On the MIMO Capacity with Multiple Power Constraints

Thuy M. Pham

Dissertation submitted in fulfilment of the requirements for candidature for degree of

Doctor of Philosophy

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

Maynooth University
Department of Electronic Engineering
CONNECT Centre

September 2019

Department Head

Prof. Dr. Ronan Farrell

Supevisors

Dr. Le-Nam Tran

Prof. Dr. Ronan Farrell

Department Head

Prof. Dr. Ronan Farrell

Referees

1. Prof. Osvaldo Simeone
King's College London

2. Dr. Arman Farhang
Maynooth University

Year of the graduation

2020

To my family

Contents

Abstract

The multiple-input multiple-output (MIMO) technology has become an essential element of modern communication systems e.g., 3G, 4G and massive MIMO technology has been recently standardized in 3GPP Rel-15 i.e., New Radio (NR) to enhance the spectral efficiency or the capacity of 5G networks. Given a digital communication system, a receiver will suffer from decoding errors if the transmission rate exceeds the capacity. Therefore, the capacity of a MIMO system is an important metric to characterize the system performance. More importantly, an efficient precoder design to achieve that capacity is of great interest.

This thesis is dedicated to this fundamental problem under multiple power constraints. From the theoretical perspective, capacity maximization is a classical problem. However efficient algorithms considering realistic scenarios or multiple power constraints, especially for massive MIMO application, are still sparse. In the thesis, the author has sought new methods of determining the capacity under two practical power constraints: 1) per-antenna power constraint (PAPC) 2) linear transmit covariance constraint (LTCC). In particular, the PAPC imposes an individual power limit on each power amplifier associated with a transmit antenna, thus is much more realistic than the traditional sum power constraint (SPC) in which all transmit antennas collaborate to satisfy a predefined total power budget. In many other practical scenarios, other power constraints can be imposed on a system, not necessarily to either SPC or PAPC. To this end, LTCCs are general enough to include those constraints. In both cases, we have proposed low-complexity approaches to the considered problems and the description of them is in the following.

For the problem of capacity maximization under PAPC, two closed-form low-complexity approaches have been developed for single-user MIMO and multi-user MIMO under different MIMO channels and precoding techniques. More specifically, the first approach is based on fixed-point-iteration to solve the problem directly in the broadcast channel (BC), whereas the other relies on alternating optimization (AO) together with successive convex optimization (SCA) to solve the equivalent problem in dual multiple access channel (MAC) domain. Interestingly, the latter approach is also applicable to the problem of computing capacity with LTCCs. For the special case of joint SPC and PAPC, we have also derived analytical solutions to this important problem. Last but not least, we have investigated the applications of machine learning to our capacity problems and presented some preliminary results.

Acknowledgments

First and foremost, I would like to give my special thanks to my supervisor, Prof. Ronan Farrell, for his tremendous support, valuable advice, comments and providing me with the financial support from Science Foundation Ireland (SFI). I am also grateful to my co-supervisor, Dr. Le-Nam Tran for his patient guidance during the last four years. I have gained a lot of knowledge and experience from his profound technical insights, visions and professional skills. He is not only a great supervisor but also a good friend with whom I do not hesitate to share my problems.

I would like to thank all my collaborators including Prof. Eryk Dutkiewicz, Dr. John Dooley, Dr. Holger Claussen, especially Dr. Diep Nguyen and Dr. Mark Flanagan for fruitful collaboration. Their critical comments and productive discussion have greatly improved the quality of the research. I am also thankful to Airrays GmBH for providing me with an internship, relatively short but valuable, which helps me appreciate the bridge between academia and industry. I also learned a lot of practical skills during my time at Airrays GmBH. As the moment of writing this dissertation, I still recall the warm support from my colleagues, especially my mentor Dr. Ulrich Walther.

My PhD journey would be extremely boring without my friends. I thank my labmates, my old and new friends, particularly the ones in Maynooth and Dublin including families of Nam and Dan, Lam and Thuy, Pavel, Ziming, to name but a few, for having fun together, sharing information, exchanging ideas and advising me. Additionally, thanks should go to Dr. Anh Huynh, Dr. Hien Ngo, Dr. Trung Ngo and Dr. Vinh Pham for counseling me on a number of problems and concerns. I also thank the secretaries of the Electronic Department, Registration Office, International Office, and Graduate Office for assisting me with administrative works. In addition, I would like to thank the committees of COST Action CA15221 and COST Action CA15104, especially Dr. Alison Farrell and Prof. Raquel Perez Leal for providing me opportunities to join a number of training schools and thank my language teachers in Maynooth and Dresden e.g., Rosaleen Gyves, Dr. Antoinette Mc Namara, for their helps. My sincere thanks also go to my landlord, Sean O'Malley, and my roommates for being kind to me.

I also would like to thank Prof. Osvaldo Simeone and Dr. Arman Farhang for taking time to participate in my viva and giving valuable comments and feedback on my thesis. Moreover, I am thankful to Dr. Martin Charlton for chairing the viva and thank the Electronic Department for arranging my defense.

Last but not least, I am deeply indebted to my family, both immediate and extended, especially my parents, Pham, Minh Hong and Nguyen, Thi Kim Loan and my two brothers Pham, Minh Quan and Pham, Minh Dan' families for their continued support and immense encouragement. Although they no longer assist me with academic knowledge, they do give me many helpful advice and comments based on their experience of life. Without support of my family, supervisors and friends, my PhD journey would be impossible, and I again express my deepest appreciation and thanks to you all.

Maynooth, 01.09.19

Nomenclature

3GPP	The 3rd Generation Partnership Project
ADMM	Alternating Direction Method of Multipliers
AEP	Asymptotic Equipartition Property
AI	Artificial Intelligence
AO	Alternating Optimization
aRRMSE	average Relative Root Mean Square Error
AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
CCP	Concave-Convex Procedure
CDF	Cumulative Distribution Function
CDI	Channel Distribution Information
CGP	Conjugate Gradient Projection
CSI	Channel State Information
DPC	Dirty Paper Coding
EVD	EigenValue Decomposition
FD	Feature Design
GDP	Gross Domestic Product
GP	Gradient Projection
IWF	Iterative Water-Filling
KKT	Karush-Kuhn-Tucker
LS	Least Square

LTCC	Linear Transmit Covariance Constraint
MAC	Multiple Access Channel
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARE	Mean Absolute Relative Error
MIMO	Multiple-Input Multiple-Ouput
MISO	Multiple-Input Single-Output
ML	Machine Learning
MMSE	Minimum Mean Square Error
MU-MIMO	Multi-User MIMO
NR	New Radio
OLS	Ordinary Least Square
PAPC	Per-Antenna Power Constraint
PCA	Principle Component Analysis
PCR	Principle Component Regression
PU	Primary User
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RMSPE	Root Mean Square Percentage Error
RRMSE	Relative Root Mean Square Error
SCA	Successive Convex Optimization
SIMO	Single-Input Multiple-Output
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
SPC	Sum Power Constraint

SRMax	Sum Rate Maximization
SU	Secondary User
SU-MIMO	Single-User MIMO
SVD	Singular Value Decomposition
SVR	Support Vector Regression
SZFDPC	Successive Zero-Forcing Dirty Paper Coding
WF	Water-Filling
WSRMax	Weighted Sum Rate Maximization
ZF	Zero-Forcing

List of Figures

Chapter 1

Introduction

In 1948, Shannon proved a fundamental result for modern communications systems in his breakthrough work: reliable communication is only guaranteed if and only if the data rate is less than a specified value, which he referred to as the channel capacity [5]. His pioneer information theoretical result has opened an important line of research in communication theory which aims to compute the capacity of communications systems. Early wireless networks were built on the single-input single-output (SISO) technology and the capacity of numerous SISO systems is well-studied. However, the SISO technology fails to accommodate the increasing demands of high data rate in modern communications networks. The solution to this challenge lies in the multiple-input multiple-output (MIMO) technology, which has been studied extensively for the last few decades. MIMO technology has proved its capability to boost the data rate of a system using the spatial resource and is an integral part of many current wireless networks [6–11].

Contrary to traditional communication systems, MIMO technology exploits multipath to increase the capacity of a system. This is surprisingly advantageous since multipath, which causes intersymbol interference at a receiver, was once considered as ‘the enemy’ of communication systems. In addition, the effect of fading, which is detrimental to the system performance, can be reduced, thus improving the reliability of the system. In particular, a transmitter sends multiple replicas of a signal over a fading channel so that a receiver can derive a good estimate of the original signal. Moreover, MIMO can not only support single-user systems (point-to-point) but also multi-user systems (point-to-multipoint) to improve the overall system performance.

In general, the capacity of MIMO systems is considered under a certain type of transmit power constraint. The sum power constraint (SPC) in which transmit antennas collaborate to satisfy a maximum power budget has drawn a lot of attention since this simple constraint, to a large extent, leads to computationally efficient algorithms. Of more practical relevance is the per-antenna power constraint (PAPC) since each transmit antenna has its own power amplifier and thus can be subject to different power constraints. In addition to SPC or PAPC, a system can also be subject to other power constraints which are altogether generalized as multiple linear transmit covariance constraints (LTCCs). Although MIMO technology has been studied and adopted for more than twenty years, the research on these mixed power constraints is still sparse despite their practical importance. Existing solutions based on high-complexity methods can be, in some particular cases, slightly modified to deal with these general capacity computation problems. However, such methods are not suitable for large-scale MIMO systems, which are the main driving force for the thesis.

The aim of this thesis is to develop efficient approaches to computing the MIMO capacity under various power constraints, including SPC, PAPC, LTCCs, and combination thereof. The computational efficiency is focused on massive MIMO. To achieve the goals we base the proposed solutions on some powerful mathematical programming frameworks: successive convex optimization (SCA), alternating optimization (AO) and fixed-point iteration. The purpose is to derive analytical solutions wherever possible. Throughout the thesis we mainly consider Gaussian MIMO channels and the channel state information is assumed to be perfectly known at both of the transmitter and receiver. The system models of interest range from single-user MIMO (SU-MIMO) to multi-user MIMO (MU-MIMO) to demonstrate the superiority of the proposed approaches.

This chapter lays the foundation to the topics considered in this thesis: Wireless communications and multiple-input multiple-output (MIMO) technology. In particular, we describe the fundamentals of a digital wireless communication system and highlight the performance boosted by the MIMO technology in Section 1.1. An overview of the thesis is presented in the next section followed by a list of publications and notation in Sections 1.3 and 1.4, respectively. The research background will be detailed in the next.

1.1. Motivation

In the following, we provide some fundamentals of two key elements of modern communication systems: Wireless digital communication system and multiple-input multiple-output technology.

1.1.1. Wireless Digital Communication Systems

A typical structure of a digital communication system is illustrated in Fig. 1.1. The information in the form of bits is transformed into bit streams by a source coder, then goes through the error control channel i.e., channel coding and modulation to transmit over a channel. The received signal is decoded at the receiver to reconstruct the original signal.

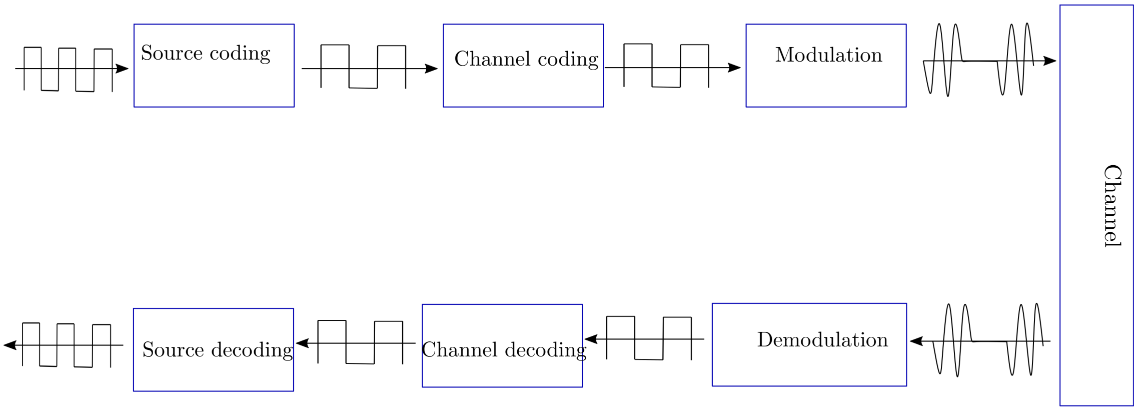


Figure 1.1.: An abstract illustration of a digital communication system.

1.1.1.1. Signal Generation

Digital signals can be transmitted without pre-processing but analog signals. Most of signals such as sound, voice etc. need to be converted into digital signals using sampling and quantization. An analog signal is first sampled following the Nyquist theorem, then the samples are quantized to obtain a finite number of levels, each of which can be easily mapped onto a sequence of bits.

In order to remove the redundancy, the sampled data is transferred to a *source coder*. The primary task of a source coder is to represent information with a minimum number of bits while preserving the specified quality of the original signal. At the receiver, source decoder reconstructs the decoded signal.

1.1.1.2. Error Control Channel

Being transmitted over wireless channels, a signal can be deteriorated due to signal impairments e.g., attenuation, distortion, or noise. To protect the signal against unwanted impairments, a process called *channel coding* is introduced to provide redundancy to the signal. Doing so, channel coding enables corrupted message to be corrected which in turn reduces probable errors in the system. Another necessary step to adapt the signal to the physical channel by means of *modulation*. In particular, the signal is converted into a proper form for physical transmission.

At the receiver, demodulation brings the signal back to baseband signal. For a simple SISO channel model, the received signal can be written as

$$y = hx + n \tag{1.1}$$

where y and x are the transmitted and received symbol, respectively and h and n denote the channel response and the noise, respectively. The signal is then fed into a channel decoder to produce a received bit stream which is expected to be identical to the one at the transmitter. The capacity of this time-invariant SISO channel or Additive White Gaussian Noise (AWGN) is given by

$$C = B\log(1 + \frac{\sigma_x^2|h|^2}{\sigma_n^2}) \tag{1.2}$$

where B is the bandwidth and $\sigma_x^2|h|^2$ and σ_n^2 are the signal power and noise power, respectively.

In the context of wireless medium, the propagation of the signal over the wireless channel is affected by the following factors:

- Shadowing: A signal can be attenuated through absorption, reflection, scattering, and diffraction due to obstacles in the signal path.

- Path loss: A signal can also be subject to path loss, which is the reduction in power density when it propagates.

- Multipath propagation: Multiple copies of an original signal can arrive at the receiver on different paths (multipath). The combined signal can be increased or decreased depending on the phase of these signals.

Based on the attenuation variation, fading can be classified as large-scale fading or small-scale fading.

- Large-scale fading: is evaluated by averaging the signal attenuation over a large area. Shadowing and path loss belong to this category since the attenuation usually occurs over a long distance. This type of fading is thus more relevant to cell planning process.

- Small-scale fading: refers to dramatic changes in signal amplitude and phase due to addition of multipath components. This fading has huge impact on the efficiency and reliability of the systems, is thus particularly of interest to design a reliable communication system.

1.1.2. MIMO Technology

In conventional SISO system, a high data rate can be achieved by increasing either transmit power or the bandwidth (cf. Eq. (1.2)). However, increasing the transmit power incurs battery and safety issues, whereas the bandwidth is precious and highly regulated. Tremendous research in the last decades has found that multiple-input multiple-output (MIMO) is the key technology to boost the data rate and to increase the reliability of wireless communication systems without increasing the transmit power or the bandwidth. More specifically, the spatial multiplexing technique enhances the throughput by exploiting multipath and the spatial diversity combats fading to improve the reliability of a communication link. In this section, we describe the key components of the MIMO technology including different deployments of multiple antennas, definitions of single- and multi-user MIMO, open- and closed-loop MIMO.

1.1.2.1. Receive Diversity

In this simple scenario, a transmitter is equipped with a single antenna and a receiver has multiple antennas (see Fig. 1.2). The receiver can find an estimate of the transmitted signal by either averaging the combined signals or selecting the signal with the highest signal-to-noise ratio (SNR).

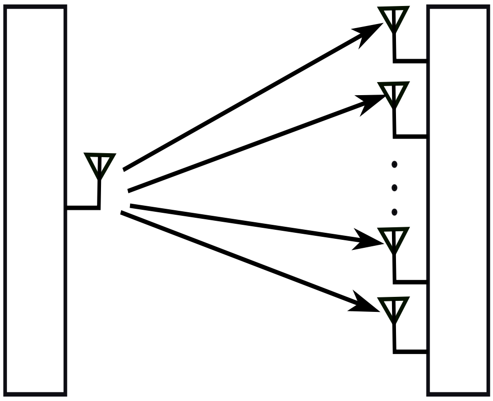


Figure 1.2.: SIMO.

1.1.2.2. Transmit Diversity

In this mode, a transmitter transmits multiple copies of a signal (see Fig. 1.3). Although transmit diversity does not boost the data rate, it increases robustness against channel fading, therefore improves the link quality.

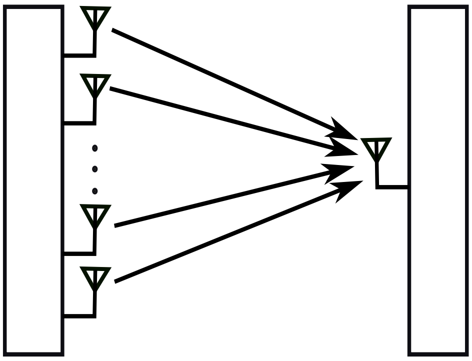


Figure 1.3.: MISO.

1.1.2.3. Spatial Multiplexing

Spatial multiplexing enhances data rate by transmitting independent data simultaneously over each transmit antenna (see Fig. 1.4). The maximum number of independent streams supported by a MIMO system is given by [6].

$$r = \min(N, M) \tag{1.3}$$

where N and M are the number of transmit and receive antennas, respectively. This parameter is in fact the multiplexing gain of a MIMO system compared to a SISO system. A rigorous explanation for the multiplexing gain can be found in Subsection 2.3.1.

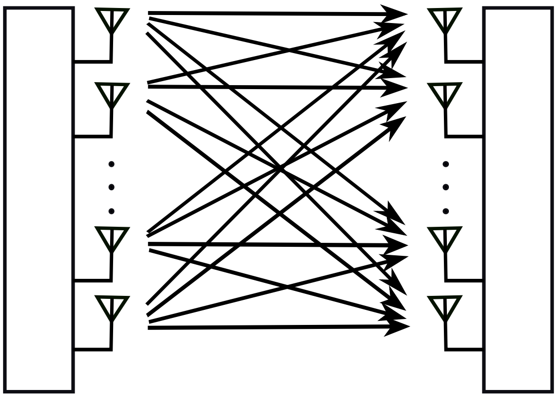


Figure 1.4.: Spatial Multiplexing.

1.1.2.4. Beamforming

Beamforming relies on antenna array to shape the radiation pattern so that the antenna gain in the user direction is maximized (see Fig. 1.5).

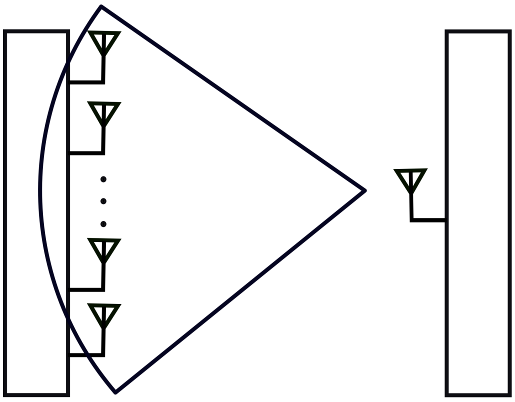


Figure 1.5.: Beamforming.

1.1.2.5. Single-user and Multi-user MIMO

As mentioned above, conventional single-user MIMO system has one transmitter and one receiver, each of which is equipped with multiple antennas. In multi-user

MIMO, a transmitter, normally, a base station communicates with many users. Note that in multi-user MIMO, the overall system will experience increased throughput but individual users. In fact, the communication in multi-user environment is more complicated than that of single-user MIMO due to the interference caused by other users to one user. The optimal non-linear dirty paper coding (DPC) can completely suppress the interference but suffers from implementation issues. Thus, non-linear precoding techniques attract more attention due to providing good balance between performance and implementation complexity. The details of mathematical formulations regarding these discussions are in the following chapter.

1.1.2.6. Open-loop and Closed-loop MIMO

In order to perform MIMO communication, a transmitter or a receiver has to know the characteristics of the channel. The channel information can be either instantaneous channel state or its distribution. The former is normally called the channel state information (CSI) while the latter is referred to as channel distribution information (CDI). Depending on the information available at the transmitter and/or the receiver, MIMO techniques can be either open-loop or closed-loop. In open-loop setting, only receiver has the knowledge of the channel, whereas a transmitter in closed-loop also has the knowledge of the channel by receiving the feedback from the receiver.

1.2. Overview of the Thesis

As mentioned previously, the capacity of single-input single-output systems is well-established but that of multi-antenna systems. In this thesis, we propose to find the capacity of MIMO systems under per-antenna power constraint (PAPC) and multiple power constraints using novel approaches which outperform existing solutions in the literature. Specifically, our contributions are summarized in the following:

- **MIMO capacity under per-antenna power constraint (PAPC)**: Per-antenna power constraint (PAPC) is more practical than SPC since each transmit antenna is connected to a different power amplifier. Despite its importance, the existing closed-form solution to single-user MIMO systems is not well-studied since the proof is incomplete, whereas the interior-point-based method for multi-user MIMO is not computationally efficient. In contrast, we proposed two approaches: The first algorithm is based on fixed-point iteration while the second one is based on alternating optimization together with successive convex optimization. Two approaches are provably convergence and have low-complexity compared to the existing solutions.

- **MIMO capacity under linear transmit covariance constraint (LTCCs)**: In practice, other power constraints can be imposed on a MIMO system in addition to either SPC or PAPC. Therefore, LTCCs are general enough to include those constraints. The current solutions to this important problem can solve this general problem properly but experience high-complexity. In light of aforementioned study, our idea is to utilize convex-concave procedure (CCP) and alternating optimization to arrive at efficient iterative algorithms. Moreover, we also take advantages of the special structure of the problem in the special case of SPC and PAPC to obtain analytical solutions. Extensive numerical and analytical results have proved the effectiveness of the proposed approach.

- **Machine learning for computing MIMO capacity**: Since an optimal solution can generate prohibitive complexity for a given channel realization, especially large-scale MIMO setting in our context, a lower complexity method is thus preferable. In this thesis, we have developed a machine learning approach which is more advantageous than optimal solutions in terms of complexity. Some initial results on this research has demonstrated the feasibility of our approach.

The structure of the thesis is as follows:

- Chapter 2 provides a literature survey of MIMO capacity. In particular, we highlight the importance of the research on PAPC and LTCCs. Then the mathematical definition of the capacity is applied to formulate the capacity of single-user MIMO and multi-user MIMO under traditional SPC. Next, we introduce the well-known water-filling (WF) solution to single-user MIMO with SPC. In this chapter, we also present the important results of MIMO capacity in broadcast channel (BC) and multiple access channel (MAC) together with non-linear and linear precoding methods. In fact, an important relation between BC and MAC, i.e., BC-MAC duality will be utilized extensively later on in the next chapters. At the end of the chapter, we discuss the open problems in computing MIMO capacity

- Our proposed approaches to MIMO capacity under PAPC are detailed in Chapter 3. Specifically, we have proposed two low-complexity approaches to the problem of interest: fixed-point iteration and AO together with SCA. The former can be adopted directly in broadcast channel (BC), whereas the latter is applied to the equivalent minimax problem in multiple access channel (MAC). In this regard, the fixed-point approach is only applicable to SU-MIMO, while the other is applicable to both SU-MIMO and MU-MIMO. We conclude the chapter by presenting and analyzing the results of the proposed iterative algorithms.

- Inspired by the approach of AO and SCA in Chapter 3, we have developed this approach to deal with the general LTCCs in Chapter 4. Similar to that of PAPC, we transform the considered problem in the BC into an equivalent problem in the MAC so that AO and CCP can be utilized to solve the obtained problem. In addition to a general framework for arbitrary power constraints, we have derived analytical solutions to the special cases of joint SPC and PAPC based on the special structure of the problem. Furthermore, we have considered both linear and nonlinear precoding methods for MU-MIMO. The numerical results have demonstrated the effectiveness of the proposed approach.

- Taking advantages of our proposed approaches, we have done some experiments regarding machine learning in Chapter 5. In particular, we introduce the fundamentals of machine learning and then apply them to estimate the maximum sum rate of successive zero-forcing dirty paper coding under PAPC. More specifically, the optimal and suboptimal solution are derived based on AO and machine learning, respectively. Since our solution relies on linear regressions, its complexity is extremely low in comparison with the optimal approach. Some interesting results have shed some light on applications of machine learning to similar problems.

- In Chapter 6, we conclude the thesis and suggest some future research directions. Particularly, we discuss about possible applications of the proposed approaches. For example, we can customize the AO and CCP approach to compute the global optimum secrecy rate of a MIMO system, which is an important problem in physical layer security. Moreover, the AO and CCP-based approach can be applicable to other minimax problem in general and we can apply the ML-based approach to solve similar problems.

1.3. Publications

Our research on MIMO capacity under multiple power constraints has resulted in a number of papers on conferences and journals below:

Journal papers

[J1] T. M. Pham, and R. Farrell, and J. Dooley, and E. Dutkiewicz, and D. N. Nguyen, and L.-N. Tran, "Efficient zero-forcing precoder design for weighted sum-rate maximization with per-antenna power constraint," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3640–3645, Apr. 2018

[J2] T. M. Pham, and R. Farrell, and L.-N. Tran, "Revisiting the MIMO capacity with per-antenna power constraint: Fixed-point iteration and alternating optimization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 388–401, Jan 2019

[J3] T. M. Pham, R. Farrell, H. Claussen, M. F. Flanagan, and L.-N. Tran, "On the MIMO capacity under multiple linear transmit covariance constraints," *IEEE Trans. Signal Process.*, 2019 submitted

Conference papers

[C1] T. M. Pham, and R. Farrell, and L.-N. Tran, "Low-complexity approaches for MIMO capacity with per-antenna power constraint," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–7

[C2] ——, "Alternating optimization for capacity region of Gaussian MIMO broadcast channels with per-antenna power constraint," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–6

[C3] T. M. Pham, R. Farrell, H. Claussen, M. F. Flanagan, and L.-N. Tran, “Weighted sum rate maximization for zero-forcing methods with general linear covariance constraints,” in *Proc. IEEE ICC*, May 2018, pp. 1–6

[C4] ——, “On the MIMO capacity with multiple linear transmit covariance constraints,” in *Proc. IEEE VTC-Spring*, Jun. 2018, pp. 1–6

[C5] T. M. Pham, R. Farrell, and L.-N. Tran, “On estimating maximum sum rate of MIMO systems with successive zero-forcing dirty paper coding and per-antenna power constraint,” in *Proc. IEEE PIMRC*, Sep. 2019

In particular, part of Chapter 2 and Chapter 5 has been appeared in the third journal paper and the fifth conference paper, respectively. Additionally, most of the content in Chapter 3 is already in the first and the second journal paper while that of Chapter 4 has been appeared in the third journal paper.

1.4. Notation

Standard notations are used in this thesis. Bold lower and upper case letters represent vectors and matrices, respectively. $\mathbf{I}_N$ defines an identity matrix of size N; $\mathbf{I}$ and $\mathbf{0}$ define identity and zero matrices respectively, of which the size can be easily inferred from the context. $\mathbb{C}^{M\times N}$ denotes the space of $M \times N$ complex matrices; $\mathbf{H}^\dagger/\mathbf{H}^H$ and $\mathbf{H}^T$ are Hermitian and ordinary transpose of $\mathbf{H}$, respectively; $\mathbf{H}_{i,j}$ is the (i, j)-entry of $\mathbf{H}$; rank$(\mathbf{H})$ is the rank of $\mathbf{H}$; $\mathcal{N}(\mathbf{H})$ is the null space of $\mathbf{H}$; $|\mathbf{H}|$ is the determinant of $\mathbf{H}$; $\lambda_{\max}(\mathbf{H})$ is the maximum eigenvalue of a Hermitian matrix $\mathbf{H}$; diag$(\mathbf{x})$ denotes the diagonal matrix having diagonal entries matching the vector $\mathbf{x}$; diag$(\mathbf{H})$, where $\mathbf{H}$ is a square matrix, is the vector of diagonal elements of $\mathbf{H}$. The notation $\mathbf{x} \odot \mathbf{y}$ denotes the Hadamard product (i.e., the entrywise product) of $\mathbf{x}$ and $\mathbf{y}$. The notation $\mathbf{A} \succeq (\succ)\mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semidefinite (definite). Furthermore, we denote the expected value of a random variable by $\mathbb{E}[.]$, and $[x]_+ = \max(x, 0)$, $\mathbf{0}_n$ and $\mathbf{1}_n$ to be a *row* vector of size n with all zeros and ones, respectively. The Euclidean and Frobenius norms are denoted by $||\cdot||_2$ and $||\cdot||_F$, respectively. The ith unit vector (i.e., its ith entry is equal to one and all other entries are zero) is denoted by $\mathbf{e}_i$.

Chapter 2

Background

This chapter provides the background for the research in this thesis. We first present a literature survey of computing MIMO capacity in Section 2.1 followed by the definition of capacity in Section 2.2. In the Section 2.3, we formulate the capacity of single-user MIMO and multi-user MIMO under conventional sum power constraint and point out some open problems related to capacity in Section 2.4. Most of the content of Section 2.1 has been published in [2] under © 2018 IEEE and [13].

2.1. Literature Survey of Computing MIMO Capacity

From a system design perspective, one of the most fundamental problems is to compute the capacity of the system of interest. For a single-user MIMO (SU-MIMO) channel, pioneer studies proved that the capacity can be achieved by Gaussian input signaling [6, 7]. For multi-user MIMO (MU-MIMO) scenarios, the seminal work of [8] showed that dirty-paper coding (DPC) in fact achieves the entire capacity region of Gaussian MIMO broadcast channel (BC). Since finding the capacity of MIMO channels is computationally expensive in general, one is also interested in near-capacity achieving transmission strategies such as successive zero-forcing DPC (SZFDPC) [9, 19] or zero-forcing (ZF) [20, 21], for which the achievable rate region is much easier to characterize.

The capacity of MIMO systems is investigated along with a certain type of constraint on the input covariance matrices. To this end, a majority of the related literature assumes a sum power constraint (SPC) as it usually leads to efficiently computational algorithms. In particular, under perfect channel state information (CSI) at both transmitter and receiver, the capacity of a SU-MIMO channel is found using the closed-form water-filling (WF) algorithm [6, 7]. In [22], Yu *et al.* presented an iterative WF (IWF) algorithm to compute the sum capacity for a Gaussian vector multiple access channel (MAC). In [23], Jindal *et al.* proposed sum power IWF to determine the sum capacity of Gaussian MIMO BCs by exploiting the MAC-BC duality. The entire capacity region of MIMO-BCs with a SPC was characterized in [24, 25], using conjugate gradient projection (CGP)- and pre-conditioned gradient projection-based approaches, respectively.

In reality, each antenna is associated with a separate power amplifier, each having a different dynamic range. As such, per-antenna power constraint (PAPC) is of more practical importance. If a sum power constraint is considered, some antennas may be allocated a power level that is beyond their dynamic range of the associated power amplifier, depending on fading situations. This will result in nonlinear distortion that has a detrimental impact on the whole system. In [8], it was shown that DPC still achieves the full capacity region of the MIMO BC under PAPC. However, finding the DPC region with PAPC is more numerically difficult than with a SPC. In fact, no closed-form design has been reported for the computation of the capacity region of the MIMO BC subject to PAPC. For this reason, numerous research endeavors have been made to understand performance limits of various suboptimal transmission strategies such as zero-forcing (ZF) beamforming, minimum mean square error (MMSE), and SZFDPC [12, 20, 21, 26–30].

The capacity of a SU-MIMO channel with PAPC was studied in [1, 3]. In particular, the author in [1, 3] proposed an iterative mode-dropping algorithm based on closed-form expressions to find the optimal input covariance. As shown in the next chapter, this algorithm still requires high computational complexity and its convergence proof is not complete. Also, the mode-dropping algorithm assumes a full-rank channel which hardly holds true in practice. To the best of our knowledge, the only attempt to characterize the entire capacity region of the MIMO BC subject to PAPC was made in [10]. Specifically, the authors established a modified duality between the MAC and BC and transformed the input optimization problem in the BC into a minimax optimization problem in the corresponding MAC. Then the resulting program is solved by a standard barrier interior-point routine. Similarly, Tran *et al.* also proposed customized interior-point methods to study the achievable rate region of SZFDPC in [29, 30]. However, the complexity of such second order optimization methods increases quadratically with the number of input dimensions, which is not practically appealing for large-scale antenna systems (also known as massive MIMO).

In practice, other types of power constraint can also be imposed on a MIMO system, not necessarily limited to SPC or PAPC separately. For example, optimal transmit covariance for MISO channels with joint SPC and PAPC was recently studied in [31,32]. In the context of cognitive networks, interference temperature constraints can be imposed on a secondary user (SU) to limit the interference generated at a primary user (PU) [33–35]. All of these constraints can be generally modeled as linear transmit covariance constraints (LTCCs) [33]. More recent research efforts have been made to characterize the capacity of Gaussian MIMO channels with joint SPC and PAPC [31, 32]. However, the work of [31] is only applicable to MISO systems, while that of [32] partially addresses general MIMO channels. For the general form of LTCCs, interior-point and subgradient methods were presented in [33, 34] to compute optimal transmit covariance matrices. However, it was demonstrated in [15] that these high-complexity methods are not useful for massive MIMO systems.

2.2. Mutual Information and Definition of Capacity

Consider a discrete random variable $X \in \{x_1, x_2, \ldots\}$. The information $I(x_i)$ associated with the event $X = x_i$ is given by

$$I(x_i) = -\log p(x_i) \tag{2.1}$$

where $p(x_i)$ denotes the probability $X = x_i$.

From the information theory, the average information denotes the entropy $H(X)$ as follows

$$H(X) = \mathbb{E}_X(I(X)) = -\sum_i p_X(x_i) \log p_X(x_i). \tag{2.2}$$

The entropy in fact measures the uncertainty of a random variable. The conditional entropy $H(X|Y)$ is defined as follows:

$$H(X|Y) = -\sum_{x,y} p(x, y) \log p_{X,Y}(y|x). \tag{2.3}$$

Recall the generic digital communication system in Chapter 1, the encoder transforms the information into a finite set of symbols $\{x_1, x_2, \cdots\}$ and pass those encoded symbols through the channel and received a set of symbols $\{y_1, y_2, \cdots\}$. Then the amount of information about the occurrence of an event $X = x_i$ given by the occurrence of the event $Y = y_i$ is called the mutual information which is

$$I(x_i; y_i) = \log(\frac{p(x_i|y_i)}{p(x_i)}). \tag{2.4}$$

The average mutual information is then given by

$$I(X; Y) = \mathbb{E}_{x,y}(I(x; y)) = \sum_{x,y} p(x, y) \log(\frac{p(x|y)}{p(x)}). \tag{2.5}$$

After some manipulations, we arrive at the following result

$$I(X; Y) = H(X) - H(X|Y). \tag{2.6}$$

The capacity C of a single-input single-output system is defined as

$$C = \max_{\{p_X(x)\}} I(X; Y) \tag{2.7}$$

where $0 \leq p_X(x) \leq 1$ and $\sum_x p_X(x) = 1$. In other words, the capacity of a communication system is the maximum amount of mutual information. Herein, we use this definition to derive the capacity of single-user and multi-user MIMO capacity in the following. Note that we assume that the channel state information is perfectly known at both the transmitter and receiver.

2.3. Single-user MIMO and Multi-user MIMO Capacity

2.3.1. Single-user MIMO Capacity

In this section, we develop the capacity of single-user MIMO based on the definition of capacity mentioned above. In particular, we examine a fundamental scenario where a transmitter has N transmit antennas and a receiver has M receive antennas. The channel matrix $\mathbf{H}$ consists a set of channel response h_{ij} between transmit antenna j to receive antenna i. The received signal is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{z} \tag{2.8}$$

where $\mathbf{s}$ and $\mathbf{z}$ are the transmitted vector and additive white circularly symmetric complex Gaussian noise vector. Extending the definition of the capacity in (2.7) to a MIMO system, we obtain

$$C = \max_{\{p_S(\mathbf{s})\}} I(S;Y) = \max_{\{p_S(\mathbf{s})\}} \{H(\mathbf{y}) - H(\mathbf{y}|\mathbf{s})\}. \tag{2.9}$$

Since $\mathbf{Hs}$ is fixed for the given channel matrix, then the uncertainty is caused by the noise i.e., $H(\mathbf{y}|\mathbf{s}) = H(\mathbf{z})$. Then (2.9) can be simplified to the following

$$C = \max_{\{p_S(\mathbf{s})\}} \{H(\mathbf{y}) - H(\mathbf{z})\}. \tag{2.10}$$

Expanding two terms in (2.10) results in the formula for MIMO capacity [6]

$$C = \log |\mathbf{I} + \frac{1}{\rho^2}\mathbf{H}\mathbf{S}\mathbf{H}^\dagger| \tag{2.11}$$

where $\mathbf{S} = \mathbb{E}(\mathbf{s}\mathbf{s}^\dagger)$ is the input covariance matrix, ρ^2 is the noise variance. Since the noise is usually normalized, (2.11) is therefore reduced to

$$C = \log |\mathbf{I} + \mathbf{H}\mathbf{S}\mathbf{H}^\dagger|. \tag{2.12}$$

When the transmitter is assumed to be subject to an average power constraint P, i.e., $\text{tr}(\mathbf{S}) \leq P.$, the capacity is the solution to the following problem

$$\begin{aligned} \underset{\mathbf{S}\succeq\mathbf{0}}{\text{maximize}} \quad & \log |\mathbf{I} + \mathbf{H}\mathbf{S}\mathbf{H}^\dagger| \\ \text{subject to} \quad & \text{tr}(\mathbf{S}) \leq P. \end{aligned} \tag{2.13}$$

A simple method to calculate the capacity is to extract the channel into multiple independent Additive White Gaussian Noise (AWGN) channels so that the capacity is the sum of AWGN channel capacity. The singular value decomposition (SVD) allows us to rewrite $\mathbf{H}$ as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger$ where $\mathbf{U}, \mathbf{V}$ are unitary matrices and $\mathbf{\Sigma}$

is a diagonal matrix with non-negative descending-order singular value i.e., $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{r_H}$where $r_H \leq \min(N, M)$ is called the rank of the channel matrix $\mathbf{H}$. The rank of a matrix determines the number of independent streams which can be multiplexed simultaneously. Thus, we can transform the system into an equivalent system by introducing new variables

$$\mathbf{s} = \mathbf{V}\tilde{\mathbf{s}} \tag{2.14}$$

$$\mathbf{y} = \mathbf{U}\tilde{\mathbf{y}} \tag{2.15}$$

thus

$$\mathbf{U}\tilde{\mathbf{y}} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger)\mathbf{V}\tilde{\mathbf{s}} + \mathbf{z} \tag{2.16}$$

or equivalently

$$\tilde{\mathbf{y}} = \mathbf{\Sigma}\tilde{\mathbf{s}} + \mathbf{U}^\dagger\mathbf{z} = \mathbf{\Sigma}\tilde{\mathbf{s}} + \tilde{\mathbf{z}}. \tag{2.17}$$

Since $\mathbb{E}(\mathbf{s}\mathbf{s}^\dagger) = \mathbb{E}(\tilde{\mathbf{s}}\tilde{\mathbf{s}}^\dagger)$ and $\mathbb{E}(\mathbf{z}\mathbf{z}^\dagger) = \mathbb{E}(\tilde{\mathbf{z}}\tilde{\mathbf{z}}^\dagger)$, the power constraint of the equivalent system is the same i.e., $\mathrm{tr}(\tilde{\mathbf{S}}) \leq P$. The output can be extracted to parallel channels as follows

$$\tilde{y}_i = \sigma_i\tilde{s}_i + \tilde{z}_i, i = 1, 2, \ldots, r_H. \tag{2.18}$$

The water-filling algorithm allocates the power to r_H channels such that

$$\sum_{i=1}^{r_H} P_i = \sum_{i=1}^{r_H}(\gamma - \frac{1}{\sigma_i^2})_+ = P \tag{2.19}$$

where γ is called the power level. Based on Eq. (1.2), the capacity of these multiple AGWN channels is given by

$$C = \sum_{i=1}^{r_H} \log(1 + P_i\sigma_i^2). \tag{2.20}$$

The principle of the water-filling is illustrated in the Fig. 2.1. Considering different containers for different eigenvalues, the water is first poured to the first largest eigenvalues, then the second largest one and so on until the water i.e., power is fully assigned.

Remark:

- At low SNR regime, i.e., $\gamma \ll \frac{1}{\sigma_i^2}$, thus the whole power is assigned to the largest eigenvalue channel.

- At high SNR regime, i.e., $\gamma \gg \frac{1}{\sigma_i^2}$, the power assigned to each channel is nearly constant, i.e., $P_i = \frac{P}{r_H}$ and an approximate capacity is given below

$$C \simeq \sum_{i=1}^{r_H} \log(1 + \frac{P}{r_H}\sigma_i^2). \tag{2.21}$$

$$\simeq r_H \log P. \tag{2.22}$$

From the equation above, we can easily see that the capacity of a MIMO system is approximately r_H times larger than that of AWGN channel at high SNR. The rank of the channel r_H is in fact the multiplexing gain which we have mentioned in the first section of Chapter 1.

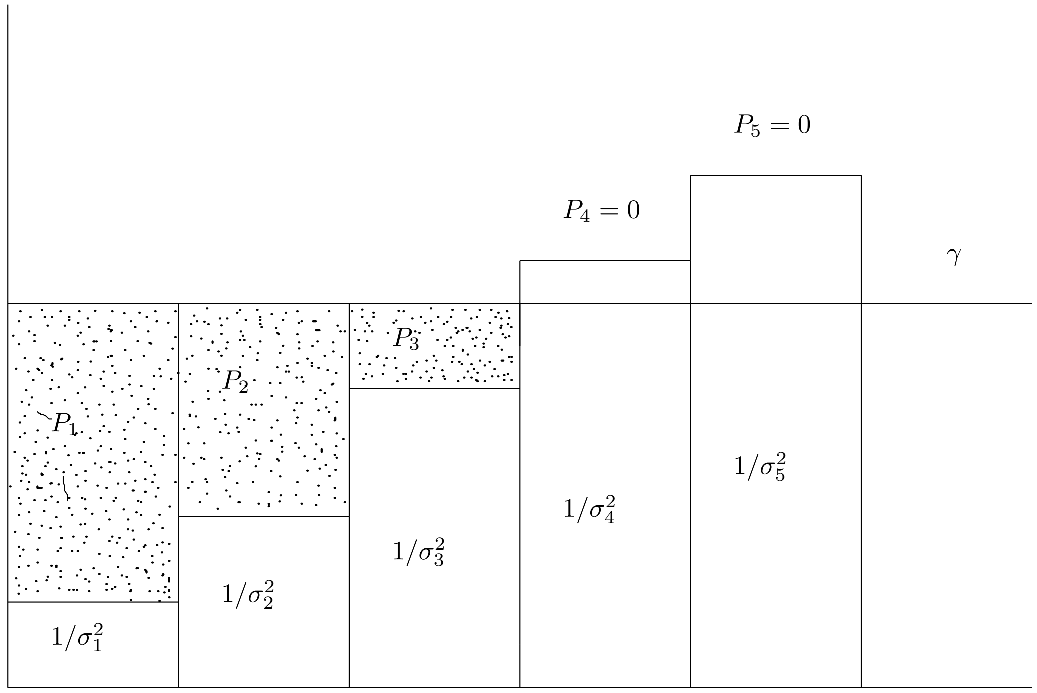


Figure 2.1.: Water-filling principle. The water is poured until fully assigned.

2.3.2. Capacity Region of Multi-user MIMO

The most important results of the capacity of multi-user MIMO Gaussian channels are summarized in this section. Specifically, we consider two basic channel models: Broadcast channel (BC) and Multiple-access channel (MAC) (see Fig. 2.2). In BC, a base station sends messages to many users. On the contrary, many users send messages to a base station in MAC scenario. In the context of cellular networks, the former is called the downlink while the latter is called the uplink. Hence, we use interchangeably BC and downlink, MAC and uplink. Recall that the capacity of a single-user MIMO is a number due to only one rate. However, in multi-user MIMO, the capacity is a region (see Fig. 2.3) since there are different rates associated with each users.

Note that in single-user MIMO, the capacity gain increase linearly with the rank of the channel matrix i.e., $\min(N, M)$ where N and M are the number of transmit and receive antennas, respectively, thus multiple antennas can be deployed at the transmitter and receiver to achieve the gain. However, in multi-user MIMO, the sum rate increases linearly with $\min(N, MK)$ where K is the number of users. Therefore, it is sufficient to deploy a large number of antennas at the base station to serve a large number of users equipped with one or a few receive antennas.

It should be noted that each user in the MAC has independent data stream and is therefore associated with a different data rate. As a result, the capacity region is K-dimension. Similarly, the transmitter in the BC has an independent message for each user and thus the capacity region is also K-dimension. The fundamental results on MAC and BC associated with SPC are summarized in the following.

2.3.2.1. Broadcast Channel

Let $\mathbf{s}$ be the transmitted signal from the base station and $\mathbf{y}_k$ denotes the received signal at user k and $\mathbf{z}_k \sim \mathcal{N}(0, \mathbf{I})$ is the noise signal with circularly symmetric complex Gaussian with an identity covariance matrix for each user. The received signal at a user k is given by

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{s} + \mathbf{z}_k \tag{2.23}$$

In general, the capacity region of the BC is unknown due to the difficulties of interference cancellation at the receivers. However, by pre-subtract multi-user interference at the transmitter, dirty paper coding (DPC) is proved to achieve the capacity region. In this technique, the rate is defined as

$$R_k \quad = \quad \log \frac{|\mathbf{H}_k(\sum\limits_{i\geq k} \mathbf{S}_i)\mathbf{H}_k^{\dagger} + \mathbf{I}|}{|\mathbf{H}_k(\sum\limits_{i>k} \mathbf{S}_i)\mathbf{H}_k^{\dagger} + \mathbf{I}|}. \tag{2.24}$$

It is worth mentioning that the rate equations are neither concave or convex, thus problem becomes nontrivial problem. In reality, the problem is transformed into an equivalent convex problem so that standard optimization methods can be utilized. In particular, by establishing the duality relationship between the BC and the MAC, the original problem is reformulated to a convex optimization problem in MAC domain. This technique will be detailed in the BC-MAC duality below.

2.3.2.2. Multiple Access Channel

Let $\bar{\mathbf{s}}_k$ be the transmitted signal associated with user k and $\mathbf{y}$ denotes the received signal and $\bar{\mathbf{z}} \sim \mathcal{N}(0, \mathbf{I})$ is the noise signal with circularly symmetric complex Gaussian with an identity covariance matrix. The received signal at the base station is given by

$$\mathbf{y} = [\mathbf{H}_1^{\dagger}, \mathbf{H}_2^{\dagger}, \ldots, \mathbf{H}_K^{\dagger}] \begin{bmatrix} \bar{\mathbf{s}}_1 \\ \bar{\mathbf{s}}_2 \\ \cdots \\ \bar{\mathbf{s}}_K \end{bmatrix} + \bar{\mathbf{z}}. \tag{2.25}$$

Each user is subject to an individual power constraint which implies $\mathrm{tr}(\bar{\mathbf{S}}_k) \leq P_k$ or $\sum\limits_{k=1}^{K} \mathrm{tr}(\bar{\mathbf{S}}_k) \leq P$.

2.3.2.3. BC-MAC Duality

Intuitively, the MAC can be achieved from the BC by reversing the arrow as shown in Fig. 2.2a. In fact, a K-user BC and MAC are dual if the following conditions hold [36, p.491]:

1. The channel response is the same for the uplink and the downlink.
2. The noise statistics of the channel in the downlink is the same as that of the uplink.
3. The sum of individual power constraints in the uplink is the power constraint in the downlink.

The duality states that there exist covariance matrices $\{\bar{\mathbf{S}}_k\}$ in MAC such that $\sum\limits_{k=1}^{K} \mathrm{tr}(\bar{\mathbf{S}}_k) = \sum\limits_{k=1}^{K} \mathrm{tr}(\mathbf{S}_k)$, and thus $\bar{R}_k = R_k$, or in other words

$$\log \frac{|\sum\limits_{i \leq k} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i \mathbf{H}_i + \mathbf{I}|}{|\sum\limits_{i < k} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i \mathbf{H}_i + \mathbf{I}|} = \log \frac{|\mathbf{H}_k (\sum\limits_{i \geq k} \mathbf{S}_i) \mathbf{H}_k^{\dagger} + \mathbf{I}|}{|\mathbf{H}_k (\sum\limits_{i > k} \mathbf{S}_i) \mathbf{H}_k^{\dagger} + \mathbf{I}|}. \tag{2.26}$$

In fact, the duality implies that the DPC region of the BC is equal to the capacity region of the MAC. In particular, for a set of weights $0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_K$, the equivalent optimization problem in the MAC is

$$\underset{\{\bar{\mathbf{S}}_k\} \in \bar{\mathcal{S}}}{\text{maximize}} \sum_{k=1}^{K} \mu_k \bar{R}_k \tag{2.27}$$

or equivalently [37]

$$\underset{\{\bar{\mathbf{S}}_k\} \succeq \mathbf{0}}{\text{maximize}} \sum_{k=1}^{K} (\mu_k - \mu_{k-1}) \log |\sum_{i=k}^{K} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i \mathbf{H}_i + \mathbf{I}| \tag{2.28}$$

$$\text{subject to} \sum_{k=1}^{K} \mathrm{tr}(\bar{\mathbf{S}}_k) \leq P. \tag{2.29}$$

It is easy to see that the equivalent problem in the MAC is convex, thus can be solved by traditional optimization techniques. If $\mu_1 = \mu_2 = \ldots = \mu_K = 1$, the problem reduces to sum rate maximization (SRMax) given by

$$\underset{\{\bar{\mathbf{S}}_k\} \succeq \mathbf{0}}{\text{maximize}} \sum_{k=1}^{K} \log |\sum_{i=k}^{K} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i \mathbf{H}_i + \mathbf{I}| \tag{2.30}$$

$$\text{subject to} \sum_{k=1}^{K} \mathrm{tr}(\bar{\mathbf{S}}_k) \leq P. \tag{2.31}$$

2.3.2.4. Near-optimal Precoding Methods

The dirty paper coding (DPC) technique has been proved to achieve the capacity of a Gaussian multiple-input multiple-output broadcast channel (BC) [8]. However, such a nonlinear coding strategy is not appealing to practical applications due to complex processing in both encoders and decoders. To this point, linear precoding methods such as zero-forcing (ZF) and successive zero-forcing dirty paper coding (SZFDPC) are promising alternatives since they provide good trade-off between performance and implementation complexity.

Consider a K-user single-cell MIMO BC where the base station (BS) and each user have N and M_k antennas, respectively. Let $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$ be the channel matrix for user k. Then, the received signal at user k is given by

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \textstyle\sum_{j\neq k}\mathbf{H}_k\mathbf{x}_j + \mathbf{z}_k \tag{2.32}$$

where $\mathbf{x}_k \in \mathbb{C}^{N\times 1}$ is the downlink signal for the kth user and $\mathbf{z}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ is the background noise. For linear precoding $\mathbf{x}_k$ can be expressed as $\mathbf{x}_k = \mathbf{R}_k\mathbf{s}_k$, where $\mathbf{R}_k \in \mathbb{C}^{N\times M_k}$ and $\mathbf{s}_k \in \mathbb{C}^{M_k\times 1}$ denote the precoding matrix and information-bearing signal, respectively. We also assume that $\mathbf{s}_k$ consists of independent zero-mean and unit energy symbols, i.e., $\mathbf{s}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. In case of ZF precoding, the inter-user interference to user k is suppressed by designing $\mathbf{R}_k$ such that $\mathbf{H}_j\mathbf{R}_k = \mathbf{0}$ for all $j \neq k$. Thus, the weighted sum rate maximization (WSRMax) problem for ZF precoding with SPC is formulated as

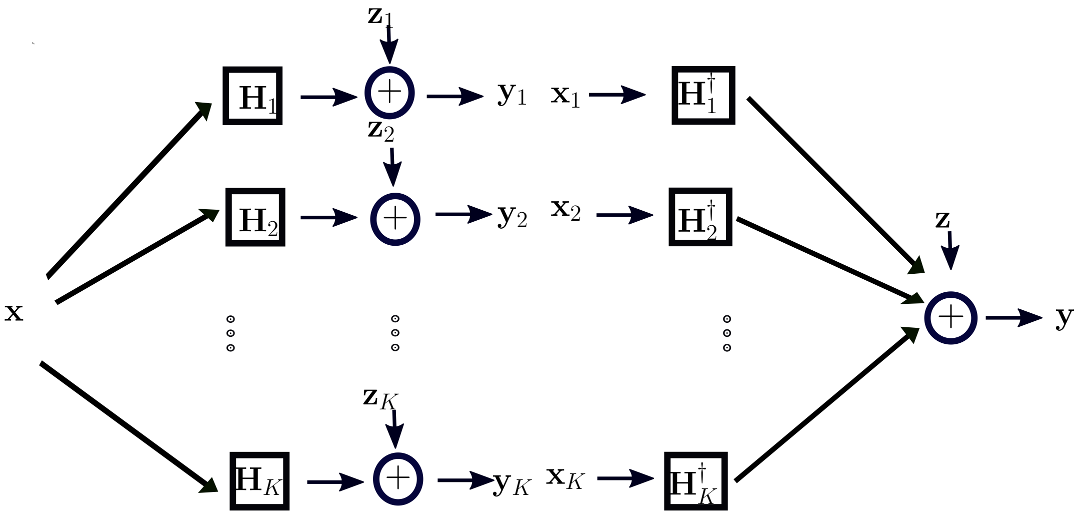
$$\begin{aligned}
&\underset{\{\mathbf{S}_k\succeq\mathbf{0}\}}{\text{maximize}} && \textstyle\sum_{k=1}^{K} w_k \log|\mathbf{I}+\mathbf{H}_k\mathbf{S}_k\mathbf{H}_k^{\dagger}| && \text{(2.33a)}\\
&\text{subject to} && \mathbf{H}_j\mathbf{S}_k\mathbf{H}_j^{\dagger} = \mathbf{0},\ \forall j\neq k && \text{(2.33b)}\\
& && \textstyle\sum_{k=1}^{K}\operatorname{tr}(\mathbf{S}_k) \le P, && \text{(2.33c)}
\end{aligned}$$

where $\mathbf{S}_k = \mathbb{E}[\mathbf{x}_k\mathbf{x}_k^{\dagger}] = \mathbf{R}_k\mathbf{R}_k^{\dagger}$ is the input covariance matrix for user k, P is the power constraint of the system, and $w_k \geq 0$ is the weight of user k.

2.4. Open Problems in Computing MIMO Capacity

The discussions so far have showed the most important results and formulations in MIMO systems under perfect channel assumption and sum power constraint. In the following, we present some of the challenges in MIMO systems, more specifically, the MIMO capacity:

1. Efficient algorithms for large-scale MIMO systems: Capacity under assumption of perfect channel state information is usually considered with tradition sum power constraint, and recently PAPC and LTCCs. Nevertheless, existing solutions in the literature rely on general high-complexity methods such as



(a) Broadcast channel **(b)** Multiple access channel

Figure 2.2.: Broadcast channel and Multiple-access channel.

interior-point method, which are in turn inapplicable to large-scale MIMO systems. Thus, the efficient algorithms for MIMO, especially large-scale MIMO, are still open problems.

2. Channel distribution information : While the capacity under the perfect channel information assumption provides useful insight, the study of the channel capacity based on channel distribution is of great practical relevance. Moreover, the channel information can change quickly over time, little results have been achieved with imperfect CSI or CDI at transmitter or receiver.

3. Non-DPC precoding methods: It is a well-known fact that DPC is optimal solution to multi-user MIMO but difficult to implement in practice due to its complexity. Thus, low-complexity precoding methods, which can trade off the implementation complexity and performance, are of more interest.

The first issue will be discussed in more details in the next chapters, which is the focus of this thesis. Specifically, Chapter 3 details two approaches to tackle the problem of computing MIMO capacity under PAPC ranging from single-user MIMO to multi-user MIMO. The approaches rely on fixed-point iteration and alternating optimization together with successive convex approximation, thus achieve low complexity. In Chapter 4, the AO approach is extended to compute the MIMO capacity under a generalized power constraint. The problem formulation, though relying on the same framework with that of PAPC, has resulted in different solutions in each iteration. In case of joint SPC and PAPC, analytical solutions have been derived, thanks to the special structures of the problems. All the aforementioned algorithms are applicable to large-scale MIMO systems due to their low complexity. In fact, the third problem is partially addressed in Chapter 3, 4, and 5 in which we have considered other precoding methods than DPC. More specifically, we have exploited

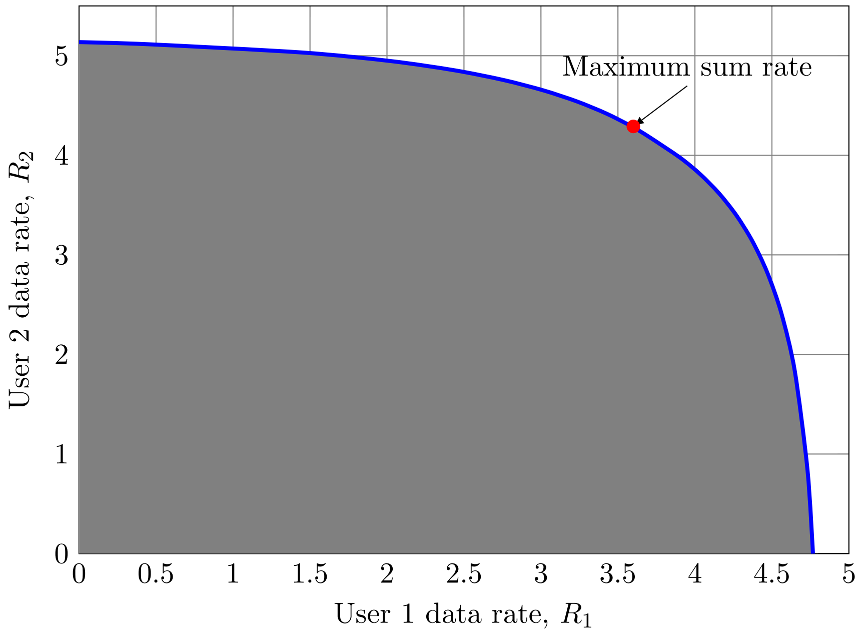


Figure 2.3.: An illustration of 2-user capacity region.

the proposed approaches to arrive at efficient solutions to ZF in Chapter 3, 4 and machine learning-based method for SZFDPC under PAPC in Chapter 5. Taking advantages of our proposed approaches, we have also outlined some possible methods to solve the capacity-related problems under imperfect channel estimation in Chapter 6.

Chapter 3

MIMO Capacity with Per-antenna Power Constraint

As mentioned in previous chapters, the majority of the related literature has investigated the capacity of MIMO systems along with SPC due to its simplicity. In reality, SPC may result in distortion since some power amplifiers connected to transmit antennas may be assigned a power level that is beyond their limits. As a result, PAPC which imposes a power limit to each power amplifier is more practical. In this chapter, we consider the problems of finding the capacity of various MIMO settings subject to PAPC, ranging from the SU-MIMO to MIMO BC. The goal is to arrive at closed-form design for considered problems which are extremely helpful for analyzing large-scale systems. In particular, our specific contributions include the following:

- For an SU-MIMO channel we proposed two fast-converging low-complexity iterative algorithms to compute the optimal input covariance matrices under PAPC. The first method is based on manipulating the optimality conditions of the considered problem and fixed-point iteration. The second one relies on the well-known MAC-BC duality, but the resulting minimax problem is solved by a novel alternating optimization (AO) algorithm. Specifically, we proposed to optimize the upper bound of the objective with respect to a coordinate, eliminating the zigzag effect likely occurring in a pure AO method. Both proposed methods are provably convergent without any specific assumption on the channel matrix. Extensive analytical and numerical results are provided to demonstrate the superior performance of the proposed method, compared to the mode-dropping algorithm in [1,3].
- By exploiting the specific structure of the weighted sum rate with ZF and PAPC, we recruit the AO-based approach to derive an iterative algorithm whose monotonic convergence is achieved. Since the subproblem at each iteration of the proposed method is solved by water-filling-like algorithms, the proposed method can be extended to deal with the ZF precoder design in large-scale MIMO systems that are beyond the capability of state-of-the-art convex solvers.

- We also characterize the entire capacity region of the MIMO BC, which was studied in [10]. For the MIMO BC, the weighted sum capacity is neither a concave nor convex function of the covariance matrices. Thus, the MAC-BC duality is invoked to obtain a convex formulation in the dual MAC, which is given in the form of a minimax optimization problem [10]. Instead of applying a standard interior-point method to find a saddle point of the resulting minimax program, we propose a closed-form design based on AO, similar to the case of SU-MIMO. The idea is to leverage the fact that the weighted sum capacity problem under a SPC can be solved by closed-form expressions in combination with a CGP method [24].

The remainder of the chapter is organized as follows. The capacity of SU-MIMO is described in Section 3.1 followed by that of ZF in Section 3.2. Section 3.3 derives closed-form expressions for the capacity region of a Gaussian MIMO BC while Section 3.4 presents the numerical results. Finally, we conclude the chapter in Section 3.5. Most of the content and results in this chapter have been published in [2, 12] under © 2019 and 2018 IEEE.

3.1. Capacity of SU-MIMO

3.1.1. System Model

Consider a SU-MIMO channel, where the transmitter is equipped with N antennas and the receiver with M antennas. The channel matrix is represented by $\mathbf{H} \in \mathbb{C}^{M\times N}$, which is assumed to be known perfectly at the transmitter. The received signal is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{z} \tag{3.1}$$

where $\mathbf{s}$ is the vector of transmitted symbols of zero-mean, and $\mathbf{z} \in \mathbb{C}^{M\times 1}$ is the background noise with distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. Let $\mathbf{S} = E\{\mathbf{s}\mathbf{s}^\dagger\}$ be the input covariance matrix for the transmitted signal. We are interested in finding the capacity of the above channel with PAPC, which is formulated as

$$\begin{aligned} &\underset{\mathbf{S}\succeq\mathbf{0}}{\text{maximize}} && \log|\mathbf{I} + \mathbf{H}\mathbf{S}\mathbf{H}^\dagger| && \text{(3.2a)} \\ &\text{subject to} && [\mathbf{S}]_{i,i} \leq P_i,\ i = 1, 2, \ldots, N && \text{(3.2b)} \end{aligned}$$

where P_i is the maximum power constraint on the ith antenna. The problem (3.2) is a convex program, and can be solved by general-purpose optimization software.[1]

[1] More specifically, (3.2) in the current form is in fact a MAXDET program [38] but can be reformulated as a semidefinite program (SDP) [39, p. 149] for which dedicated solvers are more available.

However, the computational complexity of these convex solvers, which are usually based on interior-point methods, increases rapidly with the number of transmit antennas N, thereby not suitable for large-scale MIMO systems. Herein, we propose two efficient iterative algorithms which will be numerically shown to achieve a superlinear convergence rate.

3.1.2. Proposed Algorithms

3.1.2.1. Fixed-point Iteration

We first note that the Slater condition is satisfied for (3.2) and thus strong duality holds. Now, consider the partial Lagrangian function of (3.2), which is given by

$$\mathcal{L}(\mathbf{S},\mathbf{\Lambda}) = \log|\mathbf{I}+\mathbf{H}\mathbf{S}\mathbf{H}^{\dagger}| - \operatorname{tr}(\mathbf{\Lambda}(\mathbf{S}-\mathbf{P})) \tag{3.3}$$

where $\mathbf{P}=\operatorname{diag}(P_1,P_2,\ldots,P_N)$, and $\mathbf{\Lambda}=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_N)$ is the diagonal matrix comprising the dual variables for the N power constraints in (3.2b). The dual objective of (3.2) is

$$g(\mathbf{\Lambda}) = \max_{\mathbf{S}\succeq\mathbf{0}} \mathcal{L}(\mathbf{S},\mathbf{\Lambda}). \tag{3.4}$$

To find the optimal solution of (3.2), we only need to consider the case where $\mathbf{\Lambda}\succ\mathbf{0}$, i.e, $\lambda_i>0$ for all i, otherwise $g(\mathbf{\Lambda})$ is unbounded above, which cannot be the dual optimal of (3.2). This can be easily seen by contradiction. Suppose $\lambda_i=0$ for some i. Then create a diagonal matrix $\mathbf{S}=\operatorname{diag}([0,\ldots,0,\alpha_i,0,\ldots,0]^T)$. Accordingly, we can check that $\mathcal{L}(\mathbf{S},\mathbf{\Lambda})=\log(1+\alpha_i\sum_{j=1}^{M}|\mathbf{H}_{j,i}|^2)\to\infty$ if $\alpha_i\to\infty$. Moreover, for a given $\mathbf{\Lambda}\succ\mathbf{0}$, we can solve (3.4) efficiently as described next. Let us denote $\hat{\mathbf{S}}=\mathbf{\Lambda}^{1/2}\mathbf{S}\mathbf{\Lambda}^{1/2}$. Then finding $\mathbf{S}$ to maximize $\mathcal{L}(\mathbf{S},\mathbf{\Lambda})$ amounts to solving the following problem

$$\underset{\hat{\mathbf{S}}\succeq\mathbf{0}}{\text{maximize}}\ \log|\mathbf{I}+\mathbf{H}\mathbf{\Lambda}^{-1/2}\hat{\mathbf{S}}\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}| - \operatorname{tr}(\hat{\mathbf{S}}) \tag{3.5}$$

The above problem admits the solution based on water-filling algorithm with *fixed water level* [40]. Explicitly, let $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\dagger}=\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$ be the eigenvalue decomposition (EVD) of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$, where $\mathbf{V}\in\mathbb{C}^{N\times N}$ are unitary matrix, and $\mathbf{\Sigma}\in\mathbb{C}^{N\times N}$ is a matrix of *(possibly zero) eigenvalues in decreasing order* of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$. Let $r=\operatorname{rank}(\mathbf{H}\mathbf{\Lambda}^{-1/2})$, and ρ_i, $i=1,\ldots r$, be r positive eigenvalues of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$. Then, $\hat{\mathbf{S}}$ can be found as

$$\hat{\mathbf{S}} = \mathbf{V}\operatorname{diag}([1-\frac{1}{\rho_1}]_+,\ldots,[1-\frac{1}{\rho_r}]_+,\mathbf{0}_{N-r})\mathbf{V}^{\dagger}. \tag{3.6}$$

Consequently, $\mathbf{S}$ is given by

$$\mathbf{S} = \mathbf{\Lambda}^{-1/2}\mathbf{V}\Big(\operatorname{diag}([1-\frac{1}{\rho_1}]_+,\ldots,[1-\frac{1}{\rho_r}]_+,\mathbf{0}_{N-r}\Big)\mathbf{V}^{\dagger}\mathbf{\Lambda}^{-1/2}. \tag{3.7}$$

As a closer look at (3.7), let s be the largest number such that $1 - \frac{1}{\rho_s} > 0$. Then, $\mathbf{S}$ is equivalently written as

$$\mathbf{S} = \mathbf{\Lambda}^{-1/2}\mathbf{V}\,\mathrm{diag}(1 - \frac{1}{\rho_1}, \dots, 1 - \frac{1}{\rho_s}, \mathbf{0}_{N-s})\mathbf{V}^{\dagger}\mathbf{\Lambda}^{-1/2}. \tag{3.8}$$

Since $\mathbf{V}\mathbf{V}^{\dagger} = \mathbf{I}$, $\mathbf{S}$ is further simplified as

$$\mathbf{S} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1/2}\mathbf{V}\,\mathrm{diag}(\frac{1}{\rho_1}, \dots, \frac{1}{\rho_s}, \mathbf{1}_{N-s})\mathbf{V}^{\dagger}\mathbf{\Lambda}^{-1/2}. \tag{3.9}$$

We can prove that at the optimum, $[\mathbf{S}]_{i,i} = P_i$ for all $i = 1, \dots, N$. Thus, in order to find optimal $\mathbf{S}$, we need to find $\mathbf{\Lambda}$ such that

$$\left[\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1/2}\mathbf{V}\,\mathrm{diag}(\frac{1}{\rho_1}, \dots, \frac{1}{\rho_s}, \mathbf{1}_{N-s})\mathbf{V}^{\dagger}\mathbf{\Lambda}^{-1/2}\right]_{i,i} = P_i. \tag{3.10}$$

Since $\mathbf{\Lambda}$ is a diagonal matrix, (3.10) equals to

$$\Big(\mathbf{I} - \Big[\mathbf{V}\Big(\mathrm{diag}(\frac{1}{\rho_1}, \dots, \frac{1}{\rho_s}, \mathbf{1}_{N-s}\Big)\mathbf{V}^{\dagger}\Big]_{i,i}\Big)\Big[\mathbf{\Lambda}^{-1}\Big]_{i,i} = P_i. \tag{3.11}$$

Let $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}) = \Big[\mathbf{V}\Big(\mathrm{diag}(\frac{1}{\rho_1}, \dots, \frac{1}{\rho_s}, \mathbf{1}_{N-s}\Big)\mathbf{V}^{\dagger}\Big]$. Then, we can rewrite (3.11) in the form of a nonlinear system as

$$\tilde{\boldsymbol{\lambda}} - \mathrm{diag}(\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}})) \odot \tilde{\boldsymbol{\lambda}} = \mathbf{p} \tag{3.12}$$

where $\tilde{\boldsymbol{\lambda}} \triangleq [\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_N^{-1}]^T$, $\mathbf{p} \triangleq [P_1, P_2, \dots, P_N]^T$. It is easy to see that

$$\Big[\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}})\Big]_{i,i} = \textstyle\sum_{j=1}^{N} \tilde{\rho}_j |v_{i,j}|^2 \tag{3.13}$$

where $\tilde{\rho}_j = \frac{1}{\rho_j} < 1$ for $1 \le j \le s$, and $\tilde{\rho}_j = 1$ for $s < j \le N$. Since $\sum_{j=1}^{N} |v_{i,j}|^2 = 1$, it holds that $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}) \prec \mathbf{I}$ for all $\tilde{\boldsymbol{\lambda}} \succ \mathbf{0}$, and (3.12) is thus well defined. Unfortunately, there is no analytical solution to (3.12), mostly due to the fact that $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}})$ is a nonlinear function of $\tilde{\boldsymbol{\lambda}}$. However, (3.12) already suggests a way to find $\tilde{\boldsymbol{\lambda}}$ iteratively as follows

$$\tilde{\boldsymbol{\lambda}}_{n+1} = \mathbf{p} + \mathrm{diag}(\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}_n)) \odot \tilde{\boldsymbol{\lambda}}_n \triangleq \mathfrak{J}(\tilde{\boldsymbol{\lambda}}_n). \tag{3.14}$$

In fact, (3.14) is written in a fixed-point iteration form and its convergence is stated in the following lemma.

Lemma 3.1. *The iterations in* (3.14) *converge to the unique fixed-point of* (3.12)*, thereby solving* (3.2)*.*

The proof of Lemma 3.1 is provided in Appendix A.1. The key is to show that $\mathfrak{J}(\mathbf{x})$ is a standard interference function.

We can see that the fixed-point algorithm based on (3.14) requires iteratively performing EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$. A simple way is to treat it as a new matrix at each iteration, but this is not computationally efficient. Exploiting the fact that the channel matrix $\mathbf{H}$ remains the same during the whole iterative process, we present a way to compute the EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$ more efficiently. To this end, let $\mathbf{H} = \mathbf{G}\mathbf{R}$, where $\mathbf{G}$ is unitary and $\mathbf{R}$ is upper triangular, be a QR factorization of $\mathbf{H}$. Then we can write $\mathbf{H}\mathbf{\Lambda}^{-1/2} = (\mathbf{G}\mathbf{R})\mathbf{\Lambda}^{-1/2} = \mathbf{G}(\mathbf{R}\mathbf{\Lambda}^{-1/2})$. Since $\mathbf{\Lambda}$ is diagonal, $\mathbf{R}\mathbf{\Lambda}^{-1/2}$ is also an upper triangular matrix. Now let $\mathbf{R}\mathbf{\Lambda}^{-1/2} = \mathbf{U}\bar{\mathbf{\Sigma}}\mathbf{V}^{\dagger}$ be the SVD of $\mathbf{R}\mathbf{\Lambda}^{-1/2}$. Then the EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{\Lambda}^{-1/2}$ is simply given by $\mathbf{V}\bar{\mathbf{\Sigma}}^{2}\mathbf{V}^{\dagger}$. We remark that SVD computation for an upper triangular matrix is much cheaper than for a full matrix [41, p. 492], which leads to a huge reduction in the computation cost of the proposed algorithm. The proposed algorithm based on fixed-point iteration is outlined in Algorithm 3.1.

Algorithm 3.1: Proposed Solution Based on Fixed-point Iteration.

```
Input: $\mathbf{\Lambda}_0$ diagonal matrix of positive elements, $\epsilon > 0$.
Set $n := 0$ and $\tau = 1 + \epsilon$.
Perform QR decomposition of $\mathbf{H}$: $\mathbf{H} = \mathbf{G}\mathbf{R}$, where $\mathbf{G}$ is a unitary matrix and
  $\mathbf{R}$ is an upper triangular matrix.
while $\tau > \epsilon$ do
    Perform the SVD of $\mathbf{R}\mathbf{\Lambda}_n{}^{-1/2}$: $\mathbf{R}\mathbf{\Lambda}_n{}^{-1/2} = \mathbf{U}_n\bar{\mathbf{\Sigma}}_n\mathbf{V}_n^{\dagger}$, where $\bar{\mathbf{\Sigma}}_n$ is diagonal.
      Let $\rho_i = \sigma_i^2$, $i = 1, \ldots, r$, where $\sigma_i$ is the $i$th non-zero entry of $\bar{\mathbf{\Sigma}}_n$ and
      $r = \operatorname{rank}(\mathbf{R}\mathbf{\Lambda}_n^{-1/2})$.
    $\tilde{\mathbf{\Sigma}}_n := \operatorname{diag}([1-\rho_1^{-1}]_+, ..., [1-\rho_r^{-1}]_+, \mathbf{0}_{N-r})$.
    $\boldsymbol{\Psi}_n := \mathbf{V}_n(\mathbf{I} - \tilde{\mathbf{\Sigma}}_n)\mathbf{V}_n^{\dagger}$.
    $\mathbf{S}_n := \mathbf{\Lambda}_n^{-1} - \mathbf{\Lambda}_n^{-1/2}\boldsymbol{\Psi}_n\mathbf{\Lambda}_n^{-1/2}$.
    $\tau = \sum_{i=1}^{N}[\mathbf{\Lambda}_n]_{i,i}|[\mathbf{S}_n - \mathbf{P}|]_{i,i}$.
    $\tilde{\boldsymbol{\lambda}}_{n+1} = \mathbf{p} + \operatorname{diag}(\boldsymbol{\Psi}_n) \odot \tilde{\boldsymbol{\lambda}}_n$.
    $\mathbf{\Lambda}_{n+1} = (\operatorname{diag}\tilde{\boldsymbol{\lambda}}_{n+1})^{-1}$.
    $n := n + 1$.
end
Output: $\mathbf{S}_n$.
```

Remark 1. To solve (3.2), the work of [1, 3] proposed two different algorithms for two corresponding cases: $M \geq N$ and $M < N$. Moreover, these algorithms are dedicated to full-rank channel matrices. In this regard, Algorithm 3.1 is more universal in the sense that it is applicable to channel matrices of any dimension and rank-deficiency. Another issue of the methods presented in [1, 3] is that a complete analytical proof of their convergence is sill missing. On the contrary, Algorithm 3.1 is provably convergent from an arbitrary starting point $\tilde{\boldsymbol{\lambda}}_0 > 0$. Moreover, analytical and numerical results demonstrate Algorithm 3.1 achieves lower complexity, compared to the ones in [1,3].

3.1.2.2. Alternating Optimization

The second proposed iterative method exploits an interesting result from the duality between BC and MAC [42, 43]. In fact it is shown that (3.2) is equivalent to the following minimax optimization problem [10]

$$\begin{array}{rl} \underset{\mathbf{Q}\succeq\mathbf{0}}{\min}\ \underset{\bar{\mathbf{S}}\succeq\mathbf{0}}{\max} & \log\frac{|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|}{|\mathbf{Q}|}\triangleq f(\mathbf{Q},\bar{\mathbf{S}}) \\ \text{subject to} & \operatorname{tr}(\bar{\mathbf{S}})\leq P, \operatorname{tr}(\mathbf{QP})\leq P; \mathbf{Q}:\mathsf{diagonal} \end{array} \tag{3.15}$$

where $P \triangleq \sum_{i=1}^{N} P_i$. In the above formulation we define $\log|\mathbf{Q}| = -\infty$ if $\mathbf{Q}$ is singular. For the development of the second proposed method, without loss of optimality, we assume that $||\mathbf{h}_i||_2 > 0$ where $\mathbf{h}_i$ is the ith column of the channel matrix $\mathbf{H}$, which is normally the case in practice. If $\mathbf{h}_i$ happens to be all-zero vector, the ith transmit antenna can be dropped to obtain a reduced channel matrix, to which the following proposed method is applied. As a result of Appendix A.2, the relationship between (3.2) and (3.15) is stated in the following fact.

Fact 1. *There exists a saddle point* $(\bar{\mathbf{S}}^{\star}, \mathbf{Q}^{\star})$ *for* (3.15) *such that* $\mathbf{Q}^{\star} \succ \mathbf{0}$. *Denote* $\mathbf{U\Sigma V}^{\dagger}$ *to be an SVD of* $\mathbf{H}(\mathbf{Q}^{\star})^{-1/2}$ *where* $\mathbf{\Sigma}$ *is square and diagonal. Then, the optimal solution* $\mathbf{S}^{\star}$ *to* (3.2) *can be found as*

$$\mathbf{S}^{\star} = (\mathbf{Q}^{\star})^{-1/2}\mathbf{V}\mathbf{U}^{\dagger}\bar{\mathbf{S}}^{\star}\mathbf{U}\mathbf{V}^{\dagger}(\mathbf{Q}^{\star})^{-1/2}. \tag{3.16}$$

The above result is in fact a special case of the MAC-to-BC transformation presented in [43] when applying to a single user system.

It is trivial to see that the optimality of (3.15) is not affected if the inequalities are made to be equality. To appreciate the idea behind the second proposed method, let us define $\mathcal{Q} \triangleq \{\mathbf{Q}|\mathbf{Q}:\mathsf{diagonal}, \mathbf{Q}\succeq\mathbf{0}, \operatorname{tr}(\mathbf{QP}) = P\}$ and $\mathcal{S} = \{\bar{\mathbf{S}}|\bar{\mathbf{S}}\succeq\mathbf{0}, \operatorname{tr}(\bar{\mathbf{S}}) = P\}$. Now, (3.15) can be rewritten in an abstract form as

$$\underset{\mathbf{Q}\in\mathcal{Q}}{\min}\ \underset{\bar{\mathbf{S}}\in\mathcal{S}}{\max}\ f(\mathbf{Q},\bar{\mathbf{S}}). \tag{3.17}$$

We note that $f(\mathbf{Q},\bar{\mathbf{S}})$ is concave with $\bar{\mathbf{S}}$, and convex with $\mathbf{Q}$, and twice differentiable. Thus a saddle point $(\mathbf{Q}^{*},\bar{\mathbf{S}}^{*})$ exist for (3.17) and it holds that

$$f(\mathbf{Q}^{*},\bar{\mathbf{S}}) \leq\ f(\mathbf{Q}^{*},\bar{\mathbf{S}}^{*}) \leq\ f(\mathbf{Q},\bar{\mathbf{S}}^{*}). \tag{3.18}$$

We can see that solving (3.15) boils down to finding a saddle point for (3.17). In fact, this interpretation was used in the interior-point method proposed in [10]. The minimax formulation in (3.17) also suggests a way to find a saddle point by alternatively optimizing $\mathbf{Q}$ and $\bar{\mathbf{S}}$. This method was also mentioned in [10] but note that it is not provably convergent. In fact we have very often observed that this pure method will suffer a ping-pong effect, and thus fail to converge to an optimal solution of (3.17) (cf. Fig. 3.2 for an example on this).

In the second proposed algorithm, we still capitalize on the idea of AO, but do it in a novel way to ensure strict monotonicity. Suppose at the nth iteration, we have obtained $\mathbf{Q}_n$. Then $\bar{\mathbf{S}}_n$ is found as the solution to the following problem

$$\begin{array}{rl}\text{maximize} & \log|\mathbf{Q}_n + \mathbf{H}^\dagger\bar{\mathbf{S}}\mathbf{H}|\\ \text{subject to} & \text{tr}(\bar{\mathbf{S}}) = P; \bar{\mathbf{S}} \succeq \mathbf{0}.\end{array} \tag{3.19}$$

It is well known that the above problem admits the solution based on water-filling algorithm [6,7]. More explicitly, let $\mathbf{U}_n\mathbf{\Sigma}_n\mathbf{U}_n^\dagger = \mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$ be the EVD of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, where $\mathbf{\Sigma}_n = \text{diag}(\rho_1,\ \rho_2,\ldots,\rho_r)$ is a matrix of non-negative eigenvalues of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, and $r = \text{rank}(\mathbf{H}\mathbf{Q}_n^{-1/2})$. Then, $\bar{\mathbf{S}}_n$ can be found as

$$\bar{\mathbf{S}}_n = \mathbf{U}_n\hat{\mathbf{\Sigma}}_n\mathbf{U}_n^\dagger \tag{3.20}$$

where $\hat{\mathbf{\Sigma}}_n = \text{diag}([\mu - \frac{1}{\rho_1}]_+, [\mu - \frac{1}{\rho_2}]_+, \ldots, [\mu - \frac{1}{\rho_r}]_+)$ and μ is the water-level, which is chosen to satisfy the total power constraint

$$\sum_{i=1}^{r}[\mu - \frac{1}{\rho_i}]_+ = P. \tag{3.21}$$

Note that $\bar{\mathbf{S}}_n$ in (3.20) is the *unique* solution to (3.19). To find $\mathbf{Q}_{n+1}$, we invoke the following inequality, which results from the concavity of the logdet function,

$$\log|\mathbf{Q} + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H}| \leq \log|\mathbf{\Phi}_n| + \text{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q} - \mathbf{Q}_n\big)\Big) \tag{3.22}$$

where $\mathbf{\Phi}_n = \mathbf{Q}_n + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H}$. In the second proposed algorithm, $\mathbf{Q}_{n+1}$ is found to optimize the upper bound of (3.15), i.e., $\mathbf{Q}_{n+1}$ is the solution to the following problem

$$\begin{array}{rl}\underset{\mathbf{Q}\succeq\mathbf{0}}{\text{minimize}} & \text{tr}\Big(\mathbf{\Phi}_n^{-1}\mathbf{Q}\Big) - \log|\mathbf{Q}|\\ \text{subject to} & \text{tr}(\mathbf{Q}\mathbf{P}) = P; \mathbf{Q} : \mathsf{diagonal}.\end{array} \tag{3.23}$$

We will see shortly that in the second proposed iterative algorithm, $\mathbf{Q}_n \succ \mathbf{0}$ for all iterations n, and thus $\mathbf{\Phi}_n^{-1}$ is well defined. Also, it is worth mentioning that the gradient of the objective in (3.23) with respect to $\mathbf{Q}$ is identical to that of the original objective in (3.15) when $\mathbf{Q} = \mathbf{Q}_n$. This is essentially to ensure that the first order optimality conditions of the original problem are preserved even with the use of an upper bound. To clarify this point let us write the partial derivative of $f(\mathbf{Q}, \bar{\mathbf{S}}_n)$ with respect to q_i as

$$\partial_{\mathbf{q}_i} f(\mathbf{Q}, \bar{\mathbf{S}}_n) = [(\mathbf{Q} + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H})^{-1}]_{i,i} - q_i^{-1} \tag{3.24}$$

The partial derivative of the upper bound with respect to q_i obtained at iteration n of the proposed method is

$$[\mathbf{\Phi}_n^{-1}]_{i,i} - q_i^{-1} = [(\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H})^{-1}]_{i,i} - q_i^{-1} \tag{3.25}$$

It is now clear that the partial derivatives of the original objective and the upper bound in (23) with respect to any q_i are the same when $\mathbf{Q} = \mathbf{Q}_n$. Thus when the iterates $\{(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ converge, they will satisfy the KKT conditions of the original problem, though an upper bound of the objective is minimized.

Before proceeding further we provide a remark regarding the use of the upper bound in (3.23) for updating $\mathbf{Q}$. First it is a well-known fact that if we simply alternate optimization of $\mathbf{Q}$ and $\bar{\mathbf{S}}$ as done in a pure AO method, then convergence to a saddle point is not guaranteed as monotonic convergence of the objective is not achieved. In the second proposed algorithm, the key point is to make the objective decrease after each cyclic update of $\mathbf{Q}$ and $\bar{\mathbf{S}}$. For this purpose we minimize an upper bound of the objective for updating $\mathbf{Q}$. In fact, this idea is largely inspired by successive convex approximation (SCA) principle for nonconvex optimization problems [44]. Roughly speaking, for SCA-based methods, the nonconvex objective is approximated by a convex upper bound in each iteration, which ensures monotonic decrease of the sequence of the objectives. However, the main challenge is that SCA only concerns minimization (or equivalently maximization) problems, while our considered problem is a minimax program. As such the proof for the convergence of SCA-based algorithms is not applicable to Algorithm 3.2, as shown in Appendix A.2.

Since $\mathbf{Q}$ in (3.23) is in fact diagonal, i.e., $\mathbf{Q} = \operatorname{diag}(\mathbf{q})$, we can rewrite (3.23) as

$$\begin{array}{rl} \underset{\mathbf{q}\geq 0}{\text{minimize}} & \sum_{i=1}^{N} \phi_{n,i} q_i - \log q_i \\ \text{subject to} & \sum_{i=1}^{N} P_i q_i = P \end{array} \tag{3.26}$$

where $\phi_{n,i} = \left[\mathbf{\Phi}_n^{-1}\right]_{i,i}$. Interestingly, the above problem also has a water-filling-like solution as

$$q_i = \frac{1}{\phi_{n,i} + \gamma P_i} > 0 \tag{3.27}$$

where $\gamma \geq 0$ is the solution of the equation

$$\sum_{i=1}^{N} \frac{P_i}{\phi_{n,i} + \gamma P_i} = P. \tag{3.28}$$

From (3.27) it is clear that $\mathbf{Q}_n \succ \mathbf{0}$ for all n and thus $\mathbf{\Phi}_n^{-1}$ exists as mentioned below (3.23). Further, from the definition of $\mathbf{\Phi}_n$, it holds that $\phi_{n,i} = \left[\mathbf{\Phi}_n^{-1}\right]_{i,i} \leq \left[\mathbf{Q}_n^{-1}\right]_{i,i}$. As the result, we obtain $\sum_{i=1}^{N} \frac{P_i}{\phi_{n,i}} \geq \operatorname{tr}(\mathbf{Q}_n \mathbf{P}) = P$, where the equality holds since $\mathbf{Q}_n$ is the solution to (3.23) in the previous iteration. Note that the left hand side of (3.28) is decreasing with γ, and thus (3.28) always has a unique solution, which can be found efficiently, e.g., by the bisection or Newton method. The second proposed algorithm based on AO is summarized in Algorithm 3.2. The main point of Algorithm 3.2 is the use of the inequality in (3.22) to optimize $\mathbf{Q}$ for a given $\bar{\mathbf{S}}$. This step will eliminate the ping-pong effect mentioned above and ensure the objective sequence is strictly decreasing. The convergence proof of Algorithm 3.2 is provided in Appendix A.2.

Algorithm 3.2: Proposed Solution Based on Alternating Optimization.

```
Input: $\mathbf{Q}_0$ is feasible to $\mathcal{Q}$, and $\epsilon > 0$.
1 Initialize $n := 0$, $\tau = 1 + \epsilon$.
2 while $\tau > \epsilon$ do
3     Apply water-filling algorithm (i.e., (3.20) and (3.21)) to compute
        $\bar{\mathbf{S}}_n = \underset{\bar{\mathbf{S}} \in \mathcal{S}}{\arg\max}\ \log|\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}} \mathbf{H}|$.
4     For $n \geq 1$, let $\tau = |f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) - f(\mathbf{Q}_{n-1}, \bar{\mathbf{S}}_{n-1})|$.
5     $\mathbf{\Phi}_n^{-1} := (\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H})^{-1}$.
6     Find $\mathbf{Q}_{n+1} = \underset{\mathbf{Q} \in \mathcal{Q}}{\arg\min}\ \operatorname{tr}\left(\mathbf{\Phi}_n^{-1}\mathbf{Q}\right) - \log|\mathbf{Q}|$, using (3.27) and (3.28).
7     $n := n + 1$.
8 end
Output: $\bar{\mathbf{S}}_n$ and use (3.16) to compute optimal $\mathbf{S}$.
```

We note that the error tolerance τ in line 4 of Algorithm 3.2 is only computed for $n \geq 1$. We remark that line 3 in Algorithm 3.2 involves the EVD of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, which can be computed similarly as done in Algorithm 3.1 to reduce the overall complexity. Specifically let $\mathbf{G}\mathbf{R} = \mathbf{H}$ be the QR decomposition of $\mathbf{H}$. Next we compute the SVD of the upper triangular matrix $\mathbf{R}\mathbf{Q}_n^{-1/2}$ as $\tilde{\mathbf{U}}_n \tilde{\mathbf{\Sigma}}_n \tilde{\mathbf{V}}_n^\dagger = \mathbf{R}\mathbf{Q}_n^{-1/2}$. Then the EVD of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$ is simply given by $\mathbf{U}_n \mathbf{\Sigma}_n \mathbf{U}_n^\dagger = \mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, where $\mathbf{U}_n = \mathbf{G}\tilde{\mathbf{U}}_n$ and $\mathbf{\Sigma}_n = \tilde{\mathbf{\Sigma}}_n^2$. Moreover, we note that $\bar{\mathbf{S}}_n$ needs not be computed explicitly as in (3.20) for each iteration. The reason is that the diagonal elements of $\mathbf{\Phi}_n^{-1}$ in line 5 can be found efficiently from the SVD of $\mathbf{R}\mathbf{Q}_n^{-1/2}$ as shown in the following.

Using the matrix-inversion lemma, we can write $\mathbf{\Phi}_n^{-1} = \mathbf{Q}_n^{-1/2}(\mathbf{I}+\mathbf{Q}_n^{-1/2}\mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H}\mathbf{Q}_n^{-1/2})^{-1}$ $\mathbf{Q}_n^{-1/2} = \mathbf{Q}_n^{-1/2}(\mathbf{I}+\tilde{\mathbf{V}}_n\hat{\mathbf{\Sigma}}_n\tilde{\mathbf{V}}_n^\dagger)^{-1}\mathbf{Q}_n^{-1/2}$, where the latter equality holds due to (3.20). Now let $\dot{\mathbf{\Sigma}}_n$ be the diagonal matrix containing all *strictly positive* entries of $\hat{\mathbf{\Sigma}}_n$, and $\dot{\mathbf{V}}_n$ be the corresponding singular vectors. Then we can write $(\mathbf{I} + \tilde{\mathbf{V}}_n\hat{\mathbf{\Sigma}}_n\tilde{\mathbf{V}}_n^\dagger)^{-1} = (\mathbf{I} + \dot{\mathbf{V}}_n\dot{\mathbf{\Sigma}}_n\dot{\mathbf{V}}_n^\dagger)^{-1} \overset{(a)}{=} \mathbf{I} - \dot{\mathbf{V}}_n\left(\dot{\mathbf{\Sigma}}_n^{-1} + \dot{\mathbf{V}}_n^\dagger\dot{\mathbf{V}}_n\right)^{-1}\dot{\mathbf{V}}_n^\dagger \overset{(b)}{=} \mathbf{I} - \dot{\mathbf{V}}_n\left(\dot{\mathbf{\Sigma}}_n^{-1} + \mathbf{I}\right)^{-1}\dot{\mathbf{V}}_n^\dagger$, where (a) is due to the matrix inversion lemma, and (b) holds true since $\dot{\mathbf{V}}_n^\dagger\dot{\mathbf{V}}_n = \mathbf{I}$. In summary, we have $\mathbf{\Phi}_n^{-1} = \mathbf{Q}_n^{-1} - \mathbf{Q}_n^{-1/2}\dot{\mathbf{V}}_n\left(\dot{\mathbf{\Sigma}}_n^{-1} + \mathbf{I}\right)^{-1}\dot{\mathbf{V}}_n^\dagger\mathbf{Q}_n^{-1/2}$. Since $\dot{\mathbf{\Sigma}}_n^{-1} + \mathbf{I}$ is diagonal, its inversion can be computed easily. It is also clear that, to compute $\mathbf{\Phi}_n^{-1}$, what we need is only $\tilde{\mathbf{\Sigma}}_n$ and $\tilde{\mathbf{V}}_n$ from the SVD of $\mathbf{R}\mathbf{Q}_n^{-1/2}$.

3.1.3. Complexity Analysis

In this section, we analyze the complexity of the proposed algorithms in the preceding section, counted as the number of flops. Although flop counting is a crude way to measure the actual computational complexity, it somewhat captures the order of the computation load. To this end we first assume $M \geq N$ (i.e, more receive than transmit antennas) and summarize the relevant results presented in [41] and [45] as follows. QR decomposition of an $M \times N$ matrix using Householder transformation requires $2N^2(M - N/3)$ flops for only $\mathbf{R}$, and $4M^2N - 2MN^2 + \frac{2}{3}N^3$ flops for both $\mathbf{R}$ and $\mathbf{Q}$. The computation of SVD of a full $M \times N$ matrix needs $4M^2N + 8MN^2 + 9N^3$ flops for $(\mathbf{\Sigma}, \mathbf{V}, \mathbf{U})$, $4MN^2 + 8N^3$ flops for $(\mathbf{\Sigma}, \mathbf{V})$, and $4M^2N - 8MN^2$ flops for $(\mathbf{\Sigma}, \mathbf{U})$ while that of an upper triangular matrix requires $4M^2N + 22N^3, 2M^2N + 11N^3, 4M^2N + 13N^3$, respectively. The number of flops for the water-filling algorithm with N eigenmodes is $2N^2 + 6N$. Inversion of an $N \times N$ symmetric matrix requires N^3 flops. Note that these flop counts are for a real matrix. For complex matrices, we simply treat every operation as a complex multiplication which is equal to 6 real flops [45, 46]. That is, QR decomposition of an $M \times N$ complex matrix requires $4N^2(3M - N)$ flops.

In the complexity analysis presented in the following, we only consider the main operations having the most significant complexity and ignore those contributing negligibly to the overall complexity (e.g., subtraction or addition).

3.1.3.1. Complexity of Algorithm 3.1

Algorithm 3.1 performs a QR decomposition (cf. line 2) at the first iteration and only $\mathbf{R}$ is needed, which requires $4N^2(3M - N)$ flops as explained above. In the subsequent iterations, Algorithm 3.1 involves an SVD of an upper triangular matrix (line 4), in which only $(\mathbf{\Sigma}, \mathbf{V})$ needs to be computed. This step takes $6(2MN^2 + 11N^3)$ flops. We note that other operations in Algorithm 3.1 have minor complexity, compared to QR decomposition and SVD, and thus are neglected.

3.1.3.2. Complexity of Algorithm 3.2

To reduce the complexity, Algorithm 3.2 performs a full QR decomposition in the first iteration, which takes $6(4M^2N - 2MN^2 + \frac{2}{3}N^3)$ flops. Then, the complexity incurred in line 3 of Algorithm 3.2 is due to finding $(\tilde{\mathbf{\Sigma}}_n, \tilde{\mathbf{V}}_n^\dagger)$ in the SVD of the upper triangular matrix $\mathbf{R}\mathbf{Q}_n^{-1/2}$. The flop count of the step is $6(2M^2N + 11N^3)$. The water-filling algorithm to find positive eigenmodes that meet the sum power constraint needs $6(2N^2 + 6N)$ flops. The complexity of line 5 (i.e., computing the diagonal elements of $\mathbf{\Phi}_n^{-1}$) and that of line 6 are lower compared to the remaining steps and thus can be ignored.

3.1.3.3. Complexity of the mode-dropping algorithm in [1, 3]

For comparison purpose we now present the complexity of the so-called mode-dropping algorithm proposed in [1, 3]. Specifically, this method requires an SVD of a *full* $M \times N$ matrix, in the first iteration, for which the flop count is $6(4M^2N + 8MN^2 + 9N^3)$. From the second iteration, the most complex operation of the mode-dropping algorithm is to compute an EVD which requires $6(4MN^2 + 8N^3)$ flops.

Basically, the complexity of the proposed algorithms for the case $N > M$ can be obtained by simply switching N and M in the above analytical expressions. However, for the mode-dropping algorithm, two additional matrix inversions need to be performed, resulting in an increased complexity. The per-iteration complexity comparison (after the first iteration) is summarized in Table 3.1, where the bold text refers to the algorithm with the lowest complexity, i.e., Algorithm 3.1 and 3.2. However, the total complexity of an iterative algorithm heavily depends on the number of iterations required to converge. This issue is evaluated for various numerical experiments in Section 3.4.

3.1.3.4. Complexity of interior-point methods

As mentioned earlier, problem (3.2) can be reformulated as an SDP and then solved by general-purpose optimization packages. These optimization tools are normally based on primal-dual path-following methods to solve a convex model. However, the per-iteration complexity of such interior-point solvers is $\mathcal{O}(N^6)$ [27,47], which is prohibitively high for large-scale MIMO systems. A numerical complexity comparison is shown in Fig. 3.5 to further demonstrate this point.

Table 3.1.: Per-iteration Complexity Comparison. The table is adapted from Table I in [2] under © 2019 IEEE.

Algorithms	$M \geq N$	$M < N$
Mode-dropping [1,3]	$6(4MN^2 + 8N^3)$	$6(4NM^2 + 8M^3)$ $+12(N - M)^3$
Algorithm 3.1	$\boldsymbol{6(2MN^2 + 11N^3)}$	$\boldsymbol{6(2NM^2 + 11M^3)}$
Algorithm 3.2	$\boldsymbol{6(2MN^2 + 11N^3)}$	$\boldsymbol{6(2NM^2 + 11M^3)}$

3.2. Weighted Sum Rate with ZF

3.2.1. System Model

Consider a K-user MIMO BC where the base station and each user have N and M_k antennas, respectively. Let $\mathbf{H}_k$ be the channel matrix for user k. Then, the received signal at user k is given as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{s}_k + \textstyle\sum_{j\neq k}\mathbf{H}_k\mathbf{s}_j + \mathbf{z}_k \tag{3.29}$$

where $\mathbf{s}_k$ is the downlink signal and $\mathbf{z}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ refers to the noise for the kth user. For linear ZF precoding, we can express $\mathbf{s}_k$ as $\mathbf{s}_k = \mathbf{R}_k\mathbf{x}_k$, where $\mathbf{R}_k$ and $\mathbf{x}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ denote the precoding matrix and information-bearing signal, respectively. For user k, the interference from other users in the system is suppressed by designing $\mathbf{R}_k$ such that $\mathbf{H}_j\mathbf{R}_k = 0$ for all $j \neq k$. The weighted sum rate maximization (WSRMax) problem for ZF precoding with PAPC is formulated as [48]

$$\begin{array}{ll} \underset{\{\mathbf{X}_k\succeq\mathbf{0}\}}{\text{maximize}} & \sum_{k=1}^{K} w_k \log|\mathbf{I} + \mathbf{H}_k\mathbf{X}_k\mathbf{H}_k^H| \\ \text{subject to} & \mathbf{H}_j\mathbf{X}_k\mathbf{H}_j^H = 0, \ \forall j \neq k \\ & \sum_{k=1}^{K}[\mathbf{X}_k]_{i,i} \leq P_i, \ i = 1, 2, \ldots, N \end{array} \tag{3.30}$$

where $\mathbf{X}_k = \mathbb{E}[\mathbf{s}_k\mathbf{s}_k^H] = \mathbf{R}_k\mathbf{R}_k^H$ is the input covariance matrix for user k, P_i is the power constraint on antenna i, and w_k is the positive weighting factor assigned to the kth user. In the above formulation we have omitted the rank constraint $\text{rank}(\mathbf{X}_k) \leq M_k$ but this step does not affect the optimality as proved in [48]. We also remark that this rank constraint will be automatically satisfied the proposed solution presented next.

3.2.2. Proposed Algorithm

In this section, we derive an efficient algorithm to solve (3.30) using minimax duality, AO, and SCA. Assuming that $N > \sum M_k - \min\{M_k\}$, let $\check{\mathbf{H}}_k$ be the channel matrix of all users, except for user k, i.e. $\check{\mathbf{H}}_k = [\mathbf{H}_1^H, \ldots \mathbf{H}_{k-1}^H, \mathbf{H}_{k+1}^H, \ldots \mathbf{H}_K^H]^H$, and $\mathbf{B}_k$ be a basis of the null space of $\check{\mathbf{H}}_k$. Then (3.30) reduces to the following problem

$$\begin{array}{ll} \underset{\{\tilde{\mathbf{X}}_k\succeq\mathbf{0}\}}{\text{maximize}} & \sum_{k=1}^{K} w_k \log|\mathbf{I} + \mathbf{H}_k\mathbf{B}_k\tilde{\mathbf{X}}_k\mathbf{B}_k^H\mathbf{H}_k^H| \\ \text{subject to} & \sum_{k=1}^{K}[\mathbf{B}_k\tilde{\mathbf{X}}_k\mathbf{B}_k^H]_{i,i} \leq P_i, i = 1, \ldots, N. \end{array} \tag{3.31}$$

For the special case of sum rate maximization (SRMax) problem (i.e., $w_1 = w_2 = \cdots w_K$), (3.31) becomes a MAXDET program as mentioned in [21]. We further note that for this special case, (3.31) can be recast as a semidefinite program. For the general case of WSRMax problem, the optimization package SDPT3 is a dedicated solver. However, solving (3.31) by generic convex solvers is not practically appealing for a large number of antennas N and/or a large number of users K. A closed-form solution for (3.31) was proposed in [48], but it was found by leveraging the subgradient method whose convergence rate is typically slow.

To overcome the aforementioned drawbacks, by extending Theorem 2 of [30], we first transform (3.31) into a minimax problem in the dual MAC as

$$\begin{array}{ll} \underset{\mathbf{\Lambda}\succeq\mathbf{0}}{\min} \ \underset{\{\bar{\mathbf{X}}_k\succeq\mathbf{0}\}}{\max} & \sum_{k=1}^{K} w_k \log\frac{|\mathbf{B}_k^H\mathbf{\Lambda}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H\bar{\mathbf{X}}_k\tilde{\mathbf{H}}_k|}{|\mathbf{B}_k^H\mathbf{\Lambda}\mathbf{B}_k|} \\ \text{subject to} & \sum_{k=1}^{K}\text{tr}(\bar{\mathbf{X}}_k) = P; \text{tr}(\mathbf{\Lambda}\mathbf{P}) = P, \mathbf{\Lambda} : \mathsf{diagonal} \end{array} \tag{3.32}$$

where $\tilde{\mathbf{H}}_k = \mathbf{H}_k\mathbf{B}_k$. Then the optimal solution $\tilde{\mathbf{X}}_k^*$ of (3.31) is given by

$$\tilde{\mathbf{X}}_k^* = (\mathbf{B}_k^H\mathbf{\Lambda}^*\mathbf{B}_k)^{-\frac{1}{2}}\mathbf{U}_k\mathbf{V}_k^H\bar{\mathbf{X}}_k^*\mathbf{V}_k\mathbf{U}_k^H(\mathbf{B}_k^H\mathbf{\Lambda}^*\mathbf{B}_k)^{-\frac{1}{2}} \tag{3.33}$$

where $\{\bar{\mathbf{X}}_k^*\}$ and $\mathbf{\Lambda}^*$ are a saddle-point of (3.32), and $\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^H$ is the economy-size singular value decomposition of $(\mathbf{B}_k^H\mathbf{\Lambda}^*\mathbf{B}_k)^{-1/2}\tilde{\mathbf{H}}_k^H$. A proof of this transformation is given in Appendix A.4.

The problem now is to find a saddle-point of (3.32). For a general minimax optimization, one may alternate between minimization and maximization but the convergence of such a method is not guaranteed. A more common approach to tackle (3.32) is based on Newton's method, e.g., [10]. However, the complexity of this method increases rapidly with the problem size. In the sequel, we show that (3.32) can be solved efficiently by combining AO and SCA to derive closed-form expressions.

Let $\{\bar{\mathbf{X}}_k^n\}$ be the optimal value of the following maximization in the nth iteration

$$\begin{aligned} \max \quad & \textstyle\sum_{k=1}^K w_k \log|\mathbf{B}_k^H\mathbf{\Lambda}^n\mathbf{B}_k + \tilde{\mathbf{H}}_k^H\bar{\mathbf{X}}_k\tilde{\mathbf{H}}_k| \\ \text{s.t.} \quad & \textstyle\sum_{k=1}^K \operatorname{tr}(\bar{\mathbf{X}}_k) = P; \bar{\mathbf{X}}_k \succeq \mathbf{0}, k = 1,\ldots,K \end{aligned} \tag{3.34}$$

Note that the above problem admits the water-filling solution which is skipped here for the sake of brevity.

Now, we turn our attention to the minimization of $\mathbf{\Lambda}$ for given $\{\bar{\mathbf{X}}_k^n\}$. To achieve monotonic convergence, instead of minimizing the objective of (3.32), we construct and then minimize an upper bound of it. This step is inspired by the concept of SCA which has received growing attention recently. To this end, we recall the following inequality which results from the concavity of the log-determinant function [47, p. 73]

$$\log|\mathbf{B}_k^H\mathbf{\Lambda}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H\bar{\mathbf{X}}_k^n\tilde{\mathbf{H}}_k| \leq \log|\mathbf{\Phi}_k^n| \operatorname{tr}\Big(\mathbf{B}_k\mathbf{\Phi}_k^{-n}\mathbf{B}_k^H\Big(\mathbf{\Lambda} - \mathbf{\Lambda}^n\Big)\Big) \tag{3.35}$$

where $\mathbf{\Phi}_k^n \triangleq \mathbf{B}_k^H\mathbf{\Lambda}^n\mathbf{B}_k + \tilde{\mathbf{H}}_k^H\bar{\mathbf{X}}_k^n\tilde{\mathbf{H}}_k$, and $\mathbf{\Phi}_k^{-n} \triangleq \Big(\mathbf{\Phi}_k^n\Big)^{-1}$. Thus, in the $(n+1)$th iteration of the proposed algorithm, $\mathbf{\Lambda}^{n+1}$ is the solution to the following problem

$$\begin{aligned} \min \quad & \textstyle\sum_{k=1}^K w_k\Big(\operatorname{tr}\Big(\mathbf{B}_k\mathbf{\Phi}_k^{-n}\mathbf{B}_k^H\mathbf{\Lambda}\Big) - \log|\mathbf{B}_k^H\mathbf{\Lambda}\mathbf{B}_k|\Big) \\ \text{s.t.} \quad & \operatorname{tr}(\mathbf{\Lambda}\mathbf{P}) = P, \mathbf{\Lambda}:\mathsf{diagonal}; \mathbf{\Lambda} \succeq \mathbf{0} \end{aligned} \tag{3.36}$$

We remark that the inequality in (4.55) is not entirely new. In fact it has been appeared in previous studies such as [14,15,49]. Our contributions in this regard are twofold. Firstly, the use of (4.55) allows us to analytically prove that the proposed algorithm converges monotonically to a saddle-point of (3.32). Secondly, we show that (4.56) can be solved by closed-form expressions as follows.

Since $\mathbf{\Lambda}$ is diagonal, (4.56) reduces to the following problem

$$\underset{\boldsymbol{\lambda}\in\Theta}{\text{minimize}}\ \boldsymbol{\alpha}^T\boldsymbol{\lambda} - \textstyle\sum_{k=1}^K w_k \log|\mathbf{B}_k^H \operatorname{diag}(\boldsymbol{\lambda})\mathbf{B}_k| \tag{3.37}$$

where $\boldsymbol{\alpha} = \sum_{k=1}^{K} w_k \Big(\mathrm{diag}(\mathbf{B}_k \boldsymbol{\Phi}_k^{-n} \mathbf{B}_k^H)\Big)$ and $\Theta \triangleq \{\mathbf{p}^T\boldsymbol{\lambda} = P; \boldsymbol{\lambda} \geq 0\}$. From the above, we observe that (i) Θ is a simplex, and (ii) projection onto a simplex can be computed by a water-filling-like algorithm as shown shortly. These observations lead to the proposed gradient projection method to solve (3.37), which is outlined in Algorithm 3.3.

Algorithm 3.3: The proposed gradient projection algorithm for solving (3.37).

```
Input: $\boldsymbol{\lambda}_0$ , $m = 0$, $\epsilon_1 > 0, \tau_1 = 1 + \epsilon_1$.
repeat
    Calculate the gradient
     $\tilde{\mathbf{g}}_m = -\nabla f(\boldsymbol{\lambda}_m) = \sum_{k=1}^{K} w_k \,\mathrm{diag}(\mathbf{B}_k(\mathbf{B}_k^H \,\mathrm{diag}(\boldsymbol{\lambda}_m)\mathbf{B}_k)^{-1}\mathbf{B}_k^H) - \boldsymbol{\alpha}$.
    Choose an appropriate positive scalar $\rho_m$ and create $\tilde{\boldsymbol{\lambda}}_m = \boldsymbol{\lambda}_m + \rho_m \tilde{\mathbf{g}}_m$.
    Project $\tilde{\boldsymbol{\lambda}}_m$ onto $\Theta$ to obtain $\bar{\boldsymbol{\lambda}}_m$.
    Choose appropriate step size $\nu_m$ using the Armijo rule [50] and set
     $\boldsymbol{\lambda}_{m+1} = \boldsymbol{\lambda}_m + \nu_m(\bar{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m)$.
    $m := m + 1$.
until $\tau_1 = |\nabla f(\boldsymbol{\lambda}_m)^T(\boldsymbol{\lambda}_{m+1} - \boldsymbol{\lambda}_m)| < \epsilon_1$;
Output: $\boldsymbol{\lambda}_m$ as the optimal solution to (3.37).
```

In Algorithm 3.3, the subscript m denotes the iteration index. The main operation of Algorithm 3.3 is the projection of $\tilde{\boldsymbol{\lambda}}_m$ onto Θ which can be formulated as

$$\begin{array}{rl} \text{minimize} & ||\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}_m||^2 \\ \text{subject to} & \mathbf{p}^T\boldsymbol{\lambda} = P; \boldsymbol{\lambda} \geq \mathbf{0}. \end{array} \tag{3.38}$$

It is easy to see that (3.38) can be solved efficiently by water-filling-like algorithm. Note that when an equal power constraint is considered, Θ becomes a canonical simplex for which more efficient algorithms for projection are available [51]. Moreover, Algorithm 3.3 can be easily modified into a conjugate gradient projection method. The overall algorithm to solve the WSRMax problem for ZF precoding with PAPC is summarized in Algorithm 3.4, and its convergence proof is provided in the Appendix. Note that the residual error τ_2 is only computed for $n \geq 1$.

3.2.3. Complexity Analysis

In this section, we provide the complexity analysis of the proposed algorithm in terms of the number of flops. The flop count for related operations is taken from [52] and [45]. For convenience, we assume that all the receivers have the same number of antennas i.e. $M_k = M$. To solve the SDP problem for K covariance matrices of $N \times N$ by the interior-point-based approach (e.g., [21]), the complexity is $\mathcal{O}(K^3N^6)$ [27, 47]. As explained earlier, Algorithm 3.4 performs the water-filling algorithm and eigenvalue decomposition to solve (4.54) , which needs $K(N - (K - 1)M)^3 + K(4(N - (K - 1)M)^2M - 8(N - (K - 1)M)M^2)$ flops [45]. To find $\boldsymbol{\Lambda}$,

Algorithm 3.4: Proposed algorithm for solving (3.31).

```
Input: $\mathbf{\Lambda} := \mathbf{\Lambda}^0$ , $n := 0$, $\epsilon_2 > 0, \tau_2 = 1 + \epsilon_2$
repeat
    Apply the water-filling algorithm to solve (4.54). Denote the optimal
     solution by $\{\bar{\mathbf{X}}_k^n\}$.
    For each $k$, set $\mathbf{\Phi}_k^n = (\mathbf{B}_k^H \mathbf{\Lambda}^n \mathbf{B}_k + \tilde{\mathbf{H}}_k^H \bar{\mathbf{X}}_k^n \tilde{\mathbf{H}}_k)$.
    Find $\mathbf{\Lambda}^{n+1}$ using Algorithm 3.3.
    $n := n + 1$.
until $\tau_2 = |f(\mathbf{\Lambda}^n, \bar{\mathbf{X}}^n) - f(\mathbf{\Lambda}^{n-1}, \bar{\mathbf{X}}^{n-1})| < \epsilon_2$;
Output: $\{\bar{\mathbf{X}}_k^n\}_{k=1}^K$ and apply the BC-MAC transformation to compute optimal
        $\{\tilde{\mathbf{X}}_k^n\}_{k=1}^K$.
```

Algorithm 3.3 requires $K(N-(K-1)M)^3$ flops for gradient computation (cf. line 2), while the complexity of the projection on a simplex (cf. line 4) and of other steps is negligible, and therefore is ignored. Thus, the total per-iteration complexity of Algorithm 3.4 is $\mathcal{O}(KN^3)$ flops. For the same problem, the subgradient-based method in [48] has a similar per-iteration complexity. However, the subgradient method is generally known for slow convergence, and thus potentially results in higher overall computation time that is investigated in the next section.

3.3. Capacity Region of a Gaussian MIMO BC

In this scenario we compute the capacity region of a MIMO BC. It was proved in [8] that the capacity region of a Gaussian MIMO BC is achieved by DPC. For a SPC, this problem was addressed in a number of previous studies [23, 24, 53]. The related research for PAPC is quite limited. Specifically, the capacity region can be characterized by solving the following weighted sum rate maximization

$$\begin{array}{ll} \underset{\{\mathbf{S}_k \succeq \mathbf{0}\}}{\text{maximize}} & \sum_{k=1}^K w_k \log |\frac{|\mathbf{I}+\mathbf{H}_k \sum_{i=1}^k \mathbf{S}_i \mathbf{H}_k^\dagger|}{|\mathbf{I}+\mathbf{H}_k \sum_{i=1}^{k-1} \mathbf{S}_i \mathbf{H}_k^\dagger|} \\ \text{subject to} & \sum_{k=1}^K [\mathbf{S}_k]_{i,i} \leq P_i, \ \forall i \end{array} \tag{3.39}$$

for different sets of the weights w_i. Without loss of generality, we assume that $0 < w_1 \leq w_2 \leq ... \leq w_K$ and $\sum_{k=1}^K w_k = 1$ in the following. Since (3.39) is nonconvex at hand, Algorithm 3.1 cannot be extended to solve it. Fortunately, it can be solved efficiently using the BC-MAC duality and alternating optimization as shown next. First, by the BC-MAC duality [10], (3.39) is equivalent to

$$\begin{array}{ll} \underset{\mathbf{Q}\succ\mathbf{0}}{\min} \ \underset{\{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}}{\max} & \sum_{k=1}^K \Delta_k \log |\mathbf{Q} + \sum_{i=k}^K \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i \mathbf{H}_i| - w_K \log |\mathbf{Q}| \\ \text{subject to} & \sum_{k=1}^K \operatorname{tr}(\bar{\mathbf{S}}_k) = P, \operatorname{tr}(\mathbf{Q}\mathbf{P}) = P, \mathbf{Q} : \mathsf{diagonal} \end{array} \tag{3.40}$$

where $\Delta_k = w_k - w_{k-1}$. Before proceeding further we remark for the same problem, an interior-point algorithm was proposed in [10]. The complexity of such a method does not scale favorably with the problem size, compared to our proposed approach presented in what follows, which is based on closed-form expressions.

Let $(\{\bar{\mathbf{S}}_k^n\}, \mathbf{Q}^n)$ denote the value of $(\{\bar{\mathbf{S}}_k\}, \mathbf{Q})$ after n iterations of the proposed method. In view of AO, $\{\bar{\mathbf{S}}_k^n\}$ is the solution to the following problem

$$\begin{array}{rl} \text{maximize} & \sum_{k=1}^{K} \Delta_k \log|\mathbf{Q} + \sum_{i=k}^{K} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i \mathbf{H}_i| \\ \text{subject to} & \sum_{k=1}^{K} \operatorname{tr}(\bar{\mathbf{S}}_k) = P; \{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}. \end{array} \tag{3.41}$$

Problem (3.41) can be solved by off-the-shelf convex solvers but it can be solved more efficiently by a CGP method. The motivation is that projection onto the feasible set of (3.41) can be reduced to projection onto a canonical simplex, as shown in Appendix A.3. Thus, a CGP method can be derived to find the optimal solution of (3.41) (The details are skipped due to the space limitation). We note that similar approaches were also presented in [24, 25].

For the important case of the sum capacity of the MIMO BC, (3.41) becomes

$$\begin{array}{rl} \text{maximize} & \sum_{k=1}^{K} \log|\mathbf{Q} + \sum_{i=k}^{K} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i \mathbf{H}_i| \\ \text{subject to} & \sum_{k=1}^{K} \operatorname{tr}(\bar{\mathbf{S}}_k) = P; \{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}. \end{array} \tag{3.42}$$

For the above specific problem, the sum power iterative water-filling algorithm proposed in [23] and dual decomposition based method in [40] are particularly efficient.

We now turn our attention to finding $\mathbf{Q}^{n+1}$, which can be done exactly the same as for the SU-MIMO case. Specifically, it holds that

$$\log|\mathbf{Q} + \textstyle\sum_{i=k}^{K} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i^n \mathbf{H}_i| \leq \log|\mathbf{\Phi}_k^n| + \operatorname{tr}\Big(\mathbf{\Phi}_k^{-n}\Big(\mathbf{Q} - \mathbf{Q}^n\Big)\Big) \tag{3.43}$$

where $\mathbf{\Phi}_k^n \triangleq \mathbf{Q}^n + \sum_{i=k}^{K} \mathbf{H}_i^{\dagger} \bar{\mathbf{S}}_i^n \mathbf{H}_i$. Thus, $\mathbf{Q}^{n+1}$ is the solution to the following problem

$$\begin{array}{rl} \text{minimize} & \sum_{k=1}^{K} \frac{\Delta_k}{w_K} \operatorname{tr}\Big(\mathbf{\Phi}_k^{-n} \mathbf{Q}\Big) - \log|\mathbf{Q}| \\ \text{subject to} & \operatorname{tr}(\mathbf{QP}) = P, \mathbf{Q} : \textsf{diagonal}; \mathbf{Q} \succeq \mathbf{0}. \end{array} \tag{3.44}$$

We note that the idea of using the upper bound in (3.44) to optimize $\mathbf{Q}$ follows exactly the same as that of Algorithm 3.2. The above problem has the same form as (3.23), and thus closed-form solution using (3.27) and (3.28) can be applied. The overall proposed algorithm to solve (3.39) is summarized in Algorithm 3.5. We can prove the convergence of Algorithm 3.5 using the same lines as those for Algorithm 3.2 and thus the details are omitted due to space limitation. Similar to Algorithm 3.2, τ is only calculated for $n \geq 1$.

3.4. Numerical Results

Algorithm 3.5: Proposed algorithm for the computation of the capacity region of a MIMO BC based on AO.

```
Input: $\mathbf{Q} := \mathbf{Q}^0$ diagonal matrix of positive elements, $\epsilon > 0$ .
Initialization: Set $n := 0$ and $\tau = 1 + \epsilon$.
while ($\tau > \epsilon$) do
    Solve (3.41) and denote the optimal solution by $\{\bar{\mathbf{S}}_k^n\}$
    For $n \geq 1$, let $\tau = |f^{\text{DPC}}(\mathbf{Q}^n, \{\bar{\mathbf{S}}^n\}) - f^{\text{DPC}}(\mathbf{Q}^{n-1}, \{\bar{\mathbf{S}}^{n-1}\})|$, where $f^{\text{DPC}}(\cdot)$
        denotes the objective in (3.40).
    For each $k$, compute $\mathbf{\Phi}_k^{-n} = (\mathbf{Q}^n + \sum_{i=k}^{K} \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i^n \mathbf{H}_i)^{-1}$.
    Solve (3.44) to find $\mathbf{Q}^{n+1}$.
    $n := n + 1$.
end
Output: Use the obtained $\{\bar{\mathbf{S}}_k^n\}_{k=1}^K$ and the BC-MAC transformation in [43]
        to find the optimal solution to (3.39).
```

In this section, we numerically evaluate the performance of the proposed algorithms presented in this chapter. For all iterative algorithms of comparison, we set an error tolerance of $\epsilon = 10^{-6}$ as the stopping criterion. The condition number κ is defined as the ratio between the largest singular value and the smallest one. The initial values $\mathbf{\Lambda}^0$ and $\mathbf{Q}^0$ in the corresponding proposed algorithms are set to the identity matrix for all simulations, if not mentioned otherwise. Other simulation parameters are specified for each setup. The codes are executed on a 64-bit desktop that supports 8 Gbyte RAM and Intel CORE i7.

3.4.1. Single-user MIMO

In the first numerical experiment, we demonstrate the convergence rate of Algorithms 3.1 and 3.2, and the mode-dropping algorithm in [1, 3]. In particular we consider the same channel matrix as given in [1, Eq. (26)] and a total power of 0 dBW. As can be seen in Fig. 3.1, monotonic convergence is not always achieved for Algorithm 3.1, which is expected for an iterative method based on standard interference function. For the considered scenario, Algorithm 3.2 converges much faster than other methods of comparison. It can be implied from the iteration in (3.14) that Algorithm 3.1 will attain a good convergence rate if all the diagonal entries of $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}_n)$ are much less than 1 during the whole iterative process, which is likely to occur if the singular values of $\mathbf{H}$ and/or $\mathbf{p}$ are relatively large. The same argument can also be applied to the mode-dropping method. However, this is not the case for the considered scenario, leading to slower convergence rates for Algorithm 3.1 and the mode-dropping method, compared to Algorithm 3.2. Further numerical results on this will be provided in Fig. 3.3.

In Fig. 3.2 we provide an example to show that a pure AO approach may fail to yield the optimal solution to (3.15) as briefly discussed earlier. The channel matrix

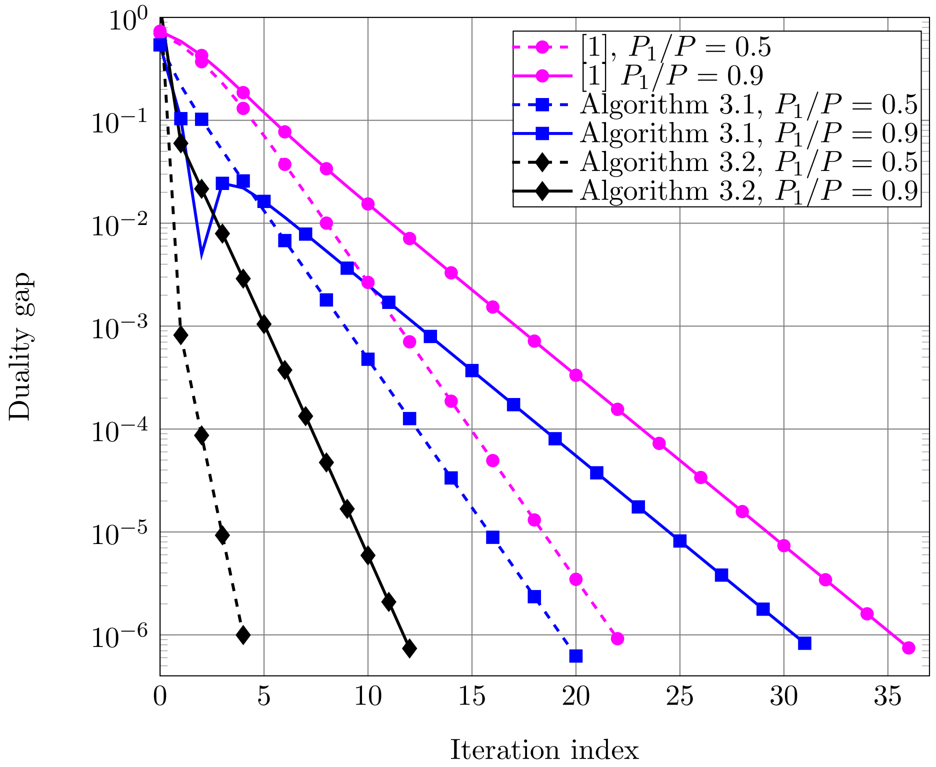


Figure 3.1.: Convergence comparison of different iterative methods for a point-to-point MIMO system with $N = 2$ and $M = 2$. The channel matrix is taken from [1]. The figure is adapted from Fig. 1 in [2] under © 2019 IEEE.

is $\mathbf{H} = [-0.0723 - 0.6116i, 0.2257 - 0.1166i; -0.1707 - 0.0212i, 0.2212 + 0.4439i]$, which is generated randomly. The other simulation parameters are the same as those for Fig. 3.1. The initial value $\mathbf{Q}^0$ is set to identity. We can easily see that the objective returned by the pure AO method is oscillating and not converging to the optimal one. On the contrary, Algorithm 3.2 always guarantees a monotonically decreasing objective sequence converging to the optimal solution.

In the next set of numerical experiments we further investigate the convergence results of the algorithms in comparison. The numbers of transmit and receive antennas are set to $N = 2$ and $M = 4$, respectively. In particular, $\mathbf{\Lambda}^0$ in Algorithm 3.1 is generated in the same way as done in the mode-dropping method [1,3]. Fig. 3.3 plots the average number of iterations as a function of P_1/P over 100 randomly generated channel realizations, and the total transmit power P is specified in the legends of the figure. In this considered setting, the channel matrix has two singular values. First, entries of $\mathbf{H}$ are generated following the i.i.d. zero mean and unit variance Gaussian, and then the smaller singular value is scaled accordingly to achieve a specific value of κ as given in Figs. 3.3a and 3.3b.

As can be seen clearly in Fig. 3.3, the convergence behavior of Algorithm 3.2 is quite consistent for different settings. On the other hand, Algorithm 3.1 and the

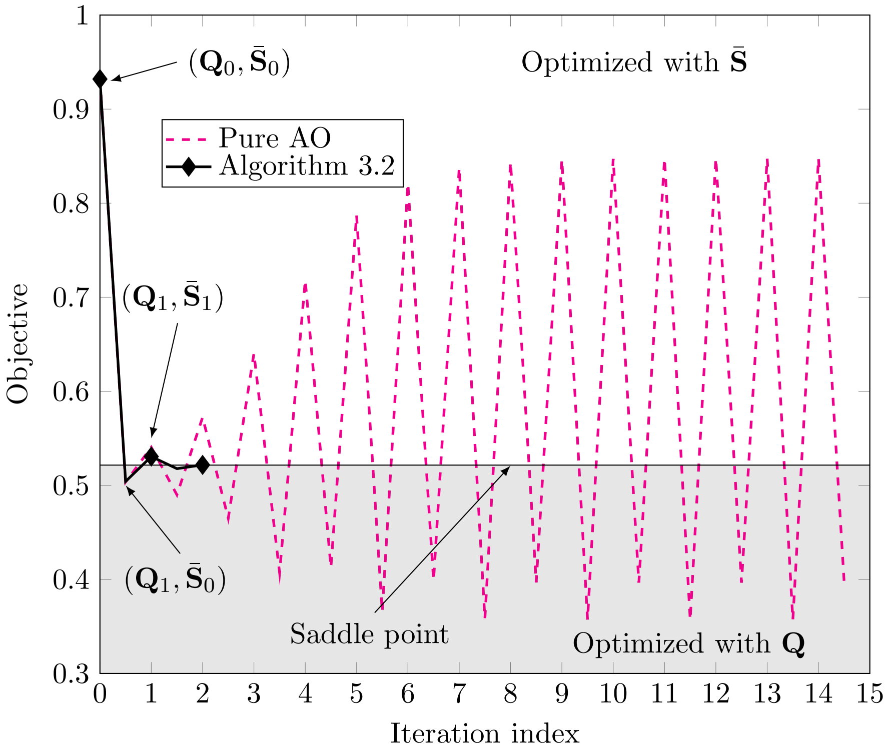


Figure 3.2.: Illustration of the ping-pong effect of the pure AO method, $P_1/P = 0.5$. The upper part of the figure plots the objective of both methods in comparison when it is optimized with $\bar{\mathbf{S}}$, while the lower part in gray color plots with $\mathbf{Q}$. The border line represents the objective of the saddle point. The figure is in [2] under © 2019 IEEE.

mode-dropping scheme obtain the same convergence rate which is sensitive to κ and $\mathbf{p}$. In particular, Algorithm 3.1 takes more iterations to converge when the channel matrix is ill-conditioned (cf. Fig. 3.3b) However, Algorithm 3.1 converges faster for well-conditioned channel matrix and large $\mathbf{p}$ (cf. Fig. 3.3a). We can also see that the convergence rate of Algorithm 3.1 becomes inferior when one of the power limits P_i is small. For such a case, one of the diagonal element of $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}_n)$ is very close to 1 for all iterations, making the fixed-point iteration converge slowly. In fact, these observations agree with what has been explained in Fig. 3.1.

As mentioned earlier, the overall complexity of an iterative algorithm depends on not only the per-iteration complexity but also the number of iterations that it takes to terminate. The overall complexity in terms of flop counts of the iterative methods in case $N \leq M$ is plotted in Fig. 3.4. As shown in the figure, Algorithm 3.2 has the lowest overall complexity. The reason is that Algorithm 3.2 has not only low per-iteration complexity but also (and more importantly) the smallest number of iterations as analyzed in Fig. 3.3. We also observe that if Algorithm 3.1 and the mode-dropping method start from the same initial point, the number of iterations to converge is identical. Thus, Algorithm 3.1 outperforms the mode-dropping method when $N < M$. However, when $N = M$, the total complexity of Algorithm 3.1 is $6N^3(13n+4/3)$ while that of mode-dropping is $6N^3(12n+21)$ where n is the number of iterations to converge. For this special case, it is possible that the total complexity of Algorithm 3.1 can be higher than the mode-dropping algorithm, depending on the number of iterations ($n \geq 20$) and vice versa.

In Fig. 3.5 we benchmark the average run time of proposed algorithms against interior-point methods. In particular the commercial interior-point-based solver MOSEK [4] is chosen for this purpose due to its recognized good performance. The results in Fig. 3.5 are averaged over 1000 channel realizations which are randomly generated using the i.i.d. channel model. It can be seen clearly that the run time of MOSEK increases quickly with the number of transmit antennas. This observation is expected and consistent with the complexity analysis of interior-point methods presented earlier in Section 3.1.3. On the contrary, other algorithms in comparison are more scalable, and Algorithms 3.1 and 3.2 still achieve better performance than the method in [3].

3.4.2. Multi-user MIMO

In the first simulation, we report the average run time for solving (3.31) by several approaches over 1000 channel realizations. As mentioned earlier, we can use generic convex solvers to solve (3.31) optimally. Here we compare Algorithm 3.4 to MOSEK [4] and SDPT3 [54] through the parser YALMIP [55]. In Table 3.2, '×' stands for either a computer crash or extremely large computation time. Table 3.2 clearly shows that Algorithm 3.4 requires the lowest computation time. Recall that in the aforementioned complexity analysis Algorithm 3.4 and [48] have similar per-iteration complexity order. However, the subgradient-based algorithm needs much more time to solve (3.31), due to slow convergence rate as illustrated in Fig. 3.1. Off-the-shelf solvers i.e. MOSEK and SDPT3 work relatively effectively for small N, but fail for large N. This result can be explained by the fact that interior-point-based solvers do not scale well with the problem size. Note that our simulation codes are built on MATLAB environment which is by no means real-time implementation. Thus the run time reported in Table 3.2 is mainly for relative benchmarking purpose. Real-time implementation of the proposed algorithm is beyond the scope of the chapter and is left as future work. However, it is normally expected that embedded implementation can speed up the efficiency of a MATLAB code by several orders of magnitude. We further remark that the coherence time in massive MIMO systems is much larger than that in conventional ones due to the channel hardening effect [56, 57]. Thus these two facts may indicate good embedded implementation of the proposed algorithm is likely suitable for real-time massive MIMO applications.

In the following simulation, the number of receive antennas M and the number of transmit antennas N are fixed to 1 and 128 antennas, respectively. The number of users K is specified for each setup and the power limit for all antennas is equal to P/N.

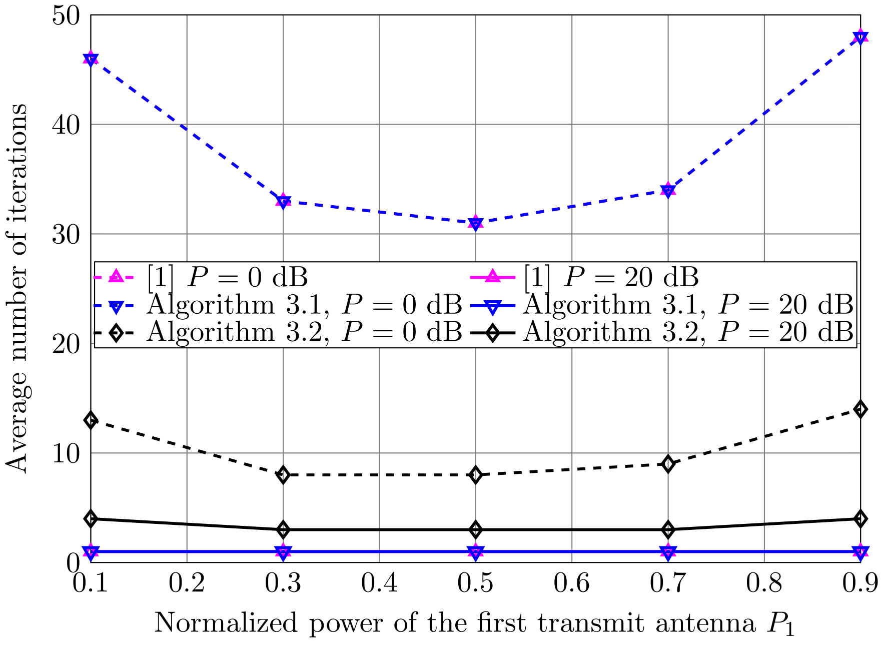
Table 3.2.: Average run time (seconds) comparison for $P = 10$ dBW, $M = 2, K = 8$. The run time is averaged over 1000 channel realizations. The table is adapted from Table I in [12] under © 2018 IEEE.

	No. of Tx. antennas N	16	32	64	128
SRMax	Algorithm 3.4	**0.097**	**3.84**	**115.79**	**175.96**
	[48]	0.85	610.92	> 1 hr	×
	SDPT3	0.32	11.36	×	×
	MOSEK	0.23	11.01	> 1 hr	×
WSRMax	Algorithm 3.4	**0.10**	**2.24**	**92.47**	**147.18**
	[48]	1.02	89.27	> 1 hr	×
	SDPT3	0.32	7.99	×	×

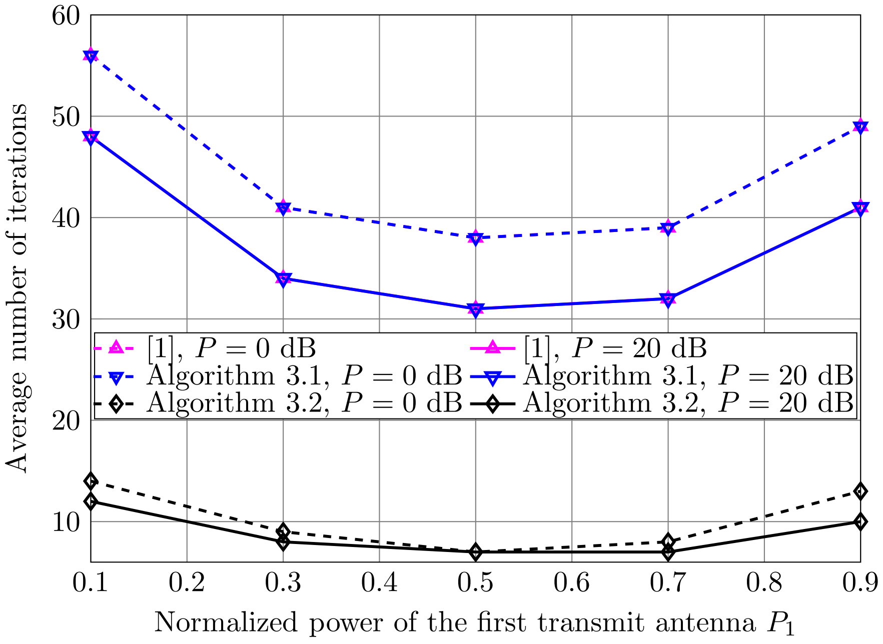
Taking the advantage that the proposed algorithms have low complexity, in the last numerical experiment we characterize the capacity region of a massive MIMO system with PAPC. In particular, we also consider achievable rate region of the well-known ZF scheme. The purpose is to understand the performance of ZF in comparison with the capacity achieving coding scheme under some realistic channel models. To this end we consider a simple urban scenario using WINNER II B1 channel model [58], where a base station, equipped with $N = 128$ antennas, is located at the center of the cell and single-antenna receivers are distributed randomly. The total power at the BS is $P = 46$ dBm and each antenna is subject to equal power constraint, i.e., $P_i = P/N$ for $i = 1, 2, \ldots, 128$. As can be seen clearly in Fig. 3.6, there is a remarkable gap between the achievable rate region of ZF and the capacity region, especially when the number of users increases. This basically implies that ZF is still far from optimal for a practical number of transmit antennas. Our observation opens research opportunities in the future to strike the balance between optimal performance by DPC and low-complexity by ZF.

3.5. Summary

We have solved the problem of computing the capacity of MIMO systems under PAPC. For a SU-MIMO system, two efficient algorithms have been proposed, one based on fixed point iteration and the other based on the MAC-BC duality together with AO. Extensive numerical experiments have been provided to demonstrate the superior performance of the two proposed algorithms over the known methods in [1, 3] in terms of computational complexity. We have also explored the capacity of multi-user MIMO systems subject to PAPC. For the optimal precoding scheme DPC, we have presented a method to compute the full capacity region. Using this low-complexity method, we have also characterized the capacity region of a single cell multiuser massive MIMO system subject to PAPC. The numerical results have demonstrated that the conventional ZF scheme still operates far from the capacity boundary for a practical number of transmit antennas.



(a) Condition number $\kappa = 10$.



(b) Condition number $\kappa = 100$.

Figure 3.3.: Average number of iterations required to converge of different iterative algorithms with $N = 2$ and $M = 4$. The figures are adapted from Fig. 3 in [2] under © 2019 IEEE.

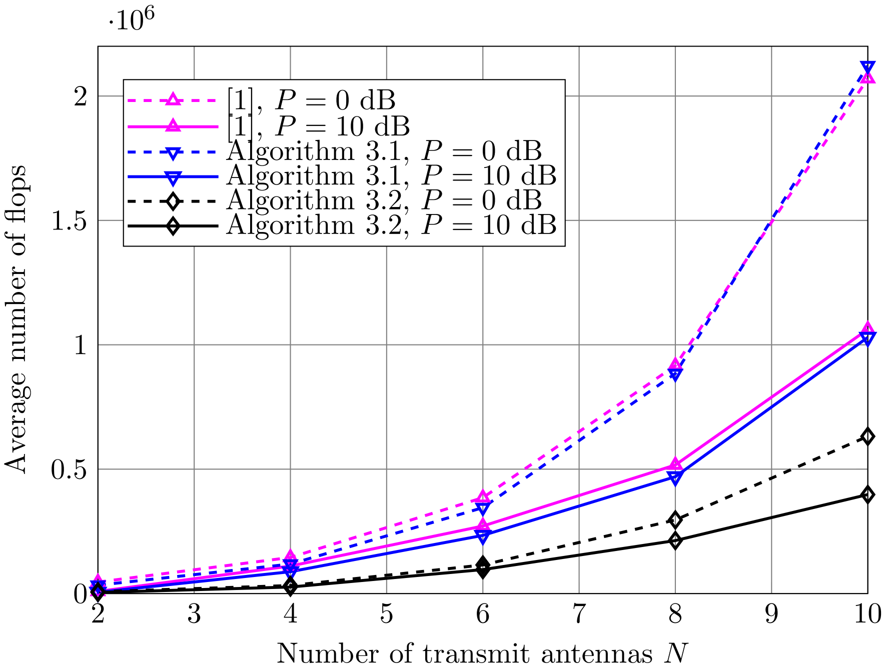


Figure 3.4.: Total complexity comparison versus the number of transmit antennas N. The number of receive antennas is taken as $M = 10$, $\kappa = 10$. The figure is adapted from Fig. 4 in [2] under © 2019 IEEE.

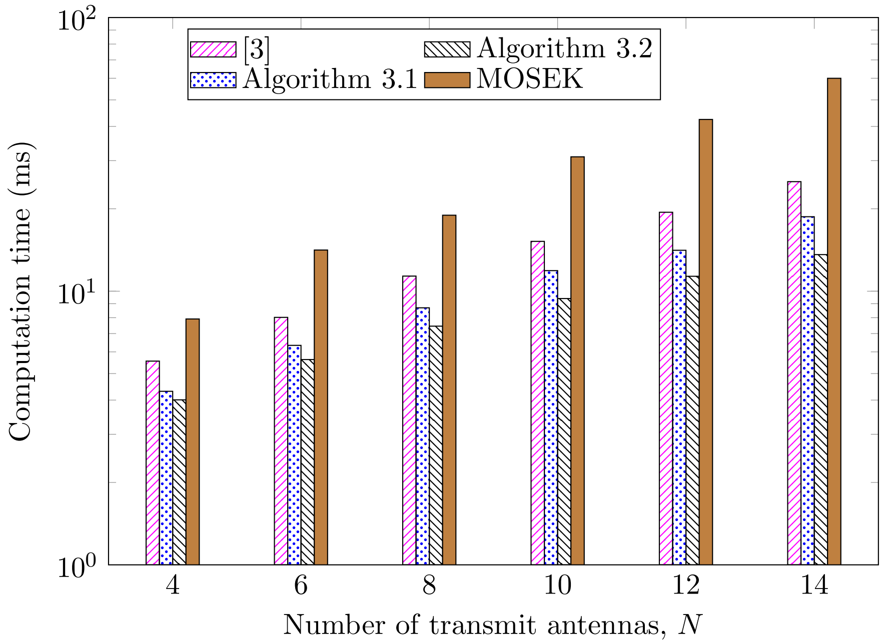
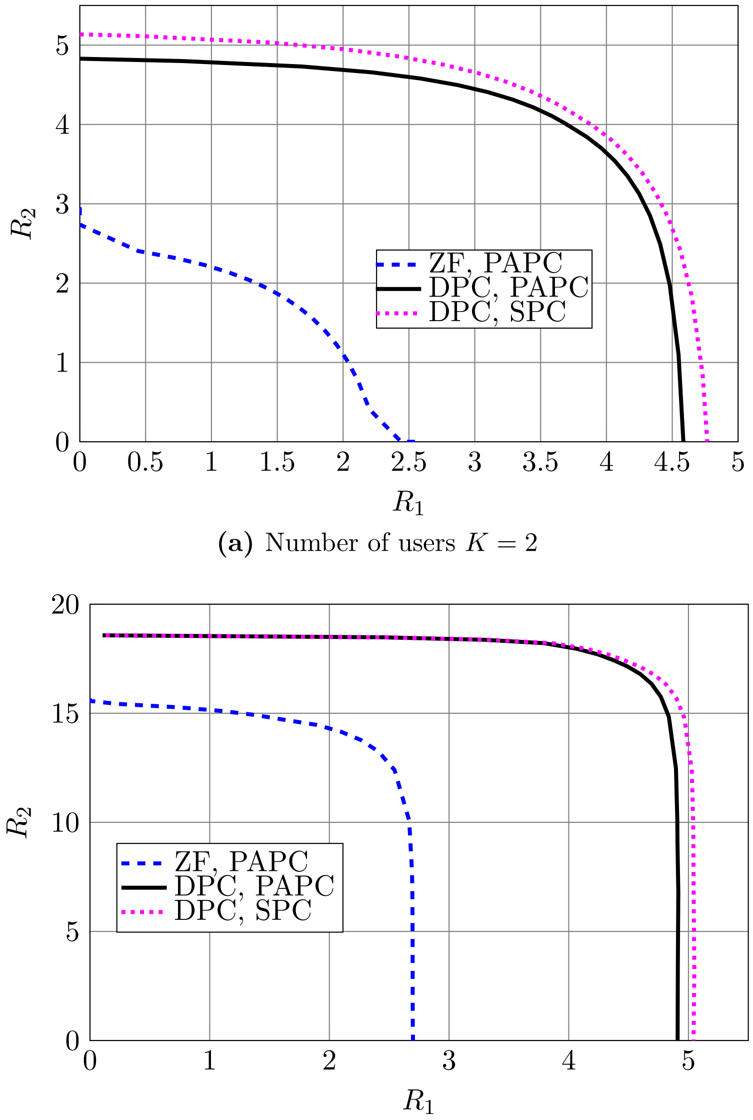


Figure 3.5.: Run time versus the number of transmit antennas N. Four methods are compared: [3], Algorithm 3.1, Algorithm 3.2, and the interior-point-based method implemented in [4]. The number of receive antennas is taken as $M = 2$. The figure is adapted from Fig. 5 in [2] under © 2019 IEEE.



(a) Number of users $K = 2$

(b) Number of users $K = 8$

Figure 3.6.: Comparison of capacity region of a massive MIMO setup with $N = 128, M = 1$. For the case $K = 8$ users, the capacity region is projected on the first two users. The figure is adapted from Fig. 6 in [2] under © 2019 IEEE.

Chapter 4

MIMO Capacity with Linear Transmit Covariance Constraints

The well-known fact that SPC can be imposed on a system to satisfy the power budget or regulations, whereas imposing PAPC is to prevent nonlinear distortions of power amplifiers associated with each transmit antenna. In fact, other power constraints can also be imposed on a MIMO system in addition to SPC or PAPC. For example, interference temperature constraints can be imposed on a secondary user (SU) to limit the interference generated at a primary user (PU) in a cognitive radio networks [33–35]. All of aforementioned constraints can be generalized as linear transmit covariance constraints (LTCCs) [33].

In this chapter we first propose an efficient approach to computing the capacity of a single-user MIMO (SU-MIMO) system under the most general form of multiple LTCCs. We then extend the proposed approach to find the capacity of Gaussian MIMO broadcast channels (BCs). The channel state information is assumed to be perfectly known at both the transmitter and the receiver(s). To this end, we first transform the considered problem in the BC into an equivalent minimax problem in the dual multiple access channel (MAC), generalizing several results on the BC-MAC duality in the previous studies of [10,33,42]. In fact, a minimax optimization approach was also considered in [10] but by interior-point algorithms, i.e., finding a saddle point of the minimax problem by a barrier method. We propose a different method in this chapter. Specifically, to find a saddle point of the considered minimax problem, we combine alternating optimization (AO) and concave-convex procedure (CCP) to arrive at an iterative algorithm, where each iteration is based on closed-form expressions. Our contributions are summarized as follows:

- For SU-MIMO, by generalizing the BC-MAC duality for an arbitrary number of LTCCs, we equivalently express the capacity of the BC with multiple LTCCs as a minimax optimization problem in the dual MAC. The objective of the minimax problem is a concave-convex function of transmit and noise covariance matrices, respectively.
- We then propose a closed-form approach to computing a saddle point of the minimax problem by efficiently combining AO and CCP. The idea is to alternately optimize the transmit and noise covariance matrices following the

general methodology of AO. For minimax problems, the convergence of a pure AO is not guaranteed in general. The novelty of our proposed method is to optimize a bound of the objective obtained from the CCP when optimizing the noise covariance matrix. The proposed algorithm is provably convergent.

- We also propose for the first time an efficient solution to compute the capacity region of a Gaussian MIMO BC, subject to multiple LTCCs. The proposed approach is also based on closed-form expressions, and thus outperforms known solutions relying on either subgradient or interior-point methods in [33, 34] in terms of complexity.
- The approach is also extended to multi-user MIMO channels with ZF methods under multiple LTCCs, the resulting minimax problem is solved by utilizing CCP and AO to find the saddle point. Each iteration of the proposed method can be solved efficiently by water-filling algorithms, leading to its fast convergence rate.
- We provide numerical results on the capacity of large-scale MIMO systems with multiple LTCCs, which have not been reported previously.

The remainder of the chapter is organized as follows. The capacity of SU-MIMO with LTCCs is described in Section 4.1 followed by the special case of joint SPC and PAPC in Section 4.2. Sections 4.3 and 4.4 derive closed-form expressions for the capacity region of a Gaussian MIMO BC and that of ZF while Section 4.5 presents the numerical results. Finally, we conclude the chapter in Section 4.6. Most of the content and results in this chapter have been appeared in [16,17] under © 2018 IEEE and submitted for publication [13] on IEEE journal.

4.1. Capacity of SU-MIMO

4.1.1. System Model

We consider a SU-MIMO model, where the transmitter and the receiver are equipped with N and M antennas, respectively. In this chapter, we assume that the channel state information is perfectly known at the transmitter and the receiver. The received signal is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{z} \tag{4.1}$$

where $\mathbf{H} \in \mathbb{C}^{M\times N}$ is the channel matrix, $\mathbf{s} \in \mathbb{C}^{N\times 1}$ is the vector of transmitted symbols, and $\mathbf{z} \in \mathbb{C}^{M\times 1}$ is the background noise with distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. Let $\mathbf{S} = \mathbb{E}[\mathbf{s}\mathbf{s}^{\dagger}]$ be the input covariance matrix for the transmitted signal. We are interested in finding the capacity of the above channel with multiple LTCCs, which is formulated as

$$\underset{\mathbf{S}\succeq\mathbf{0}}{\text{maximize}} \quad \log|\mathbf{I}+\mathbf{H}\mathbf{S}\mathbf{H}^{\dagger}| \tag{4.2a}$$

$$\text{subject to} \quad \operatorname{tr}(\mathbf{E}_i\mathbf{S}) \leq P_i,\ i=1,2,\dots,L \tag{4.2b}$$

where, for each $i=1,2,\dots,L$, $\mathbf{E}_i \succeq \mathbf{0}$ is a predefined positive semidefinite matrix and $P_i>0$ is the corresponding power constraint. Note that (4.2b) is called general linear constraints on the transmit covariance and it can include several types of transmit power constraints as special cases. Some examples are given below:

- If, for some i, $\mathbf{E}_i$ is an identity matrix, the resulting constraint (4.2b) becomes $\operatorname{tr}(\mathbf{S}) \leq P_i$, representing a SPC.
- If, for some i, $\mathbf{E}_i = \operatorname{diag}(\mathbf{e}_i)$, the constraint reduces to $[\mathbf{S}]_{i,i} \leq P_i$, denoting a maximum power constraint on the ith antenna. In this chapter, a PAPC means imposing (4.2b) for all antennas.
- If, for some i, $\mathbf{E}_i = \mathbf{G}^{\dagger}\mathbf{G}$, where $\mathbf{G}$ is the effective channel between a SU and a PU, then the resulting constraint limits the overall interference experienced by the PU [59,60]. More specifically, in this situation, the transmitter of the considered MIMO system is assumed to be a SU in the context of cognitive radio networks. Consequently, the transmitter introduces an interference term having a covariance matrix $\mathbf{G}\mathbf{S}\mathbf{G}^{\dagger}$. To reduce the interference that the transmitter causes to the PU, we may impose a constraint $\operatorname{tr}(\mathbf{G}\mathbf{S}\mathbf{G}^{\dagger}) = \operatorname{tr}(\mathbf{S}\mathbf{E}_i) \leq P_i$ [59]. Intuitively, we wish to limit the total interference power that the SU can cause to the PU. In the remainder of this chapter, we will refer to this type of constraint as an interference power constraint.

Let $\mathcal{S}$ be the feasible set of (4.2), i.e., $\mathcal{S} = \{\mathbf{S}|\mathbf{S}\succeq 0, \operatorname{tr}(\mathbf{E}_i\mathbf{S}) \leq P_i,\ i=1,2,\dots,L\}$. In this chapter we assume that $\mathbf{E}_i$'s are introduced in such a way that the feasible set $\mathcal{S}$ is compact, and thus (4.2) is solvable. We further assume that for any subset $I \subset \{1,2,\dots,L\}$ such that the matrix $\mathbf{E} = \sum_{i\in I}\mathbf{E}_i$ is singular, then $\mathcal{N}(\mathbf{E})$ is not contained in $\mathcal{N}(\mathbf{H})$. We detail this assumption as follows.

- If $\mathcal{S}$ includes $\mathbf{E}_i = \operatorname{diag}(\mathbf{e}_i)$ (i.e. the PAPC for the ith antenna), then the above assumption requires that the ith column of $\mathbf{H}$ is not an all zero-vector. That is, if all PAPCs are considered, then each of the columns of $\mathbf{H}$ is not an all-zero vector.
- If $\mathcal{S}$ also contains an interference power constraint $\mathbf{E}_i = \mathbf{G}^{\dagger}\mathbf{G}$, where $\mathbf{G}$ is the effective channel as mentioned above and $\operatorname{rank}(\mathbf{G}) < N$, then we further require that $\mathcal{N}(\mathbf{E}_i) = \mathcal{N}(\mathbf{G})$ is not contained in $\mathcal{N}(\mathbf{H})$. Note that $\mathcal{N}(\mathbf{G}) \subset \mathcal{N}(\mathbf{H})$ if and only if there exists a matrix $\mathbf{C}$ such that $\mathbf{H} = \mathbf{C}\mathbf{G}$. Considering the random nature of $\mathbf{H}$ and $\mathbf{G}$, it happens almost surely that $\mathcal{N}(\mathbf{G})$ is not contained in $\mathcal{N}(\mathbf{H})$.

The achievability of (4.2) is stated in the proposition below.

Proposition 4.1. *The capacity of a MIMO system with multiple LTCCs given in* (4.2) *can be achieved by a Gaussian distributed input.*

Proof. See Appendix B.1. □

We can pose (4.2) as a semidefinite program (SDP) and then use an off-the-shelf SDP optimization software to find the optimal transmit covariance. However, the complexity of such a method increases dramatically with the problem size, and thus is not suitable for large-scale MIMO systems. In the following we propose an efficient solution for solving (4.2) by closed-form expressions.

4.1.2. Algorithm Description

The proposed method for solving (4.2) is inspired by the one presented in [2]. In order that the current chapter is self-contained, we briefly describe the main steps of the proposed methods and refer the interested readers to [2] for further details.

The idea is to transform (4.2) in the BC into an equivalent minimax problem in the MAC for which an efficient algorithm is derived. First we note that the Slater condition holds for (4.2) due to the fact that $P_i > 0$ for all i. As a result, the duality gap is zero and (4.2) can be optimally solved in the dual domain. In this regard, let us denote by $\mathbf{q} = [q_1, q_2, \dots, q_L]^T$ the vector of the Lagrange multipliers for the constraints (4.2b) and let $\mathbf{p} = [P_1, P_2, \dots, P_L]^T$ be the corresponding power constraints. The proposed algorithm is based on the following theorem.

Theorem 4.1. *The Lagrangian dual problem of* (4.2) *is equivalent to the following minimax problem*

$$\begin{aligned} \min_{\mathbf{q}\geq 0} \max_{\bar{\mathbf{S}}\succeq \mathbf{0}} \quad & \log \frac{|\sum_{i=1}^{L} q_i\mathbf{E}_i + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|}{|\sum_{i=1}^{L} q_i\mathbf{E}_i|} \triangleq f(\mathbf{q},\bar{\mathbf{S}}) \\ \text{subject to} \quad & \mathrm{tr}(\bar{\mathbf{S}}) \leq P;\ \mathbf{p}^T\mathbf{q} \leq P; \textstyle\sum_{i=1}^{L} q_i\mathbf{E}_i \succ \mathbf{0} \end{aligned} \tag{4.3}$$

where $P = \sum_{i=1}^{L} P_i$.

Proof. See Appendix B.2. □

We remark that the above result is in fact a generalization of several results for LTCCs presented in previous studies [10, 33, 42]. Without loss of optimality, the inequalities of (4.3) can be replaced with equalities. Let us define $\mathcal{Q} \triangleq \{\mathbf{q}|\mathbf{q} \geq \mathbf{0}, \mathbf{p}^T\mathbf{q} = P, \sum_{i=1}^{L} q_i\mathbf{E}_i \succ \mathbf{0}\}$ and $\bar{\mathcal{S}} = \{\bar{\mathbf{S}}|\bar{\mathbf{S}} \succeq \mathbf{0}, \mathrm{tr}(\bar{\mathbf{S}}) = P\}$, and note that $f(\mathbf{q},\bar{\mathbf{S}})$ is concave with $\bar{\mathbf{S}}$, and convex with $\mathbf{q}$, and twice differentiable. Thus a saddle point $(\mathbf{q}^*, \bar{\mathbf{S}}^*)$ exists for the minimax problem of (4.3) and it holds that

$$f(\mathbf{q}^*,\bar{\mathbf{S}}) \leq f(\mathbf{q}^*,\bar{\mathbf{S}}^*) \leq f(\mathbf{q},\bar{\mathbf{S}}^*). \tag{4.4}$$

For this general minimax problem, pure AO which simply alternates maximization and minimization at each step is not guaranteed to converge [2, 10]. On the other hand, interior-point methods are possible but their complexity increases rapidly with the problem size, and thus they are not attractive for large-scale MIMO systems [15].

In this chapter we propose an iterative method that combines AO and CCP to solve (4.3) efficiently. The proposed method can be summarized as follows:

- For a given $\mathbf{q}$, we maximize $f(\mathbf{q}, \bar{\mathbf{S}})$ with respect to $\bar{\mathbf{S}}$, which can be solved efficiently by the classical water-filling algorithm.
- For a given $\bar{\mathbf{S}}$, we *minimize a convex upper bound* of $f(\mathbf{q}, \bar{\mathbf{S}})$ which is obtained from the CCP. The novelty of the proposed method lies in the use of a convex upper bound, which is proved to generate a decreasing sequence of objective values. This method avoids fluctuations which occur in a pure AO algorithm, and thus convergence is guaranteed.

We remark that the proposed method is entirely different from the alternating direction method of multipliers (ADMM) which is widely used in the related literature. Firstly, no augmented Lagrangian function is involved. Secondly, the dual variable update in the ADMM uses a step size equal to the parameter associated with the augmented Lagrangian function, while the $\mathbf{q}$-minimization in Algorithm 4.2 (which can be viewed as the dual update to some extent) is based on solving an approximate optimization problem. Thirdly, we maximize the objective with respect to one variable and minimize with respect to the other variable, while the update of the primal variables in the ADMM is done by only minimizing the objective. As a result, the convergence proof of the ADMM cannot be applied to our proposed method.

In fact, the approach presented in this chapter is the same as the one in [2] from an algorithmic viewpoint. The main advantage of such a method is that the maximization of $f(\mathbf{q}, \bar{\mathbf{S}})$ admits the classic water-filling algorithm regardless of the types of power constraints in the BC. However, the minimization of a convex upper bound of $f(\mathbf{q}, \bar{\mathbf{S}})$ is entirely different as can be seen shortly. More specifically, we propose efficient solutions to this step for MIMO capacity with multiple LTCCs in general and joint SPC and PAPC in particular.

In the following we provide the details of the above steps. Let $\mathbf{q}^n$ be the value of $\mathbf{q}$ at the nth iteration of the proposed method and $\mathbf{Q}_n = \sum_{i=1}^{L} q_i^n \mathbf{E}_i$. From (4.3), we can easily see that the maximization with respect to $\bar{\mathbf{S}}$ admits the water-filling solution [2, 6, 7]. Therefore we focus on the problem of finding $\mathbf{Q}_n$ which is one of our main contributions. Similar to [2], we use an upper bound of the objective for the $\mathbf{q}$-minimization. We refer the interested reader to [2] for further discussions on the use of an upper bound for the $\mathbf{q}$-minimization.

In light of the CCP, we note that $f(\mathbf{q}, \bar{\mathbf{S}})$ in (4.3) can be expressed as a difference of two convex functions. In particular, by the concavity property of the logdet function, we have

$$\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H}| \leq \log|\mathbf{\Phi}_n| + \operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}-\mathbf{Q}_n\big)\Big) \tag{4.5}$$

which produces

$$f(\mathbf{q},\bar{\mathbf{S}}_n) \leq \log|\mathbf{\Phi}_n| + \operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}-\mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}| \tag{4.6}$$

where $\mathbf{\Phi}_n = \mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H}$ and $\mathbf{Q} \triangleq \sum_{i=1}^{L} q_i\mathbf{E}_i$. The right hand side of (4.6) is a convex upper bound of the objective. To find $\mathbf{q}^{n+1}$ we solve the minimization of the upper bound given in (4.6). Since $\log|\mathbf{\Phi}_n|$ is a constant in this regard, $\mathbf{q}^{n+1}$ is in fact the solution to the following problem

$$\begin{array}{ll}\underset{\mathbf{q}\geq\mathbf{0}}{\text{minimize}} & \operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\mathbf{Q}\Big) - \log|\mathbf{Q}| \\ \text{subject to} & \mathbf{p}^T\mathbf{q} = P\end{array} \tag{4.7}$$

or equivalently,

$$\begin{array}{ll}\underset{\mathbf{q}\geq 0}{\text{minimize}} & \sum_{i=1}^{L} q_i\phi_{n,i} - \log|\sum_{i=1}^{L} q_i\mathbf{E}_i| \triangleq g(\mathbf{q}) \\ \text{subject to} & \mathbf{p}^T\mathbf{q} = P\end{array} \tag{4.8}$$

where $\phi_{n,i} = \operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\mathbf{E}_i\Big)$.

It is worth mentioning that the optimal solution $\mathbf{q}^{n+1}$ to (4.8) must satisfy $\mathbf{Q}_{n+1} = \sum_{i=1}^{L} q_i^{n+1}\mathbf{E}_i \succ 0$ for all n, assuming $\mathbf{Q}_0 \succ 0$ which can be achieved by properly choosing $\mathbf{q}^0$. This can be proved by noting that $\mathbf{\Phi}_0 \succ 0$ if $\mathbf{Q}_0 \succ 0$. Thus $\mathbf{Q}_1$ cannot be singular, otherwise the optimal value goes to infinity which is impossible. By induction we can conclude that $\mathbf{Q}_n \succ 0$ (and thus $\mathbf{\Phi}_n \succ 0$) for all n. As a result, the proposed iterative method is well defined for all iterations. Our idea of using a convex upper bound for minimizing a cost function is inspired by the successive convex approximation (SCA) framework. However, in the context of SCA, the objective to be minimized is often nonconvex. In the considered problem, $f(\mathbf{q},\bar{\mathbf{S}})$ is indeed convex with respect to $\mathbf{q}$ but an upper bound can be derived easily following the CCP. We note that other upper bounds can also be used in the proposed algorithm, as long as they meet the other conditions as well (see Property A of [44] for the details). The upper bound found in (4.6) is relatively straightforward but it results in efficient methods for solving (4.8) as shown next. In the general case of LTCCs, the gradient projection or conjugate gradient projection method can be utilized to solve (4.8) efficiently. The reason is that the feasible set of (4.8) is a simplex, and projection onto a simplex admits water-filling-like algorithms [51]. A gradient projection based algorithm for solving (4.8) is described in Algorithm 4.1. A closed-form method for solving (4.8) was proposed in [14] for the case where only a PAPC is considered. For the special case of joint SPC and PAPC, a closed-form solution for (4.8) is provided in the next subsection.

Algorithm 4.1: The Gradient Projection Algorithm for Solving (4.8).

```
Input: $\mathbf{q}_0$ , $\epsilon_1 > 0$, $m := 0$.
repeat
    Calculate the gradient $\tilde{\mathbf{u}}_m = -\nabla g(\mathbf{q}_m)$.
    Choose an appropriate positive scalar $\rho_m$ for $\tilde{\mathbf{q}}_m = \mathbf{q}_m + \rho_m \tilde{\mathbf{u}}_m$.
    Project $\tilde{\mathbf{q}}_m$ onto $\mathcal{Q}$ to obtain $\bar{\mathbf{q}}_m$.
    Choose appropriate step size $\nu_m$ using the Armijo rule [50] and set
     $\mathbf{q}_{m+1} = \mathbf{q}_m + \nu_m(\bar{\mathbf{q}}_m - \mathbf{q}_m)$.
    $m := m + 1$.
until $|\nabla g(\mathbf{q}_m)^T(\mathbf{q}_{m+1} - \mathbf{q}_m)| < \epsilon_1$;
Output: $\mathbf{q}_m$ as the solution to (4.8).
```

Algorithm 4.2: Proposed Solution Based on AO and CCP.

```
Input: $\mathbf{q}^0$ feasible to $\mathcal{Q}$, and $\epsilon_2 > 0$.
Initialize $n := 0$, $\tau = 1 + \epsilon_2$.
while $\tau > \epsilon_2$ do
    Compute $\mathbf{Q}_n = \sum_{i=1}^{L} q_i^n \mathbf{E}_i$.
    Apply water-filling algorithm to compute $\bar{\mathbf{S}}_n = \arg\max_{\bar{\mathbf{S}} \in \bar{\mathcal{S}}} \log|\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}} \mathbf{H}|$.
    If $n \geq 1$, let $\tau = |f(\mathbf{q}^n, \bar{\mathbf{S}}_n) - f(\mathbf{q}^{n-1}, \bar{\mathbf{S}}_{n-1})|$.
    Let $\mathbf{\Phi}_n^{-1} = (\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H})^{-1}$.
    Find $\mathbf{q}^{n+1} = \arg\min_{\mathbf{q} \in \mathcal{Q}} \operatorname{tr}\left(\mathbf{\Phi}_n^{-1} \mathbf{Q}\right) - \log|\mathbf{Q}|$ using Algorithm 4.1.
    $n := n + 1$.
end
Output: $\bar{\mathbf{S}}_n$ .
```

The complete algorithm to find the optimal transmit covariance matrix with multiple LTCCs is summarized in Algorithm 4.2. The convergence results of Algorithm 4.2 are stated in the following lemma.

Lemma 4.1. *Let $\{\mathbf{q}^n, \bar{\mathbf{S}}_n\}$ be the sequence generated by Algorithm 4.2. Then*

a) $f(\mathbf{q}^n, \bar{\mathbf{S}}_n) \geq f(\mathbf{q}^{n+1}, \bar{\mathbf{S}}_{n+1})$ and thus $\{f(\mathbf{q}^n, \bar{\mathbf{S}}_n)\}$ is convergent.

b) $\{\mathbf{q}^n, \bar{\mathbf{S}}_n\}$ contains at least a convergent subsequence.

c) $\mathbf{Q}^* = \sum_{i=1}^{L} q_i^* \mathbf{E}_i$, where q_i^* is a limit point of $\{\mathbf{q}^n\}$, is non-singular.

d) Every limit point of $\{\mathbf{q}^n, \bar{\mathbf{S}}_n\}$ is a saddle point of (4.3).

The proof of the lemma is provided in Appendix B.3.

4.2. Special cases of SU-MIMO with joint SPC and PAPC

The MIMO capacity with joint SPC and PAPC is important in its own right and is treated in this section. In particular, we show that more computationally efficient solutions are achievable for MIMO capacity with joint SPC and PAPC. We derive explicit closed-form expressions for solving (4.8) for the specific case of joint SPC and PAPC. We remark that previous research has made noticeable attempts to solve this problem by working directly on (4.2) [31, 32, 61, 62]. In this chapter we demonstrate that the minimax formulation in (4.3) is also particularly useful to the MIMO capacity with joint SPC and PAPC.

4.2.0.1. MIMO Capacity with Joint SPC and PAPC

The special case of MIMO capacity with joint SPC and PAPC has received growing interest recently in [61, 62]. In particular, Loyka proposed a closed-form solution for this problem in [62], which is, unfortunately, only applicable to full column rank channels, high SNR regime, and an equal power constraint on all transmit antennas. Under these conditions, a closed-form solution for the optimal covariance matrix is possible by solving the KKT conditions of (4.2) [62]. In [61] Cao *et al.* proposed an iterative method by solving a sequence of MIMO capacity problems with PAPC. However, their work was based on a closed-form solution for MIMO capacity with PAPC introduced in [63]. The issue with [63] is that their solution also assumes full column rank channels and high SNR regime, otherwise it is only suboptimal.

In this chapter we aim to fill this gap in the literature. Specifically, we make no assumptions on the MIMO channels or on the operating SNR. In our earlier work [17], we presented a nonlinear Gauss–Seidel method in combination with Lagrangian duality to solve (4.8). As an improvement, we provide herein a method to solve (4.8) by closed-form expressions, rather than using a gradient projection method as mentioned in the preceding section.

First, notice that in this case the number of constraints is $L = N+1$ and we assume $\mathbf{E}_{N+1} = \mathbf{I}_N$ which represents the SPC, and $\mathbf{E}_i = \mathrm{diag}(\mathbf{e}_i)$ for $i = 1, 2, \ldots, N$ which represents the PAPC. In this way P_i, $i = 1, 2, , \ldots, N$ is the power constraint for the ith antenna and $P_{N+1} \triangleq P_T$ is the total transmit power.

In this chapter, we only consider the non-trivial case where $\min\{P_i\} < P_{N+1} < \sum_{i=1}^{N} P_i$. If $P_{N+1} \leq \min\limits_{1\leq i\leq N}\{P_i\}$, it is easy to see that (4.2) reduces to the MIMO capacity with a single SPC. Similarly, if $P_{N+1} \geq \sum_{i=1}^{N} P_i$, the SPC can be omitted and thus (4.2) becomes the MIMO capacity with PAPC [14].

It is trivial to see that in this case, problem (4.8) may be equivalently rewritten as

$$\begin{array}{ll}\underset{\mathbf{q}\geq 0}{\text{minimize}} & q_{N+1}\phi_{n,N+1} + \sum_{i=1}^{N}\left(q_i\phi_{n,i} - \log(q_{N+1}+q_i)\right)\\ \text{subject to} & \sum_{i=1}^{N+1} P_i q_i = P.\end{array} \tag{4.9}$$

To derive a closed-form solution to the above problem, we note that in this case $\mathbf{E}_{N+1} = \sum_{i=1}^{N} \mathbf{E}_i = \mathbf{I}_N$ and thus

$$\phi_{n,N+1} = \mathrm{tr}\Big(\mathbf{\Phi}_n^{-1}\mathbf{E}_{N+1}\Big) = \mathrm{tr}\Big(\mathbf{\Phi}_n^{-1}\sum_{i=1}^{N}\mathbf{E}_i\Big) = \sum_{i=1}^{N}\phi_{n,i}. \tag{4.10}$$

By definition, we have

$$\phi_{n,i} = \mathrm{tr}\Big(\mathbf{\Phi}_n^{-1}\mathbf{E}_i\Big)[\mathbf{\Phi}_n^{-1}]_{i,i} = \Big[\Big(\mathbf{Q}_n + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H}\Big)^{-1}\Big]_{i,i} \le \frac{1}{q_i^n + q_{N+1}^n}. \tag{4.11}$$

To lighten the notation, the subscript n is to be dropped in the sequel. Accordingly, (4.9) can be rewritten as

$$\begin{array}{rl} \underset{\mathbf{q}\ge 0}{\text{minimize}} & \sum_{i=1}^{N}\left(\phi_i(q_i + q_{N+1}) - \log(q_{N+1} + q_i)\right) \\ \text{subject to} & \sum_{i=1}^{N+1} P_i q_i = P. \end{array} \tag{4.12}$$

To further simplify the problem, we make a change of variable. Specifically, let us define $q_i + q_{N+1} = x_i$ for $i = 1, 2, \dots, N$. Then the above problem is equivalent to

$$\begin{array}{rl} \text{minimize} & \sum_{i=1}^{N}\left(\phi_i x_i - \log x_i\right) \\ \text{subject to} & \sum_{i=1}^{N} P_i x_i + P_T' q_{N+1} = P. \\ & x_i - q_{N+1} \ge 0; q_{N+1} \ge 0 \end{array} \tag{4.13}$$

where $P_T' = P_T - \sum_{i=1}^{N} P_i < 0$. The optimal solution for the above problem can be found by solving the KKT conditions which are given by

$$\mu_i(x_i - q_{N+1}) = 0 \tag{4.14a}$$

$$\mu_{N+1}q_{N+1} = 0 \tag{4.14b}$$

$$\phi_i - \frac{1}{x_i} + \gamma P_i - \mu_i = 0, \ i = 1, 2, \dots, N \tag{4.14c}$$

$$\gamma P_T' - \mu_{N+1} + \sum_{i=1}^{N}\mu_i = 0. \tag{4.14d}$$

where $\mu_i \ge 0$ and γ are the KKT multipliers for the constraints $x_i - q_{N+1} \ge 0$ and $\sum_{i=1}^{N} P_i x_i + P_T' q_{N+1} = P$, respectively. We have the following proposition.

Proposition 4.2. *The solution to the KKT conditions in* (4.14) *satisfies* $q_{N+1} > 0$.

Proof. See Appendix B.4. □

We remark that from the proof of Theorem 4.1, q_{N+1} is in fact the Lagrange multiplier of the SPC constraint. Thus, Proposition 4.2 means the SPC is binding, which is not surprising.

It is now obvious that $\mu_{N+1} = 0$. For a given γ, without loss of generality, we can assume that $\frac{1}{\phi_1+\gamma P_1} \geq \frac{1}{\phi_2+\gamma P_2} \geq \ldots \geq \frac{1}{\phi_N+\gamma P_N}$. Further, we note that if $\mu_i = 0$ then

$$x_i = \frac{1}{\phi_i + \gamma P_i} \tag{4.15}$$

which leads to the following proposition.

Proposition 4.3. *If $i < j$ and $x_i > q_{N+1}$ and $x_j > q_{N+1}$ then $x_i \geq x_j$.*

It can be easily seen that $\mu_i = \mu_j = 0$; thus $x_i \geq x_j$ follows immediately from (4.15).

Proposition 4.4. *If $x_j = q_{N+1}$ for some j then $x_k = q_{N+1}$ for $k > j$.*

Suppose the contrary that $x_k > q_{N+1}$for a certain $k > j$. Then $\mu_k = 0$ and thus

$$x_k = \frac{1}{\phi_k + \gamma P_k} \leq \frac{1}{\phi_j + \gamma P_j} \leq \frac{1}{\phi_j + \gamma P_j - \mu_j} = x_j = q_{N+1} \tag{4.16}$$

which contradicts with the fact that $x_k > q_{N+1}$. From the two above propositions there exists a number k such that

$$x_1 \geq x_2 \geq \cdots \geq x_k \geq q_{N+1} \tag{4.17}$$

and

$$x_{k+1} = \cdots = x_N = q_{N+1}. \tag{4.18}$$

Using (4.14c) and (4.14d) we have

$$\gamma = -\frac{\sum_{i=k+1}^{N} \mu_i}{P'_T} = -\frac{\sum_{i=k+1}^{N} \phi_i - \frac{1}{q_{N+1}} + \gamma \sum_{i=k+1}^{N} P_i}{P'_T} \tag{4.19}$$

and thus

$$q_{N+1} = \frac{N-k}{\sum_{i=k+1}^{N} \phi_i + \gamma(P'_T + \sum_{i=k+1}^{N} P_i)}. \tag{4.20}$$

From the above derivations, we propose a bisection method to solve (4.13) as follows.

```
Step 1: Set $\gamma_{\min} = 0$ and $\gamma_{\max}$ to be sufficiently large.
Step 2: Set $\bar{\gamma} = \frac{\gamma_{\min}+\gamma_{\max}}{2}$.
Step 3: Rearrange the terms $\{\frac{1}{\phi_i+\bar{\gamma}P_i}\}$ in decreasing order.
Step 4: Find the largest $k \leq N-1$ such that
$\frac{1}{\phi_k+\bar{\gamma}P_k} \geq \frac{N-k}{\sum_{i=k+1}^{N}\phi_i+\bar{\gamma}(P'_T+\sum_{i=k+1}^{N}P_i)}$ and set $x_i = \frac{1}{\phi_i+\bar{\gamma}P_i}$ for $i \leq k$ and
$x_i = q_{N+1} = \frac{N-k}{\sum_{i=k+1}^{N}\phi_i+\bar{\gamma}(P'_T+\sum_{i=k+1}^{N}P_i)}$ for $i > k$.
Step 5: If $\sum_{i=1}^{N} P_i x_i + P'_T q_{N+1} - P > 0$ then set $\gamma_{\min} = \bar{\gamma}$, otherwise set
$\gamma_{\max} = \bar{\gamma}$.
Step 6: Repeat Step 2 until $\gamma_{\max} - \gamma_{\min} < \epsilon$ where $\epsilon$ is a predetermined error
tolerance.
```

A possible value of $\gamma_{\max}$ can be chosen as $\gamma_{\max} = \frac{N}{P} - \frac{\phi_{\min}}{P_{\max}}$. A proof that this choice is sufficient to solve (4.13) is given in Appendix B.5.

4.2.1. MISO Capacity with Joint SPC and PAPC

As mentioned earlier, this special case was recently studied in [31,32], where a closed-form solution was presented in [32]. Our purpose in this subsection is to show that a closed-form solution is also achievable using the minimax formulation in (4.2), leading to an equivalent solution to that in [32]. Again we are only interested in the nontrivial case where $\min\{P_i\} < P_{N+1} < \sum_{i=1}^{N} P_i$.

It is easy to see that for MISO channels, $\bar{\mathbf{S}}$ in (4.3) becomes *a scalar* and thus maximization of $f(\mathbf{q}, \bar{\mathbf{S}})$ with respect to $\bar{\mathbf{S}}$ is always obtained at $\bar{\mathbf{S}} = P = \sum_{i=1}^{N+1} P_i$. Thus the minimax problem is reduced to

$$\begin{array}{rl} \underset{\mathbf{q}\geq 0}{\min} & \log \frac{|\operatorname{diag}(q_{N+1}+q)+P\mathbf{H}^{\dagger}\mathbf{H}|}{|\operatorname{diag}(q_{N+1}+q)|} \\ \text{subject to} & \sum_{i=1}^{N+1} P_i q_i = P \end{array} \tag{4.21}$$

which is equivalent to (by noting that $\mathbf{H}$ is a row vector)

$$\begin{array}{rl} \underset{\mathbf{q}\geq 0}{\min} & \mathbf{H}\left(\operatorname{diag}(q_{N+1}+q)^{-1}\mathbf{H}^{\dagger} \\ \text{subject to} & \sum_{i=1}^{N+1} P_i q_i = P. \end{array} \tag{4.22}$$

To make notation clear, let us write $\mathbf{H} = [h_1, h_2, \ldots, h_N]$. Then (4.22) is explicitly written as

$$\begin{array}{rl} \underset{\mathbf{q}\geq 0}{\min} & \sum_{i=1}^{N} \frac{|h_i|^2}{q_i+q_{N+1}} \\ \text{subject to} & \sum_{i=1}^{N+1} P_i q_i = P. \end{array} \tag{4.23}$$

Since Slater's condition holds, the sufficient optimality condition is given by the KKT conditions:

$$-\frac{|h_i|^2}{\left(q_i+q_{N+1}\right)^2} + \gamma P_i - \mu_i = 0, \quad i = 1, 2, \ldots, N \tag{4.24}$$

$$-\sum_{i=1}^{N} \frac{|h_i|^2}{\left(q_i+q_{N+1}\right)^2} + \gamma P_T - \mu_{N+1} = 0 \tag{4.25}$$

$$P_T q_{N+1} + \sum_{i=1}^{N} P_i q_i - P = 0 \tag{4.26}$$

$$q_i \mu_i = 0, \quad i = 1, 2, \ldots, N \tag{4.27}$$

$$q_{N+1}\mu_{N+1} = 0. \tag{4.28}$$

Without loss of generality, let us assume $\{\frac{|h_i|}{\sqrt{P_i}}\}$ are in decreasing order, i.e., $\frac{|h_1|}{\sqrt{P_1}} \geq \frac{|h_2|}{\sqrt{P_2}} \geq \ldots \geq \frac{|h_N|}{\sqrt{P_N}}$. If $q_i > 0$ (i.e. the ith PAPC is active) then from (4.24) we have

$$\frac{|h_i|^2}{\left(q_i + q_{N+1}\right)^2} = \gamma P_i \Leftrightarrow q_i + q_{N+1} = \frac{|h_i|}{\sqrt{\gamma P_i}}. \tag{4.29}$$

Suppose all N PAPCs are active; then we have

$$-\gamma \sum P_i + \gamma P_T - \mu_{N+1} = 0. \tag{4.30}$$

If $P_T > \sum_i P_i$ then $\mu_{N+1} > 0$ and thus $q_{N+1} = 0$. That is, the SPC is inactive which can be understood clearly from considering the primal domain. The solution in this case is trivial.

Now consider the case where $P_T \leq \sum_i P_i$. In this case, $q_{N+1} > 0$ and $\mu_{N+1} = 0$. If $q_i = 0$ then

$$-\frac{|h_i|^2}{\left(q_{N+1}\right)^2} + \gamma P_i - \mu_i \Leftrightarrow q_{N+1} = \frac{|h_i|}{\sqrt{\gamma P_i - \mu_i}}. \tag{4.31}$$

Proposition 4.5. *If $q_i > 0$ and $q_j > 0$ where $i < j$ then $q_i > q_j$*

It follows immediately from (4.29).

Proposition 4.6. *If $q_j = 0$ for some j then $q_k = 0$ for $k > j$.*

Suppose the contrary that there exists $k > j$ such that $q_k > 0$. Thus we have

$$q_k + q_{N+1} = \frac{|h_k|}{\sqrt{\gamma P_k}} \leq \frac{|h_j|}{\sqrt{\gamma P_j}} \leq \frac{|h_j|}{\sqrt{\gamma P_j - \mu_j}} = q_{N+1} \tag{4.32}$$

which is impossible.

Theorem 4.1. *The solution to the dual MAC problem* (4.21) *is given by*

$$q_i = \frac{1}{\sqrt{\gamma}}\left(\frac{|h_i|}{\sqrt{P_i}} - \sqrt{\frac{\sum_{i=k+1}^{N}|h_i|^2}{\left(P_T - \sum_{i=1}^{k} P_i\right)}}\right), \quad i = 1, \ldots, k \tag{4.33}$$

$$q_i = 0, \quad i = k+1, k+2, \ldots, N \tag{4.34}$$

where k is the least solution to the following inequality:

$$\frac{\sum_{i=k+1}^{N}|h_i|^2}{P_T - \sum_{i=1}^{k} P_i} \geq \frac{|h_N|^2}{P_N} \tag{4.35}$$

and

$$\gamma = \left(\frac{\sum_{i=1}^{k}|h_i|\sqrt{P_i} + \left(P_T - \sum_{i=1}^{k} P_i\right)\sqrt{\frac{\sum_{i=k+1}^{N}|h_i|^2}{\left(P_T - \sum_{i=1}^{k} P_i\right)}}}{P}\right)^2. \tag{4.36}$$

Proof. See Appendix B.6. □

4.3. Capacity Region of a Gaussian MIMO BC

In the section, we extend the minimax duality approach in SU-MIMO to find the capacity region of Gaussian MIMO BC with multiple LTCCs. In fact, some approaches relying on either subgradient or interior-point methods have been proposed for this problem [33,34]. However, these methods were only applicable to small-scale MIMO or MISO because their computational complexity is not appealing for large-scale scenarios such as massive MIMO. Herein, we propose an efficient solution to this problem, which follows the same idea as that of the preceding section, and in which each iteration is based on closed-form expressions.

4.3.1. System Model

Consider a K-user MIMO BC where the base station and each user $k = 1, 2, \ldots, K$ are equipped with N and M_k antennas, respectively. Let $\mathbf{H}_k$ denote the channel matrix for user k, and let $\mathbf{s}$ denote the composite signal that combines the data for all users in the downlink. Then, we can express the received signal at user k as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{s} + \mathbf{z}_k \tag{4.37}$$

where $\mathbf{z}_k$ is the Gaussian noise with distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. For Gaussian input, it was proved that dirty paper coding is capacity achieving [8]. The problem of finding the capacity region is usually formulated as a weighted sum rate maximization (WSRMax), which is written as

$$\begin{array}{ll} \underset{\{\mathbf{S}_k \succeq \mathbf{0}\}}{\text{maximize}} & \sum_{k=1}^{K} w_k \log \frac{|\mathbf{I}+\mathbf{H}_k \sum_{i=1}^{k} \mathbf{S}_i \mathbf{H}_k^\dagger|}{|\mathbf{I}+\mathbf{H}_k \sum_{i=1}^{k-1} \mathbf{S}_i \mathbf{H}_k^\dagger|} \\ \text{subject to} & \sum_{k=1}^{K} \operatorname{tr}(\mathbf{E}_{ik}\mathbf{S}_k) \leq P_i, \, \forall i \end{array} \tag{4.38}$$

where $\mathbf{E}_{ik}$ and $\mathbf{S}_k$ are the ith predefined positive semidefinite matrix and input covariance matrix for the kth user, P_i is the ith power constraint, and w_k is the weighting factor assigned to user k. Without loss of generality, we assume in the rest of the chapter that $0 < w_1 \leq w_2 \leq \ldots \leq w_K$ and $\sum_{k=1}^{K} w_k = 1$.

We remark that the work of [8] proved that the capacity region given in (4.38) is achievable in the case of PAPC, i.e., in the case where $\mathbf{E}_{ik} = \operatorname{diag}(\mathbf{e}_i)$ for $k = 1, 2, \ldots K$. Following similar arguments, we can show that the capacity region for the case of multiple LTCCs is also achievable.

4.3.2. Algorithm Description

Extending the result of minimax duality above, we can equivalently rewrite (4.38) as

$$\begin{array}{ll} \min\limits_{\mathbf{q}\geq\mathbf{0}} \max\limits_{\{\bar{\mathbf{S}}_k\succeq\mathbf{0}\}} & \sum_{k=1}^{K}\Delta_k \log|\mathbf{Q}_k+\sum_{i=k}^{K}\mathbf{H}_i^{\dagger}\bar{\mathbf{S}}_i\mathbf{H}_i| - w_K\log|\mathbf{Q}_k| \triangleq f^{\mathrm{DPC}}(\mathbf{q},\{\bar{\mathbf{S}}_k\}) \\ \text{subject to} & \sum_{k=1}^{K}\mathrm{tr}(\bar{\mathbf{S}}_k)=P;\mathbf{p}^T\mathbf{q}=P \end{array} \tag{4.39}$$

where $\Delta_k = w_k - w_{k-1} \geq 0$, $\mathbf{Q}_k = \sum\limits_i q_i\mathbf{E}_{ik}$, $P=\mathrm{tr}(\mathbf{P})$; $\{\bar{\mathbf{S}}_k\}$ and $\{\mathbf{Q}_k\}$ are considered as input covariance and noise covariance matrices in the dual MAC, respectively. Note that when we only consider a PAPC, the above formulation reduces to the one in [42] [2]. We also note that the objective in (4.39) is convex with $\mathbf{q}\geq\mathbf{0}$ and concave with $\{\bar{\mathbf{S}}_k\succeq\mathbf{0}\}$. Thus, there is a saddle point for (4.39). Let $(\mathbf{q}^*,\{\bar{\mathbf{S}}_k^*\})$ be the saddle point of (4.39). Then the optimal covariance matrices that achieve the capacity region in the BC are given by

$$\mathbf{S}_k = \mathbf{M}_k^{-1/2}\mathbf{U}_k\mathbf{V}_k^{\dagger}\mathbf{B}_k^{1/2}\bar{\mathbf{S}}_k\mathbf{B}_k^{1/2}\mathbf{V}_k\mathbf{U}_k^{\dagger}\mathbf{M}_k^{-1/2} \tag{4.40}$$

where $\mathbf{U}_k,\mathbf{V}_k$ are achieved from economy-size SVD of $(\mathbf{M}_k^{-1/2}\mathbf{H}_k^{\dagger}\mathbf{B}_k^{-1/2})$; $\mathbf{M}_k=\mathbf{Q}_k+\sum_{j=k+1}^{K}\mathbf{H}_j^{\dagger}\bar{\mathbf{S}}_j\mathbf{H}_j$, $\mathbf{B}_k=\mathbf{I}+\sum_{j=1}^{k-1}\mathbf{H}_k\mathbf{S}_j\mathbf{H}_k^{\dagger}$.

We discuss three benefits of using the minimax problem in (4.39) in computing the capacity region of a Gaussian MIMO BC. Firstly, the $\mathbf{S}_k$-maximization has *the same structure* for all types of LTCCs. Secondly, *the* $\mathbf{q}$*-minimization does not scale with the number of users.* Thirdly, projection onto the feasible sets of $\mathbf{q}$ and $\{\bar{\mathbf{S}}_k\}$ can be done using closed-form expressions. We exploit these properties to derive efficient solutions to the MIMO capacity region.

The proposed method for solving (4.39) follows the approach for SU-MIMO, which is described next. Denote $(\mathbf{q}^n,\{\bar{\mathbf{S}}_k^n\})$ as the obtained values of $(\mathbf{q},\{\bar{\mathbf{S}}_k\})$ after n iterations of the proposed iterative algorithm. For a given $\mathbf{q}^n$, $\bar{\mathbf{S}}^n$ is the solution to the maximization problem under a SPC to which gradient-projection-based methods are numerically shown to be efficient (see [2,24,25] for details). Moreover, if the sum capacity is of interest, i.e. $\Delta_k=0$ for all $k\geq 2$, it is easy to see that the maximization in (4.39) admits a water-filling solution.

Turning now to the problem of finding $\mathbf{q}^{n+1}$, we solve the optimization problem below:

$$\begin{array}{ll} \underset{\mathbf{q}\geq\mathbf{0}}{\text{minimize}} & \sum_{k=1}^{K}\Delta_k\log|\mathbf{Q}_k+\sum_{i=k}^{K}\mathbf{H}_i^{\dagger}\bar{\mathbf{S}}_i\mathbf{H}_i| - w_K\log|\mathbf{Q}_k| \\ \text{subject to} & \mathbf{p}^T\mathbf{q}=P. \end{array} \tag{4.41}$$

In light of Algorithm 4.2, we choose to minimize an upper bound of the objective instead of optimizing the original objective in (4.41). To this end, by again invoking the concavity of the logdet function, we obtain the following inequality

$$\log|\mathbf{Q}_k+\sum_{i=k}^{K}\mathbf{H}_i^{\dagger}\bar{\mathbf{S}}_i^n\mathbf{H}_i| \leq \log|\mathbf{\Phi}_k^n| + \mathrm{tr}\Big(\mathbf{\Phi}_k^{-n}\Big(\mathbf{Q}_k-\mathbf{Q}_k^n\Big)\Big) \tag{4.42}$$

where $\mathbf{\Phi}_k^n = \mathbf{Q}_k + \sum_{i=k}^K \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i^n \mathbf{H}_i, \mathbf{\Phi}_k^{-n} \triangleq (\mathbf{\Phi}_k^n)^{-1}$. Thus, $\mathbf{q}^{n+1}$ is found to be the optimal solution to the following problem

$$\begin{array}{ll} \underset{\mathbf{q}\geq\mathbf{0}}{\text{minimize}} & \sum_{k=1}^K \frac{\Delta_k}{w_K} \operatorname{tr}\Big(\mathbf{\Phi}_k^{-n}\mathbf{Q}_k\Big) - \log|\mathbf{Q}_k| \\ \text{subject to} & \mathbf{p}^T\mathbf{q} = P. \end{array} \tag{4.43}$$

We remark that the problem (4.43) has a similar form to (4.7); thus a gradient-projection-based algorithm can be easily customized to apply here. The overall algorithm is summarized in Algorithm 4.3 and its convergence proof is similar to that of Algorithm 4.2.

Algorithm 4.3: Proposed Algorithm for the Computation of the Capacity Region of a MIMO BC Based on AO.

```
Input: $\mathbf{q}^0$ feasible to $\mathcal{Q}$, and $\epsilon_2 > 0$.
1 Initialization: Set $n := 0$ and $\tau = 1 + \epsilon$.
2 while ($\tau > \epsilon_2$) do
3     $\mathbf{Q}_k^n = \sum_i q_i^n \mathbf{E}_{ik}$ .
4     Solve (3.41) and denote the optimal solution by $\{\bar{\mathbf{S}}_k^n\}$ .
5     If $n \geq 1$, let $\tau = |f^{\text{DPC}}(\mathbf{q}^n, \{\bar{\mathbf{S}}^n\}) - f^{\text{DPC}}(\mathbf{q}^{n-1}, \{\bar{\mathbf{S}}^{n-1}\})|$.
6     For each $k$, compute $\mathbf{\Phi}_k^{-n} = (\mathbf{Q}_k^n + \sum_{i=k}^K \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i^n \mathbf{H}_i)^{-1}$.
7     Solve (4.43) to find $\mathbf{q}^{n+1}$.
8     $n := n + 1$.
9 end
Output: Use the obtained $\{\bar{\mathbf{S}}_k^n\}_{k=1}^K$ and a similar BC-MAC transformation to
        that in [43] to find the optimal solution to (4.38).
```

4.3.3. MIMO Capacity Region with Joint SPC and PAPC

In this subsection we customize Algorithm 4.3 to deal with the specific case of MIMO capacity region with joint SPC and PAPC. We remark that no efficient solutions have been reported for this important case previously. For the SU-MIMO case, it is possible to find closed-form solutions based on solving the KKT conditions in the BC for some specific scenarios as shown in [62]. However, such a method appears to be impossible for the MU-MIMO case.

Our main point is to demonstrate that the equivalent minimax formulation in the MAC allows for efficient solutions to this special case. In particular, for the case of joint SPC and PAPC considered in this chapter, we show that Step 7 of Algorithm 4.3 (i.e., solving (4.43)) admits a closed form-solution.

We begin by rewriting (4.43) as

$$\begin{array}{ll} \underset{\mathbf{q}\geq\mathbf{0}}{\text{minimize}} & q_{N+1}\psi_{n,N+1} + \sum_{i=1}^N (q_i\psi_{n,i} - \log(q_{N+1} + q_i)) \\ \text{subject to} & \sum_{i=1}^{N+1} P_i q_i = P \end{array} \tag{4.44}$$

where $\psi_i = [\sum_{j=1}^{K} \frac{\Delta_j}{w_K} \mathbf{\Phi}_j^{-n}]_{i,i}$.

Theorem 4.2. *The solution to* (4.43) *in the special case of joint SPC and PAPC is given by*

$$q_i = 0, \quad i = k+1, k+2, \ldots, N \tag{4.45}$$

$$q_i = \frac{1}{\psi_i + \gamma P_i} - \frac{N-k}{(\psi_{N+1} - \sum_{i=1}^{k} \psi_i) + \gamma(P_T - \sum_{i=1}^{k} P_i)}, \quad i = 1, \ldots, k \tag{4.46}$$

where k *is the largest* $k \leq N-1$ *such that*

$$\frac{1}{\psi_i + \gamma P_i} \geq \frac{N-k}{(\psi_{N+1} - \sum_{i=1}^{k} \psi_i) + \gamma(P_T - \sum_{i=1}^{k} P_i)} \tag{4.47}$$

and γ *is the solution of the equation*

$$\sum_{i=1}^{k} \frac{P_i}{\psi_i + \gamma P_i} + \frac{(N-k)(P_T - \sum_{i=1}^{k} P_i)}{(\psi_{N+1} - \sum_{i=1}^{k} \psi_i) + \gamma(P_T - \sum_{i=1}^{k} P_i)} = P. \tag{4.48}$$

Proof. See Appendix B.7. □

4.4. Weighted Sum Rate with ZF

As mentioned in the previous section, the capacity region of a Gaussian MIMO BC is achieved by DPC which requires high complexity to implement. Thus zero-forcing methods are more practically appealing as they only involve linear processing. To the best of our knowledge, the most related work to the present chapter was carried out in [34], where Huh *et al.* proposed a gradient descent algorithm with barrier functions to solve the problem of weighted sum rate maximization (WSRMax) for ZF under multiple generic LTCCs. However, the proposed method in [34] has two main drawbacks: (i) it is only applicable to MISO systems, and (ii) it converges very slowly. To overcome the above shortcomings, we again transform the considered problem into the MAC and apply the above framework to derive a computationally efficient algorithm to solve it.

4.4.1. System Model

We consider the system model as described in Section 4.3.1 and, for description purposes, rewrite (4.37) as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \textstyle\sum_{j\neq k}\mathbf{H}_k\mathbf{x}_j + \mathbf{z}_k. \tag{4.49}$$

Again, channel state information is perfectly known at both the BS and users. For linear precoding we can express $\mathbf{x}_k$ as $\mathbf{x}_k = \mathbf{R}_k\mathbf{s}_k$, where $\mathbf{R}_k \in \mathbb{C}^{N\times M_k}$ and $\mathbf{s}_k \in \mathbb{C}^{M_k\times 1}$ denote the precoding matrix and information-bearing signal, respectively. We further assume that $\mathbf{s}_k$ consists of independent zero-mean and unit energy symbols, i.e., $\mathbf{s}_k \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$. For the ZF technique, the inter-user interference to user k is suppressed by designing $\mathbf{R}_k$ such that $\mathbf{H}_j\mathbf{R}_k = \mathbf{0}$ for all $j \neq k$. Thus, the WSRMax problem for ZF precoding with LTCCs is formulated as

$$\begin{aligned}
\underset{\{\mathbf{S}_k\succeq\mathbf{0}\}}{\text{maximize}} \quad & \textstyle\sum_{k=1}^{K} w_k \log|\mathbf{I}+\mathbf{H}_k\mathbf{S}_k\mathbf{H}_k^\dagger| && \text{(4.50a)}\\
\text{subject to} \quad & \mathbf{H}_j\mathbf{S}_k\mathbf{H}_j^\dagger = \mathbf{0},\ \forall j\neq k && \text{(4.50b)}\\
& \textstyle\sum_{k=1}^{K} \operatorname{tr}(\mathbf{E}_{ik}\mathbf{S}_k) \le P_i,\ \ i=1,2,\ldots,L && \text{(4.50c)}
\end{aligned}$$

where $\mathbf{S}_k = \mathbb{E}[\mathbf{x}_k\mathbf{x}_k^\dagger] = \mathbf{R}_k\mathbf{R}_k^\dagger$ is the input covariance matrix for user k, $\mathbf{E}_{ik}$ is the ith positive semidefinite matrix of user k, P_i is the power constraint associated with $\{\mathbf{E}_{ik}\}_{k=1}^{K}$, and $w_k \geq 0$ is the weighting factor assigned to the kth user to maintain some degree of fairness. In fact we have omitted the constraint that $\operatorname{rank}(\mathbf{S}_k) \leq \min(N, M_k)$ so that the precoder $\mathbf{R}_k$ can be extracted from $\mathbf{S}_k$. However, it was proved in [48] that this relaxation does not affect the optimality. This property is also obvious from the BC-MAC duality presented later on.

Before proceeding further, let us first simplify (4.50) by eliminating the zero-interference constraints. Denote $\check{\mathbf{H}}_k$ to be the channel matrix of all users, except for user k, i.e., $\check{\mathbf{H}}_k = [\mathbf{H}_1^\dagger, \ldots \mathbf{H}_{k-1}^\dagger, \mathbf{H}_{k+1}^\dagger, \ldots \mathbf{H}_K^\dagger]^\dagger$. For ZF precoding to be feasible, it should hold that $N \geq \sum_{k=1}^{K} M_k$, which is assumed in this chapter [20]. Let $\mathbf{V}_k$ be a basis of the null space of $\check{\mathbf{H}}_k$, then (4.50) reduces to the following maximization problem

$$\begin{aligned}
\underset{\{\tilde{\mathbf{S}}_k\succeq\mathbf{0}\}}{\text{maximize}} \quad & \textstyle\sum_{k=1}^{K} w_k \log|\mathbf{I}+\mathbf{H}_k\mathbf{V}_k\tilde{\mathbf{S}}_k\mathbf{V}_k^\dagger\mathbf{H}_k^\dagger|\\
\text{subject to} \quad & \textstyle\sum_{k=1}^{K} \operatorname{tr}(\mathbf{E}_{ik}\mathbf{V}_k\tilde{\mathbf{S}}_k\mathbf{V}_k^\dagger) \le P_i, i=1,\ldots,L.
\end{aligned} \tag{4.51}$$

We note that for this general problem, (4.51) can be recast as a MAXDET program [38] and solved by a dedicated optimization package such as SDPT3 [54]. However, solving (4.51) by generic convex solvers is not computationally efficient for large-scale problems, nor does it provide useful insights into the structure of the optimal input covariance matrices. In particular, modern convex solvers are mostly based on interior-point methods whose complexity increases rapidly with the problem size, e.g., with the number of transmit antennas N and/or the number of users K in the considered context. For the special case of multi-user MISO systems, Huh *et al.* presented a gradient descent algorithm with barrier functions but it converges very slowly [34]. For large-scale MISO systems, Huh *et al.* also proposed a low-complexity solution but it cannot achieve the optimal performance. Therefore, an efficient algorithm for general MIMO systems for ZF precoding with LTCCs has remained an open problem. In the following, we propose a low-complexity method to solve this problem.

4.4.2. Algorithm Description

The proposed solution for ZF methods is similar to Algorithm 4.3. Specifically, we first transform (3.31) into an equivalent problem in MAC, then apply AO and CCP to derive an efficient algorithm. To this end we state the following theorem.

Theorem 4.3. *The equivalent problem in the dual MAC of the problem* (4.51) *in the BC is the following minimax problem*

$$\begin{array}{rl} \min\limits_{\boldsymbol{\lambda}\geq 0} \max\limits_{\{\bar{\mathbf{S}}_k\succeq\mathbf{0}\}} & \sum_{k=1}^{K} w_k \log \frac{|\mathbf{V}_k^{\dagger}\left(\sum_{i=1}^{L}\lambda_i\mathbf{E}_{ik}\right)\mathbf{V}_k+\tilde{\mathbf{H}}_k^{\dagger}\bar{\mathbf{S}}_k\tilde{\mathbf{H}}_k|}{|\mathbf{V}_k^{\dagger}\left(\sum_{i=1}^{L}\lambda_i\mathbf{E}_{ik}\right)\mathbf{V}_k|} \\ \text{subject to} & \sum_{k=1}^{K}\operatorname{tr}(\bar{\mathbf{S}}_k)=P \\ & \mathbf{p}^T\boldsymbol{\lambda}=P \end{array} \tag{4.52}$$

where $\tilde{\mathbf{H}}_k=\mathbf{H}_k\mathbf{V}_k$, $\boldsymbol{\lambda}=[\lambda_1,\lambda_2,\ldots,\lambda_L]^T$ *and* $\mathbf{p}=[P_1,P_2,\ldots,P_L]^T$. *Let* ($\boldsymbol{\lambda}^*,\{\bar{\mathbf{S}}_k^*\}$ *) be the saddle point of* (4.52). *Then, the optimal solution* $\tilde{\mathbf{S}}_k^*$ *to* (4.51) *is given by*

$$\tilde{\mathbf{S}}_k^*=(\mathbf{V}_k^{\dagger}\boldsymbol{\Lambda}_k^*\mathbf{V}_k)^{-\frac{1}{2}}\mathbf{U}_k\mathbf{X}_k^{\dagger}\bar{\mathbf{S}}_k^*\mathbf{X}_k\mathbf{U}_k^{\dagger}(\mathbf{V}_k^{\dagger}\boldsymbol{\Lambda}_k^*\mathbf{V}_k)^{-\frac{1}{2}} \tag{4.53}$$

where $\boldsymbol{\Lambda}_k^*=\sum_{i=1}^{L}\lambda_i^*\mathbf{E}_{ik}$, *and* $\mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{X}_k^{\dagger}$ *is the compact singular value decomposition of* $(\mathbf{V}_k^{\dagger}\boldsymbol{\Lambda}_k^*\mathbf{V}_k)^{-1/2}\tilde{\mathbf{H}}_k^{\dagger}$.

We remark that the above theorem is a generalization of the duality result in [33,42] and its proof can be found in [16]. Existing solutions to a minimax problem such as (3.32) are based on interior-point methods [10,34]. Unfortunately, these methods do not scale well with the problem size, and thus are not suitable for large-scale MIMO systems.

The proposed solution for ZF methods is as follows. At the nth iteration of the proposed method, $\{\bar{\mathbf{S}}_k^n\}$ is found to be the solution to the following maximization problem:

$$\begin{array}{rl} \max & \sum_{k=1}^{K} w_k \log|\mathbf{V}_k^{\dagger}\boldsymbol{\Lambda}_k^n\mathbf{V}_k+\tilde{\mathbf{H}}_k^{\dagger}\bar{\mathbf{S}}_k\tilde{\mathbf{H}}_k| \\ \text{s.t.} & \sum_{k=1}^{K}\operatorname{tr}(\bar{\mathbf{S}}_k)=P;\bar{\mathbf{S}}_k\succeq\mathbf{0},k=1,\ldots,K \end{array} \tag{4.54}$$

where $\boldsymbol{\Lambda}_k^n=\sum_{i=1}^{L}\lambda_i^n\mathbf{E}_{ik}$. That is, to find $\{\bar{\mathbf{S}}_k^n\}$ we fix $\boldsymbol{\lambda}$. The above problem is actually the one of WSRMax under a single SPC, which admits a water-filling solution [20]. We skip the details for the sake of brevity and refer the interested reader to [20] for further details.

In the next step, we fix $\{\bar{\mathbf{S}}_k^n\}$ and solve for the optimal $\boldsymbol{\lambda}$ which minimizes the objective in (3.32). Again, we minimize an upper bound of the objective. In this regard, the following inequality is in order [47]

$$\begin{aligned}\log|\mathbf{V}_k^\dagger\boldsymbol{\Lambda}_k\mathbf{V}_k+\tilde{\mathbf{H}}_k^\dagger\bar{\mathbf{S}}_k\tilde{\mathbf{H}}_k| \leq &\log|\boldsymbol{\Phi}_k^n| \\ &+\operatorname{tr}\Big(\mathbf{V}_k\boldsymbol{\Phi}_k^{-n}\mathbf{V}_k^H\Big(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_k^n\Big)\Big)\end{aligned} \tag{4.55}$$

where $\boldsymbol{\Phi}_k^n \triangleq \mathbf{V}_k^\dagger\boldsymbol{\Lambda}_k^n\mathbf{V}_k+\tilde{\mathbf{H}}_k^\dagger\bar{\mathbf{S}}_k^n\tilde{\mathbf{H}}_k$, and $\boldsymbol{\Phi}_k^{-n} \triangleq \Big(\boldsymbol{\Phi}_k^n\Big)^{-1}$. Thus, at the nth iteration of the proposed algorithm, $\boldsymbol{\Lambda}_k^{n+1}$ is the solution to the following problem

$$\begin{aligned}&\min\ \textstyle\sum_{k=1}^K w_k\Big(\operatorname{tr}\Big(\mathbf{V}_k\boldsymbol{\Phi}_k^{-n}\mathbf{V}_k^\dagger\boldsymbol{\Lambda}_k\Big)-\log|\mathbf{V}_k^\dagger\boldsymbol{\Lambda}_k\mathbf{V}_k|\Big)\triangleq g(\boldsymbol{\lambda}) \\ &\text{s.t. } \mathbf{p}^T\boldsymbol{\lambda}=P, \boldsymbol{\lambda}\succeq\mathbf{0}.\end{aligned} \tag{4.56}$$

We remark that the feasible set of (4.56), denoted by $\Theta \triangleq \{\mathbf{p}^T\boldsymbol{\lambda}=P;\boldsymbol{\lambda}\geq 0\}$, is a simplex. Since the projection on a simplex can be done efficiently [51], we can apply the gradient projection (GP) or conjugate gradient projection method to solve (4.56). The proposed algorithm to solve the WSRMax problem with ZF and LTCCs is summarized in Algorithm 4.4. Note that the GP method in line 7 follows similar procedures to those outlined in Algorithm 4.1. The convergence proof of Algorithm 4.4 in fact follows the similar arguments as those of [2, 14, 15] and Appendix B.3 of the present chapter, thus it is skipped for the sake of brevity. We also note that our proposed method can be easily modified to deal with the WSRMax problem for successive zero-forcing DPC (SZFDPC) [16] with multiple LTCCs.

Algorithm 4.4: The Proposed Algorithm for Solving (4.51).

```
  Input: $\boldsymbol{\lambda}^0$ feasible to $\Theta$, $\epsilon_2 > 0$.
1 Initialize $n := 0$, and $\tau_2 = 1+\epsilon_2$.
2 while $\tau_2 > \epsilon_2$ do
3     Compute $\boldsymbol{\Lambda}_k^n = \sum_{i=1}^L \lambda_i^n \mathbf{E}_{ik}$
4     Apply the water-filling algorithm to solve (4.54). Denote the optimal solution by
        $\{\bar{\mathbf{S}}_k^n\}$.
5     For $n \geq 1$, compute $\tau_2 = |f(\boldsymbol{\lambda}^n, \bar{\mathbf{S}}^n) - f(\boldsymbol{\lambda}^{n-1}, \bar{\mathbf{S}}^{n-1})|$.
6     For each $k$, set $\boldsymbol{\Phi}_k^n = (\mathbf{V}_k^\dagger\boldsymbol{\Lambda}_k^n\mathbf{V}_k + \tilde{\mathbf{H}}_k^\dagger\bar{\mathbf{S}}_k^n\tilde{\mathbf{H}}_k)$.
7     Solve (4.56) to find $\boldsymbol{\lambda}^{n+1}$ using GP.
8     $n := n+1$.
9 end
  Output: $\{\bar{\mathbf{S}}_k^n\}_{k=1}^K$ and apply the BC-MAC transformation to compute optimal
          $\{\tilde{\mathbf{S}}_k^n\}_{k=1}^K$.
```

4.5. Numerical Results

In this section, we evaluate the performance of the proposed algorithm by numerical experiments. More specifically, we focus on the important case of joint SPC and PAPC. For notational convenience, we denote the SPC and the sum of PAPCs as P_T and $P_A = \sum_{i=1}^{N} P_i$, respectively. Unless explicitly stated otherwise, we consider here the most common case encountered in practice, where each transmit antenna is subject to the same power constraint, i.e., $P_i = P_0 = \frac{P_A}{N}$, for $i = 1, 2, \ldots, N$. As mentioned earlier, we are interested in the nontrivial case where $\min\{P_i\} < P_T < P_A$. In fact, if $\min\{P_i\} \geq P_T$, then the PAPC can be removed without loss of optimality. Similarly, if $P_T \geq P_A$, then the SPC can be eliminated. For MU-MIMO, all users are equipped with the same number of receive antennas i.e., $M_k = M$. Unless stated otherwise, the error tolerances ϵ_1 and ϵ_2 are set to 10^{-6} for all simulations. Other relevant simulation parameters are specified for each setup. Note that each average result is based on Monte Carlo simulations over 1000 i.i.d. channel realizations. Each entry of the channel matrix is drawn from a circularly-symmetric complex Gaussian distribution with unit variance. The MATLAB code was executed on a 64-bit desktop that supports 8 GB RAM and Intel CORE i7.

4.5.1. SU-MIMO

In the first experiment, we study the convergence behavior of the objective $f(\mathbf{q}, \bar{\mathbf{S}})$ in (4.3) for different channel realizations under joint SPC, PAPC and interference power constraint. As mentioned earlier in Section 4.1, for the interference power constraint, the considered MIMO system is assumed to act as a SU. The channel between the SU and the PU, denoted by $\mathbf{G}$, is randomly generated following circularly-symmetric complex Gaussian distribution with unit variance. The received interference power at the PU is $\text{tr}(\mathbf{G}\mathbf{S}\mathbf{G}^{\dagger})$ which is constrained to stay below a predetermined threshold P_{int}. The values of P_T i.e., the SPC and P_{int} are fixed at $0.6P_A$ and $0.06P_A$, respectively where P_A is defined above. Note that the objectives are plotted including the objectives after the maximization and minimization steps of each iteration. As can be seen from Fig. 4.1, the objectives increase with respect to maximization and decrease with respect to minimization. However, our novel method in the minimization step guarantees the consistent decrease of the objective until convergence. Thus, we can obtain monotonic convergence and avoid the fluctuations which can occur in the case of pure AO.

Next, we study the convergence properties of the proposed algorithm under different settings of SPC, PAPC and interference power limit. In this experiment, we fix the interference power limit at $P_{\text{int}} = 0.1P_T$. The channel matrix associated with the interference power constraint is generated as done in the first experiment. The convergence rate of Algorithm 4.2 with different ratios of P_T/P_A and PAPC is plotted in the Fig. 4.2. The residual error is defined as the absolute difference between two

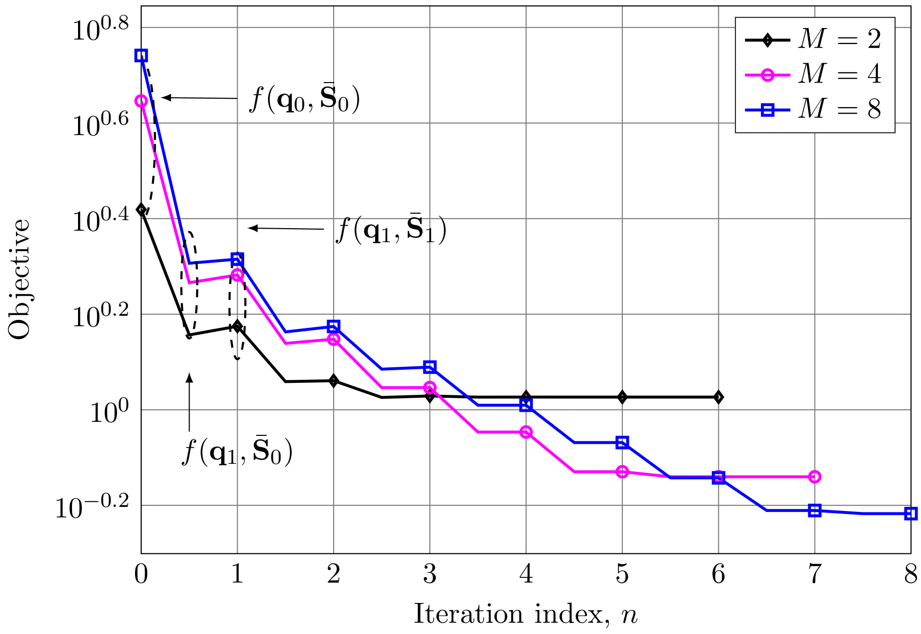


Figure 4.1.: Objective convergence behavior with respect to different number of receive antennas M with $N = 2$ transmit antennas, $P_A = 0$ dBW, $P_T = 0.6P_A$ and the interference power limit is $P_{\text{int}} = 0.06P_A$.

consecutive objectives. For the randomly generated channel realization considered in Fig. 4.2, Algorithm 4.2 can converge with less than eight iterations and even fewer iterations in high power scenarios. For a fixed P_A, the convergence results of Algorithm 4.2 seem to be insensitive to the ratio between P_T and P_A. We also notice that the rate of convergence of the proposed algorithm is also insensitive to different PAPC settings.

We now compare the capacity of the following MIMO channel

$$\mathbf{H} = \begin{bmatrix} 0.1189 + 0.1515i & 0.1238 + 0.3326i & 0.8572 + 0.1131i \\ -0.3198 - 0.3663i & -0.6491 + 0.2784i & 0.3392 - 0.1974i \\ -0.1019 + 0.6639i & 0.3663 - 0.3097i & -0.1116 - 0.1101i \end{bmatrix}$$

under joint SPC and PAPC using the proposed approach and [61,62]. Each entry of the above channel is randomly generated following a circularly-symmetric complex Gaussian distribution with unit variance. In order to execute the solution in [62], we have to set the power of each transmit antenna to be equal. Also, the closed-form solution in [62] requires that the total transmit power must be $P_T \geq 8.3137$ W and $P_0 \geq 5.1071$ W. In other words, the SNR should be larger than 9.198 dB for this method to work. In low power scenarios, the approaches in [61,62] yield an input covariance matrix that violates the positive semidefinite constraints. To achieve a fair comparison, we compensate for negative eigenvalues and scale the resulting covariance matrix by a factor so that all power constraints are met. As can be seen from Fig. 4.3a, the capacity generated by our algorithm and that of [62] are the same in high power constraint regions as expected. However, in the low power regime, our proposed solution outperforms other methods (c.f. Fig. 4.3b).

Next, we consider a MISO example with the channel matrix

$$\mathbf{H} = [-2.0819+0.9689i, 1.0171-1.2102i, -0.5338-0.1707i, 0.2299-0.0723i].$$

In addition, we choose the SPC P_T and the PAPC P_0 to be 8 dbW and 3 dBW, respectively, so that the setup is nontrivial. The purpose of this experiment is to verify the analytical solution expressed in (4.33) and (4.34). First we use MOSEK [4], which is a powerful commercial convex solver, to solve (4.23). The resulting optimal solution obtained in this way is $q = [2.3604, 1.2702, 0, 0, 1.1385]^T$. The same result is also achieved by our analytical formulas as explained as followed. We can check that the condition in (4.35) is satisfied when $k = 2$. Thus, $q_3 = q_4 = 0$ which results directly from Lemma 4.6. Moreover, substituting $k = 2$ into (4.36) immediately yields $\gamma = 0.1436$. Substituting this value of γ into (4.33) and (B.47), we obtain $q_1 = 2.3604, q_2 = 1.2702, q_5 = 1.1385$, which is in accordance with the result given by the MOSEK solver. By BC-MAC transformation, we can obtain the same solution as [32]. Since both solutions are closed-form, their computational efficiency is definitely better than that of the iterative solution in [31].

Now we investigate the performance of Algorithm 4.2 for general MIMO systems under joint SPC and PAPC. In Fig. 4.4, we compare the number of iterations of the proposed algorithm with that of the interior-point method. Specifically, the considered problem is solved by MOSEK which is a commercial interior-point-based convex solver. As can be seen, the number of iterations of Algorithm 4.2 increases nearly linearly with the number of transmit antennas N. On the other hand, the interior-point method requires fewer iterations to converge, approximately three to four times less than that of Algorithm 4.2 in large-scale scenarios. However, this does not necessarily mean that the interior-point method is more computationally efficient, as the overall complexity also depends on the per-iteration cost. The problem of the interior-point method is that its per-iteration cost is much higher than that of Algorithm 4.2. As a result, in term of overall efficiency, the interior-point method performs worse than Algorithm 4.2. This issue is further studied in the next simulation.

To obtain a more complete comparison, we report the average run time of Algorithm 4.2 along with common interior-point solvers i.e., SDPT3 [54] and MOSEK in Table 4.1. Both the solvers are executed through the parser YALMIP [55]. The ratio P_T/P_A is set to 0.8. The symbol × denotes the case where the run time is extremely high or where the solvers could not run successfully due to insufficient memory or other reasons. Note that the run time accounts for both the number of iterations and the per-iteration complexity. We recall that the per-iteration complexity of an interior-point-based method for the similar problem is $\mathcal{O}(N^6)$ [47], compared to $\mathcal{O}(N^3)$ for Algorithm 4.2. As can be seen clearly from Table 4.1, interior-point-based convex solvers are not suitable for large-scale MIMO systems because their complexity and memory requirements can increase rapidly with the problem size which results in prohibitive computation time. Meanwhile, the proposed algorithm consistently shows a low run time, which is relatively independent of P_A.

Further experiments are carried out with the average capacity of MIMO systems under different power constraint settings i.e., SPC, PAPC, and joint SPC and PAPC. As shown in Fig. 4.5, the capacity under joint SPC and PAPC is lower than that of PAPC because in this case, the maximum is achieved at a point where not all PAPCs are satisfied with equality. We also observe that when the PAPC is set to be equal for all transmit antennas, the capacity of PAPC is close to the one under SPC as previously observed in [1], [15].

Table 4.1.: Average run time (seconds) comparison with $P_T = 0.8P_A$, $M = 2$ receive antennas. The run time is averaged over 1000 channel realizations.

P_A	Algorithms/solvers	No. of transmit antennas N			
		16	32	64	128
0 dBW	Algorithm 4.2	**0.035**	**0.239**	**0.823**	**3.890**
	MOSEK	0.040	0.592	×	×
	SDPT3	0.414	2.160	×	×
10 dBW	Algorithm 4.2	**0.025**	**0.214**	**0.780**	**3.807**
	MOSEK	0.047	0.618	×	×
	SDPT3	0.419	2.221	×	×

4.5.2. MU-MIMO

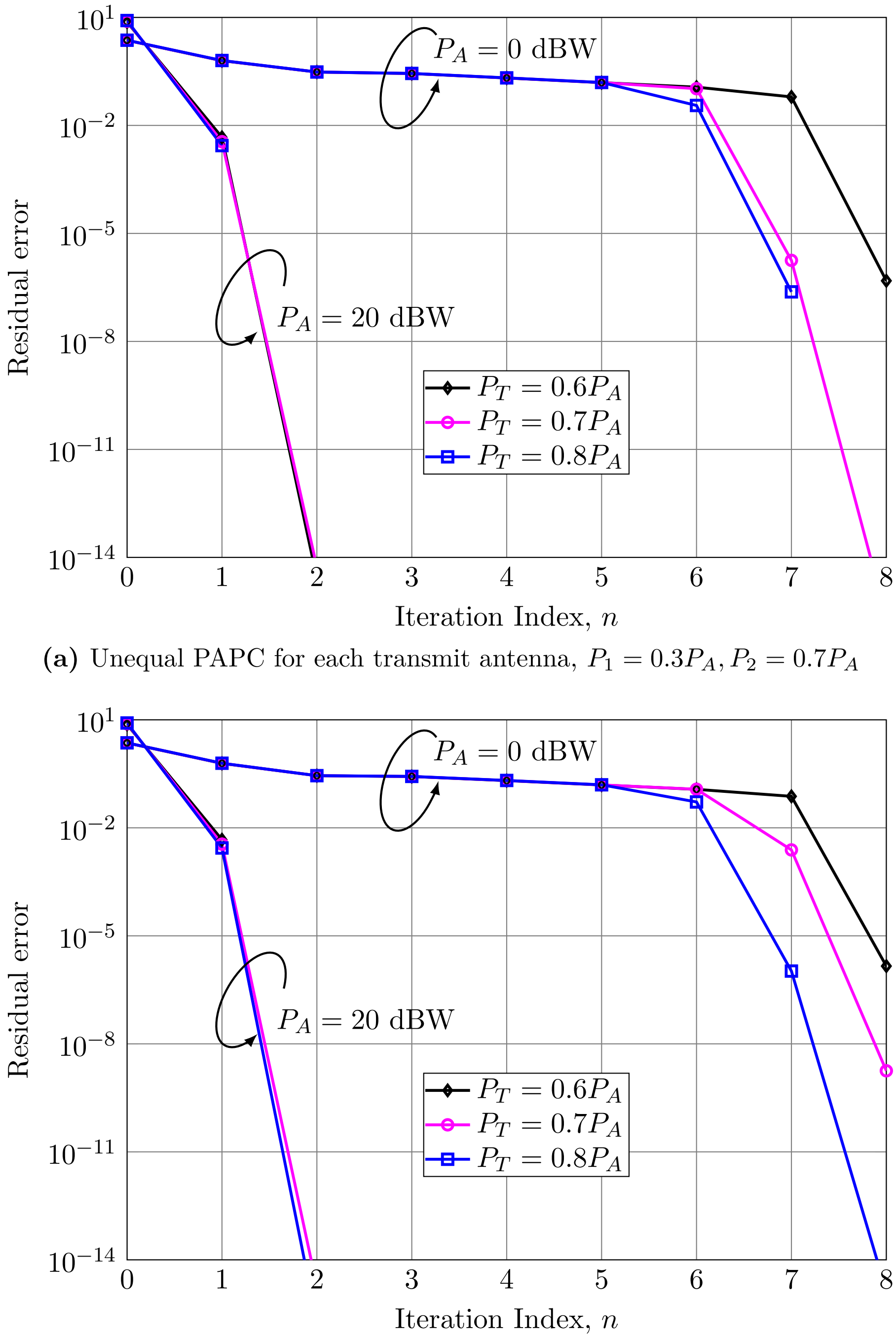
In the first simulation, we plot the cumulative distribution functions (CDF) of the number of iterations taken by Algorithm 4.3 to converge. The low and high SNR scenarios as well as different power ratios i.e., P_T/P_A are studied. The CDF for each scenario is obtained over 1000 channel realizations. We can clearly see in Fig. 4.6 that 95% of the cases, Algorithm 4.3 terminates within 30 and 10 iterations for low and high power regions, respectively.

Taking advantage of our low-complexity algorithms, we characterize the capacity region of linear precoding (ZF) and nonlinear precoding method (DPC) in a realistic massive MIMO scenario under joint SPC and PAPC. In particular, we consider the typical urban micro-cell WINNER II B1 channel model [58] where two users are distributed around a centered base station in a single cell. In addition, we only consider the path loss and ignore shadowing. The noise power is set to -94 dBm over a bandwidth of 100 MHz. The base station and each user are equipped with 128 and 2 antennas, respectively. We can see clearly that the capacity of ZF with joint SPC and PAPC is close to that of DPC in massive MIMO settings. Both are less than the capacity of DPC with PAPC and the feasible region is still bounded by the SPC.

Finally, we study the performance of the average sum rate of different precoding methods, including ZF, SZFDPC [16] and DPC under joint SPC and PAPC. For the same set of power constraints, the average sum rate of ZF is lower than that of suboptimal precoding SZFDPC, while DPC remains the optimal solution with the highest sum rate. We can also see that when the number of transmit antennas increases, the performance of ZF and SZFDPC methods approaches that of DPC.

4.6. Summary

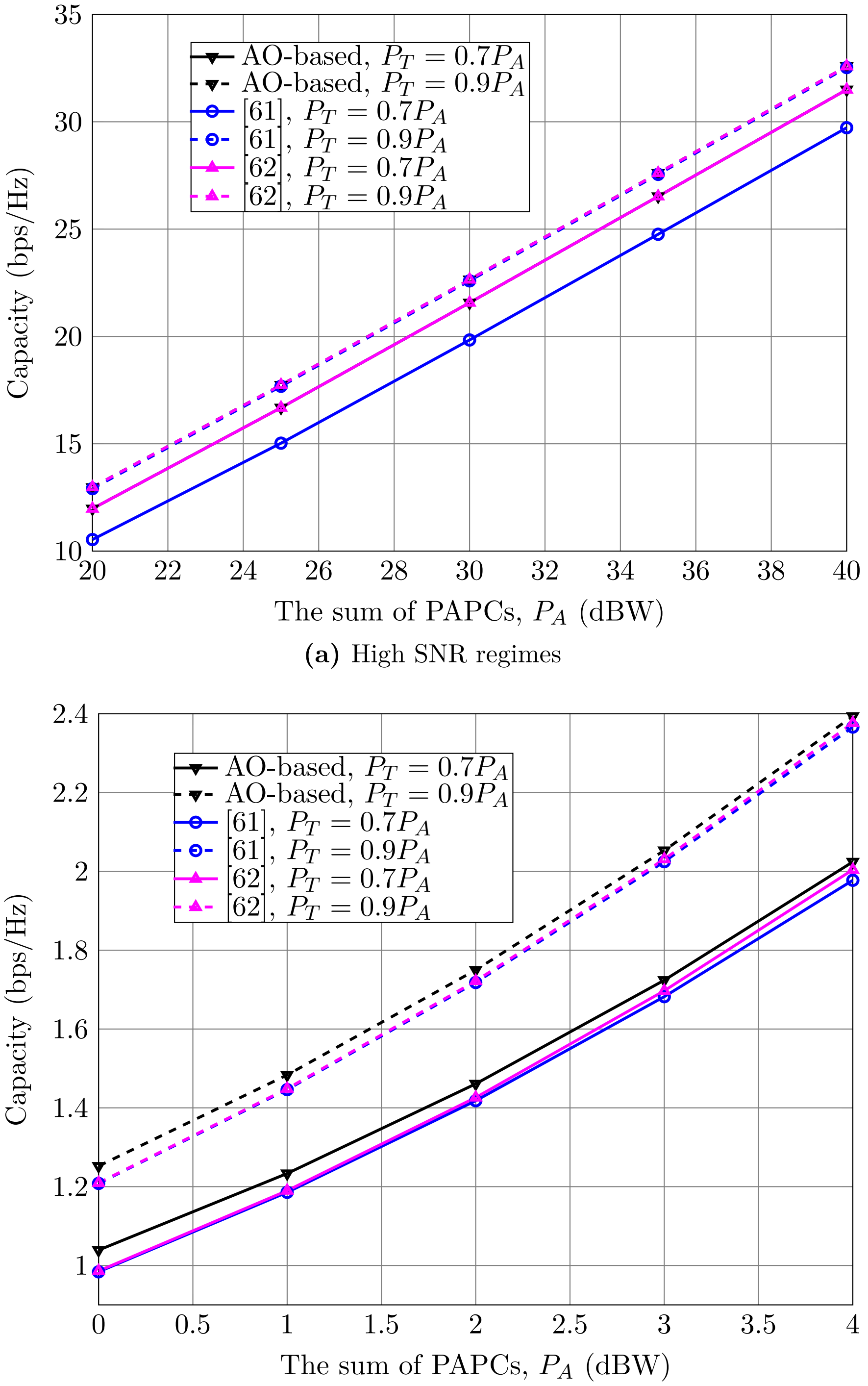
We have proposed an efficient approach to computing the MIMO capacity and characterizing the capacity region under arbitrary combination of linear transmit covariance constraints. The approach is based on minimax duality and CCP to derive water-filling-like algorithms. Interestingly, our approach can be easily extended to the MU-MIMO with DPC. For the special case of MIMO capacity with joint SPC and PAPC, we have provided analytical solutions in addition to iterative algorithms. The numerical results have supported that our proposed algorithm outperforms the well-known interior-point method in overall complexity and thus is suitable for large-scale MIMO systems. These successes may hope to tackle the computation difficulties of the previously known algorithms which are mostly relied on either subgradient or interior-point methods.



(a) Unequal PAPC for each transmit antenna, $P_1 = 0.3P_A, P_2 = 0.7P_A$

(b) Equal PAPC for each transmit antenna, $P_1 = P_2 = 0.5P_A$

Figure 4.2.: Convergence rate of the proposed algorithm for SU-MIMO under multiple power constraints i.e., SPC, PAPC and interference power constraint with $N = 2$ transmit antennas and $M = 8$ receive antennas. The interference power constraint $P_{\text{int}} = 0.1P_T$.



(a) High SNR regimes

(b) Low SNR regimes

Figure 4.3.: Capacity comparison of the proposed algorithm and existing methods under joint SPC and PAPC with $M = 3$ receive antennas and $N = 3$ transmit antennas.

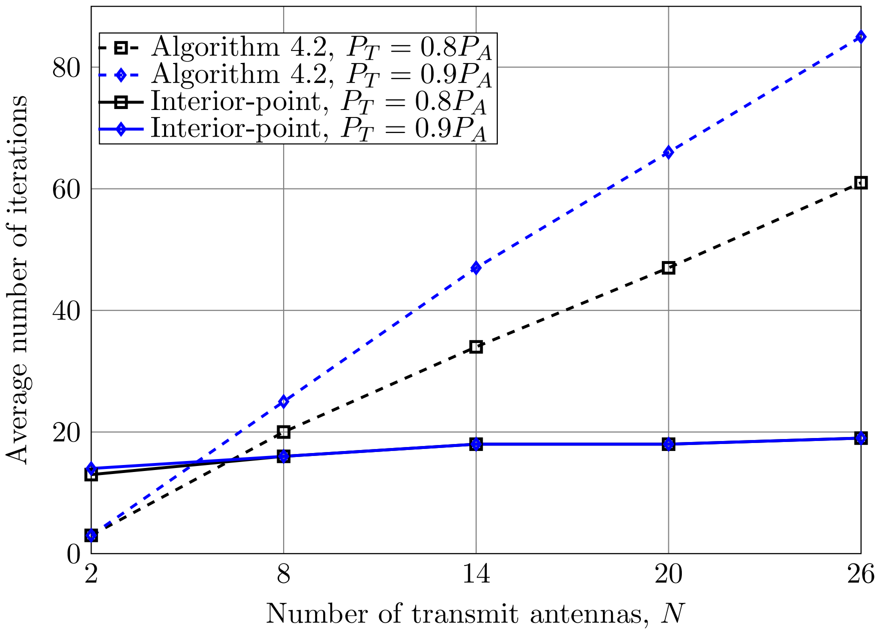


Figure 4.4.: Average number of iterations to converge under SPC and PAPC with $M = 2$ receive antennas, the sum of PAPCs is $P_A = 10$ dBW.

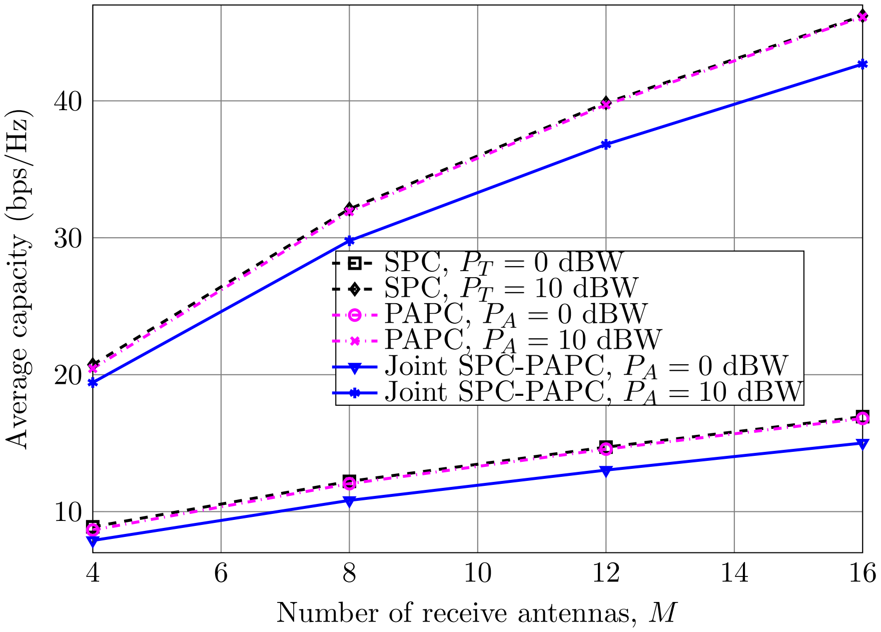
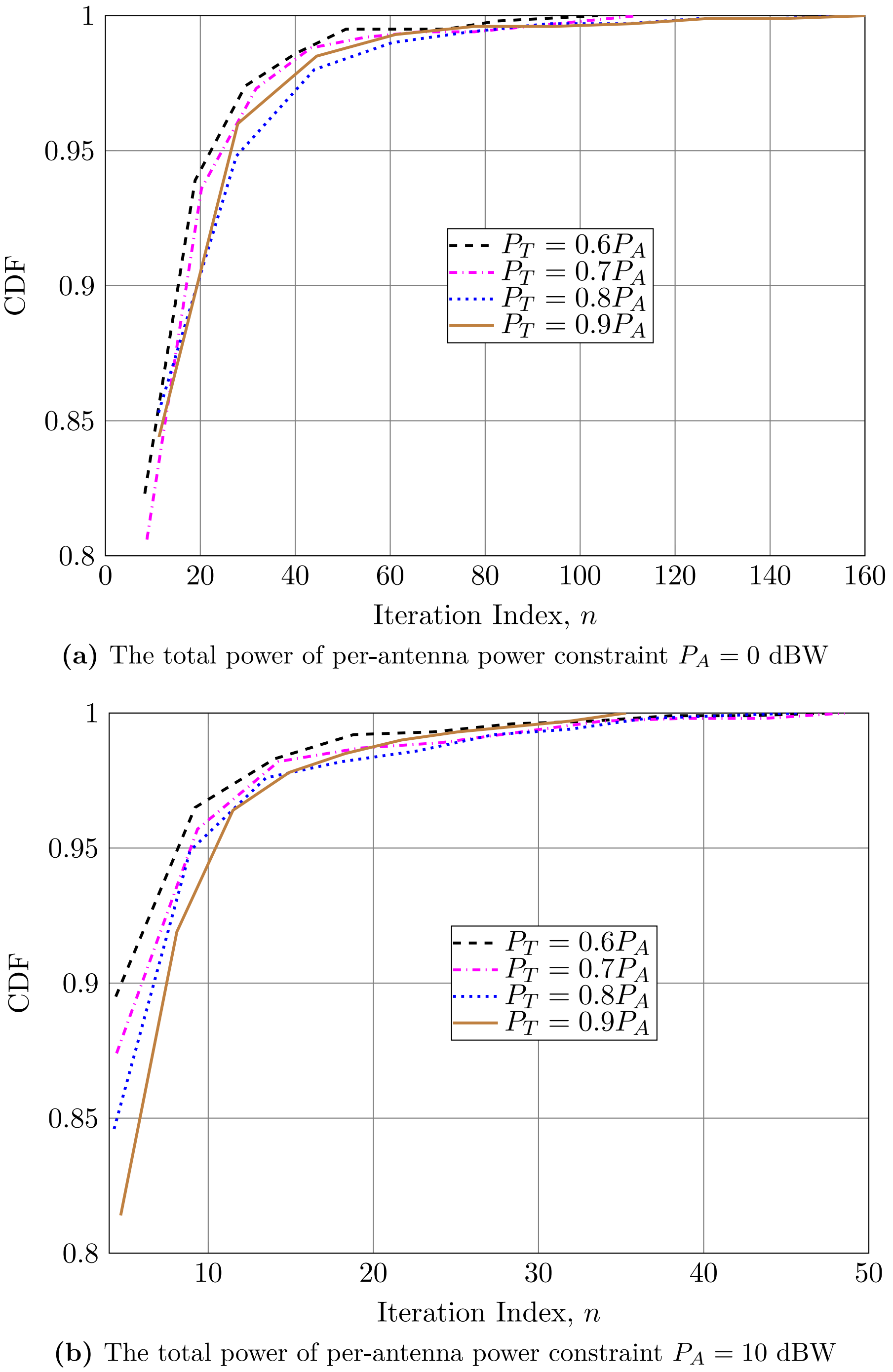


Figure 4.5.: Average capacity with $N = 16$ transmit antennas, $P_T = 0.8P_A$ for joint SPC-PAPC.



(a) The total power of per-antenna power constraint $P_A = 0$ dBW

(b) The total power of per-antenna power constraint $P_A = 10$ dBW

Figure 4.6.: Cumulative distribution function of the number of iterations to converge under joint SPC and PAPC. The number of transmit antennas $N = 4$, number of receive antennas $M = 2$ and number of users $K = 2$.

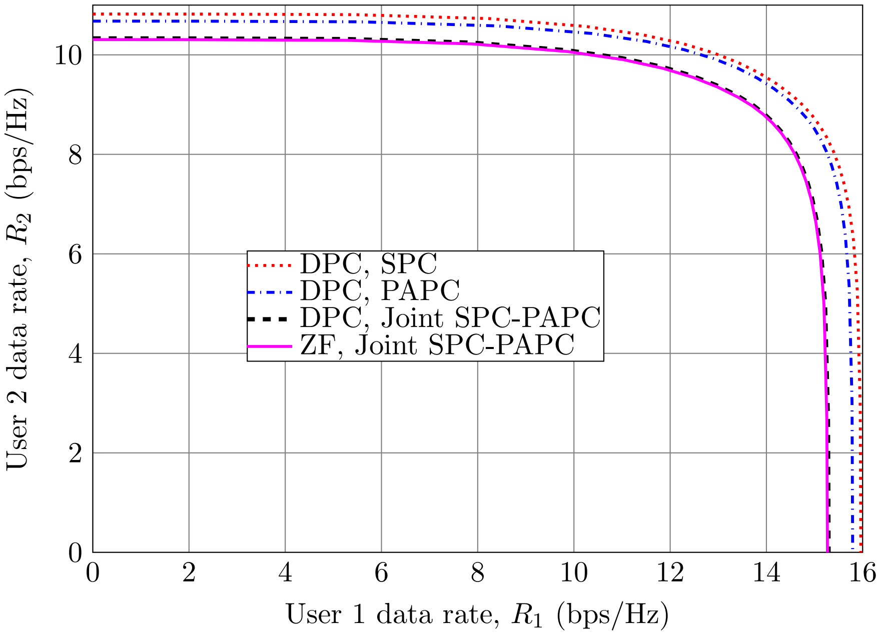


Figure 4.7.: Comparison of capacity region of a massive MIMO system with $N = 128$ transmit antennas, $M = 2$ receive antennas and $K = 2$ users. The sum power constraint is $P_T = 16$ dBW for the SPC case, the sum of PAPCs is equal to $P_A = 16$ dBW for the PAPC case, and for the case of joint SPC and PAPC we set $P_A = 16$ dBW and $P_T = 0.8P_A$.

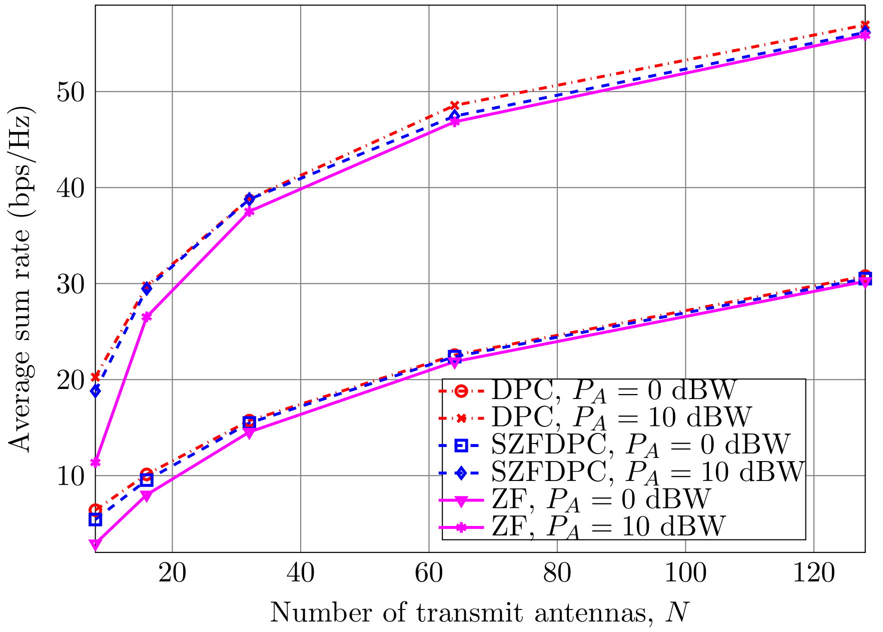


Figure 4.8.: Average sum rate of different precoding methods i.e, ZF, DPC and SZFDPC under joint SPC and PAPC with $M = 2$ receive antennas, $K = 4$ users, and the sum power constraint is $P_T = 0.8P_A$.

Chapter 5

Machine Learning and Its Applications to Capacity-related Problems

The proliferation of data-driven applications and computing resources has attracted huge attention to Artificial Intelligence (AI) and Machine Learning (ML) in recent years. The fact that ML is extremely useful for applications which are difficult to model or have no existing solutions. In addition, it can provide good trade-off between the complexity and the performance. Those features are therefore appealing to the research of wireless communications in which the majority of research cannot fully characterize a system due to the difficulties of mathematical formulations. More importantly, a proper formulation may result in high-complexity solution depending on the complexity of problems. In this chapter, we carry out some initial experiments to verify possible applications of ML to capacity-related problems. More specifically, we take advantages of ML to obtain suboptimal solution to the sum rate of successive zero-forcing dirty paper coding (SZFDPC) with PAPC.

The remainder of the chapter is organized as follows. The fundamentals of ML are described in Section 5.1. In Section 5.2, we derive a solution to compute the sum rate of a Gaussian MIMO BC with SZFDPC and PAPC using a suboptimal ML-based approach and present some numerical results followed by the conclusion in Section 5.3. Most of the content and results in Section 5.2 have been appeared in [18] under © 2019 IEEE.

5.1. Fundamentals of Machine Learning

Before going into the details, we recall what learning is. It is a well-known fact that humans can learn from experience and computers learn from data. For example, one can recognize and respond properly to a situation which once happened in the past. Human learning is therefore involved in memorization and adaptation. More importantly, it can be generalized to deal with a set of similar problems. Similarly, machine learning is designed to enable computers to adapt and generalize their actions as accurately as possible. In this section, we introduce the basis of machine learning including its applications, learning stages, different ML categories and an important learning task- regression- which will be applied to our considered capacity problem.

5.1.1. Applications of Machine Learning

According to [64, 65], ML can be applied to a number of engineering problems. In fact, it is the most feasible approach to applications of unknown or undesirable solutions due to model or algorithm deficit. On the contrary to conventional optimal engineering solutions, ML is commonly referred to as a black box, thus is suitable for tasks which do not require explicit reasoning or detailed explanation. In addition, ML is beneficial in the presence of sufficiently large training sets or the sets can be created easily. In this respect, the phenomenon or the environment of the learning task is considered stationary for a sufficiently long period. Moreover, ML can address a lot of problems of loose constraints or provide satisfactory performance to an algorithm deficit as mentioned above.

In practice, ML has broad applications including text classification, speech processing, computer vision and many other applications. The two most important learning tasks of ML are classification and regression. The former task, which have been applied extensively in objective recognition, face detection, text classification etc., is to assign a pattern to a category. On the other hand, the latter, which is usually described as fitting problem, learns a model using a training set so that it can predict an output correctly in the future.

5.1.2. Learning Stages

Here, we present the relevant processes to choose, apply and evaluate ML. Generally, different stages of ML are in the following [66, p. 10]:

- Data collection and preparation: To train and test the algorithms, the relevant data need to be collected either from scratch or assembling from available data. The inputs and the outputs are associated with the features and the targets/ responses, respectively.
- Feature selection: This process can be used to increase the robustness of ML since it is involved in identifying the most useful features for the considered problem.
- Algorithm choice: The accuracy of ML for a given data set depends on the choice of an appropriate algorithm. We will present some of fundamental algorithms in the next subsections.

- Parameters and model selection: Since ML is data-specific problem, parameters and model should be tuned to relevant problems manually or experimentally.
- Training: The training is to build a model based on given data to predict outputs of new data.
- Evaluation: We need to test the model obtained from the training before any possible deployments.

5.1.3. Types of Machine Learning

In principle, ML can be classified into four categories [66, p. 6]:

- Supervised learning: Given a training set with correct responses, an algorithm generalizes to make predictions.
- Unsupervised learning: Since correct responses are not available, an algorithm tries to categorize similar inputs.
- Reinforcement learning: This learning strategy has to explore and try out different possibilities until it knows how to get the right answer. Reinforcement learning in fact fills the gap between supervised learning and unsupervised learning.
- Evolutionary learning: This type of learning is developed based on modeling the biological evolution. Inspired by the process of natural selection, an evolutionary algorithm generally involves four important steps: initialization of population, selection, genetic operators and termination under a certain condition [67].

Most of ML-based applications rely on supervised learning which we can make predictions or classifications based on known data. In this chapter, we will focus on its key learning task - regression - which is exploited to solve our capacity problem.

5.1.4. Regression

Some of practical applications of regression are to predict stock price or Gross Domestic Product (GDP) growth rate of a country, to name a few. In the following, we present the basic methods of both linear and nonlinear regression.

5.1.4.1. Linear regression

With linear regression, we may fit data to a line in case of a scalar input or a hyperplane in other cases. Assume that we have a training set $\{\mathbf{x}_i, y_i\}_{i=1}^{M}$ where $\mathbf{x}_i$

and y_i are the ith input vector of size N and scalar output, respectively. In linear regression model, the output and input are linearly dependent:

$$y_i \approx \mathbf{x}_i^T \mathbf{w} + b \tag{5.1}$$

where $\mathbf{w}$ and b are the weight vector and the bias, respectively. Note that we use the approximately equal sign since we may not perfectly fit all the data to a line or a plane.

Denote $\hat{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}$ and $\hat{\mathbf{w}} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$, (5.1) can be rewritten as

$$y_i \approx \hat{\mathbf{x}}_i^T \hat{\mathbf{w}}. \tag{5.2}$$

Least square regression To measure the accuracy of an algorithm, we can use a cost function. A least square (LS) cost function can be formed as follows:

$$g(\hat{\mathbf{w}}) = \sum_{i=1}^{M} (\hat{\mathbf{x}}_i^T \hat{\mathbf{w}} - y_i)^2. \tag{5.3}$$

Our objective is to minimize this error over $\hat{\mathbf{w}}$ which is in turn to solve the following optimization problem

$$\underset{\hat{\mathbf{w}}}{\text{minimize}} \sum_{i=1}^{M} (\hat{\mathbf{x}}_i^T \hat{\mathbf{w}} - y_i)^2. \tag{5.4}$$

Although it is possible to utilize any convex optimization software to solve this convex problem, we are interested in deriving a closed-form solution. Specifically, taking the gradient of this problem yields

$$\nabla g(\hat{\mathbf{w}}) = 2 \sum_{i=1}^{M} \hat{\mathbf{x}}_i (\hat{\mathbf{x}}_i^T \hat{\mathbf{w}} - y_i). \tag{5.5}$$

Setting this gradient to zero, we obtain the solution to (5.4)

$$\left(\sum_{i=1}^{M} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T \right) \hat{\mathbf{w}} = \sum_{i=1}^{M} \hat{\mathbf{x}}_i y_i. \tag{5.6}$$

If we stack M vectors into a matrix, i.e., $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_M]$, then the above equation is equivalent to

$$\hat{\mathbf{X}} \hat{\mathbf{X}}^T \hat{\mathbf{w}} = \hat{\mathbf{X}} \mathbf{y} \tag{5.7}$$

where $\mathbf{y}$ is the vector of the outputs. When $\hat{\mathbf{X}} \hat{\mathbf{X}}^T$ is invertible, the weight vector can be found as

$$\hat{\mathbf{w}} = (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \hat{\mathbf{X}} \mathbf{y}. \tag{5.8}$$

The solution above is usually referred to as ordinary least square (OLS). We may employ pseudo-inverse instead of normal inverse to guarantee the feasibility of (5.8).

Ridge regression Notice that we can always guarantee the inverse of (5.1) by adding a small term $\alpha\mathbf{I}$ to $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$, in other words

$$\hat{\mathbf{w}} = (\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \alpha\mathbf{I})^{-1}\hat{\mathbf{X}}\mathbf{y}. \tag{5.9}$$

If we do not regularize $\hat{w}_0 = b$, then the considered problem is given

$$\underset{\mathbf{w},b}{\text{minimize}} \sum_{i=1}^{M} (\mathbf{x}_i^T\mathbf{w} + b - y_i)^2 + \alpha||\mathbf{w}||_2^2. \tag{5.10}$$

After some manipulation the solution to (5.10) is

$$\hat{\mathbf{w}} = (\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \alpha\hat{\mathbf{I}})^{-1}\hat{\mathbf{X}}\mathbf{y} \tag{5.11}$$

where $\hat{\mathbf{I}} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$.

Principle component regression Let $\hat{\mathbf{X}}\hat{\mathbf{X}}^T = \mathbf{U}\Sigma\mathbf{U}^T$ where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_N]$ is the matrix of eigenvectors of $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ and $\Sigma = \text{diag}(\lambda_1, \dots, \lambda_N)$ is a diagonal matrix of eigenvalues. Denote $\mathbf{Z} = \hat{\mathbf{X}}^T\mathbf{U} = [\mathbf{z}_1, \dots, \mathbf{z}_N]$, then $\mathbf{z}_i$ is referred to as the ith sample principle component of $\hat{\mathbf{X}}^T$. Note that $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$, (5.2) can be written as

$$\mathbf{y} \approx \hat{\mathbf{X}}^T\hat{\mathbf{w}} = \hat{\mathbf{X}}^T\mathbf{U}\mathbf{U}^T\hat{\mathbf{w}} = \mathbf{Z}\boldsymbol{\beta}. \tag{5.12}$$

where $\boldsymbol{\beta} = \mathbf{U}^T\hat{\mathbf{w}}$. According to the result of OLS, the solution to (5.12) is

$$\boldsymbol{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} = \boldsymbol{\Sigma}^{-1}\mathbf{Z}^T\mathbf{y}. \tag{5.13}$$

The principal component estimator is defined by

$$\hat{\mathbf{w}} = \mathbf{U}\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\Sigma}^{-1}\mathbf{Z}^T\mathbf{y}. \tag{5.14}$$

If we use all the principal components then the principle component estimator is the same as that of OLS. In reality, we retain only a subset of the important components and eliminate the others e.g., the ones with the smallest eigenvalue. As a result, the estimator can be approximated as

$$\hat{\mathbf{w}} \approx \mathbf{U}_l\boldsymbol{\Sigma}_l^{-1}\mathbf{Z}_l^T\mathbf{y}. \tag{5.15}$$

where $\mathbf{Z}_l = [\mathbf{z}_1, \dots, \mathbf{z}_l]$ and $l \leq N$.

Note that principle component regression (PCR) can be easily customized to Principle Component Analysis (PCA) to reduce the dimension of the inputs. Interested readers can refer to [68] for the details.

5.1.4.2. Nonlinear regression

Although linear model is preferable due to its simplicity and robustness, many input-output relationships are non-linear in practice. In the following, we present a typical ML solution using a classic nonlinear function

$$s(t) = \frac{1}{1+e^{-t}} \tag{5.16}$$

which is commonly referred to as logistic sigmoid function. Note that this function only generates a value between zero and one corresponding to an input. A sigmoid function-like data set satisfies

$$y_i \approx s(\mathbf{x}_i^T \mathbf{w} + b) = s(\hat{\mathbf{x}}_i^T \hat{\mathbf{w}}) \tag{5.17}$$

where $\hat{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}$ and $\hat{\mathbf{w}} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$.

A least square cost function can be formed as follows:

$$g(\hat{\mathbf{w}}) = \sum_{i=1}^{M} (s(\hat{\mathbf{x}}_i^T \hat{\mathbf{w}}) - y_i)^2. \tag{5.18}$$

The gradient of this objective is

$$\nabla g(\hat{\mathbf{w}}) = 2 \sum_{i=1}^{M} (s(\hat{\mathbf{x}}_i^T \hat{\mathbf{w}}) - y_i) s(\hat{\mathbf{x}}_i^T \hat{\mathbf{w}})(1 - s(\hat{\mathbf{x}}_i^T \hat{\mathbf{w}}))\hat{\mathbf{x}}_i. \tag{5.19}$$

Unlike the linear case, we cannot solve the first order derivation directly due to many nonlinear terms in the equation. Instead, we can use numerical methods such as gradient descent to find a local minimum. To find an optimum, we may linearize the function using the strategy in the following subsection.

5.1.4.3. Feature design for regression

The fact that ML is data-dependent, an algorithm can achieve very good or bad results depending on the data set. Thus the knowledge and the understanding of the data can help to design an efficient algorithm for a specific problem of interest. For example, in case of logistic function mention above, we can rewrite (5.17) as

$$\hat{\mathbf{x}}_i^T \hat{\mathbf{w}} = \log(\frac{y_i}{1-y_i}) = \hat{y}_i \tag{5.20}$$

which is in fact a linear function, thus the family of least-square solution is applicable.

By reasoning or background knowledge, we can also preprocess a data set for which simple or more efficient solution can be derived. In particular, we can transform the set into possibly high-dimensional feature space using non-linear features $\boldsymbol{\phi}(.)$ so that

$$y_i \approx \boldsymbol{\phi}(\mathbf{x}_i)^T \hat{\mathbf{w}}. \tag{5.21}$$

Notice that the function remains linear in the weight vector, thus linear-based algorithms can be applied. For example, the function $y_i \approx \log(\mathbf{x}_i^T \mathbf{a})\mathbf{w}$ is nonlinear with the input $\mathbf{x}_i$ while is still linear with $\mathbf{w}$. Even in case of simple linear model $y_i \approx \mathbf{x}_i^T \mathbf{w}$ we can always transfer the data to possibly higher-dimensional space

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \overset{\phi}{\rightarrow} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \overset{\hat{\mathbf{w}}}{\rightarrow} y$$

where $\boldsymbol{\phi}(\mathbf{x}) = [1, x_1, \dots, x_N]^T$. For this specific example, the new features in the transformed feature space are more advantageous than the original ones since they take bias into account.

5.1.4.4. Performance metrics

Let y_i and $\tilde{y}_i$ denote the actual value and the estimate value, respectively. We can use the following measures to evaluate an ML algorithm:

- Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \tilde{y}_i|. \tag{5.22}$$

- Mean Absolute Relative Error (MARE) or Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} |\frac{y_i - \tilde{y}_i}{y_i}|. \tag{5.23}$$

- Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2}. \tag{5.24}$$

- Root Mean Square Percentage Error (RMSPE):

$$\text{RMSPE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\frac{y_i - \tilde{y}_i}{y_i})^2} \tag{5.25}$$

- Relative Root Mean Square Error (RRMSE):

$$\text{RRMSE} = \frac{\text{RMSE}}{\bar{y}} = \frac{\sqrt{\frac{1}{N}\sum\limits_{i=1}^{N}(y_i - \tilde{y}_i)^2}}{\frac{1}{N}\sum\limits_{i=1}^{N} y_i}. \tag{5.26}$$

It is worth mentioning that when we have a limited data set, k-fold cross-validation is preferable. In particular, the data is divided into k groups so that we perform the training on $(k-1)$ groups and test the model on the rest. In fact, there are no formal criteria for selecting k, normally we can take a value of 5 or 10.

5.2. A Case Study of ML-based Approach

In this section, we apply ML to estimate the maximum sum rate for SZFDPC systems. For this specific problem, a barrier interior-point method was proposed in [29, 30]. It is well known that such a second-order approach has a computational complexity that does not scale favorably with the problem size. In the following we proposed to solve this important problem by a suboptimal solutions which is based on regression.

5.2.1. System Model

Consider a MIMO BC consisting of a base station (BS) and K users. The BS and each user k are equipped with N and M_k antennas respectively. The channel matrix for user k is denoted by $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$. Normally, a user suffers interference from all other users in the system. For user k in the SZFDPC scheme, the interference caused by users $j < k$ is canceled by DPC, while that caused by users $j > k$ is nulled out by zero-forcing technique. In this way, a MIMO BC can be decomposed into parallel interference-free channels. We refer the interested reader to [29] and references therein for a more detailed description of the SZFDPC scheme.

5.2.2. Problem Formulation

The sum rate of SZFDPC can be characterized through solving the sum rate (SR-Max) problem under PAPC which is formulated as

$$\underset{\{\mathbf{S}_k \succeq \mathbf{0}\}}{\text{maximize}} \quad \textstyle\sum_{k=1}^{K} \log|\mathbf{I} + \mathbf{H}_k\mathbf{S}_k\mathbf{H}_k^{\dagger}| \tag{5.27a}$$

$$\text{subject to} \quad \mathbf{H}_j\mathbf{S}_k\mathbf{H}_j^{\dagger} = 0, \ \forall j < k \tag{5.27b}$$

$$\textstyle\sum_{k=1}^{K}[\mathbf{S}_k]_{i,i} \leq P_i, \ i = 1, 2, \ldots, N \tag{5.27c}$$

where $\mathbf{S}_k \succeq \mathbf{0}$ is the input covariance matrix for user k. The constraint in (5.27b) is imposed to suppress the interference from users $j < k$ as mentioned above.

Due to the use of zero-forcing method, SZFDPC is a suboptimal transmission strategy compared to DPC. However, SZFDPC does not cancel multiuser inference only by zero-forcing technique since DPC is still invoked for this purpose. Thus, SZFDPC can achieve a performance close to that of DPC, which was reported in various previous studies [29, 30, 69]. We note that for SZFDPC (i.e. (5.27)) to be feasible, it should hold that $N > (K-1)M$ which is assumed in this chapter. This dimension condition basically imposes a constraint on the maximum number of users that can be supported simultaneously. When the number of demanding users increases, a user scheduling algorithm is required and this problem was studied in [69] where several efficient user selection methods were proposed for SZFDPC. We also remark that the interference cancelling process is performed sequentially after each user, and thus user ordering in SZFDPC is important. Optimal user ordering requires solving a combinatorial optimization problem but efficient user order algorithms were also proposed in [69]. In this chapter we simply assume the natural user ordering for SZFDPC and focus on the precoder design.

In order to simplify the formulation in (5.27), let $\breve{\mathbf{H}}_k = [\mathbf{H}_1^\dagger, \mathbf{H}_2^\dagger, \dots \mathbf{H}_{k-1}^\dagger]^\dagger$, $\breve{\mathbf{V}}_k = \text{null}(\breve{\mathbf{H}})$, and $\dot{\mathbf{H}}_k = \mathbf{H}_k \breve{\mathbf{V}}_k$. Intuitively, $\dot{\mathbf{H}}_k$ is called the effective channel of user k. The optimal $\mathbf{S}_k$ in (5.27) is then given by $\mathbf{S}_k = \breve{\mathbf{V}}_k \dot{\mathbf{S}}_k \breve{\mathbf{V}}_k^\dagger$, where $\dot{\mathbf{S}}_k$ is the optimal solution to the following problem

$$\begin{array}{ll} \underset{\{\dot{\mathbf{S}}_k \succeq \mathbf{0}\}}{\text{maximize}} & \sum_{k=1}^{K} \log |\mathbf{I} + \dot{\mathbf{H}}_k \dot{\mathbf{S}}_k \dot{\mathbf{H}}_k^\dagger| \\ \text{subject to} & \sum_{k=1}^{K} [\breve{\mathbf{V}}_k \dot{\mathbf{S}}_k \breve{\mathbf{V}}_k^\dagger]_{i,i} \leq P_i, \ \forall i. \end{array} \tag{5.28}$$

Inspired by the work in [12], we extend the AO approach to our considered problem. More specifically, by extending Theorem 2 of [30], we can show that (5.28) can be equivalently transformed into the following minimax problem in the dual MAC

$$\begin{array}{ll} \underset{\mathbf{Q} \succeq \mathbf{0}}{\min} \ \underset{\{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}}{\max} & \sum_{k=1}^{K} \log \frac{|\breve{\mathbf{V}}_k^\dagger \mathbf{Q} \breve{\mathbf{V}}_k + \dot{\mathbf{H}}_k^\dagger \bar{\mathbf{S}}_k \dot{\mathbf{H}}_k|}{|\breve{\mathbf{V}}_k^\dagger \mathbf{Q} \breve{\mathbf{V}}_k|} \\ \text{subject to} & \sum_{k=1}^{K} \text{tr}(\bar{\mathbf{S}}_k) = P \\ & \text{tr}(\mathbf{Q}\mathbf{P}) = P, \mathbf{Q} : \mathsf{diagonal} \end{array} \tag{5.29}$$

where $P = \text{tr}(\mathbf{P})$, $\mathbf{P} = \text{diag}([P_1, P_2, \dots, P_N]^T)$.

Since the AO-based algorithm to solve (5.29) is similar to those in [12, Algorithm 2], we refer interested readers to [12, Algorithm 2] for the details. Instead, we concentrate on our ML-based approach in the following.

5.2.3. An ML-based Approach

Following similar arguments to those in [12, Subsection III-B], the interior-point-based approach to solve the considered problem has the per-iteration complexity up to $\mathcal{O}(K^3N^6)$ while that of the AO-based algorithm is $\mathcal{O}(KN^3)$ flops. On the one hand, AO-based algorithm dominates the existing approach and reduces the complexity significantly, but on the other hand, it still experiences high complexity in case of massive MIMO settings where $\frac{N}{K} \geq 10$. In such cases, we can employ the following ML approach to obtain a suboptimal solution since this approach can adapt quickly to any changes in the systems while retaining the satisfactory performance. Regarding ML-based methods, it is also worth mentioning that deep learning has been applied recently to the relevant problems [70, 71]. However, the performance of the deep learning-based methods depends heavily on the choice of the number of hidden layers as well as the number of neurons in each layer. More importantly, the tuning of the hyperparameters is difficult. Instead, we will show shortly that we can find an appropriate estimator using simple linear regression methods which are not only tractable but also easy to implement and analyze.

Assuming that we execute the AO-based algorithm to generate optimal sum rates $\mathbf{y}$ based on $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s] \in \mathbb{C}^{p\times s}$ inputs where s is the number of samples. Note that $\mathbf{x}_i$ contains p features of the power constraints and channel coefficients. In other words, we stack the power constraints and the channel coefficients of all users into a vector, i.e., $p = N + KMN$. If we simply apply arbitrary ML algorithms, the errors will be extremely prohibitive due to the fact that the considered problem is nonlinear in nature with respect to either the power constraint or the channel matrix (c.f. Fig. 5.1). On the other hand, nonlinear ML algorithms are much more difficult to investigate since there are no available solutions to this type of optimization. Even the optimal solution mentioned above already contains many nonlinear terms. In the following, we propose a novel two-step preprocessing method to transform the inputs into another feature space to which linear regression algorithms are applicable. Herein, we will refer to this approach as feature design (FD)-based approach.

Step 1: Select a set of features $\check{\mathbf{x}}$ by customizing the principle component analysis (PCA)-based algorithm in [68]:
- Choose the number of eigenvectors whose eigenvalues are larger than 1 .
- Select the features based on l largest contribution
Step 2: Transform $\check{\mathbf{x}}$ into higher feature space by $\boldsymbol{\phi}(\check{\mathbf{x}}) = [1, \log_b(|\check{\mathbf{x}}|)]^T$.

Note that instead of choosing a number of largest eigenvalues of the covariance matrix randomly [68], we empirically choose d eigenvalues which are larger than 1. As a result, we can form a new matrix $\tilde{\mathbf{U}} = [\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_d]$ corresponding to those eigenvalues. To select the most dominant features, we first calculate the contribution measure

$$\vartheta_i = \sum_{j=1}^{d} |\tilde{u}_{i,j}| \tag{5.30}$$

where $i = 1, 2, \ldots, p$. Then we select the desired features with respect to l largest contribution ϑ_i. Again, we avoid random selection of l whose appropriate value is not easy to justify in practice. Instead, we propose to choose l based on the matrix size and the number of users:

$$l = N + Kr \tag{5.31}$$

where N and K are the number of transmit antennas and the number of users, respectively; $r = \min(M, N)$ where M is the number of receive antennas. Note that $l < p$ from (5.31) and we can therefore obtain a new matrix with reduced dimension $\check{\mathbf{X}} = [\check{\mathbf{x}}_1, \check{\mathbf{x}}_2, \ldots, \check{\mathbf{x}}_s] \in \mathbb{C}^{l \times s}$.

In fact, there are no criteria to choose a function to transform the inputs into another space where efficient algorithms can be derived. In our approach, we rely on the characteristics of the problem to propose a transform function. Specifically, recall that the considered sum rate is a logdet function, thus we can transform these features into a new feature space where linear model are possible using the following

$$\boldsymbol{\phi}(\check{\mathbf{x}}) = \begin{bmatrix} 1 \\ \log_b(|\check{\mathbf{x}}|) \end{bmatrix} \tag{5.32}$$

where b is the base of the logarithm. Under this assumption, an output i.e., an estimate sum rate is given by

$$y_i \approx \boldsymbol{\phi}(\check{\mathbf{x}}_i)^T \hat{\mathbf{w}}. \tag{5.33}$$

As a result of this formulation, we can apply any linear regression algorithms such as ordinary least square (OLS), ridge regression or principal component regression (PCR) [72, Chapter 6] to find an appropriate estimator $\hat{\mathbf{w}}$. In the numerical results, we will show the effectiveness of our proposed approach in comparison with other algorithms which do not take the feature design into account.

5.2.4. Numerical Results

In this section, we numerically evaluate the performance of the proposed algorithm. For the AO-based algorithm, we set an error tolerance of $\epsilon = 10^{-6}$ as the stopping criterion to generate the optimal sum rate. The number of transmit antennas are varied from 16 to 32 and the number of users is fixed at 4. The the PAPC ratio is chosen randomly, whereas SNR is chosen from the set SNR $= [0, 10, 20, 30, 40]$ dBW. For each MIMO setting we generated 240 samples. Also, we simply use natural log to transform the feature space. Other simulation parameters are specified for each setup. The codes are executed on a 64-bit desktop that supports 8 Gbyte RAM and Intel CORE i7.

In the first experiment, we compare the optimal and estimate sum rates of a MIMO scenario where $N = 32$ transmit antennas and $M = 2$ receive antennas. In particular, we compare the cumulative distribution functions (CDFs) of the optimal and estimate sum rates of a MIMO system with SZFDPC and different PAPC settings using linear and nonlinear regression methods. More specifically, we utilize support vector regression (SVR) with radial basis function (RBF) kernel [73] for nonlinear regression. Here, we train on 216 samples and test on 24 samples. As can be seen from the figure, conventional OLS and SVR fail to fit the data due to nonlinear nature of the problem. However, the results of the simple OLS with the feature design are very close to optimal solutions. The performance has also proved the feasibility of our approach.

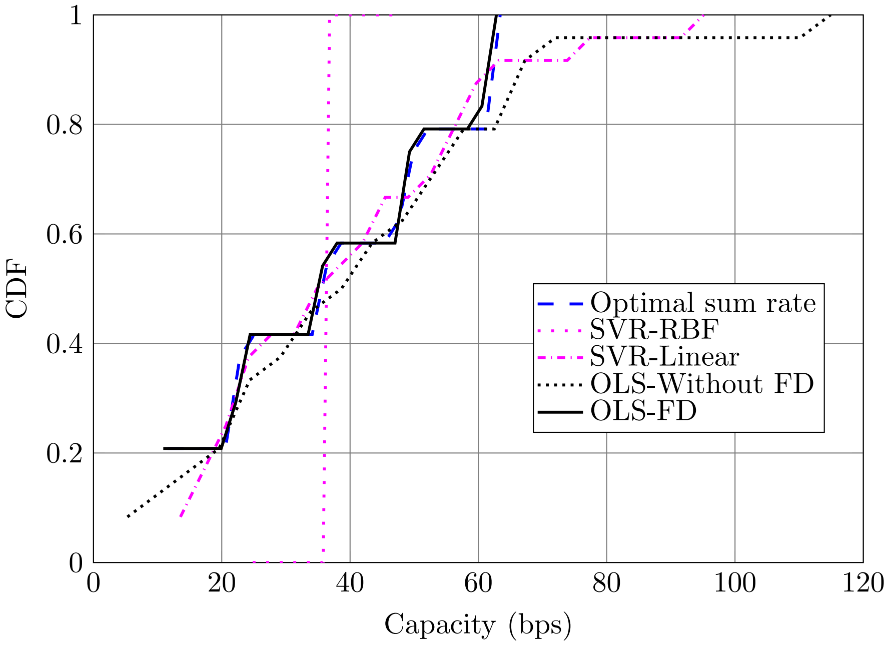


Figure 5.1.: Cumulative distribution functions of the optimal and estimate sum rates of a MIMO system with SZFDPC and random PAPC settings using linear and nonlinear regression, $N = 32$ transmit antennas, $M = 2$ receive antennas and $K = 2$ users.

In the last experiment, we consider the effectiveness of our feature-design-based approach in terms of average relative root mean square error (aRRMSE) [74] over large samples with varying number of transmit and receive antennas. In particular, we obtain the aRRMSE by executing 10-fold cross-validation using three simple linear ML algorithms: OLS, Ridge and PCR. According to [74], a learning model is considered good and excellent when 10% < aRRMSE < 20% and aRRMSE < 10%, respectively. Interestingly, the ML-based methods show sufficiently low error rates, especially when $\frac{N}{K} \geq 10$. From our observations, the training matrices are invertible and the eigenvalues are larger than 1, thus the performance of OLS and PCR is the same and has minor difference in comparison with that of ridge regression. Unsurprisingly, these observations coincide with the properties of these regression methods.

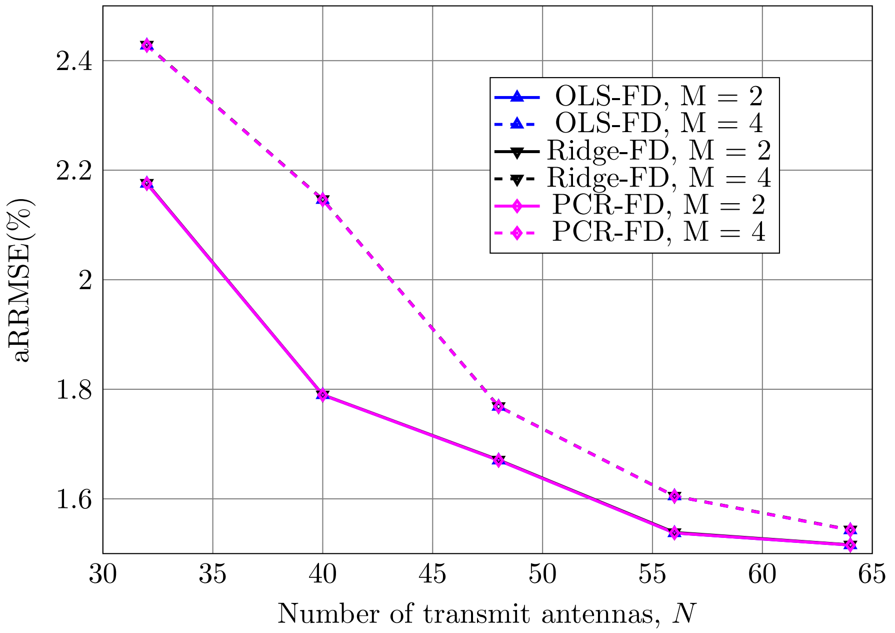


Figure 5.2.: aRRMSE of OLS, Ridge regression and PCR with feature design and $K = 2$ users.

5.3. Summary

We have presented the basis of ML and applied it to compute sum rate of MIMO systems under PAPC and SZFDPC. Our experiments using optimal solution have stated that the SZFDPC can obtain the near-capacity whereas the ZF scheme still operates far from the optimal capacity boundary for a specified number of users. The fact that the proposed optimal solution may experience high complexity in large-scale MIMO settings, a suboptimal ML-based approach is therefore more efficient. Extensive numerical results have demonstrated the superiority of the proposed algorithms over the existing interior-point method. More importantly, our ML-based approach can be applicable to a class of similar problems.

Chapter 6

Conclusions and Future Work

The capacity of MIMO systems is well-studied under SPC, however efficient algorithms to compute the MIMO capacity under per-antenna power constraint or the general form i.e., LTCCs remain open problems since the state-of-the-art solutions utilizing high-complexity methods are inapplicable. In this thesis, we have proposed novel approaches to those important problems. More importantly, our algorithms have low complexity, thus have huge impact on the development of massive MIMO systems in 5G and beyond.

For the MIMO capacity under PAPC, we proposed two efficient approaches based on fixed-point iteration and AO together with SCA. More specifically, the fixed-point solution relies on water-filling-based method and fixed-point iteration, thanks to the special structure of the problem in BC. In the second approach, the considered problem in BC is transformed into an equivalent minimax problem in the dual MAC. Since a naive minimax does not guarantee the convergence, we have proposed to optimize the upper bound of the minimization problem and this critical step in fact avoids fluctuations observed in a naive minimax algorithm. Our approaches are provably convergent and have low complexity.

When a new power constraint or multiple linear transmit covariance constraints are imposed on the system, the solution for PAPC is no longer applicable. Interestingly, the capacity problem can however be solved efficiently under the same framework for PAPC case. As a starting point, we also transformed the original maximization problems into a minimax problem. Then AO and CCP are utilized to derive both iterative low-complexity algorithms for the general power constraints and analytical solutions to the important cases of joint SPC and PAPC. To a large extent, this approach is not only efficient but also general enough to include either SPC or PAPC or joint SPC and PAPC as special cases.

In light of ML, we have developed a ML-based approach to estimate the sum rate of a multi-user MIMO system, which is specifically helpful for massive MIMO systems. Considering SZFDPC as a case study, we relied on aforementioned optimal approaches to arrive at the optimal solution. We then proposed an approach to preprocess the data generated by the optimal algorithm so that linear regression methods are feasible. The numerical results have shown that this approach strikes a good balance between the complexity and optimal rates.

It is worth noting that LTCCs considered in Chapter 4 are general enough to include PAPC or joint SPC and PAPC as special cases. From an algorithmic point of view, the second approach in Chapter 3 and the approach proposed in Chapter 4 are based on the same idea of combining alternating optimization and successive convex approximation and thus the resulting algorithms are actually the same. The differences lie on the closed-form solutions for each step of the iterative process, more specifically, the minimization utilizing the upper bound of the considered objectives. In fact, the minimax problems in Chapters 3 and 4 can be generalized as follows

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \quad g(\mathbf{x},\mathbf{y}) - h(\mathbf{y}) \tag{6.1}$$

where $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ are compact and convex sets in $\mathbb{R}^n$; $g(\mathbf{x},\mathbf{y})$ is jointly concave with $\mathbf{x}$ and $\mathbf{y}$, $g(\mathbf{x},\mathbf{y}) - h(\mathbf{y})$ is convex with $\mathbf{y}$.

The novelty of our proposed approach is to minimize an upper bound for the minimization, instead of the original objective. This is a very critical step in order to guarantee the convergence of the resulting algorithm. Otherwise, we may end up with a ping-pong situation as illustrated in Fig. 3.2 of Chapter 3. In general, different upper bounds may converge to different points. To avoid this problem, we need to impose some conditions on an upper bound to make sure that it always converges to an optimal solution. Basically an upper bound should satisfy two conditions: (i) The upper bound should be tight when the iterative procedure converges, and (ii) the KKT conditions of the optimization problem derived from the upper bound should be the same as the KKT conditions of the original problem. A more rigorous explanation is given below.

Let $f(\mathbf{y}) = g(\mathbf{x},\mathbf{y}) - h(\mathbf{y})$ be a convex function to be minimized over the convex feasible set $\mathcal{Y}$. That is, we consider the following problem

$$\begin{aligned} &\text{minimize} && f(\mathbf{y}) && (6.2)\\ &\text{subject to} && \mathbf{y} \in \mathcal{Y} && (6.3) \end{aligned}$$

We wish to develop an iterative method to solve the above problem using an upper bound. Let $\varphi(\mathbf{y};\mathbf{y}^{(n)})$ be an upper bound for this purpose where $\mathbf{y}^{(n)}$ is the obtained solution at iteration n. The two conditions mentioned above mean

- $\varphi(\mathbf{y};\mathbf{y}^{(n)}) \geq f(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$ and $\varphi(\mathbf{y}^{(n)};\mathbf{y}^{(n)}) = f(\mathbf{y}^{(n)})$. In other words, the upper bound and the original objective should be equal at $\mathbf{y}^{(n)}$. This condition guarantees that the upper bound and the original function achieve the same value when the iterative process converges.
- $\nabla_{\mathbf{y}}\varphi(\mathbf{y};\mathbf{y}^{(n)})|_{\mathbf{y}=\mathbf{y}^{(n)}} = \nabla_{\mathbf{y}}f(\mathbf{y})|_{\mathbf{y}=\mathbf{y}^{(n)}}$. That is, the gradient of the upper bound and that of the objective function evaluated at $\mathbf{y}^{(n)}$ are identical. This condition is to ensure that the optimal solution to the approximate problem is also optimal to the original problem.

We remark that for a given function $f(\mathbf{y})$, there are infinitely many upper bounds that can meet the two requirements. The choice of an upper bound also depends how the resulting problem can be solved. A good choice is one that can lead to an efficient solution. In the thesis we derive an upper bound of the objective which is based on the first order approximation of the logdet function. Thus, the upper bound can naturally satisfy the two conditions stated above. To clarify this point let us write the gradient of the objective of Eq. (4.3) with respect to $\mathbf{Q} = \sum_{i=1}^{L} q_i \mathbf{E}_i$ as

$$\nabla_{\mathbf{Q}} f(\mathbf{Q}, \bar{\mathbf{S}}) = (\mathbf{Q} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H})^{-1} - \mathbf{Q}^{-1}. \tag{6.4}$$

The gradient of the objective in iteration n of the proposed method is

$$\mathbf{\Phi}_n^{-1} - \mathbf{Q}^{-1} = (\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H})^{-1} - \mathbf{Q}^{-1}. \tag{6.5}$$

It is trivial to check that the gradients of the original objective and the upper bound in are the same when $\mathbf{Q} = \mathbf{Q}^{(n)}$. More importantly, this simple upper bound allows us to find an efficient solution.

In fact, our approaches are not only applicable to the considered problems in this thesis but also a class of similar problems. More specifically, we suggest some possible extensions to our study in what follows:

- In the context of physical security, the secrecy rate optimization is generally non-trivial since the problem is neither convex nor concave. Recently, authors in [75] proposed to transform this non-convex problem into an equivalent minimax problem. Inspired by the work of [75], the author in [76] developed a global-convergent interior-point-based algorithm for this problem under SPC or PAPC. Due to the similarity of the problem formulation to those in this thesis, our AO and CCP-based approach is promising alternative to current interior-point-based solution.
- A constrained maximization or minimization can always be considered as a minimax problem. In particular, we have to maximize/minimize a Lagrangian function with respect to (w.r.t) one or a set of primal variables, while minimizing/maximizing it w.r.t the dual variables or Lagrangian multiplier(s). If we can find a proper upper bound/lower bound for the minimization or the maximization, then the approach is still well suited.
- In a broader sense, the AO-based approach may be investigated further to apply to an arbitrary minimax problem with proper bound(s). Our conjecture is that we can choose either an upper bound for the minimization or a lower bound for the maximization or even both to solve the problem. Moreover, this methodology is also applicable to a non-convex problem if we can transform it into a minimax problem as the case of secrecy capacity.

- In practice, optimal solutions can cause high complexity in certain scenarios, therefore suboptimal solutions e.g., machine learning -based approach which trade off the complexity and performance are of more interest. Since our initial experiments with machine learning-based method only estimate the channel capacity, more research efforts can be done to not only arrive at the channel capacity but also precoding matrices to achieve that capacity. A more sophisticated technique to estimate multi-targets such as deep learning can be investigated.
- Under imperfect channel information, we may formulate a channel as a combination of a deterministic channel and an error model. Therefore, the proposed approaches can be customized to solve these problems. We may also consider imperfect channel model in light of stochastic optimization in which the algorithms mentioned above are applicable.

Appendix A

Proofs of Chapter 3

A.1. Proof of Lemma 3.1

The key to prove the convergence of the fixed-point iteration in (3.14) is to show that $\mathfrak{J}(\mathbf{x})$ is a standard interference function. That is, for all $\mathbf{x} \geq 0$ then $\mathfrak{J}(\mathbf{x})$ satisfies the following three properties

- Positivity: $\mathfrak{J}(\mathbf{x}) > 0$.
- Monotonicity: If $\mathbf{x} \geq \mathbf{y}$, then $\mathfrak{J}(\mathbf{x}) \geq \mathfrak{J}(\mathbf{y})$.
- Scalability: For all $\alpha > 1, \alpha\mathfrak{J}(\mathbf{x}) > \mathfrak{J}(\alpha\mathbf{x})$.

According to [77, Theorem 2], if a function satisfies three properties listed above, it will converge to a unique fixed point.

The positivity is obvious and the scalability can be easily shown by the following inequalities

$$\begin{aligned}
\mathfrak{J}(\alpha\mathbf{x}) &= \mathbf{p} + \alpha\,\mathrm{diag}(\mathbf{\Psi}(\alpha\mathbf{x})) \odot \mathbf{x} && \text{(A.1)}\\
&\overset{\text{(a)}}{\leq} \mathbf{p} + \alpha\,\mathrm{diag}(\mathbf{\Psi}(\mathbf{x})) \odot \mathbf{x} && \text{(A.2)}\\
&\overset{\text{(b)}}{<} \alpha(\mathbf{p} + \mathrm{diag}(\mathbf{\Psi}(\mathbf{x})) \odot \mathbf{x}) = \alpha\mathfrak{J}(\mathbf{x}) && \text{(A.3)}
\end{aligned}$$

where (a) can be proven from the definition of $\mathbf{\Psi}(\mathbf{x})$ as follows. Let $\mathbf{X} = \mathrm{diag}(\mathbf{x})$ and $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\dagger} = \mathbf{X}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{X}$ be the EVD of $\mathbf{X}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{X}$, where $\mathbf{\Sigma} = \mathrm{diag}([\rho_1, \rho_2, \ldots, \rho_r, \mathbf{0}_{N-r}])$ and $r = \mathrm{rank}(\mathbf{H}^{\dagger}\mathbf{H})$. Then it follows immediately that $\mathbf{V}\tilde{\mathbf{\Sigma}}\mathbf{V}^{\dagger} = \tilde{\mathbf{X}}\mathbf{H}^{\dagger}\mathbf{H}\tilde{\mathbf{X}}$ where $\tilde{\mathbf{X}} = \mathrm{diag}(\alpha\mathbf{x})$, $\tilde{\mathbf{\Sigma}} = \mathrm{diag}([\tilde{\rho}_1, \tilde{\rho}_2, \ldots, \tilde{\rho}_r, \mathbf{0}_{N-r}])$, and $\tilde{\rho}_i = \alpha^2\rho_i$ for $i = 1, 2, \ldots, r$. Since $\alpha > 1$, we have $\frac{1}{\tilde{\rho}_i} = \frac{1}{\alpha^2\rho_i} < \frac{1}{\rho_i}$, and thus

$$\mathbf{\Psi}(\alpha\mathbf{x}) = \mathbf{V}\,\mathrm{diag}(\frac{1}{\tilde{\rho}_1}, \cdots, \frac{1}{\tilde{\rho}_{s'}}, \mathbf{1}_{N-s'})\mathbf{V}^{\dagger} \preceq \mathbf{\Psi}(\mathbf{x}) = \mathbf{V}\,\mathrm{diag}(\frac{1}{\rho_1}, \cdots, \frac{1}{\rho_s}, \mathbf{1}_{N-s})\mathbf{V}^{\dagger} \tag{A.4}$$

where s' and s are the largest number such that $1 - \frac{1}{\rho_{s'}} > 0$ and $1 - \frac{1}{\rho_s} > 0$, respectively. Note that $s' \geq s$ and thus the above inequality is easily justified.

Consequently, $\operatorname{diag}(\boldsymbol{\Psi}(\alpha\mathbf{x})) \leq \operatorname{diag}(\boldsymbol{\Psi}(\mathbf{x}))$ which completes (a). The inequality (b) holds since $\mathbf{p} < \alpha\mathbf{p}$ for $\alpha > 1$, which results in $\mathbf{p} + \alpha \operatorname{diag}(\boldsymbol{\Psi}(\mathbf{x})) \odot \mathbf{x} < \alpha(\mathbf{p} + \operatorname{diag}(\boldsymbol{\Psi}(\mathbf{x})) \odot \mathbf{x}) = \alpha\boldsymbol{\mathfrak{I}}(\mathbf{x})$.

To prove the monotonicity of $\boldsymbol{\mathfrak{I}}(\mathbf{x})$, we need to show that for all $\mathbf{x}, \mathbf{y} \geq 0$ then $\boldsymbol{\mathfrak{I}}(\mathbf{x}) \geq \boldsymbol{\mathfrak{I}}(\mathbf{y})$ or equivalently $\operatorname{diag}(\boldsymbol{\Psi}(\mathbf{x})) \odot \mathbf{x} \geq \operatorname{diag}(\boldsymbol{\Psi}(\mathbf{y})) \odot \mathbf{y}$. Let $\mathbf{X} = \operatorname{diag}(\mathbf{x}), \mathbf{Y} = \operatorname{diag}(\mathbf{y})$. Then monotonicity proof is equivalent to showing that $\operatorname{diag}(\mathbf{X}^{1/2}\boldsymbol{\Psi}(\mathbf{x})\mathbf{X}^{1/2}) \geq \operatorname{diag}(\mathbf{Y}^{1/2}\boldsymbol{\Psi}(\mathbf{y})\mathbf{Y}^{1/2})$ for $\mathbf{X} \succeq \mathbf{Y} \succeq \mathbf{0}$.

Let us first consider the case $N \leq M$ and $\mathbf{H}$ is full column rank. Then we can write the EVD of $\boldsymbol{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\boldsymbol{\Lambda}^{-1/2}$ as

$$\underbrace{\boldsymbol{\Lambda}^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\boldsymbol{\Lambda}^{-1/2}}_{\mathbf{B}} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\dagger}. \tag{A.5}$$

For notational convenience, let $\mathbf{K} = \mathbf{H}^{\dagger}\mathbf{H}$. Note that $\mathbf{K}$ is full-rank and thus invertible. Then the above equation can be rewritten as

$$\mathbf{B}^{-1} = \boldsymbol{\Lambda}^{1/2}\mathbf{K}^{-1}\boldsymbol{\Lambda}^{1/2} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{V}^{\dagger} \tag{A.6}$$

which the results in

$$\mathbf{B}^{-1} - \mathbf{I} = \mathbf{V}\Big(\boldsymbol{\Sigma}^{-1} - \mathbf{I}\Big)\mathbf{V}^{\dagger}. \tag{A.7}$$

Let $\tilde{\boldsymbol{\Sigma}}$ be the $(N-k)$ *positive* eigenvalues of $\mathbf{B}^{-1} - \mathbf{I}$ and $\tilde{\mathbf{V}}_k$ consist of the corresponding $N-k$ eigenvectors, $\bar{\boldsymbol{\Sigma}}$ be the k *non-positive* eigenvalues of $\mathbf{B}^{-1} - \mathbf{I}$, and $\bar{\mathbf{V}}_k$ consist of the corresponding k eigenvectors, and define

$$\mathbf{A}^{+} = \tilde{\mathbf{V}}_k\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}_k^{\dagger} \tag{A.8a}$$

$$\mathbf{A}^{-} = \bar{\mathbf{V}}_k\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}_k^{\dagger}. \tag{A.8b}$$

Then it holds that

$$\mathbf{B}^{-1} - \mathbf{I} = \mathbf{A}^{+} + \mathbf{A}^{-} \tag{A.9}$$

and that $\mathbf{A}^{-}\mathbf{A}^{+} = \mathbf{0}$. Now we can write $\boldsymbol{\Psi}(\tilde{\boldsymbol{\lambda}}) = \mathbf{A}^{-} + \mathbf{I} = \mathbf{B}^{-1} - \mathbf{A}^{+}$ and thus

$$\begin{aligned}
[\boldsymbol{\Psi}(\tilde{\boldsymbol{\lambda}})\boldsymbol{\Lambda}^{-1}]_{i,i} &= [\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Psi}(\tilde{\boldsymbol{\lambda}})\boldsymbol{\Lambda}^{-1/2}]_{i,i} \\
&= [\boldsymbol{\Lambda}^{-1/2}\Big(\mathbf{B}^{-1} - \mathbf{A}^{+}\Big)\boldsymbol{\Lambda}^{-1/2}]_{i,i} \\
&= [\boldsymbol{\Lambda}^{-1/2}\mathbf{B}^{-1}\boldsymbol{\Lambda}^{-1/2}]_{i,i} - [\boldsymbol{\Lambda}^{-1/2}\mathbf{A}^{+}\boldsymbol{\Lambda}^{-1/2}]_{i,i} \\
&= [\mathbf{K}^{-1}]_{i,i} - [\boldsymbol{\Lambda}^{-1/2}\mathbf{A}^{+}\boldsymbol{\Lambda}^{-1/2}]_{i,i} \\
&= [\mathbf{K}^{-1}]_{i,i} - [\tilde{\boldsymbol{\Lambda}}^{1/2}\mathbf{A}^{+}\tilde{\boldsymbol{\Lambda}}^{1/2}]_{i,i}.
\end{aligned} \tag{A.10}$$

To proceed further we need to show that if $\mathbf{X} \succeq \mathbf{Y}$ then

$$[\mathbf{X}^{1/2}\mathbf{A}_X^{+}\mathbf{X}^{1/2}]_{i,i} \leq [\mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2}]_{i,i}. \tag{A.11}$$

Now from (A.9) we have

$$\mathbf{X}^{-1/2}\mathbf{K}^{-1}\mathbf{X}^{-1/2} - \mathbf{I} = \mathbf{A}_X^{+} + \mathbf{A}_X^{-} \tag{A.12}$$

which is equivalent to

$$\mathbf{K}^{-1} - \mathbf{X} = \mathbf{X}^{1/2}\mathbf{A}_X^{+}\mathbf{X}^{1/2} + \mathbf{X}^{1/2}\mathbf{A}_X^{-}\mathbf{X}^{1/2}. \tag{A.13}$$

The same result applies to $\mathbf{Y}$, i.e.,

$$\mathbf{K}^{-1} - \mathbf{Y} = \mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2} + \mathbf{Y}^{1/2}\mathbf{A}_Y^{-}\mathbf{Y}^{1/2}. \tag{A.14}$$

Since $\mathbf{X} \succeq \mathbf{Y} \succ \mathbf{0}$ it holds that

$$\mathbf{K}^{-1} - \mathbf{X} \preceq \mathbf{K}^{-1} - \mathbf{Y}. \tag{A.15}$$

Substituting (A.13) and (A.14) into (A.15) yields

$$\mathbf{X}^{1/2}\mathbf{A}_X^{+}\mathbf{X}^{1/2} + \mathbf{X}^{1/2}\mathbf{A}_X^{-}\mathbf{X}^{1/2} \preceq \mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2} + \mathbf{Y}^{1/2}\mathbf{A}_Y^{-}\mathbf{Y}^{1/2}. \tag{A.16}$$

We now recall the following inequality. For Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A} \succeq \mathbf{B}$, then $\mathbf{SAS}^H \succeq \mathbf{SBS}^H$ for $\mathbf{S} \succeq 0$ [78, Observation 7.7.2]. Applying this inequality to (A.16) leads to

$$\mathbf{A}_X^{+} - \mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2}\mathbf{X}^{-1/2} \preceq \mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^{-}\mathbf{Y}^{1/2}\mathbf{X}^{-1/2} - \mathbf{A}_X^{-} \tag{A.17}$$

which is then equivalent to

$$\begin{aligned}&\left(\mathbf{A}_X^{+}\right)^{1/2}\left(\mathbf{A}_X^{+} - \mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2}\mathbf{X}^{-1/2}\right)\left(\mathbf{A}_X^{+}\right)^{1/2}\\ &\preceq \left(\mathbf{A}_X^{+}\right)^{1/2}\left(\mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^{-}\mathbf{Y}^{1/2}\mathbf{X}^{-1/2} - \mathbf{A}_X^{-}\right)\left(\mathbf{A}_X^{+}\right)^{1/2} \preceq \mathbf{0}.\end{aligned} \tag{A.18}$$

The above inequality holds true since $\left(\mathbf{A}_X^{+}\right)^{1/2}\mathbf{A}_X^{-}\left(\mathbf{A}_X^{+}\right)^{1/2} = \mathbf{0}$. It is easy to see that (A.18) results in

$$\mathbf{X}^{1/2}\mathbf{A}_X^{+}\mathbf{X}^{1/2} \preceq \mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2} \tag{A.19}$$

and thus

$$[\mathbf{X}^{1/2}\mathbf{A}_X^{+}\mathbf{X}^{1/2}]_{i,i} \leq [\mathbf{Y}^{1/2}\mathbf{A}_Y^{+}\mathbf{Y}^{1/2}]_{i,i} \tag{A.20}$$

for all i. Here we have used a well-known fact that for $\mathbf{A} \succeq \mathbf{B}$, then $[\mathbf{A}]_{i,i} \geq [\mathbf{B}]_{i,i}$.

We now turn our attention to the general case where $\mathbf{K}$ is singular. This occurs when $N > M$ or $N \leq M$ but $\mathbf{H}$ is not full column rank, i.e. the columns of $\mathbf{H}$ are not linearly independent. First we add a small regularization term to both sides of (A.5) to obtain

$$\begin{aligned}\mathbf{B}_\epsilon &= \mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2} + \epsilon\mathbf{\Lambda}^{-1}\\ &= \mathbf{\Lambda}^{-1/2}\left(\mathbf{H}^\dagger\mathbf{H} + \epsilon\mathbf{I}\right)\mathbf{\Lambda}^{-1/2}\\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\dagger + \epsilon\mathbf{\Lambda}^{-1}.\end{aligned} \tag{A.21}$$

We note that $\mathbf{B}_\epsilon$ is invertible for any $\epsilon > 0$. Let $\mathbf{V}_\epsilon\mathbf{\Sigma}_\epsilon\mathbf{V}_\epsilon^\dagger$ be the EVD of $\mathbf{B}_\epsilon$ and thus

$$\mathbf{B}_\epsilon^{-1} = \mathbf{V}_\epsilon\mathbf{\Sigma}_\epsilon^{-1}\mathbf{V}_\epsilon^\dagger. \tag{A.22}$$

Applying the result for the nonsingular case, we achieve the following inequality

$$\mathbf{\Psi}_\epsilon(\mathbf{x})\mathbf{X} \succeq \mathbf{\Psi}_\epsilon(\mathbf{y})\mathbf{Y} \tag{A.23}$$

for arbitrarily small ϵ and $\mathbf{X} \succeq \mathbf{Y}$, and $\mathbf{\Psi}_\epsilon(\cdot)$ in constructed from $\mathbf{B}_\epsilon$. To complete the proof we are left to show that $\mathbf{\Psi}_\epsilon(\tilde{\boldsymbol{\lambda}})$ is continuous with ϵ, i.e., $\lim\limits_{\epsilon\to 0^+}\mathbf{\Psi}_\epsilon(\tilde{\boldsymbol{\lambda}}) = \mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}) = \mathbf{A}^- + \mathbf{I}$.

To proceed, we note that (A.9) is changed into

$$\mathbf{B}_\epsilon^{-1} - \mathbf{I} = \mathbf{A}_\epsilon^+ + \mathbf{A}_\epsilon^- \tag{A.24}$$

where $\mathbf{A}_\epsilon^+$ and $\mathbf{A}_\epsilon^-$ are defined similarly to (A.8). We will show that $\lim_{\epsilon\to 0}\mathbf{A}_\epsilon^- \to \mathbf{A}^-$. To this end let $\epsilon_{\min} = \epsilon \times \min_i\{1/\lambda_i\}$ and $\epsilon_{\max} = \epsilon \times \max_i\{1/\lambda_i\}$, where λ_i is the ith diagonal entry of $\mathbf{\Lambda}$. It is clear from from (A.21) that the following inequality holds

$$\underbrace{\mathbf{V}\left(\left(\mathbf{\Sigma} + \epsilon_{\min}\mathbf{I}\right)^{-1} - \mathbf{I}\right)\mathbf{V}^\dagger}_{\Xi_{\epsilon_{\min}}} \succ \mathbf{B}_\epsilon^{-1} - \mathbf{I} \succ \underbrace{\mathbf{V}\left(\left(\mathbf{\Sigma} + \epsilon_{\max}\mathbf{I}\right)^{-1} - \mathbf{I}\right)\mathbf{V}^\dagger}_{\Xi_{\epsilon_{\max}}}. \tag{A.25}$$

Further, the matrix $\Xi_{\epsilon_{\min}}$ can be explicitly written as

$$\begin{aligned}\mathbf{\Xi}_{\epsilon_{\min}} = \mathbf{V}\,\mathrm{diag}([\frac{1}{\rho_1 + \epsilon_{\min}} - 1, \ldots, \frac{1}{\rho_r + \epsilon_{\min}} - 1,&\\ \underbrace{\frac{1}{\epsilon_{\min}} - 1, \ldots, \frac{1}{\epsilon_{\min}} - 1}_{(N-r)\ \text{terms}}])\mathbf{V}^\dagger&\end{aligned} \tag{A.26}$$

where $r = \mathrm{rank}(\mathbf{K})$. Following (A.9), we decompose $\mathbf{\Xi}_{\epsilon_{\min}}$ as

$$\mathbf{\Xi}_{\epsilon_{\min}} = \mathbf{A}_{\epsilon_{\min}}^+ + \mathbf{A}_{\epsilon_{\min}}^- \tag{A.27}$$

where $\mathbf{A}_{\epsilon_{\min}}^+$ and $\mathbf{A}_{\epsilon_{\min}}^-$ consists of positive and non-positive eigenvalues, respectively. As $\epsilon \to 0^+$ we have $\frac{1}{\rho_i+\epsilon_{\min}} \to \frac{1}{\rho_i}$ for all $i = 1, 2, ..., r$, and $\frac{1}{\epsilon_{\min}} \gg 1$. Thus , the term

$\frac{1}{\epsilon_{\min}} - 1$ in (A.26) becomes strictly positive and thus is excluded in $\mathbf{A}^{-}_{\epsilon_{\min}}$. As a result, we have $\lim_{\epsilon\to 0^+} \mathbf{A}^{-}_{\epsilon_{\min}} = \mathbf{A}^{-}$. Following the same arguments we can also show that $\lim_{\epsilon\to 0^+} \mathbf{A}^{-}_{\epsilon_{\max}} = \mathbf{A}^{-}$. From (A.25) it is clear that $\lim\limits_{\epsilon\to 0^+} \mathbf{A}^{-}_{\epsilon} = \mathbf{A}^{-}$ and thus

$$\lim_{\epsilon\to 0^+} \mathbf{\Psi}_{\epsilon}(\tilde{\boldsymbol{\lambda}}) = \lim_{\epsilon\to 0^+} (\mathbf{A}^{-}_{\epsilon} + \mathbf{I}) = \mathbf{A}^{-} + \mathbf{I} = \mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}). \tag{A.28}$$

By the continuity property shown above, the monotonicity of Algorithm 3.1 also holds for the singular case, which completes the proof.

A.2. Convergence Proof of Algorithm 3.2

We note that the function $\log|\mathbf{Q} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$ is *jointly concave* with $\mathbf{Q}$ and $\bar{\mathbf{S}}$. Thus the following inequality holds

$$\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}| \leq \log|\underbrace{\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H}}_{\mathbf{\Phi}_n}| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}-\mathbf{Q}_n)) + \mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}-\bar{\mathbf{S}}_n)) \tag{A.29}$$

for all $\mathbf{Q} \in \mathcal{Q}$ and $\bar{\mathbf{S}} \in \mathcal{S}$. The above inequality comes from the first order approximation of $\log|\mathbf{Q} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$ around the point $(\mathbf{Q}_n, \bar{\mathbf{S}}_n)$. Substitute $\mathbf{Q} := \mathbf{Q}_{n+1}$ and $\bar{\mathbf{S}} := \bar{\mathbf{S}}_{n+1}$ into the above equality, we have

$$\log|\mathbf{Q}_{n+1}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| \leq \log|\mathbf{\Phi}_n| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)) + \mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}_{n+1}-\bar{\mathbf{S}}_n)). \tag{A.30}$$

Since $\bar{\mathbf{S}}_n = \arg\max\limits_{\bar{\mathbf{S}}\in\mathcal{S}} \log|\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$, the optimality condition results in

$$\mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}} - \bar{\mathbf{S}}_n)) \leq 0 \tag{A.31}$$

for all $\bar{\mathbf{S}} \in \mathcal{S}$. For $\bar{\mathbf{S}} = \bar{\mathbf{S}}_{n+1}$ the above inequality means

$$\mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}_{n+1} - \bar{\mathbf{S}}_n)) \leq 0 \tag{A.32}$$

which leads to

$$\log|\mathbf{Q}_{n+1} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| \leq \log|\mathbf{\Phi}_n| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1} - \mathbf{Q}_n)). \tag{A.33}$$

Subtract both sides of the above inequality by $\log|\mathbf{Q}_{n+1}|$ results in

$$\begin{aligned} f(\mathbf{Q}_{n+1}, \bar{\mathbf{S}}_{n+1}) &= \log|\mathbf{Q}_{n+1} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| - \log|\mathbf{Q}_{n+1}| \\ &\leq \log|\mathbf{\Phi}_n| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1} - \mathbf{Q}_n)) - \log|\mathbf{Q}_{n+1}|. \end{aligned} \tag{A.34}$$

Since $\mathbf{Q}_{n+1}$ solves (3.23) it holds that

$$\log|\boldsymbol{\Phi}_n| + \operatorname{tr}\Big(\boldsymbol{\Phi}_n^{-1}\big(\mathbf{Q}_{n+1} - \mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}_{n+1}| \\ \leq \log|\boldsymbol{\Phi}_n| + \operatorname{tr}\Big(\boldsymbol{\Phi}_n^{-1}\big(\mathbf{Q} - \mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}| \quad (A.35)$$

for all $\mathbf{Q} \in \mathcal{Q}$. For the special case $\mathbf{Q} := \mathbf{Q}_n$, the above inequality is reduced to

$$\log|\boldsymbol{\Phi}_n| + \operatorname{tr}\Big(\boldsymbol{\Phi}_n^{-1}\big(\mathbf{Q}_{n+1} - \mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}_{n+1}| \leq \underbrace{\log|\boldsymbol{\Phi}_n| - \log|\mathbf{Q}_n|}_{f(\mathbf{Q}_n,\bar{\mathbf{S}}_n)}. \quad (A.36)$$

Combining (A.34) and (A.36) results in $f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \geq f(\mathbf{Q}_{n+1}, \bar{\mathbf{S}}_{n+1})$.

It is easy to see that $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ is bounded below, and thus $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ is convergent. We also note that (3.22) is strict if $\mathbf{Q} \neq \mathbf{Q}_n$. Consequently, the sequence $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ is *strictly decreasing* unless it is convergent.

Let us consider the set $\mathcal{Q}_+ \triangleq \{\operatorname{tr}(\mathbf{QP}) \leq P; \mathbf{Q} \succ \mathbf{0}\}$. Note that $\mathcal{Q}_+$ is open. As mentioned previously, $\mathbf{Q}_n \in \mathcal{Q}_+$ for all n. We will prove the two following properties regarding the convergence of Algorithm 3.2:

- Algorithm 3.2 generates at least a convergent subsequence.
- Let $\mathbf{Q}^*$ be the limit point of $\{\mathbf{Q}_n\}$. Then $\mathbf{Q}^*$ is nonsingular, i.e. $\mathbf{Q}^* \in \mathcal{Q}_+$.

The first property is relatively trivial. It is easy to see that the set $\mathcal{Q}_+$ is bounded (though it is open). As $\mathcal{Q}_+$ and $\mathcal{S}$ are both bounded, Algorithm 3.2 must produce at least a convergent subsequence, due to the Bolzano-Weierstrass theorem [79, 80]. The proof for the second property is quite involved, which is done by contraction as follows.

Suppose the contrary that $\mathbf{Q}^*$ is singular, i.e., there exists $\{q_{n,i}\} \to 0$ for some i. Recall that $\bar{\mathbf{S}}_n = \arg\max \log|\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}\mathbf{H}|$, and thus replacing $\bar{\mathbf{S}}_n = \frac{P}{N}\mathbf{I}$ which is a feasible point to the maximization problem results in

$$\log|\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H}| \geq \log|\mathbf{Q}_n + \tfrac{P}{N}\mathbf{H}^\dagger\mathbf{H}|. \quad (A.37)$$

Consequently we have

$$\begin{aligned} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) &\geq \log|\mathbf{Q}_n + \tfrac{P}{N}\mathbf{H}^\dagger\mathbf{H}| - \log|\mathbf{Q}_n| \\ &= \log|\mathbf{I} + \tfrac{P}{N}\mathbf{Q}_n^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{Q}_n^{-1/2}| \\ &= \log|\mathbf{I} + \tfrac{P}{N}\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger| \\ &= \log|\mathbf{I} + \tfrac{P}{N}\sum_{l=1}^{N} q_{n,l}^{-1}\mathbf{h}_l\mathbf{h}_l^\dagger|. \end{aligned} \quad (A.38)$$

Note that $\mathbf{h}_l$ is the lth column of $\mathbf{H}$. Let $\mathbf{A}_{n,i} = \mathbf{I} + \frac{P}{N}\sum_{l\neq i}^{N} q_{n,l}^{-1}\mathbf{h}_l\mathbf{h}_l^{\dagger}$. Then we can write

$$\begin{aligned} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) &\geq \log|\mathbf{A}_{n,i} + \tfrac{P}{N}q_{n,i}^{-1}\mathbf{h}_i\mathbf{h}_i^{\dagger}| \\ &= \log|\mathbf{A}_{n,i}| + \log|\mathbf{I} + \tfrac{P}{N}q_{n,i}^{-1}\mathbf{A}_{n,i}^{-1/2}\mathbf{h}_i\mathbf{h}_i^{\dagger}\mathbf{A}_{n,i}^{-1/2}| \\ &= \log|\mathbf{A}_{n,i}| + \log(1 + \tfrac{P}{N}q_{n,i}^{-1}\mathbf{h}_i^{\dagger}\mathbf{A}_{n,i}^{-1}\mathbf{h}_i). \end{aligned} \tag{A.39}$$

Let $\upsilon_{n,i}^{\max}$ be the maximum eigenvalue of $\mathbf{A}_{n,i}$, and thus $\frac{1}{\upsilon_{n,i}^{\max}}$ is the minimum eigenvalue of $\mathbf{A}_{n,i}^{-1}$. Then we have

$$f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \geq \log(\upsilon_{n,i}^{\max}) + \log(1 + \tfrac{P}{Nq_{n,i}\upsilon_{n,i}^{\max}}||\mathbf{h}_i||_2^2) \tag{A.40}$$

where we have used the fact that all eigenvalues of $\mathbf{A}_{n,i}$ are no less than 1, and that $\mathbf{x}^{\dagger}\mathbf{B}\mathbf{x} \geq \lambda_{\min}||\mathbf{x}||_2^2$, where $\lambda_{\min}$ is the minimum eigenvalue of $\mathbf{B}$.

To proceed further we consider two cases. Specifically, if $\lim_{n\to\infty}\upsilon_{n,i}^{\max} = \infty$, then it immediately holds that $\lim_{n\to\infty} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) = \infty$. Now suppose that there exists $c < \infty$ such that $1 \leq \upsilon_{n,i}^{\max} \leq c$ for all n. In this case we obtain

$$f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \geq \log(1 + \tfrac{P}{Nc}\tfrac{1}{q_{n,i}}||\mathbf{h}_i||_2^2). \tag{A.41}$$

It is straightforward to see that $f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \to \infty$ as $q_{n,i} \to 0$, due to the fact that $||\mathbf{h}_i||_2 > 0$.

In summary we have proved that if there exists $\{q_{n,i}\} \to 0$ for some i, then $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\} \to \infty$. This contradicts the fact that $\infty > f(\mathbf{Q}_0, \bar{\mathbf{S}}_0) \geq f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)$ for all n as proved earlier. Thus it is concluded that the limit point of Algorithm 2 $\mathbf{Q}^*$ is non-singular. By the continuity of $f(\cdot)$ over $\mathcal{S}$ and $\mathcal{Q}_+$, we have $\lim\limits_{n\to\infty} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) = f(\mathbf{Q}^*, \bar{\mathbf{S}}^*)$.

Now let $\{(\mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k})\}$ be the subsequence converging to the limit point. Next we shall show that $\{(\mathbf{Q}_{n_k+1}, \bar{\mathbf{S}}_{n_k+1})\} \to (\mathbf{Q}^*, \bar{\mathbf{S}}^*)$. In fact, it is sufficient to prove that $\mathbf{Q}_{n_k+1} \to \mathbf{Q}^*$ which can be done by contradiction. Assume the contrary that $\mathbf{Q}_{n_k+1}$ does not converge to $\mathbf{Q}^*$. Consequently, there exists a $d > 0$ such that

$$d \leq d_{n_k} = ||\mathbf{Q}_{n_k+1} - \mathbf{Q}_{n_k}||, \forall k \tag{A.42}$$

where $||\cdot||$ stands for arbitrary norm. We have

$$\begin{aligned} f(\mathbf{Q}_{n_k+1}, \bar{\mathbf{S}}_{n_k+1}) &\leq F(\mathbf{Q}_{n_k+1}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && \text{(A.43)} \\ &= F(\mathbf{Q}_{n_k} + d_{n_k}\mathbf{\Gamma}_{n_k}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && \text{(A.44)} \\ &\leq F(\mathbf{Q}_{n_k} + \delta d\mathbf{\Gamma}_{n_k}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}), \forall\delta \in [0,1] \\ &\leq F(\mathbf{Q}_{n_k}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && \text{(A.45)} \\ &= f(\mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && \text{(A.46)} \end{aligned}$$

where $\mathbf{\Gamma}_{n_k} \triangleq (\mathbf{Q}_{n_k+1} - \mathbf{Q}_{n_k})/d_{n_k}$ is the normalized distance between $\mathbf{Q}_{n_k+1}$ and $\mathbf{Q}_{n_k}$, $F(\mathbf{Q}_{n_k+1}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) = \log|\mathbf{\Phi}_{n_k}| + \operatorname{tr}\Big(\mathbf{\Phi}_{n_k}^{-1}\big(\mathbf{Q}_{n_k+1} - \mathbf{Q}_{n_k}\big)\Big) - \log|\mathbf{Q}_{n_k+1}|$. Note that $||\mathbf{\Gamma}_{n_k}|| = 1$ and thus $\mathbf{\Gamma}_{n_k}$ lies in a compact set and has a limit point $\mathbf{\Gamma}^*$. Letting $k \to \infty$ (by further restricting to a subsequence converging to $\mathbf{\Gamma}^*$) leads to

$$f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) \le F(\mathbf{Q}^* + \delta d\mathbf{\Gamma}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*) \le f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) \tag{A.47}$$

or equivalently

$$f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) = F(\mathbf{Q}^* + \delta d\mathbf{\Gamma}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*), \forall \delta \in [0, 1]. \tag{A.48}$$

Furthermore

$$\begin{aligned} F(\mathbf{Q}_{n_{k+1}}; \mathbf{Q}_{n_{k+1}}, \bar{\mathbf{S}}_{n_{k+1}}) &= f(\mathbf{Q}_{n_{k+1}}, \bar{\mathbf{S}}_{n_{k+1}}) \\ &\le f(\mathbf{Q}_{n_k+1}, \bar{\mathbf{S}}_{n_k+1}) \le F(\mathbf{Q}_{n_k+1}, \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) \le F(\mathbf{Q}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}), \forall \mathbf{Q} \in \mathcal{Q}_+. \end{aligned} \tag{A.49}$$

Letting $k \to \infty$ we obtain

$$F(\mathbf{Q}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*) \le F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*), \forall \mathbf{Q} \in \mathcal{Q}_+ \tag{A.50}$$

which further implies that $\mathbf{Q}^*$ is the minimizer of $F(\cdot; \mathbf{Q}^*, \bar{\mathbf{S}}^*)$. Since $\mathbf{Q}_{n_k+1} = \underset{\mathbf{Q} \in \mathcal{Q}_+}{\arg\min} F(\mathbf{Q}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k})$ it follows that

$$F(\mathbf{Q}_{n_k+1}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) \le F(\mathbf{Q}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}), \forall \mathbf{Q} \in \mathcal{Q}_+. \tag{A.51}$$

Letting $k \to \infty$ implies

$$F(\mathbf{Q}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*) \le F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*), \forall \mathbf{Q} \in \mathcal{Q}_+. \tag{A.52}$$

That is

$$\langle \nabla_{\mathbf{Q}} F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*)|_{\mathbf{Q}=\mathbf{Q}^*}, \mathbf{Z} - \mathbf{Q}^* \rangle \ge 0, \forall \mathbf{Z} \in \mathcal{Q}_+ \tag{A.53}$$

where $\langle . \rangle$ denotes the inner product. Recall that $F(\cdot; \mathbf{Q}, \bar{\mathbf{S}})$ is the first order of $f(\mathbf{Q}, \bar{\mathbf{S}})$. Thus it is easy to see that

$$\nabla_{\mathbf{Q}} F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*)|_{\mathbf{Q}=\mathbf{Q}^*} = \nabla f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) \tag{A.54}$$

and thus (A.53) is equivalent to

$$\langle \nabla_{\mathbf{Q}} f(\mathbf{Q}^*, \bar{\mathbf{S}}^*), \mathbf{Z} - \mathbf{Q}^* \rangle \ge 0, \forall \mathbf{Z} \in \mathcal{Q}_+. \tag{A.55}$$

In the same way we can show that

$$\langle \nabla_{\bar{\mathbf{S}}} f(\mathbf{Q}^*, \bar{\mathbf{S}}^*), \mathbf{W} - \bar{\mathbf{S}}^* \rangle \le 0, \forall \mathbf{W} \in \mathcal{S}. \tag{A.56}$$

Two above inequalities imply that $(\mathbf{Q}^*, \bar{\mathbf{S}}^*)$ is a saddle point of (3.15), which completes the proof.

A.3. Projection onto the Feasible Set of (3.41)

The projection of $\{\tilde{\mathbf{S}}_k\}$ onto the feasible set of (3.41) is formulated as

$$\begin{array}{rl} \text{minimize} & \sum_{k=1}^{K} ||\bar{\mathbf{S}}_k - \tilde{\mathbf{S}}_k||_F^2 \\ \text{subject to} & \sum_{k=1}^{K} \text{tr}(\bar{\mathbf{S}}_k) = P; \{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}. \end{array} \tag{A.57}$$

Let $\mathbf{U}_k\tilde{\mathbf{D}}_k\mathbf{U}_k^\dagger = \tilde{\mathbf{S}}_k$ be the EVD of $\tilde{\mathbf{S}}_k$, where $\mathbf{U}_k$ is unitary and $\tilde{\mathbf{D}}_k$ is diagonal. Then we can write $\bar{\mathbf{S}}_k = \mathbf{U}_k\bar{\mathbf{D}}_k\mathbf{U}_k^\dagger$ for some $\bar{\mathbf{D}}_k \succeq \mathbf{0}$. Since $\mathbf{U}_k$ is unitary, it holds that $\text{tr}(\bar{\mathbf{S}}_k) = \text{tr}(\bar{\mathbf{D}}_k)$ and that $||\bar{\mathbf{S}}_k - \tilde{\mathbf{S}}_k||_F = ||\bar{\mathbf{D}}_k - \tilde{\mathbf{D}}_k||_F$. That is to say, (A.57) is equivalent to

$$\begin{array}{rl} \text{minimize} & \sum_{k=1}^{K} ||\bar{\mathbf{D}}_k - \tilde{\mathbf{D}}_k||_F^2 \\ \text{subject to} & \sum_{k=1}^{K} \text{tr}(\bar{\mathbf{D}}_k) = P; \{\bar{\mathbf{D}}_k \succeq \mathbf{0}\}. \end{array} \tag{A.58}$$

It is easy to see that $\bar{\mathbf{D}}_k$ must be diagonal to minimize the objective of (A.58). Next let $\bar{\mathbf{d}}_k = \text{diag}(\bar{\mathbf{D}}_k)$, $\tilde{\mathbf{d}}_k = \text{diag}(\tilde{\mathbf{D}}_k)$, $\bar{\mathbf{d}} = [\bar{\mathbf{d}}_1^T, \bar{\mathbf{d}}_2^T, \ldots, \bar{\mathbf{d}}_K^T]^T$, and $\tilde{\mathbf{d}} = [\tilde{\mathbf{d}}_1^T, \tilde{\mathbf{d}}_2^T, \ldots, \tilde{\mathbf{d}}_K^T]^T$. Then (A.58) can be reduced to

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}||\bar{\mathbf{d}} - \tilde{\mathbf{d}}||_2^2 \\ \text{subject to} & \mathbf{1}_{\tilde{M}}\bar{\mathbf{d}} = P; \bar{\mathbf{d}} \geq 0 \end{array} \tag{A.59}$$

where $\tilde{M} = \sum_1^K M_k$. It is now clear that (A.59) is the projection onto a canonical simplex and efficient algorithms can be found in [51].

A.4. Duality Transformation Proof

The duality transformation in (3.32) can be proved using the same arguments as those in [30]. First, we write the partial Lagrangian function of (3.31) as

$$\mathcal{L}(\{\tilde{\mathbf{X}}_k\}, \mathbf{A}) = \sum\nolimits_{k=1}^{K} (w_k \log |\mathbf{I} + \tilde{\mathbf{H}}_k\tilde{\mathbf{X}}_k\tilde{\mathbf{H}}_k^H| - \text{tr}(\mathbf{C}_k\tilde{\mathbf{X}})) + \text{tr}(\mathbf{AP}) \tag{A.60}$$

where $\mathbf{C}_k = \mathbf{B}_k^H\mathbf{A}\mathbf{B}_k$, $\mathbf{A} = \text{diag}(a_1, a_2, \ldots, a_i, \ldots, a_N)$. Let $\hat{\mathbf{X}}_k = \mathbf{C}_k^{1/2}\tilde{\mathbf{X}}_k\mathbf{C}_k^{1/2}$. Then $\mathcal{L}(\{\tilde{\mathbf{X}}_k\}, \mathbf{A})$ is equal to

$$\mathcal{L}(\{\tilde{\mathbf{X}}_k\}, \mathbf{A}) = \sum\nolimits_{k=1}^{K} (w_k \log |\mathbf{I} + \tilde{\mathbf{H}}_k\mathbf{C}_k^{-1/2}\hat{\mathbf{X}}_k\mathbf{C}_k^{-1/2}\tilde{\mathbf{H}}_k^H| - \text{tr}(\hat{\mathbf{X}}_k)) + \text{tr}(\mathbf{AP}). \tag{A.61}$$

Denote $\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^H$ to be the singular value decomposition of $\tilde{\mathbf{H}}_k\mathbf{C}_k^{-1/2}$, i.e., $\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^H = \tilde{\mathbf{H}}_k\mathbf{C}_k^{-1/2}$. By the so-called channel flipping effect, we can express the dual objective as

$$\mathcal{D}(\mathbf{A}) = \max_{\dot{\mathbf{X}}_k \succeq 0} \sum\nolimits_{k=1}^{K} (w_k \log \frac{|\mathbf{B}_k^H\mathbf{A}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H\dot{\mathbf{X}}_k\tilde{\mathbf{H}}_k|}{|\mathbf{B}_k^H\mathbf{A}\mathbf{B}_k|} - \text{tr}(\dot{\mathbf{X}}_k)) + \text{tr}(\mathbf{AP}) \tag{A.62}$$

where $\hat{\mathbf{X}}_k = \mathbf{V}_k \mathbf{U}_k^H \dot{\mathbf{X}}_k \mathbf{U}_k \mathbf{V}_k^H$. Now the dual problem of (3.31) is

$$\min_{\mathbf{A}\succeq\mathbf{0}} \max_{\{\dot{\mathbf{X}}_k\}\succeq 0} \sum\nolimits_{k=1}^{K}(w_k \log \frac{|\mathbf{B}_k^H \mathbf{A}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H \dot{\mathbf{X}}_k \tilde{\mathbf{H}}_k|}{|\mathbf{B}_k^H \mathbf{A}\mathbf{B}_k|} - \operatorname{tr}(\dot{\mathbf{X}}_k)) + \operatorname{tr}(\mathbf{A}\mathbf{P}). \quad \text{(A.63)}$$

By introducing new optimization variable $\delta > 0$, we can rewrite the above problem as

$$\begin{aligned} &\min_{\mathbf{A}\succeq\mathbf{0},\delta>0} \max_{\{\dot{\mathbf{X}}_k\}\succeq\mathbf{0}} \sum\nolimits_{k=1}^{K} w_k \log \tfrac{|\mathbf{B}_k^H \mathbf{A}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H \dot{\mathbf{X}}_k \tilde{\mathbf{H}}_k|}{|\mathbf{B}_k^H \mathbf{A}\mathbf{B}_k|} - \delta P + \operatorname{tr}(\mathbf{A}\mathbf{P}) \\ &\quad \text{subject to } \ \sum\nolimits_{k=1}^{K} \operatorname{tr}(\dot{\mathbf{X}}) \leq \delta P. \end{aligned} \quad \text{(A.64)}$$

Note that we can again change the optimization variables as

$$\bar{\mathbf{X}}_k = \frac{\dot{\mathbf{X}}_k}{\delta}; \bar{\mathbf{A}} = \frac{\mathbf{A}}{\delta}. \quad \text{(A.65)}$$

Thus, (A.64) is equivalent to

$$\begin{aligned} &\min_{\bar{\mathbf{A}}\succeq\mathbf{0}} \max_{\{\bar{\mathbf{X}}_k\}\succeq\mathbf{0}} \ \sum\nolimits_{k=1}^{K} w_k \log \tfrac{|\mathbf{B}_k^H \bar{\mathbf{A}}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H \bar{\mathbf{X}}_k \tilde{\mathbf{H}}_k|}{|\mathbf{B}_k^H \bar{\mathbf{A}}\mathbf{B}_k|} \\ &\quad \text{subject to } \ \sum\nolimits_{k=1}^{K} \operatorname{tr}(\bar{\mathbf{X}}_k) \leq P; \operatorname{tr}(\bar{\mathbf{A}}\mathbf{P}) \leq P. \end{aligned} \quad \text{(A.66)}$$

which is the form given in (3.32) and thus completes the proof.

A.5. Convergence Proof of Algorithm 3.4

Let us define $\Omega \triangleq \{\mathbf{\Lambda}|\mathbf{\Lambda} : \text{diagonal}, \mathbf{\Lambda} \succeq \mathbf{0}, \operatorname{tr}(\mathbf{\Lambda}\mathbf{P}) = P\}$ and $\mathcal{X} = \{\bar{\mathbf{X}}_k|\bar{\mathbf{X}}_k \succeq \mathbf{0}, \sum_{k=1}^{K} \operatorname{tr}(\bar{\mathbf{X}}_k) = P, k = 1, \dots, K.\}$. We note that the sets Ω and $\mathcal{X}$ are compact convex. We first show that Algorithm 3.4 yields a *decreasing* objective $f(\mathbf{\Lambda}^n, \{\bar{\mathbf{X}}_k^n\})$ following similar arguments in [14, 15]. Since $\mathbf{\Lambda}^{n+1}$ is the optimal solution to the minimization problem (4.56), the inequality below holds

$$\begin{aligned} f(\mathbf{\Lambda}^n, \{\bar{\mathbf{X}}_k^n\}) &= \sum_{k=1}^{K} w_k \Big(\log|\mathbf{\Phi}_k^n| - \log|\mathbf{B}_k^H \mathbf{\Lambda}^n \mathbf{B}_k|\Big) \\ &\geq \sum_{k=1}^{K} w_k \Big(\log|\mathbf{\Phi}_k^n| + \operatorname{tr}\Big(\mathbf{B}_k \mathbf{\Phi}_k^{-n} \mathbf{B}_k^H \Big(\mathbf{\Lambda}^{n+1} - \mathbf{\Lambda}^n\Big)\Big) - \log|\mathbf{B}_k^H \mathbf{\Lambda}^{n+1} \mathbf{B}_k|\Big). \end{aligned} \quad \text{(A.67)}$$

In addition, $\log|\mathbf{B}_k^H \mathbf{\Lambda}\mathbf{B}_k + \tilde{\mathbf{H}}_k^H \bar{\mathbf{X}}_k \tilde{\mathbf{H}}_k|$ is jointly concave with $\mathbf{\Lambda}$ and $\bar{\mathbf{X}}_k$ and note that $\bar{\mathbf{X}}_k$ is the optimal solution to (4.54), we can easily prove that

$$\begin{aligned} &\sum_{k=1}^{K} w_k \Big(\log|\mathbf{\Phi}_k^n| + \operatorname{tr}\Big(\mathbf{B}_k \mathbf{\Phi}_k^{-n} \mathbf{B}_k^H \Big(\mathbf{\Lambda}^{n+1} - \mathbf{\Lambda}^n\Big)\Big) - \log|\mathbf{B}_k^H \mathbf{\Lambda}^{n+1} \mathbf{B}_k|\Big) \\ &\qquad \geq \underbrace{\sum_{k=1}^{K} w_k \Big(\log|\mathbf{B}_k^H \mathbf{\Lambda}^{n+1} \mathbf{B}_k + \tilde{\mathbf{H}}_k^H \bar{\mathbf{X}}_k^{n+1} \tilde{\mathbf{H}}_k| - \log|\mathbf{B}_k^H \mathbf{\Lambda}^{n+1} \mathbf{B}_k|\Big)}_{f(\mathbf{\Lambda}^{n+1}, \{\bar{\mathbf{X}}_k^{n+1}\})}. \end{aligned} \quad \text{(A.68)}$$

Combining (A.67) and (A.68) results in

$$f(\mathbf{\Lambda}^n,\{\bar{\mathbf{X}}_k^n\}) \geq \sum_{k=1}^{K} w_k\Big(\log|\mathbf{\Phi}_k^n|+\mathrm{tr}\Big(\mathbf{B}_k\mathbf{\Phi}_k^{-n}\mathbf{B}_k^H\Big(\mathbf{\Lambda}^{n+1}-\mathbf{\Lambda}^n\Big)\Big)-\log|\mathbf{B}_k^H\mathbf{\Lambda}^{n+1}\mathbf{B}_k|\Big)$$
$$\overset{(a)}{\geq} f(\mathbf{\Lambda}^{n+1},\{\bar{\mathbf{X}}_k^{n+1}\}). \quad \text{(A.69)}$$

We remark that the inequality (a) is strict if $\mathbf{\Lambda}^n \neq \mathbf{\Lambda}^{n+1}$. Thus, the sequence $\{f(\mathbf{\Lambda}^n,\{\bar{\mathbf{X}}_k^n\})\}$ is *strictly decreasing* unless it is convergent. Moreover, the continuity of $f(\cdot)$ and the compactness of $\mathcal{X}$ and Ω imply $\lim_{n\to\infty} f(\mathbf{\Lambda}^n,\{\bar{\mathbf{X}}_k^n\}) = f(\mathbf{\Lambda}^*,\{\bar{\mathbf{X}}_k^*\})$.

The proof of uniqueness, i.e., $\{(\mathbf{\Lambda}^{n_i+1},\{\bar{\mathbf{X}}_k^{n_i+1}\})\} \to (\mathbf{\Lambda}^*,\{\bar{\mathbf{X}}_k^*\})$ is similar to the arguments in Proof A.2 and is thus skipped for the sake of brevity.

Appendix B

Proofs of Chapter 4

B.1. Proof of Proposition 4.1

We will adopt the notion of the matrix covariance constraint presented in [8] and adapt it into our context. Explicitly for a given covariance matrix $\mathbf{S} \succeq \mathbf{0}$, the matrix covariance constraint is denoted by $E\{\mathbf{s}\mathbf{s}^\dagger\} \preceq \mathbf{S}$. Denote by R a given rate. Let $C(n, \mathbf{S}, R, \epsilon)$ be a codebook that maps each message $m \in \{1, 2, \dots, e^{nR}\}$ into a coded transmit sequence $\mathbf{X} \in \mathbb{C}^{N \times n}$. The receiver applies maximum likelihood decoding to decode the message index with an average probability of decoding error no greater than ϵ. Moreover, the codewords satisfy

$$\mathbf{S} = \frac{1}{e^{nR}} \sum_{\mathbf{X} \in C(n, \mathbf{S}, R, \epsilon)} \mathbf{X}\mathbf{X}^\dagger. \tag{B.1}$$

A rate R is said to be achievable under the matrix covariance constraint $\mathbf{S}$, if there exists an infinite sequence of codebooks, $C(n_l, \mathbf{S}_l, R, \epsilon_l)$, with increasing lengths n_l, rate R, matrices $\mathbf{S}_l \preceq \mathbf{S}$, and decreasing probabilities of error ϵ_l, such that $\epsilon_l \to 0$ as $l \to \infty$. For the feasible set defined in (4.2b), a rate is said to be achievable if $\mathbf{S}_l \in \mathcal{S}$ for all l.

It is well known that for a covariance matrix constraint $\mathbf{S}$, the following rate is achievable [8]

$$R \leq \log |\mathbf{I} + \mathbf{H}\mathbf{S}\mathbf{H}^\dagger|. \tag{B.2}$$

The above rate can be achieved by a Gaussian code with covariance matrix $\mathbf{S}$, which is due to the fact that Gaussian distribution maximizes the differential entropy among all distributions with the same covariance [81, Theorem 8.6.5]. The proof of achievability and proof of converse utilizing similar arguments in [81, Chapter 9] and [31, Appendix A] are in what follows.

Proof of Achievability Let $\mathbf{S}_j(u) = \mathbb{E}(s_j s_j^\dagger)$ where $j = 1, 2, \dots, n$ and $u = 1, 2, \dots, 2^{nR}$ and s_j is i.i.d. Gaussian with zero-mean and covariance $\mathbf{P} - \epsilon\mathbf{I}$. The average probability of error P_e^n is given by

$$\begin{aligned} P_e^n &= Pr(E_0 \cup E_1 \cup E_2 \cup E_3 \cup \ldots E_{2^{nR}}) \\ &\leq Pr(E_0) + Pr(E_1) + \sum_{j=2}^{2^{nR}} Pr(E_j) \end{aligned} \tag{B.3}$$

where $Pr(.)$ denotes conditional probability for the given codeword, E_0 is the event that a power constraint is violated i.e., $E_0 = \{(\mathrm{tr}(\frac{1}{n}(\sum_{j=1}^{n} \mathbf{E}_i \mathbf{S}_j)) > P_i)\}$, E_1 and $E_2 \cup E_3 \cup \ldots E_{2^{nR}}$ represent the cases where the received and transmitted codeword are not jointly typical and the received sequence is jointly typical with some wrong codeword, respectively.

By the law of large number and joint asymptotic equipartition property (AEP) [81, p.196], the above inequality can be written as

$$P_e^n \leq 2\epsilon + 2^{-n(-3\epsilon + (I(\mathbf{s};\mathbf{y}) - R))} \tag{B.4}$$

Thus $P_e^n \to 0$ when $n \to \infty$ and $R < I(\mathbf{s};\mathbf{y}) - 3\epsilon$ which completes the proof of achievability.

Proof of Converse We will show that if any rate R is achievable, then $R \leq \log|\mathbf{I} + \mathbf{HSH}^\dagger|$ for some $\mathbf{S} \in \mathcal{S}$. As the rate R is achievable, there exists an infinite sequence of codebooks, $C(n_l, \mathbf{S}_l, R', \epsilon_l)$, with increasing lengths n_l, rate R, matrices $\mathbf{S}_l \in \mathcal{S}$, and decreasing probabilities of error ϵ_l, such that $\frac{1}{n_l}\sum_{i=1}^{n_l} \mathbf{S}_i \to \mathbf{S}$ and $\epsilon_l \to 0$ as $l \to \infty$. Then the following result holds

$$R \leq \frac{1}{n_l}\sum_{i=1}^{n_l} I(\mathbf{s}_i; \mathbf{y}_i) + \epsilon_l \tag{B.5}$$

$$\leq \log|\mathbf{I} + \mathbf{H}\frac{1}{n_l}\sum_{i=1}^{n_l} \mathbf{S}_i \mathbf{H}^\dagger| + \epsilon_l \tag{B.6}$$

$$= \log|\mathbf{I} + \mathbf{HSH}^\dagger| + \epsilon_l \tag{B.7}$$

Therefore $R \leq \log|\mathbf{I} + \mathbf{HSH}^\dagger|$ when $\epsilon_l \to 0$.

B.2. Proof of Theorem 4.1

In this appendix, we prove the duality transformation in (4.3). We first write the partial Lagrangian function of (4.2) as

$$\begin{aligned} \mathcal{L}(\mathbf{a}, \mathbf{S}) &= \log|\mathbf{I} + \mathbf{HSH}^\dagger| - \sum_i a_i(\mathrm{tr}(\mathbf{E}_i\mathbf{S}) - P_i) \\ &= \log|\mathbf{I} + \mathbf{HSH}^\dagger| - \mathrm{tr}(\mathbf{AS}) + \mathbf{p}^T\mathbf{a} \end{aligned} \tag{B.8}$$

where $\mathbf{A} = \sum_i a_i \mathbf{E}_i$, $\mathbf{a} = [a_1, a_2, \ldots, a_L]^T$. Note that $\mathbf{A}$ must be positive definite (i.e., invertible), otherwise $\max_{\mathbf{S}\succeq 0} \mathcal{L}(\mathbf{a}, \mathbf{S}) \to \infty$. Let $\hat{\mathbf{S}} = \mathbf{A}^{1/2}\mathbf{S}\mathbf{A}^{1/2}$, then (B.8) becomes

$$\mathcal{L}(\mathbf{a}, \mathbf{S}) = \log |\mathbf{I} + \mathbf{H}\mathbf{A}^{-1/2}\hat{\mathbf{S}}\mathbf{A}^{-1/2}\mathbf{H}^\dagger| - \operatorname{tr}(\hat{\mathbf{S}}) + \mathbf{p}^T\mathbf{a}. \tag{B.9}$$

Let $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\dagger$ be the singular value decomposition of $\mathbf{H}\mathbf{A}^{-1/2}$, we proceed with the introduction of dual objective:

$$\mathcal{D}(\mathbf{a}) = \max_{\dot{\mathbf{S}}\succeq 0} \log |\mathbf{I} + \mathbf{A}^{-1/2}\mathbf{H}^\dagger\dot{\mathbf{S}}\mathbf{H}\mathbf{A}^{-1/2}| - \operatorname{tr}(\dot{\mathbf{S}}) + \mathbf{p}^T\mathbf{a} \tag{B.10}$$

where the relationship between uplink and downlink covariance matrix is given by $\hat{\mathbf{S}} = \mathbf{V}\mathbf{U}^\dagger\dot{\mathbf{S}}\mathbf{U}\mathbf{V}^\dagger$.

By definition, the dual problem is $\min_{\mathbf{a}\geq 0} \mathcal{D}(\mathbf{a})$, or equivalently

$$\min_{\mathbf{a}\geq 0} \max_{\dot{\mathbf{S}}\succeq 0} \log \frac{|\mathbf{A} + \mathbf{H}^\dagger\dot{\mathbf{S}}\mathbf{H}|}{|\mathbf{A}|} - \operatorname{tr}(\dot{\mathbf{S}}) + \mathbf{p}^T\mathbf{a}. \tag{B.11}$$

We can introduce a new optimization variable $\delta > 0$ so that the problem above can be rewritten as

$$\begin{array}{rl} \min\limits_{\mathbf{a}\geq 0, \delta > 0} \max\limits_{\dot{\mathbf{S}}\succeq \mathbf{0}} & \log \frac{|\mathbf{A}+\mathbf{H}^\dagger\dot{\mathbf{S}}\mathbf{H}|}{|\mathbf{A}|} - \delta P + \mathbf{p}^T\mathbf{a} \\ \text{subject to} & \operatorname{tr}(\dot{\mathbf{S}}) \leq \delta P. \end{array} \tag{B.12}$$

Again, we can change the variables by

$$\hat{q}_i = \frac{a_i}{\delta}, \bar{\mathbf{S}} = \frac{\dot{\mathbf{S}}}{\delta}. \tag{B.13}$$

Substituting (B.13) into (B.12), we arrive at the following optimization problem

$$\begin{array}{rl} \min\limits_{\hat{\mathbf{q}}\geq \mathbf{0}} \max\limits_{\bar{\mathbf{S}}\succeq \mathbf{0}} & \log \frac{|\sum_i \hat{q}_i \mathbf{E}_i + \mathbf{H}^\dagger\bar{\mathbf{S}}\mathbf{H}|}{|\sum_i \hat{q}_i \mathbf{E}_i|} \\ \text{subject to} & \operatorname{tr}(\bar{\mathbf{S}}) \leq P \\ & \mathbf{p}^T\hat{\mathbf{q}} \leq P \end{array} \tag{B.14}$$

which completes the proof.

B.3. Proof of Lemma 4.1

Following the similar arguments to those in [2, Appendix B], we can prove most of Lemma 4.1 except for part (c) as follows.

Proof of part (a) We note that the function $\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$ is *jointly concave* with $\mathbf{Q}$ and $\bar{\mathbf{S}}$ where $\mathbf{Q}\triangleq\sum_{i=1}^{L}q_i\mathbf{E}_i$. Thus the following inequality holds

$$\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}| \leq \log|\underbrace{\mathbf{Q}_n+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H}}_{\mathbf{\Phi}_n}| \\ +\operatorname{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}-\mathbf{Q}_n))+\operatorname{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}-\bar{\mathbf{S}}_n)) \quad \text{(B.15)}$$

for all $\mathbf{Q}\in\mathcal{Q}$ and $\bar{\mathbf{S}}\in\bar{\mathcal{S}}$. To lighten the notation we write $\mathbf{Q}\in\mathcal{Q}$ to denote $\{\mathbf{Q}|\boldsymbol{q}\in\mathcal{Q},\mathbf{Q}=\sum_{i=1}^{L}q_i\mathbf{E}_i\}$ The above inequality comes from the first-order approximation of $\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$ around the point $(\mathbf{Q}_n,\bar{\mathbf{S}}_n)$. Substituting $\mathbf{Q}:=\mathbf{Q}_{n+1}$ and $\bar{\mathbf{S}}:=\bar{\mathbf{S}}_{n+1}$ into the above equality, we obtain

$$\log|\mathbf{Q}_{n+1}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| \leq \log|\mathbf{\Phi}_n| \\ +\operatorname{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n))+\operatorname{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}_{n+1}-\bar{\mathbf{S}}_n)). \quad \text{(B.16)}$$

Since $\bar{\mathbf{S}}_n=\arg\max_{\bar{\mathbf{S}}\in\bar{\mathcal{S}}}\ \log|\mathbf{Q}_n+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$, the optimality condition results in

$$\operatorname{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}-\bar{\mathbf{S}}_n))\leq 0 \quad \text{(B.17)}$$

for all $\bar{\mathbf{S}}\in\bar{\mathcal{S}}$. For $\bar{\mathbf{S}}=\bar{\mathbf{S}}_{n+1}$ the above inequality implies

$$\operatorname{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}_{n+1}-\bar{\mathbf{S}}_n))\leq 0 \quad \text{(B.18)}$$

which leads to

$$\log|\mathbf{Q}_{n+1}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}|\leq\log|\mathbf{\Phi}_n|+\operatorname{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)). \quad \text{(B.19)}$$

Subtracting $\log|\mathbf{Q}_{n+1}|$ from both sides of the above inequality results in

$$f(\mathbf{q}^{n+1},\bar{\mathbf{S}}_{n+1})=\log|\mathbf{Q}_{n+1}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}|-\log|\mathbf{Q}_{n+1}| \\ \leq\log|\mathbf{\Phi}_n|+\operatorname{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n))-\log|\mathbf{Q}_{n+1}|. \quad \text{(B.20)}$$

Since $\mathbf{Q}_{n+1}$ solves (4.8), it holds that

$$\log|\mathbf{\Phi}_n|+\operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}_{n+1}-\mathbf{Q}_n\big)\Big)-\log|\mathbf{Q}_{n+1}| \\ \leq\log|\mathbf{\Phi}_n|+\operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}-\mathbf{Q}_n\big)\Big)-\log|\mathbf{Q}| \quad \text{(B.21)}$$

for all $\mathbf{q}\in\mathcal{Q}$. For the special case $\mathbf{Q}:=\mathbf{Q}_n$, the above inequality is reduced to

$$\log|\mathbf{\Phi}_n|+\operatorname{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}_{n+1}-\mathbf{Q}_n\big)\Big)-\log|\mathbf{Q}_{n+1}| \\ \leq\underbrace{\log|\mathbf{\Phi}_n|-\log|\mathbf{Q}_n|}_{f(\mathbf{q}^n,\bar{\mathbf{S}}_n)}. \quad \text{(B.22)}$$

Combining (B.20) and (B.22) results in $f(\mathbf{q}^n,\bar{\mathbf{S}}_n)\geq f(\mathbf{q}^{n+1},\bar{\mathbf{S}}_{n+1})$. It is easy to see that $\{f(\mathbf{q}^n,\bar{\mathbf{S}}_n)\}$ is bounded below, and thus $\{f(\mathbf{q}^n,\bar{\mathbf{S}}_n)\}$ is convergent. We also note that (4.5) is strict if $\mathbf{q}\neq\mathbf{q}_n$. Consequently, the sequence $\{f(\mathbf{q}^n,\bar{\mathbf{S}}_n)\}$ is *strictly decreasing* unless it is convergent.

Proof of part (b) In fact, part (b) of Lemma 4.1 is trivial. It is easy to see that $\mathcal{Q}$ and $\bar{\mathcal{S}}$ are both bounded, Algorithm 4.2 must produce at least a convergent subsequence, due to the Bolzano-Weierstrass theorem.

Proof of part (c) As mentioned previously, the proof of nonsingularity of [2, Appendix B] is not applicable to our considered problem since $\mathbf{Q}$ in the present form is not a simple diagonal matrix. More specifically, part (c) is proved by contraction as follows. Let $\mathbf{q}^*$ be the limit point of $\{\mathbf{q}^n\}$. Suppose the contrary that $\mathbf{Q}^* = \sum_{i=1}^{L} q_i^* \mathbf{E}_i$ is singular. By abuse of notation, let $I = \{i \in [1, L] : q_i^* > 0\}$. It is easy to see that if $\mathbf{E}_i \succ \mathbf{0}$ for some $i \in I$, then $\mathbf{Q}^*$ is non-singular. Thus the singularity of $\mathbf{Q}^*$ implies that $\mathbf{E}_i$ is singular, $\forall i \in I$. As a result of the assumption made in the system model, there exists a vector $\boldsymbol{v} \neq 0$ such that $\mathbf{E}_i \boldsymbol{v} = \mathbf{0}$ for all $\forall i \in I$ and $||\mathbf{H}\boldsymbol{v}||_2 = c > 0$.

Recall that $\bar{\mathbf{S}}_n = \arg\max_{\bar{\mathbf{S}} \in \bar{\mathcal{S}}} \log |\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}} \mathbf{H}|$, and thus replacing $\bar{\mathbf{S}}_n = \frac{P}{M}\mathbf{I}$ which is a feasible point to the maximization problem results in

$$\log |\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H}| \geq \log |\mathbf{Q}_n + \tfrac{P}{M}\mathbf{H}^\dagger \mathbf{H}|. \tag{B.23}$$

Consequently we have

$$\begin{aligned} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) &\geq \log |\mathbf{Q}_n + \tfrac{P}{M}\mathbf{H}^\dagger \mathbf{H}| - \log |\mathbf{Q}_n| \\ &= \log |\mathbf{I} + \tfrac{P}{M}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger \mathbf{H}| \\ &\geq \log \left(1 + \tfrac{P}{M}\lambda_{\max}\left(\mathbf{Q}_n^{-1}\mathbf{H}^\dagger \mathbf{H}\right)\right). \end{aligned} \tag{B.24}$$

We have the following inequality:

$$\begin{aligned} \lambda_{\max}\left(\mathbf{Q}_n^{-1}\mathbf{H}^\dagger \mathbf{H}\right) &= \max_{\mathbf{u} \neq 0} \frac{\mathbf{u}^\dagger \mathbf{H}^\dagger \mathbf{H} \mathbf{u}}{\mathbf{u}^\dagger \mathbf{Q}_n \mathbf{u}} && \text{(B.25)} \\ &\geq \frac{\boldsymbol{v}^\dagger \mathbf{H}^\dagger \mathbf{H} \boldsymbol{v}}{\boldsymbol{v}^\dagger \mathbf{Q}_n \boldsymbol{v}} = \frac{||\mathbf{H}\boldsymbol{v}||_2}{\boldsymbol{v}^\dagger \left(\sum_{i=1}^{L} q_i^n \mathbf{E}_i\right) \boldsymbol{v}} && \text{(B.26)} \\ &= \frac{c^2}{\boldsymbol{v}^\dagger \left(\sum_{i=1}^{L} q_i^n \mathbf{E}_i\right) \boldsymbol{v}}. && \text{(B.27)} \end{aligned}$$

Let $\{q_i^{n_k}\}$ be a subsequence converging to $\mathbf{q}^*$. We have

$$\begin{aligned} \lim_{k \to \infty} \boldsymbol{v}^\dagger \left(\sum_{i=1}^{L} q_i^{n_k} \mathbf{E}_i\right) \boldsymbol{v} &= \boldsymbol{v}^\dagger \left(\sum_{i=1}^{L} q_i^{*} \mathbf{E}_i\right) \boldsymbol{v} \\ &= \boldsymbol{v}^\dagger \left(\sum_{i \in I} q_i^{*} \mathbf{E}_i\right) \boldsymbol{v} = 0. \end{aligned} \tag{B.28}$$

In summary we have proved that if $\mathbf{Q}^*$ is singular, then $\lambda_{\max}\left(\mathbf{Q}_n^{-1}\mathbf{H}^\dagger \mathbf{H}\right) \to \infty$ and thus $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\} \to \infty$. This contradicts the fact that $\infty > f(\mathbf{Q}_0, \bar{\mathbf{S}}_0) \geq f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)$

for all n as proved earlier. Thus it is concluded that the limit point of Algorithm 2 $\mathbf{Q}^*$ is non-singular. By the continuity of $f(\cdot)$ over $\bar{\mathcal{S}}$ and $\mathcal{Q}$, we have $\lim_{n\to\infty} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) = f(\mathbf{Q}^*, \bar{\mathbf{S}}^*)$.

Proof of part (d) The proof of part (d) follows similar arguments to those in [2], and thus we refer interested readers to [2, Appendix B] for the details.

B.4. Proof of Proposition 4.2

Suppose the contrary that $q_{N+1} = 0$. Then it immediately holds that $\mu_i = 0$ and $x_i = \frac{1}{\phi_i + \gamma P_i}$, for $i = 1, 2, \ldots, N$. From (4.14d) we have

$$\gamma = \frac{\mu_{N+1}}{P_T'} \leq 0. \tag{B.29}$$

As a result, the following inequality is obtained

$$\sum_{i=1}^{N} P_i x_i = \sum_{i=1}^{N} \frac{P_i}{\phi_i + \gamma P_i} \geq \sum_{i=1}^{N} \frac{P_i}{\phi_i} > \sum_{i=1}^{N} P_i (q_i^n + q_{N+1}^n). \tag{B.30}$$

Note that q_i^n and q_{N+1}^n are a solution to (4.9), and thus

$$\sum_{i=1}^{N} P_i (q_i^n + q_{N+1}^n) = \underbrace{\sum_{i=1}^{N+1} P_i q_i^n}_{P} + \left(\underbrace{\sum_{i=1}^{N} P_i - P_T}_{>0} \right) q_{N+1}^n \geq P. \tag{B.31}$$

Combining (B.30) and (B.31) yields

$$\sum_{i=1}^{N} P_i x_i > P \tag{B.32}$$

which indicates that x_i's are not feasible to (4.13) and thus completes the proof.

B.5. Proof that the Specified Choice of γ_{max} solves (4.13)

First note that (4.14c) and (4.14d) produce

$$\phi_i x_i - 1 + \gamma P_i x_i - \mu_i x_i = 0, \; i = 1, 2, \ldots, N \tag{B.33}$$

$$\gamma P_T' q_{N+1} - \underbrace{\mu_{N+1} q_{N+1}}_{0} + q_{N+1} \sum_{i=1}^{N} \mu_i = 0, \tag{B.34}$$

respectively. From (B.33), (B.34), and (4.14a) we have

$$\textstyle\sum_{i=1}^{N} \phi_i x_i - N + \gamma(\sum_{i=1}^{N} P_i x_i + P_T' q_{N+1}) = 0 \tag{B.35}$$

which is equivalent to

$$\textstyle\sum_{i=1}^{N} \phi_i x_i - N + \gamma P = 0 \tag{B.36}$$

and thus

$$\gamma P \leq N - \phi_{\min} \textstyle\sum_{i=1}^{N} x_i \tag{B.37}$$

where $\phi_{\min} = \min\limits_{1 \leq i \leq N} \{\phi_i\}$. It is easy to see that

$$P_{\max} \textstyle\sum_{i=1}^{N} x_i \geq \sum_{i=1}^{N} P_i x_i = P - P_T' q_{N+1} \geq P \tag{B.38}$$

where $P_{\max} = \max\limits_{1 \leq i \leq N} \{P_i\}$. Combining (B.37) and (B.38) yields

$$\gamma \leq \frac{N}{P} - \frac{\phi_{\min}}{P_{\max}}. \tag{B.39}$$

Thus it is sufficient to set $\gamma_{\max} = \frac{N}{P} - \frac{\phi_{\min}}{P_{\max}}$ for the bisection method to solve (4.13).

B.6. Proof of Theorem 4.1

As a result of Propositions 4.5 and 4.6, there exists a number k such that

$$q_1 \geq q_2 \geq \cdots \geq q_k > 0 \tag{B.40}$$

and

$$q_{k+1} = q_{k+2} = \cdots = q_N = 0. \tag{B.41}$$

Thus the KKT equations reduce to

$$-\frac{|h_i|^2}{(q_i + q_{N+1})^2} + \gamma P_i = 0,\ i = 1, 2, \ldots, k \tag{B.42}$$

$$-\frac{|h_i|^2}{q_{N+1}^2}+\gamma P_i-\mu_i,\ i=k+1,k+2,\ldots,N \tag{B.43}$$

$$-\sum_{i=1}^{k}\frac{|h_i|^2}{\left(q_i+q_{N+1}\right)^2}-\sum_{i=k+1}^{N}\frac{|h_i|^2}{q_{N+1}^2}+\gamma P_T=0 \tag{B.44}$$

$$P_T q_{N+1}+\sum_{i=1}^{k}P_i q_i=P. \tag{B.45}$$

The problem now boils down to finding the optimal k and γ. Combining (B.42) and (B.44) we have

$$\gamma\left(P_T-\sum_{i=1}^{k}P_i\right)=\frac{1}{q_{N+1}^2}\sum_{i=k+1}^{N}|h_i|^2 \tag{B.46}$$

and thus

$$q_{N+1}=\frac{1}{\sqrt{\gamma}}\sqrt{\frac{\sum_{i=k+1}^{N}|h_i|^2}{\left(P_T-\sum_{i=1}^{k}P_i\right)}}. \tag{B.47}$$

Furthermore, combining (B.47) with (B.42) we obtain

$$q_i=\frac{1}{\sqrt{\gamma}}\frac{|h_i|}{\sqrt{P_i}}-q_{N+1}=\frac{1}{\sqrt{\gamma}}\left(\frac{|h_i|}{\sqrt{P_i}}-\sqrt{\frac{\sum_{i=k+1}^{N}|h_i|^2}{\left(P_T-\sum_{i=1}^{k}P_i\right)}}\right). \tag{B.48}$$

We may now substitute (B.47) and (B.48) into (B.45) to yield the closed-form solution in Theorem 4.1.

B.7. Proof of Theorem 4.2

We can write the KKT conditions for the considered problem as

$$\mu_i q_i=0 \tag{B.49}$$

$$\psi_i-\frac{1}{q_{N+1}+q_i}+\gamma P_i-\mu_i=0, i=1,2,\ldots,N \tag{B.50}$$

$$\psi_{N+1} - \sum\nolimits_{i=1}^{N} \frac{1}{q_{N+1} + q_i} + \gamma P_T - \mu_{N+1} = 0, i = N+1. \tag{B.51}$$

From (B.49), if $q_i > 0$ for $i = 1, 2, \ldots, N$ then the corresponding $\mu_i = 0$ which results in

$$\mu_{N+1} = \gamma(P_T - \sum\nolimits_{i=1}^{N} P_i). \tag{B.52}$$

In this chapter, we only consider the case where the sum power constraint is less than the total power of PAPC i.e., $P_T < \sum_{i=1}^{N} P_i$, therefore $\mu_{N+1} = 0, q_{N+1} > 0$.

Without loss of generality, we can sort $\{\frac{1}{\psi_i + \gamma P_i}\}$ in decreasing order. Following similar arguments to those in Propositions 4.3, 4.4, 4.5 and 4.6, we can find an integer k such that

$$q_1 \geq q_2 \geq \cdots \geq q_k > 0 \tag{B.53}$$

and

$$q_{k+1} = q_{k+2} = \cdots = q_N = 0. \tag{B.54}$$

Based on these results, combining (B.50) and (B.51) results in

$$q_{N+1} = \frac{N-k}{(\psi_{N+1} - \sum_{i=1}^{k} \psi_i) + \gamma(P_T - \sum_{i=1}^{k} P_i)} \tag{B.55}$$

$$q_i = \frac{1}{\psi_i + \gamma P_i} - \frac{N-k}{(\psi_{N+1} - \sum_{i=1}^{k} \psi_i) + \gamma(P_T - \sum_{i=1}^{k} P_i)}. \tag{B.56}$$

Substituting these values of q_i into the power constraint $\sum_{i=1}^{N+1} P_i q_i = P$, we obtain

$$\sum_{i=1}^{k} \frac{P_i}{\psi_i + \gamma P_i} + \frac{(N-k)(P_T - \sum_{i=1}^{k} P_i)}{(\psi_{N+1} - \sum_{i=1}^{k} \psi_i) + \gamma(P_T - \sum_{i=1}^{k} P_i)} = P \tag{B.57}$$

whose value of γ can be solved easily by the Newton method or bisection method.

Bibliography

[1] M. Vu, “MIMO capacity with per-antenna power constraint,” in *Proc. IEEE GLOBECOM*, Dec. 2011, pp. 1 – 5.

[2] T. M. Pham, and R. Farrell, and L.-N. Tran, “Revisiting the MIMO capacity with per-antenna power constraint: Fixed-point iteration and alternating optimization,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 388–401, Jan 2019.

[3] M. Vu, “The capacity of MIMO channels with per-antenna power constraint,” *CoRR*, vol. abs - 1106 - 5039, 2011.

[4] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).*, 2015.

[5] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[6] E. Telatar, “Capacity of multi-antenna Gaussian channels,” *Eur. Trans. Telecommun*, vol. 10, pp. 585–598, Nov. 1999.

[7] G. J. Foschini and M. J. Gans, “On limits of wireless communications in a fading environment when using multiple antennas,” *Wireless Pers.Commun*, vol. 6, pp. 311–335, Mar. 1998.

[8] H. Weingarten, Y. Steinberg, and S. S. Shamai, “The capacity region of the Gaussian multiple-input multiple-output broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936 – 3964, Sep. 2006.

[9] G. Caire and S. Shamai, “On the achievable throughput of a multiantenna Gaussian broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[10] W. Yu and T. Lan, “Transmitter optimization for the multi-antenna downlink with per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.

[11] H. Q. Ngo, E. Larsson, and T. Marzetta, “Energy and spectral efficiency of very large multiuser MIMO systems,” *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[12] T. M. Pham, and R. Farrell, and J. Dooley, and E. Dutkiewicz, and D. N. Nguyen, and L.-N. Tran, “Efficient zero-forcing precoder design for weighted sum-rate maximization with per-antenna power constraint,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3640–3645, Apr. 2018.

[13] T. M. Pham, R. Farrell, H. Claussen, M. F. Flanagan, and L.-N. Tran, "On the MIMO capacity under multiple linear transmit covariance constraints," *IEEE Trans. Signal Process.*, 2019 submitted.

[14] T. M. Pham, and R. Farrell, and L.-N. Tran, "Low-complexity approaches for MIMO capacity with per-antenna power constraint," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–7.

[15] ——, "Alternating optimization for capacity region of Gaussian MIMO broadcast channels with per-antenna power constraint," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–6.

[16] T. M. Pham, R. Farrell, H. Claussen, M. F. Flanagan, and L.-N. Tran, "Weighted sum rate maximization for zero-forcing methods with general linear covariance constraints," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[17] ——, "On the MIMO capacity with multiple linear transmit covariance constraints," in *Proc. IEEE VTC-Spring*, Jun. 2018, pp. 1–6.

[18] T. M. Pham, R. Farrell, and L.-N. Tran, "On estimating maximum sum rate of MIMO systems with successive zero-forcing dirty paper coding and per-antenna power constraint," in *Proc. IEEE PIMRC*, Sep. 2019.

[19] A. Dabbagh and D. Love, "Precoding for multiple antenna Gaussian broadcast channels with successive zero-forcing," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3837–3850, Jul. 2007.

[20] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[21] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409 – 4418, Sep. 2008.

[22] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[23] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[24] J. Liu, Y. T. Hou, and H. D. Sherali, "On the maximum weighted sum-rate of MIMO Gaussian broadcast channels," in *Proc. IEEE ICC*, May 2008, pp. 3664 – 3668.

[25] R. Hunger, D. A. Schmidt, M. Joham, and W. Utschick, "A general covariance-based optimization framework using orthogonal projections," in *Proc. IEEE SPAWC*, Jul. 2008, pp. 76 – 80.

[26] T. E. Bogale and L. Vandendorpe, “Weighted sum rate optimization for downlink multiuser MIMO systems with per antenna power constraint: Downlink-uplink duality approach,” in *Proc. IEEE ICASSP*, Mar. 2012, pp. 3245 – 3248.

[27] B. Li, C. Z. Wu, H. H. Dam, A. Cantoni, and K. L. Teo, “A parallel low complexity zero-forcing beamformer design for multiuser MIMO systems via a regularized dual decomposition method,” *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4179 – 4190, Aug. 2015.

[28] S. Shi, M. Schubert, and H. Boche, “Per-antenna power constrained rate optimization for multiuser MIMO systems,” in *Proc. WSA*, Feb. 2008, pp. 270 – 277.

[29] L.-N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, “Weighted sum rate maximization for MIMO broadcast channels using dirty paper coding and zero-forcing methods,” *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2362–2373, Jun. 2013.

[30] ——, “Beamformer designs for MISO broadcast channels with zero-forcing dirty paper coding,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1173–1185, Mar. 2013.

[31] P. L. Cao, T. J. Oechtering, R. F. Schaefer, and M. Skoglund, “Optimal transmit strategy for MISO channels with joint sum and per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4296 – 4306, Aug. 2016.

[32] S. Loyka, “The capacity of Gaussian MIMO channels under total and per-antenna power constraints,” *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1035 – 1043, Mar. 2017.

[33] L. Zhang, R. Zhang, Y. C. Liang, Y. Xin, and H. V. Poor, “On Gaussian MIMO BC-MAC duality with multiple transmit covariance constraints,” *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 2064 – 2078, Apr. 2012.

[34] H. Huh, H. C. Papadopoulos, and G. Caire, “Multiuser MISO transmitter optimization for intercell interference mitigation,” *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4272 – 4285, Aug. 2010.

[35] Y. Yang, G. Scutari, P. Song, and D. P. Palomar, “Robust MIMO cognitive radio systems under interference temperature constraints,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2465–2482, Nov. 2013.

[36] A. Goldsmith, *Wireless Communications.* New York, NY, USA: Cambridge University Press, 2005.

[37] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684 – 702, Jun. 2003.

[38] L. Vandenberghe, S. Boyd, and S.-P. Wu, “Determinant maximization with linear matrix inequality constraints,” *SIAM J. on Matrix Anal. and Appl.*, vol. 19, no. 2, pp. 499–533, 1998.

[39] A. Ben-Tal and A. Nemirovskiaei, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications.* Philadelphia, PA, USA: SIAM, 2001.

[40] W. Yu, “Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition,” *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 754 –759, Feb. 2006.

[41] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th ed. The John Hopkins Univ. Press, 2013.

[42] W. Yu, “Uplink-downlink duality via minimax duality,” *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361–374, Feb. 2006.

[43] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.

[44] A. Beck, A. Ben-Tal, and L. Tetruashvili, “A sequential parametric convex approximation method with applications to nonconvex truss topology design problems,” *Journal of Global Optimization*, vol. 47, no. 1, pp. 29–51, 2010.

[45] Z. Shen, R. Chen, J. Andrews, J. Heath, R.W., and B. Evans, “Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization,” *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658–3663, Sep. 2006.

[46] L.-N. Tran, M. Bengtsson, and B. Ottersten, “Iterative precoder design and user scheduling for block-diagonalized systems,” *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3726–3739, Jul. 2012.

[47] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[48] R. Zhang, “Cooperative multi-cell block diagonalization with per-base-station power constraints,” *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1435–1445, Dec. 2010.

[49] Chris T. K. Ng and H. Huang, “Linear precoding in cooperative MIMO cellular networks with limited coordination clusters,” *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Dec. 2010.

[50] L. Armijo, “Minimization of functions having Lipschitz continuous first partial derivatives.” *Pacific J. Math.*, vol. 16, no. 1, pp. 1 – 3, 1966.

[51] L. Condat, “Fast projection onto the simplex and the ℓ_1 ball,” *Mathematical Programming, Series A*, vol. 158, no. 1, pp. 575 – 585, Jul. 2016.

[52] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The John Hopkins Univ. Press, 1996.

[53] J. Liu, Y. T. Hou, S. Kompella, and H. D. Sherali, “Conjugate gradient projection approach for MIMO Gaussian broadcast channels,” in *Proc. IEEE ISIT*, June 2007, pp. 781 – 785.

[54] K. C. Toh, M. J. Todd, and R. Tutuncu, “SDPT3— a Matlab software package for semidefinite programming,” *Optimization Methods and Software*, Nov. 1999.

[55] J. Löfberg, “YALMIP : A toolbox for modeling and optimization in MATLAB,” in *Proc. the CACSD Conference*, 2004.

[56] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, “Multiple-antenna channel hardening and its implications for rate feedback and scheduling,” *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893 – 1909, Sep. 2004.

[57] H. Q. Ngo and E. G. Larsson, “No downlink pilots are needed in massive MIMO,” *CoRR*, vol. abs - 1606 - 02348, 2016.

[58] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V.-M. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen, “Winner II channel models,” *tech. rep. D1.1.2 V1.2, IST-4-027756 WINNER II*, 2007.

[59] V. Nguyen, L.-N. Tran, T. Q. Duong, O. Shin, and R. Farrell, “An efficient precoder design for multiuser MIMO cognitive radio networks with interference constraints,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3991–4004, May 2017.

[60] L. Zhang, Y. Xin, and Y. Liang, “Weighted sum rate optimization for cognitive radio MIMO broadcast channels,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 2950–2959, Jun. 2009.

[61] P. L. Cao and T. J. Oechtering, “Optimal transmit strategy for MIMO channels with joint sum and per-antenna power constraints,” in *Proc. IEEE ICASSP*, Mar. 2017, pp. 3569–3573.

[62] S. Loyka, “On the capacity of Gaussian MIMO channels under the joint power constraints,” *IEEE Commun. Lett.*, vol. 8, no. 2, pp. 332–335, Apr. 2019.

[63] C. Xing, Z. Fei, Y. Zhou, and Z. Pan, “Matrix-field water-filling architecture for MIMO transceiver designs with mixed power constraints,” in *Proc. IEEE PIMRC*, Aug. 2015, pp. 392–396.

[64] O. Simeone, “A very brief introduction to machine learning with applications to communication systems,” *CoRR*, vol. abs/1808.02342, 2018.

[65] E. Brynjolfsson and T. Mitchell, “What can machine learning do? Workforce implications,” *Science*, vol. 358, no. 6370, pp. 1530–1534, 2017.

[66] S. Marsland, *Machine Learning: An Algorithmic Perspective*, 1st ed. Chapman & Hall/CRC, 2009.

[67] K.-C. Wong, *Evolutionary algorithms: Concepts, designs, and applications in bioinformatics.* IGI Global, Nov. 2015, vol. 1-3, pp. 111–137.

[68] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in *Proc. IEEE ICSEM*, vol. 1, Nov. 2010, pp. 27–30.

[69] L.-N. Tran and E.-K. Hong, "Multiuser diversity for successive zero-forcing dirty paper coding: Greedy scheduling algorithms and asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3411–3416, Jun. 2010.

[70] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *arXiv e-prints*, p. arXiv:1902.02647, Feb. 2019.

[71] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[72] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 1st ed. Orlando, FL, USA: Academic Press, Inc., 2015.

[73] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[74] M.-F. Li, X.-P. Tang, W. Wu, and H.-B. Liu, "General models for estimating daily global solar radiation for different solar radiation zones in mainland china," *Energy Conversion and Management*, vol. 70, pp. 139 – 148, 2013.

[75] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas - part II: The MIMOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5515–5532, Nov. 2010.

[76] S. Loyka and C. D. Charalambous, "An algorithm for global maximization of secrecy rates in Gaussian MIMO wiretap channels," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2288–2299, Jun. 2015.

[77] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.

[78] R. A. Horn and C. R. Johnson, *Matrix Analysis.* New York, NY, USA: Cambridge University Press, 1986.

[79] R. G. Bartle and D. R. Sherbert, *Introduction to real analysis*, 4th ed. John Wiley and Sons, Inc., 2011.

[80] P. M. Fitzpatrick, *Advanced calculus*, 2nd ed. Thomson Brooks/Cole, 2006.

[81] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, 2006.