Our work focuses on optimizing the pilot variable $\underline{\mathbf{S}}$ given the weights α_i in (5.25), regardless of how each α_i is determined.

5.2.2 Iterative Optimization by Matrix Fractional Programming

In light of Theorem 11 in Section 2.5.2, we can reformulate the weighted sum-MSE minimization problem (5.25) as follows.

Proposition 11. *The weighted sum MSE problem (5.25) is equivalent to*

$$\underset{\underline{\mathbf{S}},\,\underline{\mathbf{Y}}}{\text{maximize}} \quad \sum_{i=1}^{L} \alpha_i \mathrm{tr}\Big(\mathbf{W}_i\big(2\Re\big\{\mathbf{P}_{ii}\mathbf{S}_i^H\mathbf{Y}_i\big\} - \mathbf{Y}_i^H\mathbf{D}_i\mathbf{Y}_i\big)\Big) \tag{5.26a}$$

$$\text{subject to} \quad \|\mathbf{s}_{ik}\|^2 \le \rho_{ik},\ \forall (i,k) \tag{5.26b}$$

$$\mathbf{Y}_i \in \mathbb{C}^{\tau\times K},\ \forall i, \tag{5.26c}$$

where $\mathbf{Y}_i$ *is the auxiliary variable.*

Proof. The new objective function (5.26a) is derived by treating $\mathbf{S}_i\mathbf{P}_{ii}^H$ and $\mathbf{D}_i$ respectively as $\mathbf{A}_i$ and $\mathbf{B}_i$ in Theorem 11 with $f_i(\mathbf{R}) = \alpha_i \mathrm{tr}(\mathbf{W}_i\mathbf{R})$. □

In the remainder of this section, we use $\mathbf{y}_{ik}$ to denote the kth column of $\mathbf{Y}_i$. To solve problem (5.26) in Proposition 11, we propose to optimize $\underline{\mathbf{S}}$ and $\underline{\mathbf{Y}}$ alternatively. When $\underline{\mathbf{S}}$ is held fixed, each $\mathbf{Y}_i$ can be optimally determined by completing the square for $\mathbf{Y}_i$ in (5.26a), *i.e.*,

$$\mathbf{Y}_i = \mathbf{D}_i^{-1}\mathbf{S}_i\mathbf{P}_{ii}. \tag{5.27}$$

Next, we optimize $\underline{\mathbf{S}}$ for fixed $\underline{\mathbf{Y}}$. The key step is to rewrite the objective function (5.26a) in a quadratic form with respect to each $\mathbf{s}_{ik}$, as specified in the following proposition.

Proposition 12. *The new objective function (5.26a) can be rewritten as*

$$f(\underline{\mathbf{S}},\underline{\mathbf{Y}}) = \sum_{(i,k)} \xi_{ik} + c(\underline{\mathbf{Y}}), \tag{5.28}$$

where

$$\xi_{ik} = 2\Re\{w_{ik}\beta_{i,ik}\mathbf{s}_{ik}^H\mathbf{y}_{ik}\} - \mathbf{s}_{ik}^H\left(\sum_{j=1}^{L}\beta_{j,ik}\mathbf{Y}_j\mathbf{W}_j\mathbf{Y}_j^H\right)\mathbf{s}_{ik} \tag{5.29}$$

and

$$c(\underline{\mathbf{Y}}) = \sigma^2\sum_{i=1}^{L}\mathrm{tr}\big(\mathbf{Y}_i^H\mathbf{Y}_i\big). \tag{5.30}$$

Now, combining Proposition 12 with the power constraint (5.26b), we arrive at a Lagrangian function for problem (5.26):

$$\mathcal{L}(\underline{\mathbf{S}},\underline{\mathbf{Y}},\underline{\lambda}) = \sum\nolimits_{(i,k)}\Big(\xi_{ik} - \lambda_{ik}\big(\|\mathbf{s}_{ik}\|^2 - \rho_{ik}\big)\Big) + c(\underline{\mathbf{Y}}), \tag{5.31}$$

Algorithm 11: Coordinated Nonorthogonal Pilot Design (CNPD)

1 Initialize all the variables to feasible values;
2 **repeat**
3 Update the auxiliary variable $\underline{\mathbf{Y}}$ by (5.27);
4 Update the pilot variable $\underline{\mathbf{S}}$ by (5.32) along with the Lagrangian multiplier λ_{ik} in (5.33), or by simultaneous scaling as stated in Remark 3 when $\sigma^2 = 0$;
5 **until** *the variables* $(\underline{\mathbf{S}}, \underline{\mathbf{Y}})$ *converge*;

where each λ_{ik} is a Lagrangian multiplier for constraint (5.26b). By completing the square in (5.31), the optimal $\mathbf{s}_{ik}$ is given by

$$\mathbf{s}_{ik} = \left(\sum_{j=1}^{L} \beta_{j,ik} \mathbf{Y}_j \mathbf{W}_j \mathbf{Y}_j^H + \lambda_{ik} \boldsymbol{I}_\tau \right)^{-1} w_{ik} \beta_{i,ik} \mathbf{y}_{ik}, \tag{5.32}$$

where λ_{ik} is determined by the complementary slackness, *i.e.*,

$$\lambda_{ik} = \begin{cases} 0, \text{ if } \|\mathbf{s}_{ik}\|^2 \leq \rho_{ik} \text{ already;} \\ \lambda_{ik}^\star > 0 \text{ such that } \|\mathbf{s}_{ik}\|^2 = \rho_{ik}, \text{ otherwise.} \end{cases} \tag{5.33}$$

For the second case of (5.33) wherein $\|\mathbf{s}_{ik}\|^2 = \rho_{ik}$, the optimal $\lambda_{ik}^\star$ can be computed efficiently by a bisection search because $\|\mathbf{s}_{ik}\|^2$ in (5.32) is monotonically decreasing with $\lambda_{ik} > 0$. When $\underline{\mathbf{Y}}$ is held fixed, maximizing the objective function (5.28) over $\underline{\mathbf{S}}$ is a convex problem, so $(\underline{\mathbf{S}}, \underline{\lambda})$ obtained from (5.32) and (5.33) are jointly optimal from a Lagrangian dual theoretic perspective.

Remark 2. The above matrix-FP reformulating procedure mimics that of [71] until (5.27). But a further reformulation in Proposition 12 is needed here in order to carry out the process of completing the square. This is due to the difference in how $\mathbf{S}_j$ enters the objective function.

Remark 3. We show that the computation of $\mathbf{s}_{ik}$ can be further simplified in a special case. When $\mathbf{Z}_i$ is negligible (i.e., $\sigma^2 = 0$), as shown in [67], the sum of weighted MSEs remains the same if each pilot $\mathbf{s}_{ik}$ is multiplied by the same nonzero factor α. Thus, we can enforce the power constraint (5.26b) by scaling all the $\mathbf{s}_{ik}$'s (assuming that $\lambda_{ik} = 0$) simultaneously with a sufficiently small positive factor, instead of going through the computation of λ_{ik}.

Algorithm 11 summarizes the overall approach. The following proposition analyzes its convergence.

Proposition 13. *The sum of weighted MSEs is monotonically nonincreasing per iteration in CNPD; the pilot variable* $\underline{\mathbf{S}}$ *converges to a stationary point of problem (5.25).*

Proof. The iterative update in CNPD can be interpreted as a sequence of MM steps [30, 31]. Specifically, when $\underline{\mathbf{Y}}$ is held fixed, $\underline{\mathbf{S}}$ in (5.32) is the globally optimal solution that maximizes $f(\underline{\mathbf{S}}, \underline{\mathbf{Y}})$ in (5.28), namely the *maximization* phase; when $\underline{\mathbf{S}}$ is held fixed, $f(\underline{\mathbf{S}}, \underline{\mathbf{Y}})$ is less than or

equal to the original objective in (5.26a), where the equality holds iff $\underline{\mathbf{Y}}$ meets (5.27), namely the *minorization* phase. The above proposition then directly follows by the property of the MM algorithm [30, 31]. □

We next compare CNPD with the GSRTM algorithm proposed in [67]. The main idea behind GSRTM is to optimize one row of the matrix $\mathbf{S}_i$ at a time while fixing all other rows. Because the rows of $\mathbf{S}_i$ are not optimized jointly in GSRTM, this greedy method is prone to being trapped in a local optimum. Furthermore, it can be shown that CNPD has a computational complexity scaling of $O(\tau^3 LKT)$ where T is the number of iterations, while GSRTM has computational complexity scaling of $O(\tau^3 L^2 K + \tau^4 L)$, so that GSRTM is more sensitive[4] to τ and L. Moreover, the convergence property of GSRTM is difficult to analyze, whereas CNPD is guaranteed to converge to a stationary-point solution.

5.2.3 Discrete Pilot Sequence Design

Arbitrary complex-valued pilot sequences may be difficult to implement in practice. In this section, we restrict the choice of each pilot symbol to a 4-quadrature amplitude modulation (QAM) constellation $\mathcal{C} = \{\varepsilon(1+\mathrm{j}), \varepsilon(1-\mathrm{j}), \varepsilon(-1+\mathrm{j}), \varepsilon(-1-\mathrm{j})\}$ with a power control factor $\varepsilon = \sqrt{\rho/2\tau}$. Such sequences are referred to as discrete pilot sequences.

To design optimal discrete pilot sequences, we maximize the objective function $f(\underline{\mathbf{S}}, \underline{\mathbf{Y}})$ in (5.31) for fixed $\underline{\mathbf{Y}}$ (which has been updated by (5.27)) over the QAM-constellation as follows:

$$\mathbf{s}'_{ik} = \arg\min_{\mathbf{s}' \in \mathcal{C}^\tau} \left\| \left(\sum_{j=1}^{L} \beta_{j,ik} \mathbf{Y}_j \mathbf{W}_j \mathbf{Y}_j^H \right)^{\frac{1}{2}} (\mathbf{s}' - \mathbf{s}_{ik}) \right\|, \tag{5.34}$$

where $\mathbf{s}_{ik}$ has the same form as (5.32) but with $\lambda_{ik} = 0$ (since there is no power constraint on $\mathbf{s}_{ik}$ in this case); $\mathcal{C}^\tau$ refers to a Cartesian power of set $\mathcal{C}$. The projection of $\mathbf{s}_{ik}$ onto $\mathcal{C}^\tau$ may be computationally complex in practice as the size of $\mathcal{C}^\tau$ grows exponentially with the pilot length τ. Thus, we propose a suboptimal solution of simply rounding each $s_{ik}[t]$ to $\mathcal{C}$, *i.e.*,

$$s'_{ik}[t] = \varepsilon \cdot \mathrm{sgn}(\Re\{s_{ik}[t]\}) + \mathrm{j}\,\varepsilon \cdot \mathrm{sgn}(\Im\{s_{ik}[t]\}) \tag{5.35}$$

where $\mathrm{sgn}(\cdot)$ is the sign function. Observe that the heuristic in (5.35) amounts to $\mathbf{s}'_{ik} = \arg\min_{\mathbf{s}' \in \mathcal{C}^\tau} \|\mathbf{s}' - \mathbf{s}_{ik}\|$. As compared to (5.34), this heuristic is in essence assuming that

$$\sum_{j=1}^{L} \beta_{j,ik} \mathbf{Y}_j \mathbf{W}_j \mathbf{Y}_j^H \approx \theta \boldsymbol{I}_\tau \tag{5.36}$$

for some positive scalar $\theta > 0$. According to our simulations, the above approximation is not tight in general, but the resulting discrete pilot sequences are quite effective in reducing the

[4] Although T may increase with τ and L, we can stop CNPD early because of its monotonic improvement as stated in Proposition 13.

channel estimation error.

5.2.4 Achievable Data Rates

Due to the nonorthogonal pilots used for channel estimation, the conventional achievable rate expression does not hold true in our massive MIMO system. Before proceeding to the further optimization, we introduce three new achievable rate expressions that specialize in the nonorthogonal pilot case:

(i) Instantaneous Ergodic Rate

$$R_{ik} = \mathbb{E}\left[\log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{i,ik}\right\|^4 \tilde{\rho}_{ik}}{\sum_{(j,\ell)\neq(i,k)} \left|\hat{\mathbf{h}}_{i,ik}^H \mathbf{h}_{i,j\ell}\right|^2 \tilde{\rho}_{j\ell} + \sigma^2\left\|\hat{\mathbf{h}}_{i,ik}\right\|^2 + \left|\hat{\mathbf{h}}_{i,ik}^H(\mathbf{h}_{i,ik} - \hat{\mathbf{h}}_{i,ik})\right|^2 \tilde{\rho}_{ik}}\right)\right]. \tag{5.37}$$

(ii) Closed-Form Rate

$$\tilde{R}_{ik} = \log_2\left(1 + \frac{M^2\mu_{ik}^2\tilde{\rho}_{ik}}{M\mu_{ik}\sum_{(j,\ell)}\beta_{i,j\ell}\tilde{\rho}_{j\ell} + M^2\beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\cdot\mathbf{F}_i(\underline{\tilde{\rho}})\cdot\mathbf{D}_i^{-1}\mathbf{s}_{ik} + M\mu_{ik}\sigma^2 - M^2\mu_{ik}^2\tilde{\rho}_{ik}}\right). \tag{5.38}$$

where

$$\mu_{ik} = \beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\mathbf{s}_{ik} \tag{5.39}$$

and

$$\mathbf{F}_i(\underline{\tilde{\rho}}) = \sum_{j=1}^{L}\left(\mathbf{S}_j\mathbf{P}_{ij}^2\cdot\mathrm{diag}\left[\tilde{\rho}_{j1},\tilde{\rho}_{j2},\ldots,\tilde{\rho}_{jK}\right]\cdot\mathbf{S}_j^H\right). \tag{5.40}$$

(iii)Asymptotic Rate

$$\tilde{R}_{ik,\infty} = \log_2\left(1 + \frac{\mu_{ik}^2\tilde{\rho}_{ik}}{\beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\cdot\mathbf{F}_i(\underline{\tilde{\rho}})\cdot\mathbf{D}_i^{-1}\mathbf{s}_{ik} - \mu_{ik}^2\tilde{\rho}_{ik}}\right),\ \text{when } M\to\infty. \tag{5.41}$$

The derivations of the above three rate expressions are relegated to Appendix C. The closed-form rate is used to approximate the ergodic rate, while The asymptotic rate is a simplification of the closed-form rate when the number of antennas M at each BS is huge.

5.2.5 Rate-Aware Setting of MMSE Weights

This section aims to find a set of MSE weights w_{ik} in (5.24) that account for data rates. We use the asymptotic rate in (5.41) throughout the discussion.

We first explore the relation between MSE minimization and rate maximization. The primary idea here is to rewrite the rate expression (5.41) in a weighted MSE form. By the Lagrangian

dual transform in Theorem 8, the weighted sum rates maximization, with $\eta_{ik} \geq 0$ being the rate weight of user (i,k), is recast into a new form:

$$\underset{\underline{\mathbf{S}}}{\text{maximize}} \quad \sum_{(i,k)} \eta_{ik} \tilde{R}_{ik,\infty} \iff \underset{\underline{\mathbf{S}},\underline{\gamma}}{\text{maximize}} \quad \sum_{(i,k)} \eta_{ik} T_{ik}, \tag{5.42}$$

where

$$T_{ik} = \log(1+\gamma_{ik}) - \gamma_{ik} + \frac{(1+\gamma_{ik})\mu_{ik}^2 \tilde{\rho}_{ik}}{\beta_{i,ik}^2 \mathbf{s}_{ik}^H \mathbf{D}_i^{-1} \cdot \mathbf{F}_i\big(\underline{\tilde{\rho}}\big) \cdot \mathbf{D}_i^{-1} \mathbf{s}_{ik}}. \tag{5.43}$$

The auxiliary variable γ_{ik} can be interpreted as the signal-to-interference-and-noise ratio (SINR) of user (i,k). It turns out that the optimal γ_{ik} in (5.42) coincides with the real SINR in (5.41).

To make it tractable, we further assume that the data signal is stronger than any individual interfering signal, *i.e.*,

$$\beta_{i,ik}\tilde{\rho}_{ik} \geq \beta_{i,j\ell}\tilde{\rho}_{j\ell}, \ \forall (i,k) \text{ and } (j,\ell). \tag{5.44}$$

This is a reasonable assumption for a massive MIMO system with proper power control. Now, assume the use of the channel inversion power control [72], the desired signals received at each particular BS would be of the same strength, *i.e.*,

$$\beta_{i,ik}\tilde{\rho}_{ik} = \varrho_i, \text{ for some } \varrho_i \geq 0. \tag{5.45}$$

We then obtain an upper bound on $\mathbf{F}_i\big(\underline{\tilde{\rho}}\big)$ as

$$\begin{aligned} \mathbf{F}_i\big(\underline{\tilde{\rho}}\big) &= \sum_{j=1}^{L} \Big(\mathbf{S}_j \mathbf{P}_{ij}^2 \cdot \mathrm{diag}\big[\tilde{\rho}_{j1}, \tilde{\rho}_{j2}, \ldots, \tilde{\rho}_{jK}\big] \cdot \mathbf{S}_j^H \Big) \\ &\preceq \varrho_i \sum_{j=1}^{L} \Big(\mathbf{S}_j \mathbf{P}_{ij} \mathbf{S}_j^H \Big) \\ &= \varrho_i \mathbf{D}_i, \end{aligned} \tag{5.46}$$

which further leads to a lower bound on T_{ik}:

$$\begin{aligned} T_{ik} &\geq \log(1+\gamma_{ik}) - \gamma_{ik} + \frac{(1+\gamma_{ik})\mu_{ik}^2 \tilde{\rho}_{ik}}{\varrho_i \beta_{i,ik}^2 \mathbf{s}_{ik}^H \mathbf{D}_i^{-1} \mathbf{s}_{ik}} \\ &= \log(1+\gamma_{ik}) - \gamma_{ik} + \left(\frac{1+\gamma_{ik}}{\beta_{i,ik}} \right) \mu_{ik}. \end{aligned} \tag{5.47}$$

If the optimal auxiliary variables $\gamma_{ik}^\star$ are already determined, then the weighted sum-rate maximization problem in (5.42) can be convert to

$$\underset{\underline{\mathbf{S}}}{\text{maximize}} \quad \sum_{(i,k)} \eta_{ik} \left(\frac{1+\gamma_{ik}^\star}{\beta_{i,ik}} \right) \mu_{ik} \tag{5.48}$$

by using the lower bound (5.47) to approximate T_{ik}. Contrasting (5.25) and (5.48) gives the following strategy for setting the appropriate MSE weights:

$$w_{ik} = \eta_{ik}\left(\frac{1+\gamma_{ik}^{\star}}{\beta_{i,ik}}\right). \tag{5.49}$$

Now, the auxiliary variable $\gamma_{ik}^{\star}$, which represents the optimal SINR of user (i,k), is unknown *a priori* in general. To resolve this issue, we suggest some heuristic methods, e.g., (i) set $\gamma_{ik}^{\star}$ to some target SINR; (ii) update $\gamma_{ik}^{\star}$ iteratively with $\underline{\mathbf{S}}$.

We further argue that in many cases, the exact $\gamma_{ik}^{\star}$ is not required. As shown in [72], if the pilot contamination has been suppressed effectively, then the users from the same cell would achieve similar data rates under the *channel inversion power control* $\tilde{\rho}_{ik} = \delta/\beta_{i,ik}$ for some positive constant $\delta > 0$. We further argue that this similarity in achievable rate holds throughout the massive MIMO system, *i.e.*, $\gamma_{ik}^{\star} \approx \gamma_{j\ell}^{\star}$, for any (i,j,k,ℓ), so long as every cell has a similar setup (e.g., cell size, channel condition, and user distribution). In this case, the MSE weight in (5.49) is equivalent to

$$w_{ik} = \frac{\eta_{ik}}{\beta_{i,ik}}. \tag{5.50}$$

Hence, the normalized MMSE scheme with $w_{ik} = 1/\beta_{i,ik}$ as suggested in [70] is suitable for maximizing the sum rates under the channel inversion power control.

5.2.6 Numerical Results

We validate the performance of the proposed method in a 7-cell wrapped-around network. Each cell consists of a 100-antenna BS located at the center and 9 single-antenna user terminals uniformly distributed in a hexagonal area. The BS-to-BS distance is 1000 meters. Let $\tau = 16$ and let $\rho_{ik} = 1$. Following [66, 67], we assume that the background noise is negligible and that $\beta_{j,ik} = \varphi_{j,ik}/(d_{j,ik})^3$ where $\varphi_{j,ik}$ is an i.i.d. log-normal random variable according to $\mathcal{N}(0,\zeta^2)$ with $\zeta = 8$ dB and $d_{j,ik}$ is the distance between user (i,k) and BS i. In addition to the GSRTM algorithm with a random dictionary (see [67]), we introduce two baseline methods as follows:

- *Orthogonal Method:* Pick nine out of sixteen fixed orthogonal pilots randomly per cell.
- *Random Method:* Generate the i.i.d. pilot symbols according to Gaussian distribution.

The orthogonal method is used to initialize CNPD. For data signals, we adopt the channel inversion power control in [72], *i.e.*, $\tilde{\rho}_{ik} = \delta/\beta_{i,ik}$ and we set $\delta = 1$ without loss of generality.

Fig. 5.9 compares the sum of MSEs for the various methods. According to the figure, the coordinated approach in CNPD reduces the sum of MSEs sharply as compared to the conventional orthogonal method. Furthermore, around 75% of the sum-MSE reduction is obtained after just 10 iterations. It can also be seen that the discrete pilot strategy already improves

upon the baseline methods and GSRTM, albeit not by as much as the infinite precision coordinated approach. Fig. 5.10 takes a closer look at the CDF of the MSE; the coordinated approach outperforms all the other techniques by giving the smallest MSE in all the percentiles.

We now consider minimizing the sum of weighted MSEs across the network. Because the absolute value of MSE is proportional to the channel magnitude, weighting MSEs equally would give preference to the users with strong channels. To provide some measure of fairness, a possible heuristic [70] is to weight the MSEs by $w_{ik} = 1/\beta_{i,ik}$.

Fig. 5.11 shows a scatter plot of MSE vs. channel strength for this weighted coordinated approach as compared to the sum-MSE version of CNPD and the orthogonal method. Although the sum-MSE coordinated approach has considerable advantage over the orthogonal method in minimizing the overall MSEs as shown in the previous results, its performance in the weak-channel region (e.g., when $\beta_{i,ik} < -75$ dB) is close to or even slightly worse than that of the orthogonal method as shown in Fig. 5.11. The reason is that using the sum of MSEs as the objective does not take into account the difference in channel strengths among the users, while the weighted coordinated approach is able to improve the MSE for the cell-edge users (which are more vulnerable to pilot contamination) at a slight cost to the cell-center users. Indeed, the weighted coordinated approach is inferior to the orthogonal method when the channel strength is very strong ($\beta_{i,ik} > -68$ dB), but only a very small portion of users have such strong channels. Thus, there is an overall benefit for the weighted coordinated approach.

Furthermore, we compare the data rates achieved by the different pilot strategies. We use (5.37) to compute the instantaneous ergodic rate for each of the algorithms, and further use (5.38) to obtain another achievable rate for the weighted CNPD. Since the spectrum bandwidth is normalized in simulations, we also refer to data rate as spectral efficiency. Fig. 5.12 shows the CDF of data rates. As shown in the figure, the weighted CNPD outperforms the other methods significantly, especially in the low-rate regime. For instance, at the 10th percentile point, the spectral efficiency of the weighted CNPD is at most five times higher than that of any other algorithm. Observe also that the proposed closed-form rate is close to the instantaneous ergodic rate. Note that for the weighted CNPD, these two types of achievable rates both have the cumulative distribution curved as staircase. This is because under the channel inversion power control, the users in the same cell ought to have similar SINRs if the pilot contamination is sufficiently low, and then the cumulative distribution of data rates have 7 "jumps" across the 7 cells; this observation agrees with the conclusion of [72]. This staircase phenomenon however does not occur to the other methods because their channel estimations are not as accurate. Observe also that the orthogonal method achieves higher throughput than the random method and GSRTM, even though it has the worst performance in minimizing the sum of MSEs according to Fig. 5.9. Hence, in terms of the data rate objective, the sum of MSEs may not be a suitable metric for pilot design. Finally, we comment that the MSE weights η_{ik} are set as $1/\beta_{i,ik}$. It possible to use (5.49) then iteratively update γ_{ik} in setting the weights, but such a strategy does not give appreciable further rate improvement.

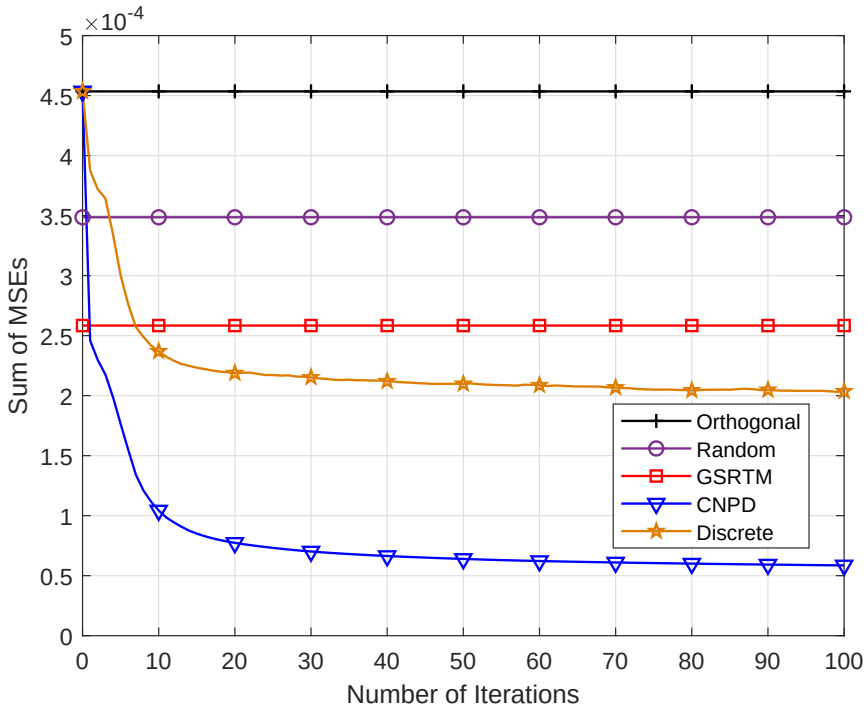


Figure 5.9: Sum of MSEs after each iteration.

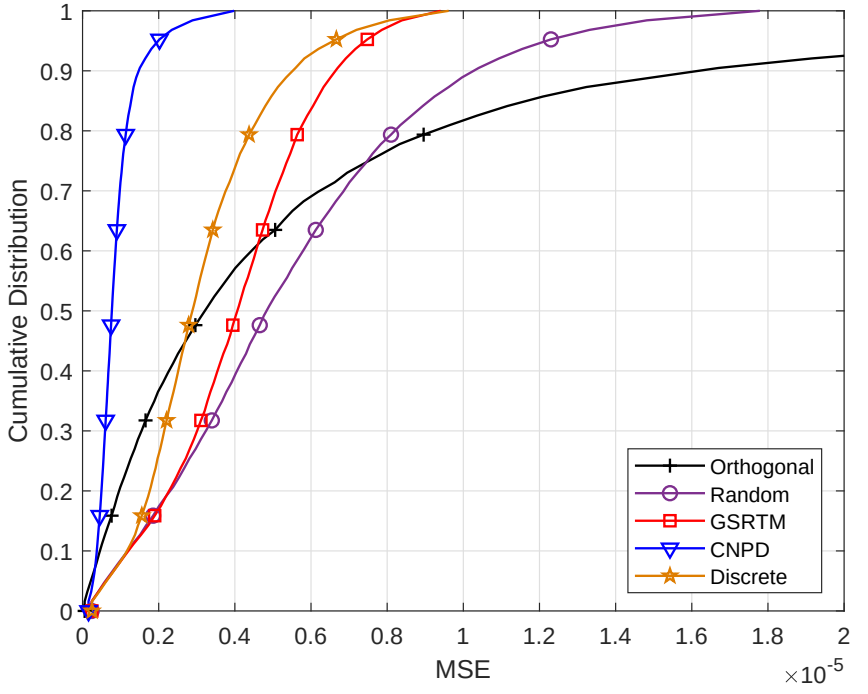


Figure 5.10: Cumulative distribution of MSEs.

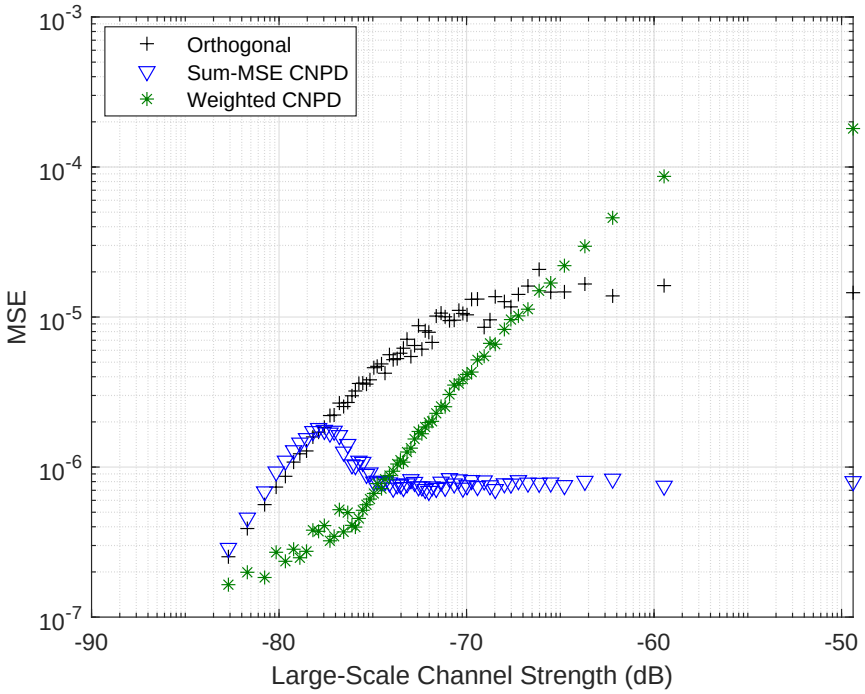


Figure 5.11: MSE vs. Large-scale channel strength $\beta_{i,ik}$.

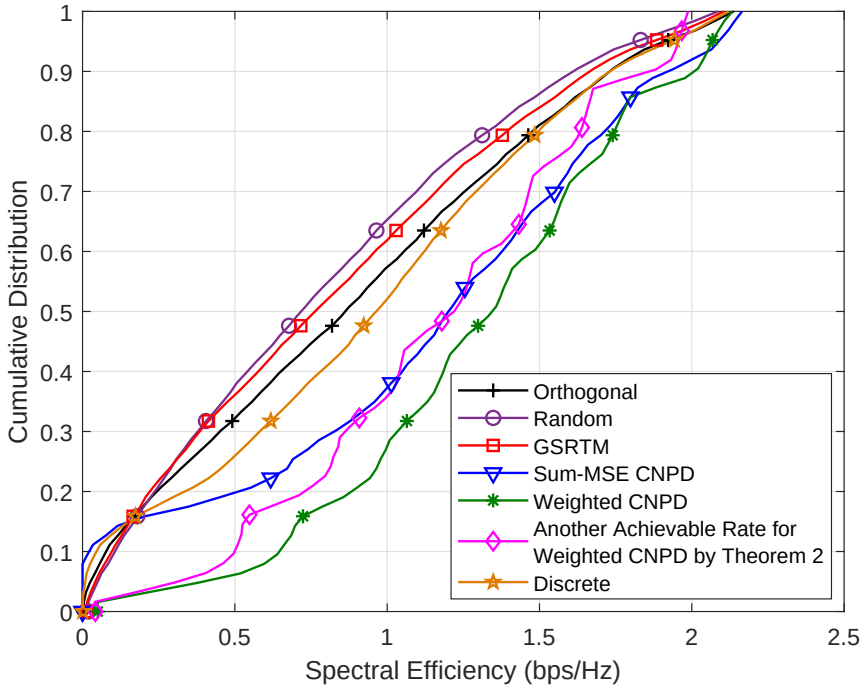


Figure 5.12: Cumulative distribution of data rates.

5.3 Summary

This chapter provides two typical examples of utilizing the matrix FP in the communication system design. The first example is about interference-aware spectrum sharing in wireless D2D networks. The problem setup differs from the joint scheduling and beamforming in Chapter 4 in that it assumes multi-layer transmissions on the same link and thus the SINR term has a matrix form. This work proposes the so-called FPLinQ strategy to coordinate the scheduling decisions along with beamforming and power control across the wireless D2D links. The key step is to treat the weighted sum-rate maximization as a matrix FP problem and to use a sequence of matrix FP transforms to allow iterative optimization of scheduling and beamforming. In contrast to the existing methods, *i.e.*, FlashLinQ, ITLinQ, and ITLinQ+, the proposed method does not involve tuning of design parameters and does not suffer from the premature turning-off problem. The second example studies the uplink pilot design for massive MIMO and utilizes the matrix FP differently. Instead of the conventional orthogonal pilot sequences, we advocate using the nonorthogonal pilot sequences to mitigate pilot contamination in a multi-cell massive MIMO network. Based on the matrix FP, the proposed algorithm optimizes the pilots iteratively in closed form, guaranteeing a monotonic reduction of the sum of weighted MSEs of channel estimation throughout the network. Furthermore, we show a relation between rate maximization and MSE minimization whereby the MSE weights can be properly chosen according to the data rate objective.

Chapter 6

Conclusion

This thesis is intended to provide a unifying treatment of FP in communication system design. While a series of recent works focus on using the classic Dinkelbach's method for the single-ratio problem such as the energy efficiency maximization, we propose a new FP technique named the quadratic transform that is valid for a broad range of multi-ratio problems, along with a Lagrangian dual transform devised for the logarithmic ratio problems. Furthermore, we generalize the conventional scalar FP to a multidimensional space wherein the fractional term is of a matrix form. A theoretical insight is that the proposed FP technique amounts to constructing the so-called surrogate functions from an MM perspective. Equipped with these new tools, we can examine many issues of communication system design in a new way, most of which have not previously been examined from an FP perspective. The application subjects considered in the thesis include the continuous optimization problems, discrete optimization problems, and the matrix optimization problems. The continuous case is concerned with some typical applications of FP for optimizing continuous variables like transmit powers and beamforming vectors in SINR. We choose the quadratic transform over the classic technique because of the multiple ratios nested in logarithm. Different ways of using FP is discussed. In particular, the proposed closed-form power control algorithm leads to a fixed-point iteration with provable convergence, whereas the existing ones in the prior works cannot guarantee convergence in general. The subsequent examples in the discrete part are less straightforward. These applications of FP for uplink scheduling build on a novel idea of using the quadratic transform in conjunction with the Lagrangian dual transform to recast the highly complicated integer program into a weighted bipartite matching. A remarkable fact about this approach is that it encompasses the well-known WMMSE algorithm. We show that WMMSE can actually be recognized as a particular way of ratio decoupling, and yet the way we advocate is more suited for discrete optimization. Next comes the matrix part that further extends the continuous and discrete optimizations to higher dimensions by assuming matrix-form ratios. The two application examples of this part both involve multiple antennas: (i) Joint link scheduling and beamforming for multi-data-stream transmission in a D2D network; (ii) nonorthogonal pilot sequence design for massive MIMO.

Appendices

Appendix A

Uniqueness of Quadratic Transform

We aim to show that the form of $g(\mathbf{x}, y)$ in (2.9) is necessary and sufficient when C4 is strengthened to require that $\partial^2 g/\partial y^2$ is independent of y. First, under the strengthened C4 and by C1, function g must be of the form:

$$g(\mathbf{x}, y) = f(A(\mathbf{x}))(\alpha_2 y^2 + \alpha_1 y + \alpha_0) + h(B(\mathbf{x}))(\beta_2 y^2 + \beta_1 y + \beta_0) \tag{A.1}$$

for some parameters α_i and β_i such that

$$\frac{\partial^2 g(\mathbf{x}, y)}{\partial y^2} = 2\alpha_2 f(A(\mathbf{x})) + 2\beta_2 h(B(\mathbf{x})) \leq 0. \tag{A.2}$$

For ease of notation, we omit the function arguments of $A(\mathbf{x})$ and $B(\mathbf{x})$ in the rest of the proof. First, note that $\partial^2 g(\mathbf{x}, y)/\partial y^2$ cannot be zero, as otherwise $\max_y g(\mathbf{x}, y) = \infty$ and thus C3 cannot be satisfied. Given a particular $\mathbf{x}$, the maximum value of $g(\mathbf{x}, y)$ over y can now be obtained in closed form as

$$\max_y g(\mathbf{x}, y) = \alpha_0 f(A) + \beta_0 h(B) - \frac{(\alpha_1 f(A) + \beta_1 h(B))^2}{4(\alpha_2 f(A) + \beta_2 h(B))}. \tag{A.3}$$

As required by C3, we must have $\max_y g(\mathbf{x}, y) = A/B$. One way to satisfy this relation is to have $\alpha_0 = 0, \beta_0 = 0, \alpha_1 = 2, \beta_1 = 0, \alpha_2 = 0, \beta_2 = 1$, $f(A) = \sqrt{A}$, and $h(B) = B$. This gives the proposed quadratic transform (2.8). The remainder of the proof aims to show that a more general form of this solution (2.9) is the unique solution satisfying the above.

The main idea is to determine functions f and h as well as parameters α_i and β_i by substituting different (A, B) pairs in (A.3). First, put $A = 0$ (so $A(\mathbf{x})$ is a zero constant function) then $\max_y g = A/B = 0$ for any B, *i.e.*,

$$(4\beta_0\beta_2 - \beta_1^2)h^2(B) + (4\alpha_2\beta_0 + 4\alpha_0\beta_2 - 2\alpha_1\beta_1)f(0)h(B) + (4\alpha_0\alpha_2 f^2(0) - \alpha_1^2 f^2(0)) = 0. \tag{A.4}$$

For this to hold for any B, we must have

$$4\beta_0\beta_2 - \beta_1^2 = 0. \tag{A.5}$$

In this case, the expression (A.3) reduces to

$$\max_y g(\mathbf{x}, y) = \frac{C}{D}, \tag{A.6}$$

where

$$C = (4\alpha_0\alpha_2 - \alpha_1^2)f^2(A) + (4\alpha_0\beta_2 + 4\alpha_2\beta_0 - 2\alpha_1\beta_1)f(A)h(B) \tag{A.7}$$

and

$$D = 4(\alpha_2 f(A) + \beta_2 h(B)). \tag{A.8}$$

Second, consider the case that $B \to 0_+$, then $\max_y g(\mathbf{x}, y) = A/B = \infty$ for any $A \neq 0$. For this to happen, we need $D \to 0$ for any A, whenever $B \to 0_+$. This means that the first term in D, which is a function of A only, must be zero, or

$$\alpha_2 = 0. \tag{A.9}$$

Third, consider the case that $A \to 0_+$, then $\max_y g(\mathbf{x}, y) = A/B = 0$ for any B. For this to happen, we need $C \to 0$ for any B, whenever $A \to 0_+$. This means that the second term in C, which is a function of B must be zero. Since $f(A)$ cannot be a constant zero, we must have

$$4\alpha_0\beta_2 + 4\alpha_2\beta_0 - 2\alpha_1\beta_1 = 4\alpha_0\beta_2 - 2\alpha_1\beta_1 = 0. \tag{A.10}$$

The $\max_y g(\mathbf{x}, y)$ expression now becomes

$$\max_y g(\mathbf{x}, y) = -\frac{\alpha_1^2 f^2(A)}{4\beta_2 h(B)}. \tag{A.11}$$

It can be readily seen that for it to be equal to A/B, we must have

$$f(A) = s_1\sqrt{A} \tag{A.12}$$

and

$$h(B) = s_2 B \tag{A.13}$$

for some nonzero (s_1, s_2) such that

$$-\alpha_1^2 s_1^2 = 4\beta_2 s_2. \tag{A.14}$$

Summarizing, $g(\mathbf{x}, y)$ must have this form:

$$g(\mathbf{x}, y) = s_1(\alpha_1 y + \alpha_0)\sqrt{A(\mathbf{x})} + s_2(\beta_2 y^2 + \beta_1 y + \beta_0)B(\mathbf{x}) \tag{A.15}$$

subject to (A.5), (A.10) and (A.14). Using (A.5), (A.10) and (A.14), *i.e.*,

$$\begin{cases} \beta_1^2 = 4\beta_0\beta_2 \\ 2\alpha_0\beta_2 = \alpha_1\beta_1 \\ -\alpha_1^2 s_1^2 = 4\beta_2 s_2, \end{cases} \tag{A.16}$$

we obtain

$$\beta_2 = -\frac{\alpha_1^2 s_1^2}{4s_2}, \quad \beta_1 = -\frac{\alpha_1\alpha_0 s_1^2}{2s_2}, \quad \beta_0 = -\frac{\alpha_0^2 s_1^2}{4s_2}. \tag{A.17}$$

With the above identities substituted in (A.1) to get rid of β_i's, the reformulation $g(\mathbf{x}, y)$ becomes

$$g(\mathbf{x}, y) = s_1(\alpha_1 y + \alpha_0)\sqrt{A(\mathbf{x})} - \frac{s_1^2(\alpha_1 y + \alpha_0)^2}{4}B(\mathbf{x}). \tag{A.18}$$

The above form of $g(\mathbf{x}, y)$ can be rewritten as (2.9) by defining two new parameters: $t_1 = s_1\alpha_1/2$ and $t_2 = s_1\alpha_0/2$. Finally, we note that $g(\mathbf{x}, y)$ in (2.9) satisfies the strengthened C1-C4 when $t_1 \neq 0$. This form of $g(\mathbf{x}, y)$ is therefore necessary and sufficient for this set of conditions.

Appendix B

Pseudoconvex Function

The definition of pseudoconvex function is presented here:

Definition 3. A differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be pseudoconvex on the convex compact constraint set $\mathcal{C}$ if

$$f(\mathbf{x}) < f(\mathbf{y}) \text{ implies } \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) < 0, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{C}. \tag{B.1}$$

Moreover, $-f(\mathbf{x})$ is said to be pseudoconcave provided that $f(\mathbf{x})$ is pseudoconvex.

In general, the above pseudoconvex condition is less strict than the convex condition, and yet stricter than the quasiconvex condition, *i.e.*,

$$\text{Convex Functions} \subset \text{Pseudoconvex Functions} \subset \text{Quasiconvex Functions}.$$

From an optimization perspective, the following critical property of the pseudoconvex function, first shown in [73], is what distinguishes it from the quasiconvex function:

Theorem 15. *Some point $\mathbf{x}^\star$ is a local minimum of the pseudoconvex function $f(\mathbf{x})$ if and only if $\mathbf{x}^\star$ is a stationary point.*

Roughly speaking, the gradient of some pseudoconvex function would not vanish unless at the local (or global) optimum. This is in contrast to the quasiconvex function whose stationary point is not necessarily a local (or global) optimum. Fig. B.1 displayed on the next page gives an example to illustrate this point.

Proposition 14. *A concave-convex single-ratio objective function $A(\mathbf{x})/B(\mathbf{x})$ is pseudoconcave.*

Proof. Recall that the concave-convex condition implies that $A(\mathbf{x})$ is a concave function while $B(\mathbf{x})$ is a convex function. Hence, given any two points $\mathbf{x}_1, \mathbf{x}_2$, we have

$$A(\mathbf{x}_1) \leq A(\mathbf{x}_2) + \nabla A(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2), \tag{B.2a}$$

$$B(\mathbf{x}_1) \geq B(\mathbf{x}_2) + \nabla B(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2). \tag{B.2b}$$

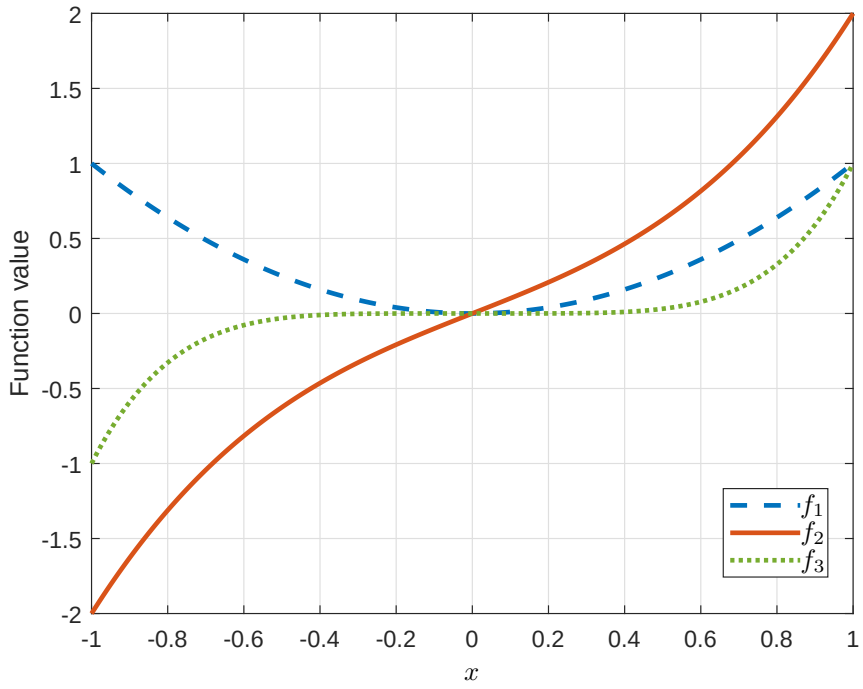


Figure B.1: Convex function $f_1(x) = x^2$, pseudoconvex function $f_2(x) = x+x^3$, and quasiconvex function $f_3(x) = x^5$. Observe that $(0, 0)$ is a stationary point of f_3 but not its local minimum, namely inflection point.

Without loss of generality, let us assume that

$$\frac{A(\mathbf{x}_1)}{B(\mathbf{x}_1)} \leq \frac{A(\mathbf{x}_2)}{B(\mathbf{x}_2)}. \tag{B.3}$$

We then have

$$\left(\nabla \frac{A(\mathbf{x}_2)}{B(\mathbf{x}_2)}\right)^\top (\mathbf{x}_1 - \mathbf{x}_2) = \left(\frac{\nabla A(\mathbf{x}_2) \cdot B(\mathbf{x}_2) - A(\mathbf{x}_2) \cdot \nabla B(\mathbf{x}_2)}{B^2(\mathbf{x}_2)}\right)^\top (\mathbf{x}_1 - \mathbf{x}_2) \tag{B.4a}$$

$$\overset{(a)}{\geq} \frac{B(\mathbf{x}_2)(A(\mathbf{x}_1) - A(\mathbf{x}_2)) + A(\mathbf{x}_2)(B(\mathbf{x}_2) - B(\mathbf{x}_1))}{B^2(\mathbf{x}_2)} \tag{B.4b}$$

$$= \frac{B(\mathbf{x}_2)A(\mathbf{x}_1) - A(\mathbf{x}_2)B(\mathbf{x}_1)}{B^2(\mathbf{x}_2)} \tag{B.4c}$$

$$\overset{(b)}{>} 0, \tag{B.4d}$$

where (a) follows by (B.2) and (b) follows by (B.3). The pseudoconcavity is thus verified. □

Appendix C

Data Rates of Massive MIMO with Nonorthogonal Pilots

The conventional rate expression for massive MIMO does not apply to our case in Section 5.2 because of the nonorthogonal pilots. Our achievability analysis rests on two special assumptions. First, the data signal symbol x_{ik} of each user (i,k) has an i.i.d. Gaussian distribution $\mathcal{CN}(0,1)$. Second, each BS i multiplies the received signal of user (i,k) with the complex conjugate of its channel estimate $\hat{\mathbf{h}}_{i,ik}$, namely maximum-ratio combining (MRC). Let $\tilde{\rho}_{ik}$ be the transmit power of the data signal of user (i,k), so the received signal at the target BS i after MRC is

$$\begin{aligned}\tilde{v}_{ik} &= \hat{\mathbf{h}}_{i,ik}^H \left(\sum_{(j,\ell)} \sqrt{\tilde{\rho}_{j\ell}} \mathbf{h}_{i,j\ell} x_{j\ell} + \tilde{\mathbf{z}}_{ik} \right) \\ &= \sqrt{\tilde{\rho}_{ik}} \|\hat{\mathbf{h}}_{i,ik}\|^2 x_{ik} + \hat{\mathbf{h}}_{i,ik}^H \left(\sum_{(j,\ell)\neq(i,k)} \sqrt{\tilde{\rho}_{j\ell}} \mathbf{h}_{i,j\ell} x_{j\ell} \right) \\ &\quad + \hat{\mathbf{h}}_{i,ik}^H \Big(\tilde{\mathbf{z}}_{ik} + \sqrt{\tilde{\rho}_{ik}} (\mathbf{h}_{i,ik} - \hat{\mathbf{h}}_{i,ik}) x_{ik} \Big), \end{aligned} \tag{C.1}$$

where $\tilde{\mathbf{z}}_{ik} \sim \mathcal{CN}(0, \sigma^2 \boldsymbol{I}_M)$ is an i.i.d. additive Gaussian noise of the data transmission phase.

Moreover, a simplified form of $\hat{\mathbf{h}}_{i,ik}$ is show in the following lemma.

Lemma 1. *The MMSE estimate of channel* $\mathbf{h}_{i,ik}$ *can be rewritten as*

$$\hat{\mathbf{h}}_{i,ik} = \beta_{i,ik} \mathbf{V}_i \bar{\mathbf{D}}_i^{-1} \bar{\mathbf{s}}_{ik}. \tag{C.2}$$

Proof. The right-hand side of (5.21) can be rewritten as

$$\begin{aligned} \big(\mathbf{P}_{ii}\mathbf{S}_i^H \otimes \boldsymbol{I}_M\big)\big(\mathbf{D}_i \otimes \boldsymbol{I}_M\big)^{-1} \mathrm{vec}\big(\mathbf{V}_i\big) &\overset{(a)}{=} \big(\mathbf{P}_{ii}\mathbf{S}_i^H \otimes \boldsymbol{I}_M\big)\big(\mathbf{D}_i^{-1} \otimes \boldsymbol{I}_M\big) \mathrm{vec}\big(\mathbf{V}_i\big) \\ &\overset{(b)}{=} \Big(\big(\mathbf{P}_{ii}\mathbf{S}_i^H \boldsymbol{I}_M\big) \otimes \boldsymbol{I}_M\Big) \cdot \mathrm{vec}\big(\mathbf{V}_i\big) \\ &\overset{(c)}{=} \mathrm{vec}\Big(\mathbf{V}_i \big(\mathbf{P}_{ii}\mathbf{S}_i^H \mathbf{D}_i^{-1}\big)^\top\Big), \end{aligned} \tag{C.3}$$

where (a) follows as $(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$, (b) follows since $(\mathbf{A}\otimes\mathbf{B})\cdot(\mathbf{A}'\otimes\mathbf{B}')=(\mathbf{A}\mathbf{A}')\otimes(\mathbf{B}\mathbf{B}')$, and (c) is a result of $(\mathbf{A}^{\top}\otimes\mathbf{B})\cdot\mathrm{vec}(\mathbf{C})=\mathrm{vec}(\mathbf{A}\mathbf{C}\mathbf{B})$. The identity in (C.2) is thus established. □

C.1 Instantaneous Ergodic Rate

By treating interference as noise, we obtain an achievable data rate as

$$R_{ik}=\mathbb{E}\left[\log_2\left(1+\frac{\left\|\hat{\mathbf{h}}_{i,ik}\right\|^4\tilde{\rho}_{ik}}{\sum_{(j,\ell)\neq(i,k)}\left|\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,j\ell}\right|^2\tilde{\rho}_{j\ell}+\sigma^2\left\|\hat{\mathbf{h}}_{i,ik}\right\|^2+\left|\hat{\mathbf{h}}_{i,ik}^H(\mathbf{h}_{i,ik}-\hat{\mathbf{h}}_{i,ik})\right|^2\tilde{\rho}_{ik}}\right)\right] \tag{C.4}$$

for each user (i,k), where the expectation is taken over the random fadings $\{\mathbf{h}_{i,j\ell},\forall(i,j,\ell)\}$ for a large number of coherence intervals with independent small-scale fading. The pilots affect R_{ik} through $\hat{\mathbf{h}}_{i,ik}$. Following [74], we refer to the above R_{ik} as the *instantaneous ergodic rate*. As pointed out in [74], the instantaneous ergodic rate is hard to interpret because of the expectation outside the logarithm, e.g., it is hard to tell whether an increment of $\tilde{\rho}_{ik}$ can enhance the sum rates or not.

C.2 Closed-Form Rate

We now derive a more interpretable achievable rate expression that does not include this expectation over time, that is

$$\tilde{R}_{ik}=\log_2\left(1+\frac{M^2\mu_{ik}^2\tilde{\rho}_{ik}}{M\mu_{ik}\sum_{(j,\ell)}\beta_{i,j\ell}\tilde{\rho}_{j\ell}+M^2\beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\cdot\mathbf{F}_i(\underline{\tilde{\rho}})\cdot\mathbf{D}_i^{-1}\mathbf{s}_{ik}+M\mu_{ik}\sigma^2-M^2\mu_{ik}^2\tilde{\rho}_{ik}}\right) \tag{C.5}$$

with

$$\mu_{ik}=\beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\mathbf{s}_{ik} \tag{C.6}$$

and

$$\mathbf{F}_i(\underline{\tilde{\rho}})=\sum_{j=1}^{L}\left(\mathbf{S}_j\mathbf{P}_{ij}^2\cdot\mathrm{diag}\left[\tilde{\rho}_{j1},\tilde{\rho}_{j2},\ldots,\tilde{\rho}_{jK}\right]\cdot\mathbf{S}_j^H\right). \tag{C.7}$$

We give a step by step procedure to derive the above rate expression.

Step 1: Artificial Channel Gain

Note that we have to average the rate expression in (5.37) because the effective channel gain $\|\hat{\mathbf{h}}_{i,ik}\|^2$ used in (C.1) depends on each realization of $\mathbf{h}_{i,ik}$ which varies over time. The main step of our approach is to replace $\|\hat{\mathbf{h}}_{i,ik}\|^2$ with an artificial channel gain

$$\tilde{h}_{i,ik}=\mathbb{E}\left[\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,ik}\right], \tag{C.26}$$

which reflects the average effective channel gain when the MMSE estimate $\hat{\mathbf{h}}_{i,ik}$ is used; observe that $\tilde{h}_{i,ik}$ is fixed over time. It can be shown that the randomness caused by the small-scale fading can all be encompassed in $\tilde{h}_{i,ik}$, so the resulting data rate is a deterministic function of the large-scale fading, as stated in the following theorem.

Inspired by the decoding method (under the orthogonal scheme) in [75], we rewrite the received signal after the MRC processing as

$$\tilde{v}_{ik} = \sqrt{\tilde{\rho}_{ik}}\tilde{h}_{i,ik}x_{ik} + \Delta_{ik}, \tag{C.27}$$

where

$$\Delta_{ik} = \sum_{(j,\ell)} \sqrt{\tilde{\rho}_{j\ell}}\hat{\mathbf{h}}_{i,ik}^H \mathbf{h}_{i,j\ell}x_{j\ell} + \hat{\mathbf{h}}_{i,ik}^H \tilde{\mathbf{z}}_{ik} - \sqrt{\tilde{\rho}_{ik}}\tilde{h}_{i,ik}x_{ik}. \tag{C.28}$$

In (C.27), we express the received signal $\tilde{v}_{ik}$ as if x_{ik} passed through the known channel $\tilde{h}_{i,ik}$. (Note that the BS knows $\tilde{h}_{i,ik}$, although it is not aware of every realization of $\mathbf{h}_{i,ik}$.). Further, due to the fact that $\mathbb{E}[(\tilde{h}_{ik}x_{ik})^H\Delta_{ik}] = 0$ with the expectation taken over $(\underline{\mathbf{H}}, \underline{\mathbf{Z}}, \underline{x}, \underline{\tilde{\mathbf{z}}})$, Δ_{ik} can be recognized as an uncorrelated noise added to the desired signal $\sqrt{\tilde{\rho}_{ik}}\tilde{h}_{i,ik}x_{ik}$. The early work [76] shows that the worst-case uncorrelated additive noise under a variance constraint has a Gaussian distribution. Henceforth, with respect to (C.27), a lower bound on the achievable data rate of user (i,k) is

$$\tilde{R}_{ik} = \log_2\left(1 + \frac{\tilde{\rho}_{ik}\mathbb{E}\big[|\tilde{h}_{i,ik}x_{ik}|^2\big]}{\mathbb{E}\big[|\Delta_{ik}|^2\big]}\right), \tag{C.29}$$

where the expectation of $|\tilde{h}_{i,ik}x_{ik}|^2$ is taken over x_{ik} while the expectation of $|\Delta_{ik}|^2$ is taken over $(\underline{\mathbf{H}}, \underline{\mathbf{Z}}, \underline{x}, \underline{\tilde{\mathbf{z}}})$. It remains to compute $\mathbb{E}\big[|\tilde{h}_{i,ik}x_{ik}|^2\big]$ and $\mathbb{E}[|\Delta_{ik}|^2]$, both of which depend on the MMSE channel estimation $\hat{\mathbf{h}}_{i,ik}$.

Step 2: Computation of $\mathbb{E}\big[|\tilde{h}_{i,ik}x_{ik}|^2\big]$ in (C.29)

Note that the virtual channel $\tilde{h}_{i,ik}$ is deterministic. Using the simplified form of $\hat{\mathbf{h}}_{i,ik}$ in Lemma 1, we can evaluate $\tilde{h}_{i,ik}$ as

$$\begin{aligned}
\tilde{h}_{i,ik} &= \mathbb{E}\big[\hat{\mathbf{h}}_{i,ik}^H \mathbf{h}_{i,ik}\big] \\
&= \mathbb{E}\big[\beta_{i,ik}\mathbf{s}_{ik}^\top \bar{\mathbf{D}}_i^{-1}\mathbf{V}_i^H \mathbf{h}_{i,ik}\big] \\
&= \mathbb{E}\left[\beta_{i,ik}\mathbf{s}_{ik}^\top \bar{\mathbf{D}}_i^{-1}\left(\sum_{j=1}^{L}\bar{\mathbf{S}}_j\mathbf{H}_{ij}^H + \mathbf{Z}_i^H\right)\mathbf{h}_{i,ik}\right] \\
&= M\beta_{i,ik}^2 \mathbf{s}_{ik}^\top \bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik} \\
&= M\beta_{i,ik}^2 \mathbf{s}_{ik}^H \mathbf{D}_i^{-1}\mathbf{s}_{ik} \\
&= M\mu_{ik},
\end{aligned} \tag{C.30}$$

where the expectation is taken over $\underline{\mathbf{H}}$. We then commutate the numerator of the SINR term in (C.29) by taking expectation over $\underline{x}$:

$$\mathbb{E}\big[|\tilde{h}_{i,ik}x_{ik}|^2\big] = |\tilde{h}_{i,ik}|^2 \cdot \mathbb{E}\big[|x_{ik}|^2\big] = M^2\mu_{ik}^2. \tag{C.31}$$

Step 3: Computation of $\mathbb{E}\big[|\Delta_{ik}|^2\big]$ in (C.29)

Recall that the interference-plus-noise strength $\mathbb{E}\big[|\Delta_{ik}|^2\big]$ in (C.29) is the expectation over the random variables $(\underline{\mathbf{H}}, \underline{\mathbf{Z}}, \underline{x}, \underline{\tilde{z}})$. We expand $\mathbb{E}\big[|\Delta_{ik}|^2\big]$ as follows:

$$\begin{aligned}\mathbb{E}\big[|\Delta_{ik}|^2\big] &= \mathbb{E}\left[\left|\sum_{(j,\ell)} \sqrt{\tilde{\rho}_{j\ell}}\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,j\ell}x_{j\ell} + \hat{\mathbf{h}}_{i,ik}^H\tilde{\mathbf{z}}_{ik} - \sqrt{\tilde{\rho}_{ik}}\tilde{h}_{i,ik}x_{ik}\right|^2\right] \\ &\overset{(a)}{=} \sum_{(j,\ell)} \tilde{\rho}_{j\ell}\mathbb{E}\big[|\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,j\ell}|^2\big] + \sigma^2\mathbb{E}\big[\|\hat{\mathbf{h}}_{i,ik}\|^2\big] - \tilde{\rho}_{ik}\mathbb{E}\big[|\tilde{h}_{i,ik}x_{ik}|^2\big], \end{aligned} \tag{C.32}$$

where (a) follows by taking expectation over $(\underline{x}, \underline{\tilde{z}})$. The first term of (C.32), we start with each expected channel strength after the MRC processing, that is

$$\begin{aligned}\mathbb{E}\big[|\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,j\ell}|^2\big] &= \mathbb{E}\big[\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,j\ell}\mathbf{h}_{i,j\ell}^H\hat{\mathbf{h}}_{i,ik}\big] \\ &= \mathbb{E}\big[\beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1}\mathbf{V}_i^H\mathbf{h}_{i,j\ell}\mathbf{h}_{i,j\ell}^H\mathbf{V}_i\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik}\big] \\ &= \beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1} \cdot \mathbb{E}\big[\mathbf{V}_i^H\mathbf{h}_{i,j\ell}\mathbf{h}_{i,j\ell}^H\mathbf{V}_i\big] \cdot \bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik} \\ &= \beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1} \cdot \mathbb{E}\left[\sum_{j'=1}^{L}\bar{\mathbf{S}}_{j'}\mathbf{H}_{ij'}^H\mathbf{h}_{i,j\ell}\mathbf{h}_{i,j\ell}^H\mathbf{H}_{ij'}\mathbf{S}_{j'}^\top + M\beta_{i,j\ell}\boldsymbol{I}_\tau\right] \cdot \bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik}, \end{aligned} \tag{C.33}$$

where the expectation part in the middle can be further computed as

$$\begin{aligned}&\mathbb{E}\left[\sum_{j'=1}^{L}\bar{\mathbf{S}}_{j'}\mathbf{H}_{ij'}^H\mathbf{h}_{i,j\ell}\mathbf{h}_{i,j\ell}^H\mathbf{H}_{ij'}\mathbf{S}_{j'}^\top + M\beta_{i,j\ell}\boldsymbol{I}_\tau\right] \\ &= M\beta_{i,j\ell}\left(\sum_{j'=1}^{L}\bar{\mathbf{S}}_{j'}\mathbf{P}_{ij'}\mathbf{S}_{j'}^\top + \boldsymbol{I}_\tau\right) + M^2\bar{\mathbf{S}}_j \cdot \mathrm{diag}\big[\underbrace{0,\ldots,0}_{(\ell-1)\text{ zeros}}, \beta_{i,j\ell}^2, \underbrace{0,\ldots,0}_{(K-\ell)\text{ zeros}}\big] \cdot \bar{\mathbf{S}}_j^\top \\ &= M\beta_{i,j\ell}\bar{\mathbf{D}}_i + M^2\bar{\mathbf{S}}_j \cdot \mathrm{diag}\big[0,\ldots,0,\beta_{i,j\ell}^2,0,\ldots,0\big] \cdot \mathbf{S}_j^\top. \end{aligned} \tag{C.34}$$

The substitution of (C.33) and (C.34) into the first term of (C.32) yields

$$\begin{aligned}&\sum_{(j,\ell)}\tilde{\rho}_{j\ell}\mathbb{E}\big[|\hat{\mathbf{h}}_{i,ik}^H\mathbf{h}_{i,j\ell}|^2\big] \\ &= \sum_{(j,\ell)} M\beta_{i,ik}^2\beta_{i,j\ell}\tilde{\rho}_{j\ell}\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik} + \sum_{j=1}^{L}\Big(M^2\beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{S}}_j\mathrm{diag}\big[\beta_{i,j1}^2\tilde{\rho}_{j1},\ldots,\beta_{i,jK}^2\tilde{\rho}_{jK}\big]\mathbf{S}_j^\top\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik}\Big)\end{aligned}$$

$$= \sum_{(j,\ell)} M\beta_{i,ik}^2 \beta_{i,j\ell}\tilde{\rho}_{j\ell} \mathbf{s}_{ik}^H \mathbf{D}_i^{-1}\mathbf{s}_{ik} + \sum_{j=1}^{L}\Big(M^2\beta_{i,ik}^2 \mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\mathbf{S}_j \mathrm{diag}\big[\beta_{i,j1}^2\tilde{\rho}_{j1},\ldots,\beta_{i,jK}^2\tilde{\rho}_{jK}\big]\mathbf{S}_j^H\mathbf{D}_i^{-1}\mathbf{s}_{ik}\Big)$$
$$= M\mu_{ik}\sum_{(j,\ell)}\beta_{i,j\ell}\tilde{\rho}_{j\ell} + M^2\beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\cdot\mathbf{F}_i\big(\underline{\tilde{\rho}}\big)\cdot\mathbf{D}_i^{-1}\mathbf{s}_{ik}. \tag{C.35}$$

The second term of (C.32) can be computed as

$$\begin{aligned}\sigma^2\mathbb{E}\big[\|\hat{\mathbf{h}}_{i,ik}\|^2\big] &= \sigma^2\mathbb{E}\big[\hat{\mathbf{h}}_{i,ik}^H\hat{\mathbf{h}}_{i,ik}\big]\\ &= \sigma^2\mathbb{E}\big[\beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1}\mathbf{V}_i^H\mathbf{V}_i\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik}\big]\\ &= \sigma^2 M\beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1}\left(\sum_{j=1}^{L}\bar{\mathbf{S}}_j\mathbf{P}_{ij}\mathbf{S}_j^\top + \sigma^2\boldsymbol{I}_\tau\right)\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik}\\ &= \sigma^2 M\beta_{i,ik}^2\mathbf{s}_{ik}^\top\bar{\mathbf{D}}_i^{-1}\bar{\mathbf{s}}_{ik}\\ &= M\mu_{ik}\sigma^2.\end{aligned} \tag{C.36}$$

Observe that the last term of (C.32) is already obtained from (C.31). Finally, combining (C.32), (C.36), and (C.35) along with (C.31) establishes the achievability of the proposed data rate in (5.38).

C.3 Asymptotic Closed-Form Rate

Recall that

$$\tilde{R}_{ik,\infty} = \log_2\left(1 + \frac{\mu_{ik}^2\tilde{\rho}_{ik}}{\beta_{i,ik}^2\mathbf{s}_{ik}^H\mathbf{D}_i^{-1}\cdot\mathbf{F}_i\big(\underline{\tilde{\rho}}\big)\cdot\mathbf{D}_i^{-1}\mathbf{s}_{ik} - \mu_{ik}^2\tilde{\rho}_{ik}}\right),\ \text{when } M\to\infty. \tag{C.37}$$

The nonorthogonality of pilots comes into (C.37) through $\mathbf{D}_i^{-1}\cdot\mathbf{F}_i\big(\underline{\tilde{\rho}}\big)\cdot\mathbf{D}_i^{-1}$ which would be a diagonal matrix if the pilots are orthogonal. Furthermore, (5.41) reduces to well-known results if $\tau \geq K$ and the same set of orthogonal pilots $\{\mathbf{s}_1^\perp, \mathbf{s}_2^\perp, \ldots, \mathbf{s}_K^\perp\}$ is reused in each cell. In this case, the above $\tilde{R}_{ik,\infty}$ reduces to

$$\tilde{R}_{ik,\infty} = \log_2\left(1 + \frac{\beta_{i,ik}^2\tilde{\rho}_{ik}}{\sum_{j\neq i}\beta_{i,\kappa(j,ik)}^2\tilde{\rho}_{\kappa(j,ik)}}\right), \tag{C.38}$$

where $\kappa(j,ik)$ outputs the index of the user in cell j assigned the same pilot sequence as user (i,k). We remark that the asymptotic rate in (C.38) is a well-known result in the massive MIMO literature [69, 77].

Bibliography

[1] J. von Neumann, "Über ein ökonomisches gleichgewichtssystem und eine verallgemeinerung des brouwerschen fixpunktsatzes," *Ergebnisse eines Mathematischen Kolloquiums*, vol. 8, pp. 73–83, 1937.

[2] S. Schaible, "Fractional programming," *Zeitschrift f ur Operations Research*, vol. 27, pp. 39–54, Oct. 1982.

[3] I. M. Stancu-Minasian, *Fractional Programming: Theory, Methods and Applications*, Kluwer Academic Publishers, 1992.

[4] E. B. Bajalinov, *Linear-Fractional Programming: Theory, Methods, Applications and Software*, Kluwer Academic Publishers, 2003.

[5] A. Charnes and W. W. Cooper, "Programming with linear fractional functionals," *Naval Research Logistics (NRL)*, vol. 9, no. 3, pp. 181–186, Dec. 1962.

[6] S. Schaible, "Parameter-free convex equivalent and dual programs of fractional programming problems," *Zeitschrift für Operations Research*, vol. 18, no. 5, pp. 187–196, Oct. 1974.

[7] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 133, no. 7, pp. 492–498, Mar. 1967.

[8] H. P. Benson, "On the global optimization of sums of linear fractional functions over a convex set," *J. Optim. Theory Appl.*, vol. 121, no. 1, pp. 19–39, 2004.

[9] T. Kuno, "A branch-and-bound algorithm for maximizing the sum of several linear ratios," *J. Global Optimization*, vol. 22, pp. 155–174, 2002.

[10] N. T. H. Phuong and H. Tuy, "A unified monotonic approach to generalized linear fractional programming," *J. Global Optimization*, vol. 22, pp. 229–259, 2003.

[11] J. G. Carlsson and J. Shi, "A linear relaxation algorithm for solving the sum-of-linear ratios problem with lower dimension," *Operations Research Lett.*, vol. 41, no. 4, pp. 381–389, 2013.

[12] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations Trends Commun. Inf. Theory*, vol. 11, no. 3, pp. 185–396, June 2015.

[13] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.

[14] A. Zappone, E. Björnson, L. Sanguinetti, and E. Jorswieck, "Globally optimal energy-efficient power control and receiver design in wireless networks," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2844–2859, June 2017.

[15] K. T. K. Cheung, S. Yang, and L. Hanzo, "Achieving maximum energy-efficiency in multi-relay OFDMA cellular networks: A fractional programming approach," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2746–2757, July 2013.

[16] J.-P. Crouzeix, "Algorithms for generalized fractional programming," *Mathematical Programming*, vol. 52, no. 1, pp. 191–207, May 1991.

[17] R. W. Freund and F. Jarre, "Solving the sum-of-ratios problem by an interior-point method," *J. Global Optimization*, vol. 19, no. 1, pp. 83–102, 2001.

[18] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allcoation in downlink multicell OFDMA networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 6, pp. 2835–2848, July 2012.

[19] H. Dahrouj, W. Yu, and T. Tang, "Power spectrum optimization for interference mitigation via iterative function evaluation," *EURASIP J. Wireless Commun. Netw.*, Aug. 2012.

[20] W. Yu, "Multiuser water-filling in the presence of crosstalk," in *Inf. Theory Appl. Workshop (ITA)*, Jan. 2007.

[21] K. Shen and W. Yu, "Load and interference aware joint cell association and user scheduling in uplink cellular networks," in *IEEE Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, July 2016.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[23] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, Mar. 2018.

[24] H. P. Benson, "Solving sum of ratios fractional programs via concave minimization," *J. Optim. Theory Appl.*, vol. 135, no. 1, pp. 1–17, 2007.

[25] J. Mairal, "Incremental majorization-minimization optimization with application to large-scale machine learning," *SIAM J. Optim.*, vol. 25, no. 2, pp. 829–855, 2015.

[26] S. Schaible, "Fractional programming. II, On Dinkelbach's algorithm," *Manage. Sci.*, vol. 22, no. 8, pp. 868–873, Apr. 1976.

[27] V. P. Sreedharan, "ϵ-subgradient projection algorithm," *J. Approximation Theory*, vol. 51, no. 1, pp. 27–46, Sept. 1987.

[28] T. T. Wu and K. Lange, "The MM alternative to EM," *Statistical Sci.*, vol. 25, no. 4, pp. 492–505, 2010.

[29] A. S. Lewis, "Derivatives of spectral functions," *Math. Oper. Res.*, vol. 21, no. 3, pp. 576–588, Aug. 1996.

[30] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[31] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[32] J. Song, P. Babu, and D. P. Palomar, "Optimization methods for designing sequences with low autocorrelation sidelobes," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3998–4009, Apr. 2015.

[33] K. Shen and W. Yu, "Fractional programming for communication systems—Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, Mar. 2018.

[34] L. P. Qian, Y. J. Zhang, and J. W. Huang, "MAPEL: Achieving global optimality for a non-convex wireless power control problem," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1553–1563, Mar. 2009.

[35] M. Chiang, "Geometric programming for communication systems," *Foundations Trends Commun. Inf. Theory*, vol. 2, no. 1, pp. 1–154, Aug. 2005.

[36] H. Boche and M. Schubert, "A general theory for SIR balancing," *EURASIP J. Wireless Commun. Netw.*, vol. 2006, no. 2, pp. 1–18, Apr. 2006.

[37] H. Boche and M. Schubert, "The structure of general interference functions and applications," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4980–4990, Nov. 2008.

[38] W. Yu, T. Kwon, and C. Shin, "Multicell coordination via joint scheduling, beamforming and power spectrum adaptation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 1–14, June 2013.

[39] J. Papandriopoulos and J. S. Evans, "SCALE: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3711–3724, July 2009.

[40] S. S. Christensen, R. Argawal, E. de Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 1–7, Dec. 2008.

[41] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[42] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Signal Process.*, vol. 11, no. 9, pp. 3292–3304, Sept. 2012.

[43] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-efficient power control: A look at 5G wireless technologies," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1668–1683, Apr. 2016.

[44] J. Xu and L. Qiu, "Energy efficiency optimization for MIMO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 690–701, Feb. 2013.

[45] K. Shen and W. Yu, "Interference management in full-duplex wireless cellular networks via fractional programming," in *IEEE Veh. Technol. Conf. (VTC)*, June 2018.

[46] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Trans. Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[47] J. Håstad, "Clique is hard to approximate within $n^{1-\epsilon}$," *Acta Mathematica*, vol. 182, pp. 105–142, 1996.

[48] D. P. Bertsekas, "The auction algorithm: A distributed relaxation method for the assignment problem," *Ann. Operations Research*, vol. 14, no. 1, pp. 105–123, Dec. 1988.

[49] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quart.*, vol. 2, no. 1, pp. 83–97, Mar. 1955.

[50] D. B. West, *Introduction to graph theory (2nd edition)*, Pearson, 2000.

[51] J. Schwartz, A. Steger, and A. Weißl, "Fast algorithms for weighted bipartite matching," in *Proc. Int. Conf. Experimental Efficient Algorithms*, 2005, pp. 476–487.

[52] W. Cui, K. Shen, and W. Yu, "Spatial deep learning for wireless scheduling," in *IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018.

[53] W. Cui, K. Shen, and W. Yu, "Spatial deep learning for wireless scheduling," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1248–1261, June 2019.

[54] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynmaic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[55] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communnications of ACM*, vol. 18, no. 9, pp. 509–517, Sept. 1975.

[56] G. Sharma, R. R. Mazumdar, and N. B. Shroff, "On the complexity of scheduling in wireless networks," in *ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, Sept. 2006, pp. 227–238.

[57] A. Goldsmith, *Wireless Communications*, Cambridge University Press, 2005.

[58] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.

[59] N. Naderializadeh and A. S. Avestimehr, "ITLinQ: A new approach for spectrum sharing in device-to-device communication systems," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1139–1151, June 2014.

[60] X. Yi and G. Caire, "Optimality of treating interference as noise: A combinatorial perspective," *IEEE Trans. Inf. Theory*, vol. 62, no. 8, pp. 4654–4673, June 2016.

[61] C. Geng, N. Naderializadeh, A.S. Avestimehr, and S.A. Jafar, "On the optimality of treating interference as noise," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1753–1767, Feb. 2015.

[62] K. Shen and W. Yu, "FPLinQ: A cooperative spectrum sharing strategy for device-to-device communications," in *IEEE Int. Symp. Inf. Theory (ISIT)*, June 2018, pp. 2323–2327.

[63] Z. Zhou and D. Guo, "1000-cell global spectrum management," in *ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, July 2017.

[64] Z. Zhou and D. Guo, "A centralized metropolitan-scale radio resource management scheme," *arXiv.org*, Aug. 2018. [Online]. Available: https://arxiv.org/pdf/1808.02582.pdf. [Accessed Aug. 09, 2018].

[65] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, July 2004.

[66] S. S. Ioushua and Y. C. Eldar, "Pilot contamination mitigation with reduced RF chains," in *IEEE Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, July 2017.

[67] S. S. Ioushua and Y. C. Eldar, "Pilot contamination mitigation with reduced RF chains," [Online]. Available: https://arxiv.org/abs/1801.05483, 2018.

[68] H. Al-Salihi, T. Van Chien, T. A. Le, and M. R. Nakhai, "A successive optimization approach to pilot design for multi-cell massive MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1086–1089, May 2018.

[69] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[70] T. E. Bogale and L. B. Le, "Pilot optimization and channel estimation for multiuser massive MIMO systems," in *Ann. Conf. Inf. Sci. Sys. (CISS)*, Mar. 2014.

[71] K. Shen, W. Yu, L. Zhao, and D. P. Palomar, "Optimization of MIMO device-to-device networks via matrix fractional programming: A minorization-maximization approach," *IEEE/ACM Trans. Netw.*, Nov 2019.

[72] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[73] O. L. Mangasarian, "Pseudo-convex functions," *Journal of the Society of Indstrial and Applied Mathematics Series A Control*, vol. 3, no. 2, pp. 281–290, 1965.

[74] T. Marzetta, E. G. Larsson, H. Yang, and H. Ngo, *Fundamentals of Massive MIMO*, Cambridge University Press, Cambridge, U.K., 2016.

[75] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[76] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[77] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov 2010.