On Optimization of Multiuser Multiple Input Multiple Output Communication Systems

ON OPTIMIZATION OF MULTIUSER MULTIPLE INPUT MULTIPLE OUTPUT COMMUNICATION SYSTEMS

BY

PETER HE

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2009) McMaster University

(Computational Engineering & Science) Hamilton, Ontario, Canada

TITLE:	On Optimization of Multiuser Multiple Input Multiple Output Communication Systems
AUTHOR:	Peter He
SUPERVISOR:	Dr. Tim Davidson
NUMBER OF PAGES:	1, 99

Abstract

This thesis considers the Shannon capacity of multiuser multiple input multiple output (MIMO) wireless communication systems. That is, the fundamental limit on the rates at which data can be reliably communicated. The focus is on scenarios in which the channel has long coherence times and perfect channel state information is available to both transmitters and receivers. The thesis considers two important design problems in multiuser MIMO wireless communication systems: the design of the sum-rate optimal input distribution for the MIMO multiple access channel (MIMO MAC), and the design of the sum-rate optimal input distribution for the MIMO broadcast channel (MIMO BC).

The thesis considers algorithms for solving these design problems that are based on the principle of iterative water-filling. The contributions of the thesis are twofold. First, a correct and rigorous proof of convergence of the family of water-filling algorithms is derived. This proof overcomes weaknesses in the previous attempts of others to prove convergence. Second, an efficient algorithm is presented for the water-filling procedure that lies at the heart of the iterative water-filling algorithm. This algorithm will open the door for further efficient utilization of the iterative water-filling algorithm. This novel algorithm is based on the principle of Fibonacci search, and

since the iterative water-filling algorithm involves repeated water-filling procedures, the impact of this efficient algorithm is magnified.

The outcomes of this research are that the iterative water-filling algorithms are mathematically validated for the above-mentioned design problems in multiuser MIMO wireless communication systems, and that the implementation of these algorithms is made more efficient through the application of the efficient Fibonacci search method for the underlying water-filling procedure.

Acknowledgements

The thesis is written under the guidance and with the help of my supervisor, Dr. Tim Davidson, whose valuable advices and extended knowledge help me all along. Not only is he an expert on signal processing, communications and control, but he also a specialist on optimization. I learn from him the optimization theory and methods, and how they combine with and apply to signal processing, communications and control. I would like to express my sincere gratitude to my supervisor Dr. Tim Davidson. I am also grateful to my friends at the Advanced Optimization Lab and teachers and classmates at the School of Computational Engineering and Science. Their company makes my two years at McMaster very enjoyable. I appreciate the great aid and support from all the members of the Advanced Optimization Laboratory and the McMaster School of Computational Engineering and Science.

My special thanks go to the members of the examination committee: Professor Christopher Anand, Professor Jun Chen and Professor Tim Davidson.

Finally, I am indebted to thank my family for their patience, understanding and continuous support.

Notation and Acronyms

Notation:

$A(\cdot)$	Mapping
$\mathbb{C}(H,P)$	Channel capacity
$\mathbb{C}^n$	Unitary space with dimension n
$\mathbb{C}^{m\times n}$	The set of $m \times n$ complex matrices
$\det(\cdot)$	Determinant operation
$E[\cdot]$	Expected value
f_ξ	The probability density function of random variable ξ
H	Channel matrix
$H^\dagger$	The conjugate transpose of matrix H
H_i	The channel matrix of the i-th user
$H_i^\dagger$	The conjugate transpose of matrix H_i
H_i^T	The transpose of matrix H_i
$\mathbb{H}(\xi)$	Differential entropy
$\mathbb{H}(\xi\|\eta)$	Conditional entropy

Notation:

I_r	Identity matirx with demension r
$\mathbb{I}(\xi;\eta)$	Mutual imformation
log	Natural logarithm
$M \succeq 0$	Matrix M being positive semidefinite
$\mathbb{N}$	The set of natural numbers
n_s	Dimension of vector valued variable for the generalized mathematical models
P	The upper bound of signal power
P_i	The upper bound of signal power for the i-th user
$P_{\widehat{\xi}}$	The probability distribution function of the random variable (or vector) $\widehat{\xi}$
	Similarly understanding the others
$\mathbb{P}_i$	The (optimization) operator or mapping over the i-th coordinate block
Q	Covariance matrix
$\mathbb{R}^n$	Euclidean space with demension n
$\mathbb{R}^{m\times n}$	The set of $m \times n$ real matrices
S	Covariance matrix
$\mathrm{Tr}(M)$	The trace of matrix M
x	The input vector of the channel or a vector
$x^\dagger$	The conjugate transpose of vector x
x_i	The i-th entry (scalar) or the i-th block of vector x
x^i	The input vector of the i-th user
$(x)_i$	the i-th entry (scalar) of vector x

Notation:

$x_i^\dagger$	The conjugate transpose of vector x_i
y	The output vector
y_{dmac}	The output vector of the dual uplink channel
$y^\dagger$	The conjugate transpose of vector y
z	Gaussian noise vector
Z	Generalized objective function
μ	Expected value
ξ	Random noise
$\lfloor\ \rfloor$	Floor function
$\lceil\ \rceil$	Ceiling function
Σ_i	Covariance matrix
$\tilde{\Sigma}_i$	Covariance matrix
$\Longleftrightarrow$	Optimality equivalence

Acronyms:

BC	Broadcast channel
BCAA	Block Coordinate Ascent Algorithm
DBCAA	Diagonal Block Coordinate Ascent Algorithm
IWFA	Iterative water-filling algorithm
IWFAwFIS	Iterative water-filling algorithm with Fibonacci search
KKT	Karush-Kuhn-Tucker
MAC	Multiple access channel
MIMO	Multiple input multiple output
WFA	Water-filling algorithm

Contents

Chapter 1

Introduction

The use of multiple antennas at the transmitter and receiver, i.e., Multiple-Input Multiple-Output (MIMO) technology, constitutes a breakthrough [4, page 1] in the design of wireless communication systems, and MIMO technology is now at the core of several existing and emerging wireless standards [4, page 18]. Exploiting multipath scattering, MIMO techniques have delivered significant performance enhancements in terms of data transmission rate and interference reduction on point-to-point links. In this thesis, we focus our attention on multiuser MIMO systems. In particular, we consider the design of multiuser systems so as to enable operation at rates approaching the fundamental limits.

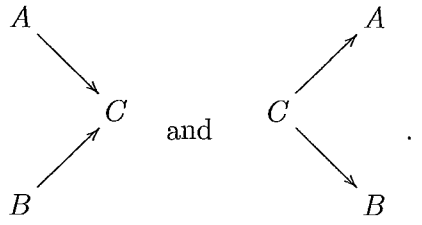
We consider two multiuser MIMO systems:

(i) the MIMO multiple access channel (MAC) in which a number of users wish to send messages to a single destination, and

(ii) the MIMO broadcast channel (BC) in which a single source wishes to send

independent messages to different destinations.

These two classes are illustrated for the case of two users in the following diagram.



2-user MAC 2-user BC

The MAC model is used in the study of the cellular uplink (from user to base-station) and the BC model is used to study the cellular downlink.

In the case of the MAC and BC channels, the fundamental limit on the rates at which data can be reliably communicated is the capacity region. That is, the region of rate vectors for which there exists a coding strategy such that the probability of error goes to zero as the block length of the code increases. In this thesis, we will focus on one point on the boundary of the region, namely the maximum sum-rate point. That is, the point at which the sum of the rates is maximized. More specially, we will consider scenarios in which the channel coherence times are long, precise channel state information is available to both the transmitter and receivers, and the additive noise at the receivers is Gaussian. For this scenario, a popular algorithm for designing input covariance(s) that maximize the sum rate is the iterative water-filling algorithm [42, 20]. In this thesis, two important contribution will be made to the generic iterative water-filling algorithm. The first contribution is to point out that there are weaknesses in the existing attempts to prove that the water-filling algorithm

converges. We also show that by taking a somewhat different approach, a rigorous proof of convergence of the algorithm can be obtained. We develop that proof for the MAC and BC models. The second contribution is the development of an efficient algorithm for the water-filling procedure that is performed in each steps of the iterative water-filling algorithm (IWFA). This algorithm is based on the Fibonacci search technique and reduces the complexity of the water-filling step from $O(n)$ to $O(\log(n))$, where n is the number of columns in the channel matrix. Since this procedure must be implemented during each iteration of the IWFA, this complexity reduction can have a significant impact in practice.

As an aside, we point out that iterative water-filling forms the basis of a recent patent application [18] for power control in wireless communication systems.

1.1 Structure of the Thesis

The rest of the thesis has the following structure.

In Chapter 2, the single-user MIMO wireless communication channel is considered. For multiuser systems, the iterative water-filling procedure involves considering one user at a time and treating the signals from the other users as noise. As such, the single-user MIMO model lies at the core of the iterative water-filling algorithms. Therefore, this is a natural framework in which to introduce the proposed efficient water-filling procedure that is based on the Fibonacci search. The optimality of the obtained solution to the problem is proved and reduction in the computation cost for the problem is evaluated.

In Chapter 3, the multiuser MIMO multiple access channel (MAC) is considered. The sum rate optimization problem is formulated, and we demonstrate how the proposed Fibonacci search can be applied in this case. Then we point out the weaknesses in the existing attempts to prove that the iterative water-filling algorithm converges, and subsequently we develop a somewhat different proof that rigorously establishes convergence of the algorithm.

In Chapter 4, the multiuser MIMO broadcast channel (BC) is considered, along with its dual MIMO MAC model. Due to the equivalence between the sum rate of the MIMO BC and that of the dual MIMO MAC, and because the latter can be much more easily solved than the former, the iterative water-filling algorithm with Fibonacci search for optimizing the sum rate of the dual MIMO MAC is presented. Then we point out the weaknesses in the existing attempts to prove that the iterative water-filling algorithm converges. Finally, in Section 4.4, a rigorous proof of convergence of the algorithm is provided.

In Chapter 5, we summarize our results. We may conclude that, under the assumption that the channel matrix is constant and known, the iterative water-filling algorithms are convergent and that the proposed Fibonacci search procedure reduces their computational cost.

Chapter 2

Single-User MIMO Channel

In this chapter, we consider the single-user MIMO system, and present an efficient implementation of the water-filling algorithm for optimizing the input distribution that is based on Fibonacci search method. The application of the Fibonacci search in this context is new, and offers a substantial reduction in the computational cost compared to that of the often employed approach for the WFA. Indeed, the reduction of this component of the algorithm is by a factor of about $\log(n)/n$, where n denotes the number for the columns of the channel gain matrix. The discussion of the single-user MIMO system also serves as the preparation for that of the multiple-user MIMO systems in Chapters 3 and 4.

2.1 Model of the Single-User MIMO Channel

A single-user Gaussian channel with multiple transmitting and/or receiving antennas is considered as follows. We denote the number of transmitting antennas by t and the number of receiving antennas by r. We restrict our discussion to a linear model

in which the received vector $y \in \mathbb{C}^r$ depends on the transmitted vector $x \in \mathbb{C}^t$ via

$$y = Hx + z, \tag{2.1}$$

where H is an $r \times t$ complex channel gain matrix and $z \in \mathbb{C}^r$. We assume that vector z is a zero mean circular complex Gaussian noise vector; cf. [27, 26, 35] and Appendix-I. For any complex matrix or vector, its superscript $\dagger$ denotes the conjugate transpose of the matrix or the conjugate transpose of the vector. Without loss of generality, we assume $E\left[zz^\dagger\right] = I_r$, where I_r is an identity matrix with order r and $E[\cdot]$ denotes the expectation operation. That is, the noises corrupting the different receivers are independent. The average power of the transmitter is bounded by P, i.e.,

$$E\left[x^\dagger x\right] \leq P.$$

Equivalently,

$$\mathrm{Tr}\left(E\left[xx^\dagger\right]\right) \leq P. \tag{2.2}$$

This second form of the power constraint will prove to be more useful than the first form in the upcoming discussions.

In a generic wireless communications set up, there are three scenarios [35] for the matrix H:

1. H is deterministic,

2. H is a random matrix, which is chosen according to a probability distribution, and

3. H is a random matrix, but is fixed once it is chosen.

The focus of this thesis is on the first of these cases. This case is normally referred to as the static channel case with full channel state information at transmitter and receiver sides; e.g., [4] [17]. In this chapter, we will develop a water-filling algorithm with Fibonacci search for this system. This algorithm will also be utilized in the subsequent chapters.

2.2 Channel Capacity of the Single-User MIMO Channel

In this section we will discuss the channel capacity of the single-user MIMO system in (2.1), with perfect channel state information at the transmitter (see, e.g., [35]).

Given the model in (2.1), the channel capacity C is defined as $C \triangleq \max_{p_x} \mathcal{I}(x;y)$, where $\mathcal{I}(x;y)$ is the mutual information between x and y, and p_x is the probability density function of x. Let $S \triangleq E\left[xx^\dagger\right]$. As shown by (2.2), the input power constraint can be written as that $\mathrm{Tr}(S) \leq P$. The corresponding channel capacity $C(H,P)$ is expressed as:

$$C(H,P) = \max_{p_x} \left\{ \mathcal{I}(x;y) \,|\, S \succeq 0, \mathrm{Tr}(S) \leq P \right\}. \tag{2.3}$$

Using the argument in Appendix-II, it can be shown that for the model in (2.1) in which the channel H is deterministic, and the additive noise is Gaussian and, without loss of generality, has a unit variance, the optimal input distribution for the input x is zero-mean and Gaussian and hence the mutual information can be written as $\log\left(\det\left(I_r + HSH^\dagger\right)\right)$. Since a zero-mean Gaussian distribution is completely

specified by its covariance, the expression in (2.3) can be simplified to

$$C(H,P) = \max_{S} \left\{ \log\left(\det\left(I_r + HSH^\dagger\right)\right) | S \succeq 0, \mathrm{Tr}(S) \leq P \right\}. \tag{2.4}$$

The expression on the right hand side of (2.4) is a nonlinear semidefinite optimization problem. A direct and effective algorithm is presented in the following section.

2.3 Simplification of the Problem of the Optimal Input Covariance

To efficiently solve the problem in (2.4) of the optimal input covariance for the single-user MIMO channel, we first simplify the problem; e.g., [35]. Using the singular value decomposition (SVD) of the H, $H = U\Sigma V^\dagger$ and the properties of the determinant, we have that

$$\log\det\left(I_r + HSH^\dagger\right) = \log\det\left(I_r + \Sigma\widehat{S}\Sigma^\dagger\right),$$

where $\widehat{S} = V^\dagger SV$. Using Hadamard's Determinant inequality, it can be shown that we can restrict attention to diagonal $\widehat{S}$, say Γ, and any optimal input covariance takes the form $S = V\Gamma V^\dagger$, where $\Gamma = \mathrm{Diag}(\gamma_i^*)$ and

$$\{\gamma_i^*\}_{i=1}^t = \arg\max_{\{\gamma_i\}_{i=1}^t} \left\{ \sum_{i=1}^t \log(1+\lambda_i\gamma_i) \,|\, \gamma_i \geq 0, \forall i; \sum_{i=1}^t \gamma_i \leq P \right\}, \tag{2.5}$$

where γ_i is the i-th element of $\Sigma^\dagger\Sigma$. The remaining challenge is to obtain an efficient algorithm for solving (2.5). The problem in (2.5) can be solved using a generic water-filling procedure. However, the generic procedure involves enumeration over

the number of the diagonal elements of Γ that are to be non zero. The following algorithm uses a Fibonacci search to determine this number of active subchannels, and hence is significantly more computationally efficient that the enumerative algorithm.

Since the water-filling procedure is repeated many times in iterative water-filling algorithms, the reduction in complexity can have a significant impact in practice. As far as we are aware, the following algorithm is the first time in which Fibonacci search has been used in water-filling.

The water-filling algorithm with Fibonacci search is stated as follows:

Algorithm: Water-Filling Algorithm with Fibonacci Search

Step 1: Pre-Processing. Compute the unitary matrix $U \in \mathbb{C}^{t\times t}$ by the SVD:

$$\Lambda = U^{\dagger}H^{\dagger}HU = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right).$$

Let $\{\lambda_i\}_{i=1}^{t}$ be ordered in the monotonically decreasing order; Let

$$\widehat{i} \triangleq \max\left\{i|\lambda_i > 0\right\} \leq \min\left\{t, r\right\}.$$

Step 2: Water-Filling with Fibonacci Search for (2.5). For $k = 1, 2, \ldots, \widehat{i}$, let

$$S_k \triangleq \frac{1}{k}\left\{P - \left[(k-1)\frac{1}{\lambda_k} - \sum_{i\in\{1,\cdots,k-1\}\cap\{k\geq 2\}} \frac{1}{\lambda_i}\right]\right\}.$$

Now search for

$$k^* = \max\left\{k|S_k > 0, 1 \leq k \leq \widehat{i}\right\}, \tag{2.6}$$

using the Fibonacci search method with approximation ratios $\frac{1}{3}$ and $\frac{2}{3}$. More specifically,

1st Step. Assume that $a = 1$ and $b = \widehat{i}$.

2nd Step. If $a = b$, then $k^* = a$ and go to **Step 3**.

Else, $a_1 = \lfloor a + \frac{1}{3}(b-a) \rfloor$, $b_1 = \lceil a + \frac{2}{3}(b-a) \rceil$.

3rd Step. If $S_{a_1} \leq 0$, then $b = a_1 - 1$ and go to the **2nd Step**;

If $S_{b_1} > 0$, then $a = b_1$ and go to the **2nd Step**;

If $S_{a_1} > 0$ and $S_{b_1} \leq 0$, then $a = a_1$, $b = b_1 - 1$ and go to the **2nd Step**.

Step 3: Finding Optimal Solution to (2.5).

- Compute $S^* \in \mathbb{C}^{t \times t}$ as follows:

$$\begin{aligned} &S^*_{ii} = \tfrac{1}{\lambda_{k^*}} - \tfrac{1}{\lambda_i} + S_{k^*}, 1 \leq i \leq k^*; \\ &S^*_{ii} = 0, k^* < i \leq t; \\ &S^*_{ij} = 0, i \neq j. \end{aligned} \tag{2.7}$$

- Compute $US^*U^\dagger$, as the optimal solution to the model (2.4). $US^*U^\dagger$ is proved to be the optimal solution in Section 1.5.

Remark 2.3.1. *In Step 2 of the above water-filling algorithm, the Fibonacci search method is used to find k^*. This can reduce the computational cost of computing*

$$\max\left\{k | S_k > 0, 1 \leq k \leq \widehat{i}\right\},$$

compared with the regular searching method of enumeration. Indeed, the ratio of the computation burden from the Fibonacci search to that from the enumeration method

is about $\log(t)/t$. *The impact of this reduction is amplified by the fact in iterative water-filling schemes, where the water-filling procedure is repeated many times. As anecdote, we point out that the water level obtained in the water-filling procedure is* $\frac{1}{\lambda_{k^*}} + S_{k^*}$.

2.4 Optimality and Complexity

For the channel capacity problem of the single-user MIMO channel, the proof of optimality for $US^*U^\dagger$ found by the water-filling algorithm with Fibonacci search is presented in this section.

Theorem 2.4.1. *Let* $H = U\Sigma V^\dagger$ *denote the singular value decomposition of the* H. *Further let* k^* *and* S^* *be given as in Step 2 and 3 of the Water-Filling Algorithm with Fibonacci search; see (2.6) and (2.7). Then* $US^*U^\dagger$ *is an optimal input covariance for the problem in (2.4).*

Before proving this new theorem, we present some lemmas and introduce a remark.

Lemma 2.4.2. *For the channel* H, *there is a unitary matrix* U *such that* $U^\dagger H^\dagger H U = diag(\lambda_1, \cdots, \lambda_t)$ *(a diagonal matrix) and*

$$\max\left\{\log\left(\det\left(I_r + HSH^\dagger\right)\right) | S \succeq 0, Tr(S) \leq P\right\} =$$

$$\max\left\{\log\left(\det\left(I_t + diag(\lambda_1, \cdots, \lambda_t)\, S\right)\right) | S \succeq 0, Tr(S) \leq P\right\}$$

and $U^\dagger S_l U = S_r$, *where* S_l *and* S_r *are two optimal solutions of the two optimization problems mentioned above, respectively.*

Lemma 2.4.3. *For the channel H, there is a unitary matrix U such that $U^\dagger H^\dagger H U = \Lambda$ (a diagonal matrix) and*

$$\max\left\{\log\left(\det\left(I_r + HSH^\dagger\right)\right) | S \succeq 0, Tr(S) \leq P\right\}$$
$$= \max\left\{\log\left(\det\left(I_t + \Lambda^{\frac{1}{2}} S \Lambda^{\frac{1}{2}}\right)\right) | S \succeq 0, Tr(S) \leq P\right\},$$

and $U^\dagger S_l U = S_r$, where S_l and S_r are two optimal solutions of the two optimization problems mentioned above, respectively.

The proofs of both these lemmas are provided in Appendix III.

For simplification of the optimization model, which is

$$\max\left\{\log\left(\det\left(I_t + \Lambda^{\frac{1}{2}} S \Lambda^{\frac{1}{2}}\right)\right) | S \succeq 0, \mathrm{Tr}\,(S) \leq P\right\},$$

the Hadamard Determinantal Inequality is introduced.

The Hadamard Determinantal Inequality [25] is: if $A = (a_{ij})_{n\times n}$ is a real (or Hermitian) positive semidefinite matrix, then

$$\det(A) \leq a_{11} \cdots a_{nn}.$$

Then we present our proof for the problem (2.5) as follows.

Proof of Theorem 2.4.1. When pre-processed, as the first step of the water-filling algorithm,

$$\max\left\{\log\left(\prod_{i=1}^{t}(1+\lambda_i S_{ii})\right) | S_{ii} \geq 0, \forall i, \sum_{i=1}^{t} S_{ii} \leq P\right\}$$

is equivalent to

$$\max\left\{\log\left(\prod_{i=1}^{\widehat{i}}(1+\lambda_i S_{ii})\right)\Big|S_{ii}\geq 0,\forall i,\sum_{i=1}^{\widehat{i}}S_{ii}=P\right\},$$

under the meaning of the equivalence for two optimization models.

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_{\widehat{i}}\ \text{ and }\ \frac{1}{\lambda_1}\leq\frac{1}{\lambda_2}\leq\cdots\leq\frac{1}{\lambda_{\widehat{i}}}.$$

Formulation

$$\max_{\{S_{ii}\}}\left\{\log\left(\prod_{i=1}^{\widehat{i}}(1+\lambda_i S_{ii})\right)\Big|S_{ii}\geq 0,\forall i,\sum_{i=1}^{\widehat{i}}S_{ii}=P\right\},\tag{2.8}$$

is equivalent to (2.4) even although it has an equality constraint and the linear inequality constraints.

Below, the set $\left\{S_{ii}^*, 1\leq i\leq\widehat{i}\right\}$ is proved to be the optimal solution to the problem in (2.8).

The Lagrangian function of the problem in (2.8) is

$$L\left(\{S_{ii}\}_{i=1}^{\widehat{i}},\mu,\{\sigma_i\}_{i=1}^{\widehat{i}}\right)=-\sum_{i=1}^{\widehat{i}}\log\left(\frac{1}{\lambda_i}+S_{ii}\right)-\mu\left(P-\sum_{i=1}^{\widehat{i}}S_{ii}\right)-\sum_{i=1}^{\widehat{i}}\sigma_i S_{ii},$$

where μ and $\{\sigma_i\}_{i=1}^{\widehat{i}}$ are the Lagrange multipliers. Therefore, the **KKT** conditions

of the problem in (2.8) are

$$\begin{cases} \frac{1}{\frac{1}{\lambda_i}+S_{ii}} - \mu + \sigma_i = 0, 1 \leq i \leq \widehat{i} \\ S_{ii} \geq 0,\ \sigma_i S_{ii} = 0,\ \sigma_i \geq 0,\ \forall i \\ \sum_{i=1}^{\widehat{i}} S_{ii} = P,\ \mu \in \mathbb{R}. \end{cases}$$

Now we choose $S_{ii} = S^*_{ii}, 1 \leq i \leq \widehat{i}$. Therefore,

$$\frac{1}{\frac{1}{\lambda_1} + S^*_{11}} = \cdots = \frac{1}{\frac{1}{\lambda_{k^*}} + S^*_{k^*k^*}}.$$

Now also choose $\mu = \frac{1}{\frac{1}{\lambda_1}+S^*_{11}}$ and choose $\sigma_i = 0, i = 1, \ldots, k^*$. By simple substitution, it can be shown that these values solve the **KKT** conditions.

To show that this point is optimal, we now argue that, for the problem in (2.8), the **KKT** conditions are sufficient as well as very necessary for optimality. First the Hessian matrix of the objective, $\log\left(\prod_{i=1}^{\widehat{i}}(1+\lambda_i S_{ii})\right)$, is clearly negative definite and hence the objective is concave. Furthermore, the constraints are linear and hence the feasible set is convex.

Therefore, the problem in (2.8) is a convex optimization problem. Since only there are the linear constraints for the problem in (2.8), the constraint qualification is satisfied and the **KKT** conditions are both necessary and sufficient.

Since

$$\max\left\{\log\left(\prod_{i=1}^{t}(1+\lambda_i S_{ii})\right) \middle| S_{ii} \geq 0, \forall i, \sum_{i=1}^{t} S_{ii} \leq P\right\}$$

is equivalent to

$$\max\left\{\log\left(\prod_{i=1}^{\widehat{i}}(1+\lambda_i S_{ii})\right)\,|S_{ii}\geq 0,\forall i,\sum_{i=1}^{\widehat{i}}S_{ii}=P\right\},$$

then $US^*U^\dagger$ is an optimal solution of the model (2.4).

Q.E.D.

Remark 2.4.4. *Although there are other proofs available (e.g. [41]), the above proof was obtained independently and is somewhat simpler than that in [41].*

Chapter 3

Sum Capacity of the MIMO MAC

In this chapter, we consider a multiuser multi-input multi-output (MIMO) system. In particular, we consider the MIMO multiple access channel (MAC). We consider the important problem of finding the set of input covariances that maximizes the sum rate of the MIMO MAC, and we develop an efficient algorithm for solving this problem.

In the first section of this chapter, we provide a general mathematical model for the MIMO MAC, and we state the sum capacity of that model. Then an iterative water-filling algorithm based on a Fibonacci search is introduced. Subsequently, we point out the weaknesses in the existing attempts to prove that the IWFA converges for the MIMO MAC. Finally, a rigorous proof of convergence of the iterative water-filling algorithm is derived.

3.1 Models for the MIMO MAC and Its Sum Capacity

We consider the MIMO MAC illustrated on the left hand side of the figure on Page 2, in which the base-station has m antennas and there are K mobile stations, each of which has n antennas. When the information flows from the mobile stations to the base-station, we have a MAC channel or uplink channel.

With appropriate synchronization, the MAC can be described as [17]

$$y_{\text{mac}} = \sum_{i=1}^{K} H_i^{\dagger} x^i + z,$$

where y_{mac} is the signal received at the base-station, and $H_i^{\dagger} \in \mathbb{C}^{m \times n}, i = 1, 2, \cdots, K$, denotes the matrix of channel gains from each antenna at the i-th mobile station to each antenna at the base-station. Without loss of generality, we will assume that $H_i \neq 0$, $\forall i$. The x^i's are $n \times 1$ complex input vectors, and $z \in \mathbb{C}^m$ is an additive Gaussian noise vector and, without loss of generality, has identity covariance. We will let $S_i \triangleq E\left[x^i \left(x^i\right)^{\dagger}\right]$ denote the covariance matrix of x^i. For the convenience of later discussion and without loss of generality, the MAC is described as $y_{\text{mac}} = \sum_{i=1}^{K} H_i^{\dagger} x^i + z$ instead of $y_{\text{mac}} = \sum_{i=1}^{K} H_i x^i + z$ [17]. Notation-wise, in this chapter, we will use different superscripts of vectors to denote vectors corresponding to different users and different subscripts of a vector to denote different entries of a vector.

The sum capacity of the MIMO MAC is the fundamental limit on the sum of the rates at which reliable communication can be achieved. When the channels are known

and deterministic and the noise is Gaussian, the optimal input distribution at each mobile station is a zero-mean Gaussian random vector with covariance S_i. Therefore, the problem maximizing the sum rate can be written as (see Appendix II and, e.g., [4])

$$\max_{\{S_i\}_{i=1}^K} \left\{ \log \left(\det \left(I + \sum_{i=1}^K H_i^\dagger S_i H_i \right) \right) \middle| S_i \succeq 0, \mathrm{Tr}\left(S_i\right) \leq P_i, \forall i \right\}, \tag{3.1}$$

where $S_i \succeq 0$ denotes that S_i is a Hermitian positive semidefinite matrix with complex entries.

Remark 3.1.1. *Since it allows for complex representations, the expression (3.1) is slightly more general than the real valued expressions used in [42], and better matches the way in which communication systems are constructed in practice.*

3.2 Iterative Water-Filling Algorithm with Fibonacci Search under Individual Power Constraints

The iterative water-filling algorithm [42] is an algorithm for finding the input covariances that maximize the sum rate of a MIMO MAC. The principle behind the algorithm is to iteratively select each user and optimize that user's input distribution using a single user water-filling algorithm in which the interference from other users is treated as noise. This procedure is continued until a convergence criterion is met. The basic algorithm can be formulated as follows.

Algorithm 3.1: **Iterative water-filling algorithm (IWFA)**

Input: precision constant $0 < \varepsilon < 1$, counter $n = 0$, variable $n_{max}, J = 0, J_1 = \infty$; channel matrices H_i, power constraints P_i, covariance matrices $S_i = 0, \forall i$.

Begin

While $|J_1 - J| > \varepsilon \times J_1$ **and** $n \leq n_{max}$ **Do**

$J = J_1$.

For $i = 1$ to K

Compute $G_i = I_m + \sum_{j=1, j\neq i}^{K} H_j^{\dagger} S_j H_j$,

Compute $G_i = H_i \left(G_i\right)^{-\frac{1}{2}}$,

Compute S_i, the optimal input covariance for the single user channel G_i with power constraint P_i using a single user water-filling technique.

End

Compute $J_1 = \log\left(\det\left(I_m + \sum_{i=1}^{K} H_i^{\dagger} S_i H_i\right)\right)$.

$n = n + 1$.

End

End

The water-filling algorithm with Fibonacci search described in Section 2.3 can be used in the single-user water-filling step of the above algorithm, as we now show.

Algorithm 3.2: **Iterative water-filling algorithm with Fibonacci search (IWFAwFIS)**

Input: precision constant $0 < \varepsilon < 1$, counter $n = 0$, variable $J = 0, J_1 = 1$; channel matrices H_i, power constraints P_i, covariance matrices $S_i = 0, \forall i$.

Begin

While $|J_1 - J| > \varepsilon \times J_1$ **Do**

If $n \neq 0$, then $J = J_1$.

For $i = 1$ to K

Compute $G_i = I_m + \sum_{j=1, j\neq i}^{K} H_j^{\dagger} S_j H_j$,

Compute $G_i = H_i (G_i)^{-\frac{1}{2}}$,

Compute S_i, the optimal input covariance for the single user channel G_i with power constraint P_i using the single user Water-Filling Algorithm with Fibonacci Search in Section 2.3.

End

Let $J_1 = \log\left(\det\left(I_m + \sum_{i=1}^{K} H_i^{\dagger} S_i H_i\right)\right)$.

$n = n + 1$.

End

End

Remark 3.2.1. *It is seen from Section 2.3 that the purpose of Step 3 of Algorithms 3.1 and 3.2 is equivalent to finding the optimal input covariance of a modified single-user MIMO channel. Algorithm 3.1 is defined by a description method from the formation of its point sequence. This descriptive way to define an algorithm is popular among the engineering fields. An alternative method to define an algorithm is to regard the algorithm as mapping (see [44], pp. 83). We will use these two methods interchangeably in this chapter. The former is used when the convenience of implementing the detailed algorithm in computer programs is emphasized; the latter is more suitable for rigorous analysis of the algorithm.*

3.3 Weaknesses of the Existing Research Regarding Convergence of the IWFA on Multiuser MIMO MAC

Many recent advances on the signalling schemes for the Gaussian Multi-Input Multi-Output (MIMO) Multiple Access Channel (MAC) are based on the important paper [42]. In the following sections, we first point out the a weakness of the proof of convergence of the iterative water-filling algorithm for the Gaussian MAC proposed in [42].

In the first subsection below, the iterative water-filling algorithm and the convergence theorem of the capacity of the MIMO MAC with the individual power constraint are stated. Then we discuss in detail the weakness of the proof of convergence of the algorithm (see Theorem 2 of [42] and its proof therein).

3.3.1 The Algorithm

In [42], the iterative water-filling algorithm is used to compute the optimal input distributions that maximize the sum rate of a Gaussian MAC with vector inputs, a vector output, and real-valued matrices. Notation-wise, in this section, we will follow [42] and use the symbol $|\cdot|$ of [42] to denote the determinant.

As suggested by (3.1), the sum rate problem for a Gaussian MAC with real channel

matrices can be written as

$$\begin{array}{lll} \text{maximize} & \frac{1}{2}\log|\sum_{i=1}^{K} H_i S_i H_i^T + S_z| - \frac{1}{2}\log|S_z| & \\ \text{subject to} & \text{tr}(S_i) \leq P_i & i = 1, \ldots, K \cdot \\ & S_i \succeq 0, & i = 1, \ldots, K \end{array}$$

Notation-wise, the symbol T of H_i^T mentioned above means the transposition operation of the matrix.

The iterative water-filling algorithm with the individual power constraints [42] is essentially the same as that in Algorithm 3.1 and can be stated as follows. Notation-wise, we will follow the convention in [42] and label the iterative water-filling algorithm with the individual power constraints as Algorithm 3.3.

Algorithm 3.3: Iterative water-filling algorithm in [42]

Initialization $S_i = 0,\ i = 1, \ldots, K.$

repeat

for $i = 1$ to K

$S_z' = \sum_{j=1, j\neq i}^{K} H_j S_j H_j^T + S_z;$

$S_i = \arg\max_S \frac{1}{2}\log|H_i S H_i^T + S_z'|;$

end

until the sum rate converges.

The sum rate converges is understood here as the difference between the current sum rate and the previous sum rate satisfies the permitted computational error.

3.3.2 Theorems Regarding the Iterative Water-filling Algorithm

A significant part of [42] deals with convergence of the algorithm. Theorem 1 and Theorem 2 of [42] attempt to present convergence of the algorithm and they are quoted as follows.

Theorem 1 in [42]. "*In a $K-$user multiple-access channel, $\{S_i\}$ is an optimal solution to the rate-sum maximization problem*

$$\begin{array}{lll} \text{maximize} & \frac{1}{2}\log|\sum_{i=1}^{K} H_i S_i H_i^T + S_z| - \frac{1}{2}\log|S_z| & \\ \text{subject to} & \text{tr}\,(S_i) \leq P_i & i = 1, \ldots, K \\ & S_i \succeq 0, & i = 1, \ldots, K \end{array}$$

if and only if S_i is the single-user water-filling covariance matrix of the channel H_i, with $S_z + \sum_{j=1, j\neq i}^{K} H_j S_j H_j^T$ as noise, for all $i = 1, 2, ..., K$."

Theorem 2 in [42]. "*In the iterative water-filling algorithm, the sum rate converges to the sum capacity, and $\{S_1, S_2, ..., S_K\}$ converges to an optimal set of input covariance matrices for the Gaussian vector multiple-access channel.*"

The proof of Theorem 2 in [42] is quoted below with the important passages italicized.

"At each step, the iterative water-filling algorithm finds the single-user water-filling covariance matrix for each user while regarding all other users signals as additional noise. Since the single-user rate objective differs from

the multiuser rate-sum objective by only a constant, the rate-sum objective is nondecreasing with each water-filling step. The rate-sum objective is bounded above, so the sum rate converges to a limit.

The convergence matrices $S_1, ..., S_K$ also converge to a limit [sentence 1]. *Because the single-user water-filling matrix is unique, each water-filling step in the iterative algorithm must either yield a strict increase of the sum rate or keep the covariance matrices the same* [sentence 2]. *At the limit, all S_is are simultaneously the single-user water-filling covariance matrices of user i with all other users signals regarded as additional noise* [sentence 3]. Then, by Theorem 1, this set of $(S_1, ..., S_K)$ must achieve the sum capacity of the Gaussian vector multiple-access channel."

3.3.3 Regarding the Proof in [42]

The proof of Theorem 2 in [42] has embedded a circular reasoning. This is stated as follows.

As quoted above, the proof of Theorem 2 in [42] that was proposed in that paper utilizes an argument (see sentence 3) that at the limit, all S_is are simultaneously the single-user water-filling covariance matrices of user i with all other users' signals regarded as additional noises, in order to arrive at the conclusion of Theorem 2. The conclusion of Theorem 2 is that in the iterative water-filling algorithm, the sum rate converges to the sum capacity, and the set $\{S_1, S_2, ..., S_K\}$ converges to an optimal set of input covariance matrices for the Gaussian vector multiple-access channel.

According to the conclusion of Theorem 1 in [42] that S_i is an optimal solution to the rate-sum maximization problem iff S_i is the single-user water-filling covariance matrices of the channel H_i, with $S_z + \sum_{j=1, j\neq i}^{K} H_j S_j H_j^T$ as noise, for all $i = 1, 2, ..., K$, and, at the same time, according to the argument quoted in the proof of Theorem 2 that, at the limit, all the S_is are simultaneously the single-user water-filling covariance matrices of user i with all other users' signals regarded as additional noises, so the argument quoted in the proof of Theorem 2 is equivalent to the statement that, $\{S_i\}_{i=1}^{K}$, as the limit point, is an optimal solution to the rate-sum maximization problem. Thus, the quoted argument is equivalent to the statement that, in the iterative water-filling algorithm, the $\{S_1, S_2, ..., S_K\}$ converges to an optimal set of input covariance matrices for the Gaussian vector multiple-access channel, and the sum rate converges to the sum capacity.

If the quoted argument were proved, the proof of Theorem 2 would be correct. However, the quoted argument for proving Theorem 2 is also the conclusion of Theorem 2, i.e., what [42] intends to prove has been utilized for proving what [42] intends to prove. In addition, the arguments used in the proof of Theorem 2 in [42] do not guarantee that the set $\{S_1, S_2, ..., S_K\}$ converges. The preceding observations have highlighted the need for and have paved the way for our subsequent developments.

3.4 Convergence of the Algorithms

3.4.1 Fixed Point Theory, Continuity and Convergence of the Algorithms

In this section, we will develop a rigorous proof of convergence of the iterative water-filling algorithm. The proof is based on the interpretation of the iterative water-filling algorithm as a Block Coordinate Ascent Algorithm (BCAA), and on the representation of BCAA algorithms as a mapping. Although an alternative proof can be constructed from arguments in [2], the proof here explicitly exposes the relationship between the fixed point and the optimal point. As a result, a direct link is established between the KKT conditions and the variational inequality of convex optimization. This allows further algorithm developments under a unified mathematical framework. Furthermore, the proof includes a proof of the continuity of the optimization operator BCAA, which, in itself, is a contribution to optimization theory.

We begin by observing that the problem in (3.1) can be written in the following more general form:

$$\max_{x} \{f(x)|x \in V\}, \text{ where } V = \otimes_{i=1}^{K} V_i, \quad V_i \subset \mathbb{R}^{n_i}, \tag{3.2}$$

where $\sum_{i=1}^{K} n_i = n_s, n_i \in \mathbb{N}$, $V \subset \mathbb{R}^{n_s}$ is convex and closed and V is a Cartesian product. Furthermore, $f(x_1, x_2, \ldots, x_K)$ is concave and differentiable, where $x_i \in V_i, \forall i$. As we will show in Proposition 3.4.1 below, the problem in (3.1) is a special case of that in (3.2), but we will find the abstract form in (3.2) to be more convenient in the development of the proof.

Proposition 3.4.1. *The problem in (3.1) is a special case of the problem in (3.2).*

Proof. If matrix $S_i, \forall i$, in the problem in (3.1) is decomposed into its real part and imaginary part matrices, i.e., $S_i = \mathbb{R}e\left(S_i\right) + \sqrt{-1}\ \mathbb{I}m\left(S_i\right)$, then matrix $\left(\mathbb{R}e\left(S_i\right)\ \mathbb{I}m\left(S_i\right)\right)$ can be vectorized (see [25]) into x_i in the problem (3.2). Furthermore, a matrix S_i is positive semidefinite, $S_i \succeq 0$ iff the principal minors of S_i are all non-negative definite. The non-negative principal minors of the Hermitian matrix S_i and $\mathrm{Tr}\left(S_i\right) \leq P_i$ form constraints on $\left(\mathbb{R}e\left(S_i\right)\ \mathbb{I}m\left(S_i\right)\right)$ that result in a feasible set V_i for x_i. It is easily shown that V_i is convex and closed.

Because the objective function of (3.1) is a function of $S_i, i = 1, \ldots, K$, it can easily be written as a function of $\left(\mathbb{R}e\left(S_i\right), \mathbb{I}m\left(S_i\right)\right), i = 1, 2, \cdots, K$. This is to say that the objective function is a function of vector $x_i, i = 1, 2, \cdots, K$. This objective function is also concave over $\prod_{i=1}^{K} V_i$ due to both the concavity of $\log\det\left(\cdot\right)$ (refer to [6], page 74) and the existence of isomorphism between S_i and $\left(\mathbb{R}e\left(S_i\right), \mathbb{I}m\left(S_i\right)\right), \forall i$.

Therefore, the problem in (3.1) is a special case of the problem in (3.2). □

A block coordinate ascent algorithm (BCAA) for the generalized formulation in (3.2) is formally defined as follows. Because the iterative water-filling algorithm is a special case of the BCAA, the BCAA is introduced here to help develop a rigorous proof of convergence of the IWFA.

BCAA: **Block Coordinate Ascent Algorithm**

Step 1. Choice of Initial Point:

An initial point $z^0 \in V$.

Step 2. Definition of Operator $\mathbb{P}_i$:

Given $z^{k-1} \in V$, $V^*\left(z^{k-1}\right)$ is defined as the optimal solution set of

$$\max_{x \in V} \left\{ f(x) \,\middle|\, (x)_j = \left(z^{k-1}\right)_j, j = n_1 + 1, \ldots, n_s \right\}.$$

$(x)_j \in \mathbb{R}$ denotes the j-th entry (scalar) of vector x in this section.

Let operator $\mathbb{P}_1$ be defined by $\mathbb{P}_1\left(z^{k-1}\right) = V^*\left(z^{k-1}\right)$. Let $z^{k-1,1}$ denote an arbitrary element of $\mathbb{P}_1\left(z^{k-1}\right)$, i.e., $z^{k-1,1} \in \mathbb{P}_1\left(z^{k-1}\right)$. The second superscript of $z^{k-1,1}$ means the optimal point over $V_1 \times \left\{\left(z^{k-1}\right)_{n_1+1}\right\} \times \cdots \times \left\{\left(z^{k-1}\right)_{n_s}\right\}$. Similar understanding applies throughout this section. If the operators,

$$\left\{\mathbb{P}_i\right\}_{i=1}^{\ell}, 0 < \ell < K,$$

are defined, $V^*\left(z^{k-1,\ell}\right)$ is defined as the optimal solution set of

$$\max_{x \in V} \left\{ f(x) \,\middle|\, (x)_j = \left(z^{k-1,\ell}\right)_j, j = 1, \ldots, \sum_{i=1}^{\ell} n_i, \left(\sum_{i=1}^{\ell+1} n_i\right) + 1, \ldots, n_s \right\}.$$

Hence, operator $\mathbb{P}_{\ell+1}$ is defined by $\mathbb{P}_{\ell+1}\left(z^{k-1,\ell}\right) = V^*\left(z^{k-1,\ell}\right)$. Let $z^{k-1,\ell+1}$ denote an arbitrary element of $\mathbb{P}_{\ell+1}\left(z^{k-1,\ell}\right)$, i.e., $z^{k-1,\ell+1} \in \mathbb{P}_{\ell+1}\left(z^{k-1,\ell}\right)$. If $\ell + 1 = K, z^k = z^{k-1,\ell+1}$ and $z^k \in \mathbb{P}_K\left(z^{k-1,K-1}\right)$.

Step 3. Definition of Mapping A:

Mapping A is defined by

$$A \triangleq \mathbb{P}_K \mathbb{P}_{K-1} \ldots \mathbb{P}_1. \text{ Thus } z^k \in A\left(z^{k-1}\right).$$

In [44] a formal mathematical definition of an algorithm in terms of a mapping was provided. Following that approach, mapping A will also be known as an algorithm called the block coordinate ascent algorithm. For proving convergence of the BCAA, the maximization mapping is defined as follows.

Definition 3.4.2 (Maximization Mapping). *For the objective function, $Z : \widetilde{V} \longrightarrow \mathbb{R}$, where $\widetilde{V}$ is the set of feasible solutions, let $O : \widetilde{V}_1 \left(\subset \widetilde{V}\right) \longrightarrow \widetilde{V}$ be a mapping that projects a feasible point in $\widetilde{V}_1$ to the unique maximum point. Then O is called the maximization mapping.*

For the proposes of this thesis, the objective function Z over the abstract feasible solution set in Definition 3.4.2 can be assumed to be continuous. As an example, if the operator $\mathbb{P}_\ell$ of the block coordinate ascent algorithm is a point-to-point mapping, then it is also the maximization mapping. In particular, the objective function $Z \triangleq f$ and the set $\widetilde{V}_1 = \widetilde{V}\left(\triangleq V\right)$ are the terms associated with the maximization mapping $\mathbb{P}_\ell$.

Given the definition of the maximization mapping, we have the following proposition to reveal the relationship between the accumulation point and the fixed point of the BCAA.

Proposition 3.4.3. *Let the Cartesian product mapping A be defined as*

$$A \triangleq O_\ell \cdots O_2 O_1,$$

where $\ell \leq K$, and determine an algorithm that given a point z^0 generates the sequence $\left\{z^k\right\}_{k=0}^{\infty}$ with $z^{k+1} = A\left(z^k\right), \forall k$. Suppose

1. *All points z^k are in a compact set $X \subset \widetilde{V}_1 \left(\subset \widetilde{V}\right)$,*
2. *$O_1 : \widetilde{V}_1 \longrightarrow \widetilde{V}$ and $O_m : \widetilde{V} \longrightarrow \widetilde{V}$, $\forall m > 1$, are the maximization mappings, and*
3. *the maximization mapping $O_m, \forall m$, is continuous over its domain.*

Then any accumulation point of $\left\{z^k\right\}_{k=0}^{\infty}$, z^∞, is a fixed point, i.e., $z^\infty = O_m\left(z^\infty\right)$, $\forall m$.

Proof. Applying Condition 1, there must be a set

$$\kappa \subset \mathbb{N} \cup \{0\}$$

and a convergent subsequence such that $z^k \longrightarrow z^\infty$ for $k \in \kappa$. Using Condition 2, we see that

$$Z\left(z^{k+1}\right) \geq Z\left(z^k\right), \forall k \in \mathbb{N} \cup \{0\}.$$

So, $\left\{Z\left(z^k\right)\right\}_{k=0}^{\infty}$ is a monotonically increasing sequence. According to the limit property of the monotonic sequence, if the limit of some subsequence of the sequence exists and the sequence is monotonically increasing, then the limit of the sequence exists and the limit of the sequence is equal to the limit of the subsequence. Thus,

$$\lim_{k \to \infty} Z\left(z^k\right)$$

exists and

$$\lim_{k\to\infty} Z\left(z^k\right) = \lim_{k\in\kappa} Z\left(z^k\right).$$

Since Z is continuous,

$$\lim_{k\to\infty} Z\left(z^k\right) = Z\left(z^\infty\right). \tag{3.3}$$

Since $\left\{z^{k+1}\right\}_{k\in\kappa} \subset X$, where X is compact, $\exists\kappa^1 \subset \kappa$ such that $\lim_{k\in\kappa^1} z^{k+1}$ exists. We will write that limit as y^∞, i.e., $\lim_{k\in\kappa^1} z^{k+1} = y^\infty$. Similar to the derivation mentioned above,

$$\lim_{k\to\infty} Z\left(z^k\right) = Z\left(y^\infty\right). \tag{3.4}$$

From (3.3), (3.4) and Condition 3 of Proposition 3.4.3,

$$Z\left(z^\infty\right) = Z\left(y^\infty\right) = Z\left(\lim_{k\in\kappa^1} A\left(z^k\right)\right) = Z\left(A\left(z^\infty\right)\right).$$

Therefore,

$$Z\left(z^\infty\right) = Z\left(A\left(z^\infty\right)\right).$$

As $m = 1$,

$$Z\left(z^\infty\right) = Z\left(A\left(z^\infty\right)\right) \geq Z\left(O_1\left(z^\infty\right)\right) \geq Z\left(z^\infty\right).$$

Then

$$Z\left(z^\infty\right) = Z\left(O_1\left(z^\infty\right)\right).$$

Because z^∞ is the feasible point related to O_1 and O_1 is the mapping from a point to the unique maximum point, corresponding to the definition of mapping, $O_1\left(z^\infty\right) =$

z^{∞}.

Assume that, as $1 \leq m < \ell$, $O_m(z^{\infty}) = z^{\infty}$. Because

$$Z(z^{\infty}) = Z(A(z^{\infty}))$$

$$\geq Z(O_{m+1}O_m \cdots O_1(z^{\infty})) \geq Z(z^{\infty})$$

and, due to the induced assumption,

$$Z(O_{m+1}O_m \cdots O_1(z^{\infty})) = Z(O_{m+1}(z^{\infty})),$$

and

$$Z(z^{\infty}) = Z(O_{m+1}(z^{\infty})).$$

Because z^{∞} is the feasible point related to O_{m+1} and O_{m+1} is a mapping from a feasible point to the unique maximum point, $O_{m+1}(z^{\infty}) = z^{\infty}$.

Therefore for the maximization mapping O_m, $\forall m$, the accumulation point, z^{∞}, is a fixed point, i.e., $z^{\infty} = O_m(z^{\infty})$. □

The conclusion of Proposition 3.4.3 implies that z^{∞} is also a fixed point for the mapping A, because

$$A(z^{\infty}) = O_{\ell} \cdots O_2 O_1(z^{\infty}) = O_{\ell} \cdots O_2(z^{\infty}) = \cdots = O_{\ell}(z^{\infty}) = z^{\infty}.$$

Therefore, based on the conclusions of Proposition 3.4.3, we can introduce the following theorem on convergence of the block coordinate ascent algorithm.

Theorem 3.4.4. *Consider the abstract formulation in (3.2) and, assume that f is concave and differentiable, that V is convex and that either V is compact or the superlevel set $\{x|f(x) \geq f(z^0)\}$ is bounded. Now, if the mapping $\mathbb{P}_\ell$ is continuous over V, $\forall \ell$, then the limit, z^∞, of any convergent subsequence of $\{z^k\}_{k=0}^\infty$ generated by the BCAA is an optimal point of (3.2), and $\{f(z^k)\}_{k=0}^\infty$ approaches to the optimal value.*

Proof. The block coordinate ascent algorithm is a product mapping,

$$\mathbb{P}_K \cdots \mathbb{P}_2 \mathbb{P}_1.$$

Assume that the block coordinate ascent algorithm generates the sequence $\{z^k\}_{k=0}^\infty$ and z^∞ is the limit of a convergent subsequence of the sequence. We will prove z^∞ to be an optimal point below.

Due to the compactness of V or the boundness of set

$$\{x|f(x) \geq f(z^0)\},$$

Condition 1 of Proposition 3.4.3 is satisfied.

Since

$$\mathbb{P}_m : V \longrightarrow V, \forall m,$$

is continuous, $\mathbb{P}_m$ is then a point-to-point mapping. Also due to the definition of the operator $\mathbb{P}_m$, $\mathbb{P}_m$ is the maximization mapping. Hence, Condition 2 of Proposition 3.4.3 is satisfied by $\mathbb{P}_m$.

Because the continuity of mapping $\mathbb{P}_\ell, \forall \ell$, is assumed, Condition 3 of Proposition 3.4.3 is satisfied.

Therefore, according to Proposition 3.4.3, the accumulation point, z^∞, is a fixed point of the maximization mapping $\mathbb{P}_m$, i.e.,

$$z^\infty = \mathbb{P}_m\left(z^\infty\right), \forall m.$$

According to the optimality condition of convex programming, if f_{x_m} denotes the transposition of the gradient of f in the variable (vector) x_m,

$$f_{x_m}\left(z_m^\infty\right)\left(z_m - z_m^\infty\right) \leq 0, \ \forall m, \ \forall z_m \in V_m.$$

So

$$\sum_{m=1}^{K} f_{x_m}\left(z_m^\infty\right)\left(z_m - z_m^\infty\right) \leq 0.$$

Thus,

$$f_x\left(z^\infty\right)\left(z - z^\infty\right) \leq 0, \forall z \in V.$$

Therefore, z^∞ is the optimal solution, and, due to the monotonicity of sequence $\left\{f\left(z^k\right)\right\}_{k=0}^\infty$, $\left\{f\left(z^k\right)\right\}_{k=0}^\infty$ approaches to the optimal value. □

Although a similar result on convergence of the BCAA can be found in [2], the

proof presented here explicitly exposes the relationship between the fixed point and the optimal point. As a result of this proof, a direct link is established between the KKT condition and the variational inequality of convex optimization. This result is proven for the first time to our knowledge. It provides an opportunity for further algorithm developments under a unified mathematical framework.

It is meaningful to question how to guarantee that operator

$$\mathbb{P}_m : V \longrightarrow V, \ \forall m,$$

satisfies the continuity over V, as a required condition of Theorem 3.4.4. It is obvious that if operator $\mathbb{P}_m, \forall m$, has a unique optimal solution, then operator $\mathbb{P}_m, \forall m$, is the maximization mapping, i.e., it is the mappings from a point to a point, and operator $\mathbb{P}_m, \forall m$, is called satisfying uniqueness. In fact, it can be proved (see below) that if the uniqueness holds, then mapping $\mathbb{P}_m, \forall m$, also satisfies the continuity, i.e., it is continuous over V. First, the uniqueness is formally introduced and then the continuity is proved.

Definition 3.4.5 (Uniqueness Condition). *$\forall z^k \in V$, if the optimal solution set*

$$\arg\left(\max_{x\in V}\left\{f(x)\,\middle|\, x_j = z_j^k, j = 1,\ldots,\sum_{i=1}^{\ell-1} n_i, \left(\sum_{i=1}^{\ell} n_i\right)+1,\ldots,n_s\right\}\right)$$

is a single point set, $\forall \ell$, then $f(x)$ is said to satisfy the uniqueness condition over set V.

Based on Definition 3.4.5, we have the following lemma on the continuity of the

block coordinate ascent algorithm.

Lemma 3.4.6. *Suppose that $f(x)$ is continuous and satisfies the uniqueness condition over set V in (3.2), and suppose that V is closed. If V is bounded or set $\{x|f(x) \geq f(z^0)\}$ is bounded, then mapping $\mathbb{P}_m, \forall m$, is continuous over V.*

Proof. $\forall \overline{z}, \overline{z} \in V$. Assume that $\left\{z^k\right\} \subset V$ and $\lim_{k\to\infty} z^k = \overline{z}$.

$$\forall \mathbb{P}_\ell,\ 0 < \ell \leq K, x^k \triangleq \mathbb{P}_\ell\left(z^k\right),\ \text{ and } \overline{y} \triangleq \mathbb{P}_\ell\left(\overline{z}\right).$$

$\left\{x^k\right\}$ has an accumulation point denoted by $\overline{x}$. If $\left\{x^k\right\}$ does not converge, a subsequence $\left\{x^{k_r}\right\}$, which converges to $\overline{x}$ as r tends to infinity, can be selected from $\left\{x^k\right\}$. Hence, two convergent subsequences $\left\{x^{k_r}\right\}$ and $\left\{z^{k_r}\right\}$ can be acquired. So, without loss of generality, assume that $\left\{x^k\right\}$ converges to $\overline{x}$. From the definition of $\mathbb{P}_\ell$, it holds that

$$\overline{x}_j = \lim_{k\to\infty} x_j^k = \lim_{k\to\infty} z_j^k = \overline{z}_j, j = 1, \ldots, \sum_{i=1}^{\ell-1} n_i, \left(\sum_{i=1}^{\ell} n_i\right) + 1, \ldots, n_s.$$

Since V is closed and $\left\{x^k\right\} \subset V$, $\overline{x} \in V$. Due to $\overline{y} \triangleq \mathbb{P}_\ell\left(\overline{z}\right)$ and

$$\overline{x}_j = \overline{z}_j = \overline{y}_j, j = 1, \ldots, \sum_{i=1}^{\ell-1} n_i, \left(\sum_{i=1}^{\ell} n_i\right) + 1, \ldots, n_s,$$

$$f\left(\overline{y}\right) \geq f\left(\overline{x}\right). \tag{3.5}$$

On the other hand, construct a sequence as follows.

$y^k \in \mathbb{R}^{n_s}, \forall k$, is defined by following two steps.

- $y_j^k \triangleq \overline{y}_j$, for $j = \left(\sum_{i=1}^{\ell-1} n_i\right) + 1, \ldots, \sum_{i=1}^{\ell} n_i$
- $y_j^k \triangleq z_j^k$, for $j = 1, \ldots, \sum_{i=1}^{\ell-1} n_i, \left(\sum_{i=1}^{\ell} n_i\right) + 1, \ldots, n_s$.

Thus, $y^k \in \mathbb{R}^{n_s}, \forall k$, is obtained.

It is immediate that $\{y^k\} \subset V$ and $\lim_{k\to\infty} y^k = \overline{y}$. Hence,

$$f(\overline{x}) \geq f(\overline{y}). \tag{3.6}$$

From (3.5) and (3.6),

$$f(\overline{x}) = f(\overline{y}) = f(\mathbb{P}_\ell(\overline{z})).$$

When the consideration of the uniqueness condition is added into the above equalities, $\overline{x} = \overline{y}$ holds, i.e.,

$$\lim_{k\to\infty} \mathbb{P}_\ell(z^k) = \lim_{k\to\infty} x^k = \overline{x} = \overline{y} = \mathbb{P}_\ell(\overline{z}).$$

Thus, $\mathbb{P}_\ell$ is continuous over V, $\forall \ell$. □

The following corollary summarizes the above analysis and states the convergence theorem for BCAA applied to the generalized problem in (3.2).

Corollary 3.4.7. *Consider the abstract formulation in (3.2). Assume that f is concave and differentiable, and satisfies the uniqueness condition over set V, that V is convex, and that either V is compact or the superlevel set $\{x|f(x) \geq f(z^0)\}$ is bounded. Then the limit, z^∞, of any convergent subsequence of $\{z^k\}_{k=0}^\infty$ is an optimal point of (3.2) and $\{f(z^k)\}_{k=0}^\infty$ approaches to the optimal value.*

3.4.2 Convergence of IWFA under the Individual Power Constraints

To simplify the discussion of convergence of the BCAA and IWFA for the problem in (3.1), and check the conditions of Corollary 3.4.7, we define

$$f\left(S_1, S_2, \cdots, S_K\right) \triangleq \log\left(\det\left(I + \sum_{i=1}^{K} H_i^{\dagger} S_i H_i\right)\right)$$

and

$$V_i = \left\{S_i \succeq 0, \mathrm{Tr}\left(S_i\right) \leq P_i\right\}, \forall i.$$

Here, the block coordinate ascent algorithm is chosen as the algorithm for finding the relevant optimal matrices. The number of entries of S_i is equivalent to n_i in the definition of the algorithm. Let us define

$$G_i \triangleq H_i\left(I_m + \sum_{t=1, t\neq i} H_t^{\dagger} S_t H_t\right)^{-\frac{1}{2}}.$$

It is known that there is a unitary matrix

$$U_i \text{ such that } U_i^{\dagger}\left(G_i G_i^{\dagger}\right) U_i \text{ is a diagonal matrix.}$$

Without loss of generality, we assume that $H_i \neq 0, \forall i$. This diagonal matrix is written as Λ_i.

$$\Lambda_i \triangleq \mathrm{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_n\right) \in \mathbb{R}^{n\times n} \subset \mathbb{C}^{n\times n},$$

where

$$\lambda_\ell \geq \lambda_{\ell+1} \ (\forall \ell), \text{ the set } \{\ell | \lambda_\ell > 0, \ell \in \{1, 2, \ldots, n\}\}$$

is not empty and

$$\ell_{max} \triangleq \max\{\ell | \lambda_\ell > 0, \ell \in \{1, 2, \ldots, n\}\}.$$

Given U_i, $U_i^\dagger V_i U_i \triangleq \left\{U_i^\dagger S_i U_i | S_i \in V_i\right\}$ and $V_{i_1} \triangleq U_i^\dagger V_i U_i$. It is obvious that V_{i_1} is convex, $V_{i_1} \subset V_i$ and a one-to-one mapping, $V_i \longrightarrow V_{i_1}$, is introduced based on the definition of V_{i_1}.

$$V_{i_2} \triangleq \left\{ \overline{S}_i \in V_{i_1} \middle| \begin{array}{c} \left(\overline{S}_i\right)_{s,t} = 0 \text{ for } s \neq t \text{ and } s,t \in \{1,2,\ldots,n\}; \\ \left(\overline{S}_i\right)_{s,s} = 0 \text{ for } s \in \{\ell_{max}+1, \ell_{max}+2, \ldots, n\}. \end{array} \right\}.$$

It is obvious that V_{i_2} is convex and $V_{i_2} \subset V_{i_1}$.

It is known that $f(S_1, S_2, \cdots, S_K)$ is concave and continuously differentiable, due to the concavity of $\log\det(\cdot)$ (refer to [6], page 74). The existence of isomorphism between S_i and $(\mathbb{R}e(S_i), \mathbb{I}m(S_i)), \forall i$, is also straightforward, and V_i $(\forall i)$ is convex. Hence the first condition of Corollary 3.4.7 is satisfied.

$$\forall S_i, S_i = E\left[x^i \left(x^i\right)^\dagger\right] \in V_i$$

implies $(S_i)_{j,j} \leq P_i$ and thus

$$\begin{aligned} |(S_i)_{s,t}| &= |E\left[\left(x_s^i\right)\left(\overline{x_t^i}\right)\right]| \\ &\leq \sqrt{E\left[|x_s^i|^2\right] E\left[|x_t^i|^2\right]} \text{ (Cauchy-Schwartz Inequality)} \\ &\leq \sqrt{(S_i)_{s,s}\,(S_i)_{t,t}}. \end{aligned}$$

This implies $|(S_i)_{s,t}| \leq P_i$ $(\forall i, j, s \text{ and } t)$ and further $\| S_i \|_F \leq nP_i$. Hence V_i is

bounded. According to the definition of V_i, V_i is obviously closed. Then V_i is compact. Another condition of Corollary 3.4.7 is now also satisfied. We will now define a function f_i to denote the simplified objective function, whose domain is V_{i_2} mentioned above. And it corresponds to the case when the i-th user is being optimized while other users are kept unchanged.

$$f_i\left(\left(\overline{S}_i\right)_{1,1},\left(\overline{S}_i\right)_{2,2},\cdots,\left(\overline{S}_i\right)_{\ell_{max},\ell_{max}}\right) \triangleq \sum_{\ell=1}^{\ell_{max}} \log\left(1+\lambda_\ell\left(\overline{S}_i\right)_{\ell,\ell}\right) \quad (\forall i),$$

Thus, with the differences between the definitions of V_i, V_{i_1} and V_{i_2},

$$\max_{S_i\in V_i}\left\{\log\left(\det\left(I_m+\sum_{t=1,t\neq i}^{K} H_t^\dagger S_t H_t + H_i^\dagger S_i H_i\right)\right)\,\middle|\, \begin{array}{c} S_t\in V_t\,(\text{for } t\neq i) \\ \text{is given}\end{array}\right\} \tag{3.7}$$

is equivalent to

$$\max_{\overline{S}_i\in V_{i_1}} \log\left(\det\left(I_n+\Lambda_i^{\frac{1}{2}}\overline{S}_i\Lambda_i^{\frac{1}{2}}\right)\right), \tag{3.8}$$

and then it is equivalent to

$$\max_{\overline{S}_i\in V_{i_2}} \sum_{l=1}^{\ell_{max}} \log\left(1+\lambda_l\left(\overline{S}_i\right)_{\ell,\ell}\right). \tag{3.9}$$

Because the Hessian matrix of

$$f_i = \sum_{l=1}^{\ell_{max}} \log\left(1+\lambda_\ell\left(\overline{S}_i\right)_{\ell,\ell}\right)$$

is

$$-\text{diag}\left(\frac{\lambda_1^2}{\left(1+\lambda_1\left(\overline{S}_i\right)_{1,1}\right)^2},\frac{\lambda_2^2}{\left(1+\lambda_2\left(\overline{S}_i\right)_{2,2}\right)^2},\cdots,\frac{\lambda_{\ell_{max}}^2}{\left(1+\lambda_{\ell_{max}}\left(\overline{S}_i\right)_{\ell_{max},\ell_{max}}\right)^2}\right),$$

it is a negative definite matrix, and the cardinality of the optimal solution set

$$\arg\left(\max_{\overline{S}_i\in V_{i_2}}\sum_{\ell=1}^{\ell_{max}}\log\left(1+\lambda_\ell\left(\overline{S}_i\right)_{\ell,\ell}\right)\right)$$

is one, i.e.,

$$\left(\arg\left(\max_{\overline{S}_i\in V_{i_2}}\sum_{l=1}^{\ell_{max}}\log\left(1+\lambda_\ell\left(\overline{S}_i\right)_{\ell,\ell}\right)\right)\right)^{\sharp}=1.$$

Therefore, it is easily obtained that the cardinality of the optimal solution set of

$$\max_{S_i\in V_i}\left\{\log\left(\det\left(I_m+\sum_{t=1,t\neq i}^{K}H_t^\dagger S_t H_t+H_i^\dagger S_i H_i\right)\right)\middle|\begin{array}{c}\text{given } S_t\\ \in V_t\,(t\neq i)\end{array}\right\}$$

is equal to one, corresponding to the equivalence (3.9) of the optimization problems and the one-to-one mapping of $V_{i_1}\longrightarrow V_i$. An important consequence is that the uniqueness condition of Corollary 3.4.7 is obtained.

The final optimization problem (3.9) of the equivalent optimization problems (3.7), (3.8) and (3.9) can easily solved by some optimization algorithms, for instance, as well known, the water-filling algorithm. According to Corollary 3.4.7, convergence of the algorithm is acquired by the block coordinate ascent algorithm. If $\exists H_i=0$ (the null matrix), convergence of the algorithm is still obtained from the above derivation

and the compactness of the feasible set.

Corollary 3.4.8. *For the problem in (3.1), the iterative water-filling algorithm with or without Fibonacci search is convergent. As the number of iterative steps increases, the corresponding objective value approaches to the optimal objective value.*

We also obtain a sufficient and necessary condition of the optimal solution to the problem in (3.1) as follows.

Corollary 3.4.9. *For the problem in (3.1) with the K users, S_i is an optimal solution iff S_i is obtained by the single-user water-filling algorithm, with or without Fibonacci search, for the channel H_i with noise covariance $I + \sum_{j=1,j\neq i}^{K} H_j^{\dagger} S_j H_j$, for all $i = 1, 2, ..., K$.*

Chapter 4

Sum Capacity of the MIMO BC

In this chapter, we consider the MIMO broadcast channel (BC), in which a single base-station sends independent messages to several users. We consider the important problem of finding the set of input covariances that maximize the sum rate of the MIMO BC, and we study efficient algorithms for solving this problem.

In the first section of this chapter, we provide a general mathematical model for the MIMO BC. We then explain that there is a so-called dual MIMO MAC for the MIMO BC, and that this dual is equivalent from the sum rate perspective, in the sense that the sum rates are equal and the optimal input covariances can be computed from each other, e.g., [20]. The dual MIMO MAC is of significant interest because the sum rate optimization problem is convex, whereas that for the MIMO BC is not. Furthermore, there is the potential to extend the iterative water-filling algorithm for the conventional MIMO MAC to this dual of the MIMO BC. However, that extension is not straightforward (e.g., [20]) because in the case of the dual of the MIMO BC, the power constraint couples the stages of the iterative water-filling algorithm. These

stages are not coupled in the MAC case. In the second section of the chapter we describe two modified iterative water-filling algorithms for the MIMO BC [20], and we show how the Fibonacci search technique discussed in Chapter 2 can be applied to these algorithms. In third section, we point out weaknesses in the existing attempts to prove convergence of one of these algorithms, and in the fourth section we provide a rigorous proof.

4.1 Models for the MIMO BC

We consider the MIMO BC illustrated on the right hand side of the figure on Page 2, in which the base-station has m antennas and there are K mobile stations, each of which has n antennas. When the information flows from the base-station to the mobile stations, we have a BC channel or downlink channel.

With appropriate synchronization, the BC can be described as [20]

$$y_i = H_i x + z_i,$$

where y_i is the signal received at the i-th mobile station, and $H_i \in \mathbb{C}^{n\times m}, i = 1, 2, \cdots, K$, denotes the matrix of channel gains from each antenna at the base-station to each antenna at the i-th mobile station. As in Chapter 3, without loss of generality we will assume $H_i \neq 0, \forall i$. The vector x is an $m \times 1$ complex input vector, and $z_i \in \mathbb{C}^n$ is an additive Gaussian noise vector that, without loss of generality, has identity covariance. To simplify the derivation of the sum capacity of the MIMO BC, the dual MIMO MAC is introduced. As mentioned above, the reasons for introducing this

dual are that the sum capacity of the MIMO BC is equal to that of the dual MIMO MAC, i.e., $C_{\text{BC}}\left(H_1, \cdots, H_K, P\right) = C_{\text{dmac}}\left(H_1^\dagger, \cdots, H_K^\dagger, P\right)$ [20], and that the latter can be much more easily determined than the former [20].

The dual MIMO MAC can be described as

$$y_{\text{dmac}} = \sum_{i=1}^{K} H_i^\dagger x^i + z,$$

which is the same as that of MIMO MAC considered in Chapter 3 (see [38]). If $Q_i \triangleq E\left[x^i \left(x^i\right)^\dagger\right]$, then the mathematical formulation [4, 20, 38] of the sum capacity of the dual MIMO MAC is:

$$\mathbb{C}_{\text{dmac}}\left(H_1^\dagger, \cdots, H_K^\dagger, P\right) =$$

$$\max_{\{Q_i\}_{i=1}^K} \left\{\log\left(\det\left(I + \sum_{i=1}^K H_i^\dagger Q_i H_i\right)\right) | Q_i \succeq 0, \sum_{i=1}^K \text{Tr}\left(Q_i\right) \leq P\right\}. \quad (4.1)$$

The key difference between (4.1) and the problem formulated in (3.1) for the MAC is the power constraint. In (4.1) we have a sum power constraint, $\sum_{i=1}^K \text{Tr}\left(Q_i\right) \leq P$, whereas in (3.1) we have individual power constraints, $\text{Tr}\left(Q_i\right) \leq P_i$. As mentioned earlier, this couples the iterative water-filling stages and requires different treatment from that in Chapter 3.

We observe that the problem in (4.1) can be written in the following abstract form, reminiscent of that in Chapter 3,

$$\max\left\{f\left(x_1, x_2, \ldots, x_K\right) | x_i \in \mathbb{R}^{n_i}, \left(x_1, x_2, \ldots, x_K\right) \in V, \quad \forall i \in \{1, 2, \ldots, K\}\right\}, \quad (4.2)$$

where $f(x_1, x_2, \ldots, x_K)$ is concave and differentiable, n_i is a natural number, $\sum_{i=1}^{K} n_i = n_s$ and $V \subset \mathbb{R}^{n_s}$ is convex and closed. We would like to draw attention to the difference between the feasible set V here and the feasible set of the formulation for the MIMO MAC in (3.2). The problem in (4.2) will be used to construct a rigorous proof of convergence of the iterative water-filling algorithm for the problem in (4.1) that will be discussed in Section 4.4. Utilizing a similar derivation to that in Proposition 3.4.1, it is simple to show that the problem in (4.1) is a special case of the generalized mathematical problem (4.2).

4.2 Iterative Water-Filling under a Sum Power Constraint

Given the successful application of iterative water-filling to the MIMO MAC (with individual power constraints; cf. Chapter 3), a natural question to ask is whether the iterative water-filling algorithm can be modified in such a way that it will generate an optimal solution to the dual MIMO MAC of the MIMO BC (in which there is a sum power constraint), and hence generate an optimal solution to the sum rate optimization problem for the MIMO BC. In [20], two such modifications were proposed. The first is based on the principles of the block coordinate ascent algorithm, but requires a significant amount of memory. The second is a modification of the first that requires less memory, but deviates from the principles of block coordinate ascent. The algorithms are as follows.

Algorithm 4.1 (Algorithm 1 in [20]):

Input: Channel matrices H_i, the initial covariances $Q_i^{(\tilde{n})}$ assigned to arbitrary positive semidefinite matrices, $\tilde{n} = -(K-2), \ldots, 0$ and $i = 1, \ldots, K$.

1) Generate effective channels

$$G_i^{(\tilde{n})} = H_i \left(I + \sum_{j=1}^{K-1} H_{[i+j]_K}^{\dagger} Q_{[i+j]_K}^{(\tilde{n}-K+j)} H_{[i+j]_K} \right)^{-\frac{1}{2}},$$

$i = 1, \ldots, K$, where $[x]_K \triangleq mod\left((x-1), K\right) + 1$.

2) Treating these effective channels as parallel, noninterfering channels, obtain the new covariance matrices $\left\{Q_i^{(\tilde{n})}\right\}_{i=1}^{K}$ by water-filling over a virtual block diagonal channel matrix with diagonal blocks $G_i^{(\tilde{n})}$ under the sum power P constraint. That is,

$$\left\{Q_i^{(\tilde{n})}\right\}_{i=1}^{K} = \arg\left(\max_{\{Q_i\}_{i=1}^{K}} \left\{ \sum_{i=1}^{K} \log\left(\det\left(I + \left(G_i^{(\tilde{n})}\right)^{\dagger} Q_i G_i^{(\tilde{n})} \right) \right) \middle| \begin{array}{l} Q_i \succeq 0, \\ \sum_{i=1}^{K} \mathrm{Tr}\left(Q_i\right) \leq P \end{array} \right\} \right). \tag{4.3}$$

Set $\tilde{n} = \tilde{n} + 1$ **and then go to 1)**.

Remark 4.2.1. *Each set of K iterations in Algorithm 4.1 constitutes one iteration of the corresponding block coordinate ascent algorithm (BCAA). Therefore, Algorithm 4.1 can be viewed as a submapping of the BCAA under the sum power constraint, and hence it converges (to the global optimum). This conclusion stems directly from the rigorous proof of the convergence theorem in Section 3.4.*

The above statement is a clean proof of convergence of Algorithm 4.1, but this algorithm bears a heavy memory burden [20], and the following modification was proposed in [20].

Algorithm 4.2 (Algorithm 2 in [20]):

Input: Channel matrices H_i, the initial covariances $Q_i^{(0)}$ assigned to arbitrary positive semidefinite matrices, $i = 1, \ldots, K$.

1) Generate effective channels

$$G_i^{(\tilde{n})} = H_i \left(I + \sum_{j \neq i} H_j^{\dagger} Q_j^{(\tilde{n}-1)} H_j \right)^{-\frac{1}{2}},$$

$i = 1, \ldots, K$.

2) Treating these effective channels as parallel, noninterfering channels, the new covariance matrices $\left\{ Q_i^{(\tilde{n})} \right\}_{i=1}^{K}$ by water-filling over a virtual block diagonal channel matrix with diagonal blocks $G_i^{(\tilde{n})}$ under the sum power P constraint. That is,

$$\left\{ Q_i^{(\tilde{n})} \right\}_{i=1}^{K} = \arg \left(\max_{\{Q_i\}_{i=1}^{K}} \left\{ \sum_{i=1}^{K} \log \left(\det \left(I + \left(G_i^{(\tilde{n})} \right)^{\dagger} Q_i G_i^{(\tilde{n})} \right) \right) \middle| \begin{array}{l} Q_i \succeq 0, \\ \sum_{i=1}^{K} \mathrm{Tr}\left(Q_i\right) \leq P \end{array} \right\} \right). \tag{4.4}$$

3) Compute the updated covariance matrices $Q_i^{(\tilde{n})}$, as

$$Q_i^{(\tilde{n})} = \frac{1}{K} Q_i^{(\tilde{n})} + \frac{K-1}{K} Q_i^{(\tilde{n}-1)}, \; i = 1, \ldots, K.$$

Set $\tilde{n} = \tilde{n} + 1$ **and then go to 1).**

Remark 4.2.2. *As was the case for the conventional MIMO MAC discussed in Chapter 3, in the water-filling steps of both Algorithm 1 and Algorithm 2 (step 2 in both algorithms) one can employ the Fibonacci search algorithm developed in Chapter 2. However, because of the block diagonal nature of the virtual channel, the details of that algorithm need to be modified for this case. The modifications to find* $Q_i^{(\tilde{n})}$ *are divided into the following three steps:*

Step 1: Pre-Processing.

Compute the unitary matrix $U_i^{(\tilde{n})} \in \mathbb{C}^{n\times n}$ *by calculating the eigenvalue decomposition satisfying*

$$\Lambda_i = \left(U_i^{(\tilde{n})}\right)^{\dagger} G_i^{(\tilde{n})} \left(G_i^{(\tilde{n})}\right)^{\dagger} U_i^{(\tilde{n})} = diag\left((\Lambda_i)_{1,1}, \cdots, (\Lambda_i)_{n,n}\right).$$

Assume that $(\Lambda_i)_{\ell,\ell} \geq (\Lambda_i)_{\ell+1,\ell+1}, \forall \ell, i$. *Let* $\left\{(\Lambda_1)_{1,1}, \cdots, (\Lambda_K)_{n,n}\right\}$ *be ordered monotonically decreasing into* $\{\lambda_t\}_{t=1}^{K\times n}$.

$$j(i) \triangleq \max\left\{j \middle| (\Lambda_i)_{j,j} > 0\right\}$$

and let

$$\overline{n} \triangleq \sum_{i=1}^{K} j(i) \leq \min\{Kn, Km\}.$$

Step 2: Water-Filling with Fibonacci Search.

Let

$$S_k \triangleq \frac{1}{k}\left\{P - \left[(k-1)\frac{1}{\lambda_k} - \sum_{i\in\{1,\cdots,k-1\}\cap\{k\geq 2\}} \frac{1}{\lambda_i}\right]\right\}.$$

Search

$$k^* = \max\{k | S_k > 0, 1 \leq k \leq \overline{n}\}$$

and

$$k_i^* = \max\left\{j | (\Lambda_i)_{j,j} \geq \lambda_{k^*}, 1 \leq k \leq j\,(i)\right\}.$$

Here, the Fibonacci approximation ratio $\frac{1}{3}$ *and* $\frac{2}{3}$ *are used for searching* k^* *and* k_i^* *and this method is also called the Fibonacci search. Searching* k^* *is similar to that in Algorithm 3.1. The searching of* k_i^* *is presented as follows.*

1st Step. *Assume that* $a = 1$ *and* $b = j\,(i)\,, \forall i$.

2nd Step. *If* $a = b$, *then* $k_i^* = a$ *and go to* ***Step 3.***

Else, $a_1 = \lfloor a + \frac{1}{3}(b-a) \rfloor$, $b_1 = \lceil a + \frac{2}{3}(b-a)\ \rceil$.

3rd Step. *If* $(\Lambda_i)_{a_1,a_1} < \lambda_{k^*}$, *then* $b = a_1 - 1$ *and go to the* ***2nd step;***

If $(\Lambda_i)_{b_1,b_1} \geq \lambda_{k^*}$, *then* $a = b_1$ *and go to the* ***2nd step;***

If $(\Lambda_i)_{a_1,a_1} \geq \lambda_{k^*} > (\Lambda_i)_{b_1,b_1}$, *then* $a = a_1$, $b = b_1 - 1$ *and go to the* ***2nd step.***

Step 3: Find Optimal Solution of *(4.3).*

1st Step. *Compute* $S^* \in \mathbb{C}^{n \times n}$ *as follows:*

$$\begin{aligned} &(S_i^*)_{t,t} = \tfrac{1}{\lambda_{k^*}} - \tfrac{1}{(\Lambda_i)_{t,t}} + S_{k^*}, 1 \leq t \leq k_i^*; \\ &(S_i^*)_{t,t} = 0, k_i^* < t \leq n; \\ &(S_i^*)_{s,t} = 0, s \neq t. \end{aligned} \tag{4.5}$$

2nd Step. *Compute* $Q_i^{(\tilde{n})} = U_i^{(\tilde{n})} S_i^* \left(U_i^{(\tilde{n})}\right)^{\dagger}$.

4.3 Weaknesses of the Existing Research Regarding Convergence of Algorithm 4.2

As discussed in Remark 4.2.1, Algorithm 4.1 (Algorithm 1 in [20]) is based on the principles of block coordinate ascent and by modifying the convergence argument in [20] to include the rigorous convergence proof in Chapter 3, a rigorous proof of convergence of Algorithm 4.1 can be established. Algorithm 4.2 (Algorithm 2 in [20]) was developed to avoid the heavy memory burden carried by Algorithm 4.1, but the modifications that were made to obtain Algorithm 4.2 result in deviations from the principles of block coordinate ascent, and hence convergence of this algorithm needs to be examined separately. This was attempted in [20], but as we will discuss intuitively below, that attempt contains a weakness, due to an invalid application of Zangwill's Convergence Theorem B (see [44], page 128). Interestingly, a similar weakness can be found in [23]. In Section 4.4 we will provide a rigorous proof of convergence of Algorithm 4.2 based on fixed point theory.

4.3.1 Inapplicability of Zangwill's Convergence Theorem B to Algorithm 2 of [20]

Let us use the symbol S_1 to denote steps 1) and 2) of Algorithm 4.2 and let SS denote step 3), which is sometimes called the spacer step. Algorithm 4.2 iterates in the following order: $S_1, SS, S_1, SS, S_1, SS, \cdots$.

The cyclic coordinate ascent algorithm with the spacer step iterates by this order: $CCAA, SS, CCAA, SS, CCAA, SS, \cdots$ where $CCAA$ denotes the cyclic coordinate

ascent algorithm. Therefore, to apply Zangwill's Convergence Theorem B, which applies to cyclic coordinate ascent algorithm, to Algorithm 4.2 it must be shown that S_1 is a $CCAA$ step. However, it is clear from Algorithm 4.2 that S_1 is not cyclic. Indeed it is only a submapping of a $CCAA$ step.

Without further ado, the difference here is easily seen. A counter example can be easily constructed so that a sub-mapping of $CCAA$, e.g. S_1, is not guaranteed to converge under Zangwill's Convergence Theorem B. Hence, Zangwill's Convergence Theorem B cannot be used to guarantee convergence of Algorithm 4.2. Therefore, the convergence proof in [20], which relies on Zangwill's Convergence Theorem B, does not guarantee convergence of Algorithm 4.2.

4.4 Convergence of Algorithm 4.2

We will provide a rigorous proof of convergence of Algorithm 4.2.

4.4.1 Fixed Point Theory, Continuity and Convergence of the Algorithms

In this subsection, we will first propose a definition of the average function to present a new algorithm that we will denoted by DBCAA, which stands for the diagonal block coordinate ascent algorithm. Convergence of the DBCAA will be proved rigorously, and this will lead to a rigorous proof of convergence of the iterative water-filling algorithm for the MIMO BC, with or without Fibonacci search.

To clearly express the process of finding the solution to the problem in (4.1), function F will be defined as follows,

$$\forall x \in \mathbb{R}^{Kn_s}, x \triangleq (x_{1,1}, x_{1,2}, \cdots, x_{1,K}; \cdots; x_{K,1}, x_{K,2}, \cdots, x_{K,K}),$$

$$(x_{1,1}, x_{2,2}, \cdots, x_{K,K}) \in V, x_{l,j} \in \mathbb{R}^{n_j} \ (\forall l), j = 1, 2, \ldots, K.$$

$$F(x) \triangleq \frac{1}{K} \sum_{l=1}^{K} f(x_{l,1}, x_{l,2}, \cdots, x_{l,K}), \tag{4.6}$$

where we call the function F the average function. We also define the sets

$$V_D \triangleq \left\{ z \middle| \begin{array}{l} z = (z_{1,1}, \cdots, z_{1,K}; \cdots; z_{K,1}, \cdots, z_{K,K}); \\ z_{l,j} \in \mathbb{R}^{n_j}, \forall l, j; (z_{1,1}, z_{2,2}, \cdots, z_{K,K}) \in V. \end{array} \right\} \subset \mathbb{R}^{Kn_s},$$

$$V_{D1} \triangleq$$

$$\left\{ \left(\overbrace{z_1, \cdots, z_K; \cdots; z_1, \cdots, z_K}^{K^2} \right) \in V_D \middle| z_l \in \mathbb{R}^{n_l}, l \in \{1, 2, \cdots, K\} \right\}.$$

It can be easily seen that V_D and V_{D1} are convex, and $V_{D1} \subset V_D$.

The diagonal block coordinate ascent algorithm for the generalized mathematical problem in (4.2), which is an abstract framework for Algorithm 4.2, is formally defined as follows.

DBCAA: **Diagonal Block Coordinate Ascent Algorithm**

Step 1. Choice of Initial Point:

An initial point

$$z^0 \in V_{D1} \left(\subset V_D \subset \mathbb{R}^{Kn_s}\right).$$

$$z^0 = (z_1^0, \ldots, z_K^0; \ldots; z_1^0, \ldots, z_K^0),$$

where $(z_1^0, \ldots, z_K^0) \in V \subset \mathbb{R}^{n_s}$ and $z_l^0 \in \mathbb{R}^{n_l}, \forall l \in \{1, 2, \ldots, K\}$.

Step 2. Definition of Operator B:

Given

$$z^{k-1} \in V_{D1} \left(\subset V_D \subset \mathbb{R}^{Kn_s}\right), \forall k-1 \in \kappa \subset \mathbb{N} \cup \{0\},$$

where $\mathbb{N}$ is the set of natural numbers. Here, $\kappa \triangleq \{0, 2, 4, 6, \ldots, 2m, \ldots\}$, where m is a non-negative integer, i.e., the set of non-negative even numbers.

Operator B is obtained as follows. It is a mapping from z^{k-1} to the optimal solution set of

$$\max_{x \in V_D} \left\{ F(x) \,\middle|\, \begin{array}{c} (x_{1,1}, x_{2,2}, \cdots, x_{K,K}) \in V; \\ x_{i,j} = z_j^{k-1}, \text{ for } i \neq j \text{ and } 1 \leq i, j \leq K \end{array} \right\},$$

an optimal solution belongs to $B\left(z^{k-1}\right) (\subset V_D)$ and we denote by z^k the optimal solution. Obviously, operator B is an algorithm over V_D. Thus it is called the block diagonal coordinate ascent algorithm. Note the difference in terminology between the block diagonal coordinate ascent algorithm and the diagonal block coordinate ascent algorithm.

Step 3. Definition of the Average Mapping:

Given $z^k \in B\left(z^{k-1}\right) \subset V_D, k-1 \in \kappa$, i.e., $k \notin \kappa$,

$$z_{[1:K]}^{k+1} \triangleq \left(\frac{1}{K}\sum_{j=1}^{K} z_{j,1}^k, \frac{1}{K}\sum_{j=1}^{K} z_{j,2}^k, \ldots, \frac{1}{K}\sum_{j=1}^{K} z_{j,K}^k\right) \in V \subset \mathbb{R}^{n_s}.$$

Thus, $z_{[1:K]}^{k+1} =$

$$\left(\frac{1}{K}z_{1,1}^k + \frac{K-1}{K}z_1^{k-1}, \frac{1}{K}z_{2,2}^k + \frac{K-1}{K}z_2^{k-1}, \ldots, \frac{1}{K}z_{K,K}^k + \frac{K-1}{K}z_K^{k-1}\right).$$

Furthermore,

$$z^{k+1} \triangleq \left(\overbrace{z_{[1:K]}^{k+1}, \ldots, z_{[1:K]}^{k+1}}^{K}\right) \in V_{D1}\,(\subset V_D),$$

where a mapping from z^k to z^{k+1} is formed naturally and it is called the average mapping and is written as Av. This step will be called the spacer step.

Step 4. Definition of Algorithm DBCAA:

Algorithm DBCAA is defined as the mapping product $Av \cdot B$.

Since f is concave,

$$\begin{aligned} F\left(z^k\right) &= \frac{1}{K}\sum_{\ell=1}^{K} f\left(z_{\ell,1}^k, z_{\ell,2}^k, \cdots, z_{\ell,K}^k\right) \\ &= \frac{1}{K}\left\{f\left(z_{1,1}^k, z_2^{k-1}, \cdots, z_K^{k-1}\right) + \cdots + f\left(z_1^{k-1}, \cdots, z_{K-1}^{k-1}, z_{K,K}^k\right)\right\} \\ &\leq f\left(\frac{1}{K}z_{1,1}^k + \frac{K-1}{K}z_1^{k-1}, \ldots, \frac{1}{K}z_{K,K}^k + \frac{K-1}{K}z_K^{k-1}\right) \\ &\leq \frac{1}{K}\sum_{\ell=1}^{K} f\left(\frac{1}{K}z_{1,1}^k + \frac{K-1}{K}z_1^{k-1}, \ldots, \frac{1}{K}z_{K,K}^k + \frac{K-1}{K}z_K^{k-1}\right) \\ &\leq F\left(z^{k+1}\right). \end{aligned}$$

Remark 4.4.1. *In Step 3 of the diagonal block coordinate ascent algorithm, the average mapping, which has already been denoted by Av, can be extended into* $Av: V_D \longrightarrow V_D$. *If,* $\forall x, x \in V_D$, *then* $Av(x) \in V_D$ *and* $F(x) \leq F(Av(x))$, *due to the definitions of* V_D *and* F.

Because Proposition 3.4.3 is only applicable to the block coordinate ascent algorithm (BCAA), a more general proposition than that in Proposition 3.4.3 will be introduced to help to establish convergence of the diagonal block coordinate ascent algorithm (DBCAA). This more generalized proposition will unify the foundation for convergence of Algorithm 3.1 and Algorithm 4.2.

The maximization mapping was defined previously by Definition 3.4.2. For example, algorithm B of the diagonal block coordinate ascent algorithm is the maximization mapping, when its optimal solution is unique. In detail, algorithm B, as a maximization mapping, is defined with the objective function $Z \triangleq F$, $\widetilde{V}_1 \triangleq V_{D1}$ and $\widetilde{V} \triangleq V_D$ in the diagonal block coordinate ascent algorithm.

To simplify the discussion of convergence of the DBCAA and its extension, the monotonically increasing mapping is defined as follows.

Definition 4.4.2 (Monotonically Increasing Mapping). *For a continuous function* $Z : \widetilde{V} \longrightarrow \mathbb{R}$, *if the mapping* $M : \widetilde{V} \longrightarrow \widetilde{V}$ *satisfies*

$$Z\left(M(v)\right) \geq Z\left(v\right),$$

then M *is said to be a monotonically increasing mapping.*

For instance, the identity mapping I and the average mapping, Av, of the diagonal block coordinate ascent algorithm are monotonically increasing mappings.

Given the definition of a monotonically increasing mapping, we have the following proposition to reveal the relationship between the accumulation point and the fixed point for more generalized algorithms.

Proposition 4.4.3. *Let the product mapping* $A \triangleq M_\ell O_\ell \cdots M_2 O_2 M_1 O_1$, *where the natural number* $\ell \leq K$, *determine an algorithm that given a point* z^0 *generates the sequence* $\left\{z^k\right\}_{k=0}^{\infty}$ *with* $z^{k+1} = A\left(z^k\right), \forall k$. *Suppose*

1. *All points* z^k *are in a compact set* $X \subset \widetilde{V}_1 \left(\subset \widetilde{V}\right)$,
2. $O_1 : \widetilde{V}_1 \longrightarrow \widetilde{V}$ *and* $O_m : \widetilde{V} \longrightarrow \widetilde{V}$, $\forall m > 1$, *are the maximization mappings, and each* $M_m : \widetilde{V} \longrightarrow \widetilde{V}$, $\forall m$, *is a monotonically increasing mapping,*
3. *For any accumulation point,* z^∞, *of* $\left\{z^k\right\}_{k=0}^{\infty}$, $M_m\left(z^\infty\right) = z^\infty$, $m \in \{1, 2, \cdots, l-1\}$, *and*

4. *the maximization mappings O_m and the monotonically increasing mappings $M_m, \forall m$, are continuous over their domains respectively.*

Then for the maximization mapping O_m, the accumulation point, z^∞, is a fixed point, i.e., $z^\infty = O_m\left(z^\infty\right), m = 1, 2, \ldots, \ell$.

Proof. Observing Condition 1, there must be a $\kappa \subset \mathbb{N} \cup \{0\}$ and a convergent subsequence such that $z^k \longrightarrow z^\infty$, for $k \in \kappa$. Using Condition 2, we see that

$$Z\left(z^{k+1}\right) \geq Z\left(z^k\right), \forall k \in \mathbb{N} \cup \{0\}.$$

So, $\left\{Z\left(z^k\right)\right\}_{k=0}^{\infty}$ is a monotonically increasing sequence. According to the limit property of the monotonic sequence, if the limit of some subsequence of the sequence exists and the sequence is monotonically increasing, then the limit of the sequence exists and the limit of the sequence is equal to the limit of the subsequence. Thus, $\lim_{k\to\infty} Z\left(z^k\right)$ exists and

$$\lim_{k\to\infty} Z\left(z^k\right) = \lim_{k\in\kappa} Z\left(z^k\right).$$

Since Z is continuous,

$$\lim_{k\to\infty} Z\left(z^k\right) = Z\left(z^\infty\right). \tag{4.7}$$

Due to the fact that $\left\{z^{k+1}\right\}_{k\in\kappa} \subset X$, where X is compact, $\exists \kappa^1 \subset \kappa$ such that $\lim_{k\in\kappa^1} z^{k+1}$ exists and the limit is written as y^∞, i.e., $\lim_{k\in\kappa^1} z^{k+1} = y^\infty$. Similar to the derivation mentioned above,

$$\lim_{k\to\infty} Z\left(z^k\right) = Z\left(y^\infty\right). \tag{4.8}$$

From (4.7), (4.8) and Condition 4,

$$Z\left(z^{\infty}\right)=Z\left(y^{\infty}\right)=Z\left(\lim_{k\in\kappa^{1}}A\left(z^{k}\right)\right)=Z\left(A\left(z^{\infty}\right)\right) \text{ implies } Z\left(z^{\infty}\right)=Z\left(A\left(z^{\infty}\right)\right).$$

For $m=1$, $Z\left(z^{\infty}\right)=Z\left(A\left(z^{\infty}\right)\right)\geq Z\left(O_{1}\left(z^{\infty}\right)\right)=Z\left(z^{\infty}\right)$. Then $Z\left(z^{\infty}\right)=Z\left(O_{1}\left(z^{\infty}\right)\right)$. Because z^{∞} is the feasible point related to O_1 and O_1 is the mapping from a point to a maximum point, corresponding to the definition of mapping, $O_{1}\left(z^{\infty}\right)=z^{\infty}$.

Assume that, as $1\leq m<\ell, O_{m}\left(z^{\infty}\right)=z^{\infty}$. Because

$$Z\left(z^{\infty}\right)=Z\left(A\left(z^{\infty}\right)\right)$$

$$\geq Z\left(O_{m+1}M_{m}O_{m}\cdots M_{1}O_{1}\left(z^{\infty}\right)\right)\geq Z\left(z^{\infty}\right)$$

and, due to the induced assumption and Condition 3,

$$Z\left(O_{m+1}M_{m}O_{m}\cdots M_{1}O_{1}\left(z^{\infty}\right)\right)=Z\left(O_{m+1}\left(z^{\infty}\right)\right), \text{ and } Z\left(z^{\infty}\right)=Z\left(O_{m+1}\left(z^{\infty}\right)\right).$$

Because z^{∞} is the feasible point related to O_{m+1} and O_{m+1} is a mapping from a feasible point to a maximum point, $O_{m+1}\left(z^{\infty}\right)=z^{\infty}$.

Therefore for the maximization mapping O_m, $\forall m$, the accumulation point, z^{∞}, is a fixed point, i.e., $z^{\infty}=O_{m}\left(z^{\infty}\right)$. □

Based on Proposition 4.4.3, we can introduce the following theorem on convergence of the diagonal block coordinate ascent algorithm.

Theorem 4.4.4. *Consider the abstract formulation in (4.2) and, assume that f is concave and differentiable, that V is convex and that either V is compact or the superlevel set $\{x|f(x) \geq f(z^0)\}$ is bounded. Now, if the mapping B is continuous over V_{D1}, then the confined preceding K block components, $z^\infty|_{[1:K]}$, of the limit of any convergent subsequence of $\{z^k\}_{k=0}^\infty$ generated by the diagonal block coordinate ascent algorithm is an optimal point of (4.2) and $\{f(z^k)\}_{k=0}^\infty$ approaches to the optimal value.*

Proof. Assume that the diagonal block coordinate ascent algorithm produces the sequence $\{z^k\}_{k=0}^\infty$ and z^∞ is the limit of a convergent subsequence of the sequence.

Due to the compactness of V or the boundedness of set $\{x|f(x) \geq f(z^0)\}$, Condition 1 of Proposition 4.4.3 is satisfied. If $M_1 : V_D \longrightarrow V_D$ is the average mapping and mapping O_1 is defined by mapping $B : V_{D1} \longrightarrow V_D$, then the diagonal block coordinate ascent algorithm, as an algorithm, is the same as the product mapping $A(= M_1O_1)$ and Condition 2 of Proposition 4.4.3 is satisfied. Condition 3 of Proposition 4.4.3 is satisfied due to $\ell = 1$. Because the average mapping is continuous over V_D, the continuity of mapping B is assumed, Condition 4 of Proposition 4.4.3 is satisfied. Then for maximization mapping B, the accumulation point, z^∞, is a fixed point, i.e., $z^\infty = B(z^\infty)$. According to the optimality condition of convex programming,

$$\left(F_{x_{1,1}}(z^\infty), F_{x_{2,2}}(z^\infty), \cdots, F_{x_{K,K}}(z^\infty)\right) \begin{pmatrix} z_{1,1} - z_1^\infty \\ z_{2,2} - z_2^\infty \\ \vdots \\ z_{K,K} - z_K^\infty \end{pmatrix} \leq 0, \tag{4.9}$$

where, $\forall (z_{1,1}, z_{2,2}, \cdots, z_{K,K}), (z_{1,1}, z_{2,2}, \cdots, z_{K,K}) \in V$.

So,

$$\left(f_{x_1}\left(z^{\infty}|_{[1:K]}\right), f_{x_2}\left(z^{\infty}|_{[1:K]}\right), \cdots, f_{x_K}\left(z^{\infty}|_{[1:K]}\right)\right) \begin{pmatrix} x_1 - z_1^{\infty} \\ x_2 - z_2^{\infty} \\ \vdots \\ x_k - z_K^{\infty} \end{pmatrix} \leq 0, \tag{4.10}$$

where, $\forall (x_1, x_2, \cdots, x_K), (x_1, x_2, \cdots, x_K) \in V$.

Hence,

$$f_x\left(z^{\infty}|_{[1:K]}\right)\left(x - z^{\infty}|_{[1:K]}\right) \leq 0, \forall x \in V.$$

Therefore, $z^{\infty}|_{[1:K]}$ is an optimal solution, corresponding to the assumption of convex programming from the generalized mathematical problem in (4.2), and, due to the monotonicity of sequence $\left\{f\left(z^k\right)\right\}_{k=1}^{\infty}$, the sequence approaches to the optimal value.

□

It is a meaningful question how to guarantee that algorithm $B : V_{D1} \longrightarrow V_D$ in the definition of the diagonal block coordinate ascent algorithm is a mapping and satisfies the continuity over V_{D1}. It is immediate that if the maximization programming, corresponding to algorithm B, has a unique optimal solution, then algorithm B is a mapping, i.e., it is a mapping from a point to a point, and algorithm B satisfies the uniqueness condition. In fact, it can be proved that if the uniqueness holds, then mapping B also satisfies the continuity, i.e., it is continuous over V_{D1}. Below, we will

first show the uniqueness of algorithm B. Then the continuity of algorithm B will be proved.

Definition 4.4.5 (Diagonal Uniqueness Condition). $\forall z^{k-1} \in V_{D1}$, *If the optimal solution set*

$$arg\left(\max_{x \in V} \left\{ \frac{1}{K} \sum_{\ell=1}^{K} f\left(z_1^{k-1}, \ldots, z_{\ell-1}^{k-1}, x_\ell, z_{\ell+1}^{k-1}, \ldots, z_K^{k-1}\right) \right\} \right) \subset V$$

is a single point set, then the average sum function $F(x)$ in (4.6) of $f(x)$ is said to satisfy the diagonal uniqueness condition over set V_{D1}.

The name "Diagonal Uniqueness Condition" stems from positions of the optimized block variables, which match with the diagonal blocks of a blocked matrix.

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{K,1} & x_{K,2} & \cdots & x_{K,K} \end{pmatrix}.$$

The meaning of an entry $x_{\ell,j}$ $(\forall \ell, j)$, as a block of the blocked matrix, is the same as in (4.6).

Based on Definition 4.4.5, we have the following lemma on the continuity of the diagonal block coordinate ascent algorithm.

Lemma 4.4.6. *Suppose that $f(x)$ is continuous and $F(x)$ satisfies the diagonal uniqueness condition over set V_{D1}, and suppose that V is closed. If V is bounded*

or set

$\{x \in V | f(x) \geq f(z^0)\}$ *is bounded, then the algorithmic mapping* B *is continuous.*

Proof. $\forall \overline{z}, \overline{z} \in V_{D1}$. Assume that $\left\{z^k\right\} \subset V_{D1}$ and $\lim_{k\to\infty} z^k = \overline{z}$. $v^k \triangleq B\left(z^k\right), \overline{w} \triangleq B\left(\overline{z}\right)$. $\left\{v^k\right\}$ has an accumulation point denoted by $\overline{v}$ due to the boundedness of V or set $\{x \in V | f(x) \geq f(z^0)\}$. If $\left\{v^k\right\}$ does not converge, a subsequence $\left\{v^{k_r}\right\}$, which converges to $\overline{v}$ as r tends to infinity, can be selected from $\left\{v^k\right\}$. Hence, two convergent subsequences $\left\{v^{k_r}\right\}$ and $\left\{z^{k_r}\right\}$ can be acquired. So, without loss of generality, assume that $\left\{v^k\right\}$ converges to $\overline{v}$. From the definition of B, it holds that $\overline{v}_{\ell,j} = \lim_{k\to\infty} v^k_{\ell,j} = \lim_{k\to\infty} z^k_{\ell,j} = \overline{z}_{\ell,j}$ $(\ell, j \in \{1, \ldots, K\} \cap \{\ell \neq j\})$. Since V is closed and $\left\{v^k\right\} \subset V_D, \overline{v} \in V_D$. Due to $\overline{w} \triangleq B\left(\overline{z}\right)$ and $\overline{v}_{\ell,j} = \overline{z}_{\ell,j} = \overline{w}_{\ell,j}$, $(\ell, j \in \{1, \ldots, K\} \cap \{\ell \neq j\})$,

$$F(\overline{w}) \geq F(\overline{v}). \tag{4.11}$$

On the other hand, construct a sequence as follows.

$w^k \in \mathbb{R}^{Kn_s}, \forall k$, is defined by following two steps.

- $w^k_{\ell,j} \triangleq \overline{w}_{\ell,j}$, for $\ell = j$
- $w^k_{\ell,j} \triangleq z^k_{\ell,j}$, for $\ell \neq j$

Thus, $w^k \in V_D, \forall k$, is obtained.
It is easily seen that $\left\{w^k\right\} \subset V_D$ and $\lim_{k\to\infty} w^k = \overline{w}$. Hence,

$$F(\overline{v}) \geq F(\overline{w}). \tag{4.12}$$

From (4.11) and (4.12), $F(\overline{v}) = F(\overline{w}) = F(B\left(\overline{z}\right))$. When the consideration of

the diagonal uniqueness condition is being added into the above equalities, $\overline{v} = \overline{w}$ holds, i.e., $\lim_{k\to\infty} B\left(z^k\right) = \lim_{k\to\infty} v^k = \overline{v} = \overline{w} = B\left(\overline{z}\right)$. Thus, B is continuous over V_{D1}. □

Corollary 4.4.7. *Consider the abstract formulation in (4.2). Assume that f is concave and differentiable, that the average sum function $F(x)$ in (4.6) satisfies the diagonal uniqueness condition over set V_{D1}, that V is convex and that either V is compact or the superlevel set $\{x \in V | f(x) \geq f(z^0)\}$ is bounded. Then the confined preceding K block components, $z^\infty|_{[1:K]}$, of the limit of any convergent subsequence of $\left\{z^k\right\}_{k=0}^{\infty}$ generated by the diagonal block coordinate ascent algorithm is an optimal point of (4.2) and $\left\{f\left(z^k\right)\right\}_{k=0}^{\infty}$ approaches to the optimal value.*

Remark 4.4.8. *Proposition 4.4.3 offers a unified structure for the block coordinate ascent algorithm and the diagonal block coordinate ascent algorithm. Under this structure, the block coordinate ascent algorithm and the diagonal block coordinate ascent algorithm are considered as two cases of the same theoretic framework.*

4.4.2 Convergence of IWFA under the Sum Power Constraint

To simplify the discussion of convergence of the DBCAA and Algorithm 4.2, for the problem in (4.1), we define

$$f\left(Q_1, Q_2, \cdots, Q_K\right) \triangleq \log\left(\det\left(I + \sum_{i=1}^{K} H_i^\dagger Q_i H_i\right)\right),$$

$$V = \left\{\left(Q_1, Q_2, \cdots, Q_K\right) \middle| Q_i \succeq 0, \forall i, \sum_{i=1}^{K} \mathrm{Tr}\left(Q_i\right) \leq P\right\},$$

and $V_{D1} =$

$$\left\{ \overbrace{(Q_1, Q_2, \cdots, Q_K; Q_1, Q_2, \cdots, Q_K; \ldots; Q_1, Q_2, \cdots, Q_K)}^{K^2} \middle| (Q_1, Q_2, \cdots, Q_K) \in V \right\}.$$

Here, the diagonal block coordinate ascent algorithm with the spacer step, as applied to the problem in (4.1), corresponds to Algorithm 4.2. The number of entries of Q_i is equivalent to n_i in the definition of the DBCAA. Let us define

$$\forall \left(\overline{Q}_1, \overline{Q}_2, \cdots, \overline{Q}_K\right) \in V, x|_{[1:K]} \triangleq \left(\overline{Q}_1, \overline{Q}_2, \cdots, \overline{Q}_K\right),$$

where $\overline{Q}_t$ denotes the t-th input covariance at the previous iteration, and

$$x = \overbrace{\left(x|_{[1:K]}, \cdots, x|_{[1:K]}\right)}^{K} \in V_{D1}.$$

$$G_i \triangleq H_i \left(I_m + \sum_{t=1, t\neq i} H_t^\dagger \overline{Q}_t H_t \right)^{-\frac{1}{2}}.$$

We know that there is a unitary matrix U_i such that

$$U_i^\dagger \left(G_i G_i^\dagger \right) U_i$$

is a diagonal matrix, written as Λ_i, and

$$\Lambda_i \triangleq \operatorname{diag}\left((\Lambda_i)_{1,1}, (\Lambda_i)_{2,2}, \cdots, (\Lambda_i)_{n,n} \right) \in \mathbb{R}^{n\times n} \subset \mathbb{C}^{n\times n},$$

where

$$(\Lambda_i)_{\ell,\ell} \geq (\Lambda_i)_{\ell+1,\ell+1} \, (\forall \ell).$$

For some i, the set

$$\left\{\ell |\, (\Lambda_i)_{\ell,\ell} > 0, \ell \in \{1, 2, \ldots, n\}\right\}$$

is not empty due to the assumption $H_i \neq 0$ made just before the definition of Algorithm 4.1 and, $\forall i$,

$$\ell_{max}(i) \triangleq \max\left\{\ell |\, (\Lambda_i)_{\ell,\ell} > 0, \ell \in \{1, 2, \ldots, n\}\right\}.$$

Given $\{U_i\}_{i=1}^K$,

$$V^1 \triangleq \left\{\left(U_1^\dagger Q_1 U_1, U_2^\dagger Q_2 U_2, \cdots, U_K^\dagger Q_K U_K\right) \mid (Q_1, Q_2, \cdots, Q_K) \in V\right\},$$

and it is easy to see that V^1 is convex and $V^1 \subset V$.

$$V^2 \triangleq \left\{ (Q_1, Q_2, \cdots, Q_K) \in V \middle| \begin{array}{c} (Q_i)_{s,t} = 0, \text{ for } s \neq t; \\ (Q_i)_{s,s} = 0, \forall s \in \{\ell_{max}(i) + 1, \ldots, n\}; \forall i \end{array} \right\}$$

and it is true that V^2 is convex and $V^2 \subset V^1$. The objective

$$F(x) = \frac{1}{K} \sum_{i=1}^{K} \log\left(\det\left(I_m + \sum_{l=1, \ell \neq i} H_l^\dagger \overline{Q}_\ell H_\ell + H_i^\dagger Q_i H_i\right)\right),$$

where $x = \left(Q_1, \overline{Q}_2, \cdots, \overline{Q}_K; \cdots; \overline{Q}_1, \cdots, \overline{Q}_{K-1}, Q_K\right)$ and $(Q_1, Q_2, \cdots, Q_K) \in V$. It is easy to see that the assumed $f(Q_1, Q_2, \cdots, Q_K)$ (the definition of f is referred to in the beginning of this subsection) is concave and continuously differentiable due to the concavity of $\log\det(\cdot)$ (refer to [6], page 74). It is also easy to see the existence of isomorphism between S_i and $(\mathbb{R}e(S_i), \mathbb{I}m(S_i)), \forall i$, and that V_{D1} and V are convex

and closed. Hence, the first condition of Corollary 4.4.7 is satisfied. The boundedness of V is to be proven as follows.

Let $(Q_1, Q_2, \cdots, Q_K) \in V$ hold. $\forall Q_i, Q_i = E\left[x^i \left(x^i\right)^\dagger\right]$ implies $(Q_i)_{j,j} \leq P$. Thus, using the Cauchy-Schwartz inequality,

$$\begin{aligned} \left|\left(Q_i\right)_{s,t}\right| &= \left|E\left[\left(x_s^i\right)\left(\overline{x_t^i}\right)\right]\right| \\ &\leq \sqrt{E\left[|x_s^i|^2\right] E\left[|x_t^i|^2\right]} \\ &\leq \sqrt{\left(Q_i\right)_{s,s}\left(Q_i\right)_{t,t}}. \end{aligned}$$

So $\left|\left(Q_i\right)_{s,t}\right| \leq P$ $(\forall i, j, s \text{ and } t)$.

Hence V is bounded, and hence compact. Another condition of Corollary 4.4.7 is now also satisfied. We will now define a function f_i to be the simplified objective function that corresponds to the case when the i-th user is being optimized while other users are kept unchanged under the sum power constraint.

$$f_i\left(\left(Q_i\right)_{1,1}, \left(Q_i\right)_{2,2}, \cdots, \left(Q_i\right)_{\ell_{max}(i),\ell_{max}(i)}\right) \triangleq \sum_{\ell=1}^{\ell_{max}(i)} \log\left(1 + \left(\Lambda_i\right)_{\ell,\ell}\left(Q_i\right)_{\ell,\ell}\right) \quad (\forall i).$$

Thus,

$$\max_{(Q_1,\cdots,Q_K)\in V} \left\{ \frac{1}{K}\sum_{i=1}^{K} \log\left(\det\left(I_m + \sum_{t=1,t\neq i}^{K} H_t^\dagger \overline{Q}_t H_t + H_i^\dagger Q_i H_i\right)\right) \,\middle|\, \begin{array}{c} \overline{Q}_t \text{ is} \\ \text{given} \end{array} \right\} \quad (4.13)$$

$$\Longleftrightarrow \quad \max_{(Q_1,\cdots,Q_K)\in V^1} \sum_{i=1}^{K} \log\left(\det\left(I_n + \Lambda_i^{\frac{1}{2}} Q_i \Lambda_i^{\frac{1}{2}}\right)\right) \tag{4.14}$$

$$\Longleftrightarrow \quad \max_{(Q_1,\cdots,Q_K)\in V^2} \sum_{i=1}^{K} \sum_{\ell=1}^{\ell_{max}(i)} \log\left(1 + (\Lambda_i)_{\ell,\ell} (Q_i)_{\ell,\ell}\right). \tag{4.15}$$

Because the Hessian matrix of $\sum_{i=1}^{K} f_i$, i.e., $\sum_{i=1}^{K} \sum_{\ell=1}^{l_{max}(i)} \log\left(1 + (\Lambda_i)_{\ell,\ell} (Q_i)_{\ell,\ell}\right)$, is

$-\text{diag}$

$$\left(\frac{(\Lambda_1)_{1,1}^2}{\left(1 + (\Lambda_1)_{1,1} (Q_1)_{1,1}\right)^2}, \cdots, \frac{(\Lambda_K)_{\ell_{max}(K),\ell_{max}(K)}^2}{\left(1 + (\Lambda_K)_{\ell_{max}(K),\ell_{max}(K)} (Q_K)_{\ell_{max}(K),\ell_{max}(K)}\right)^2} \right)$$

$$\in \mathbb{R}^{\left(\sum_{i=1}^{K} \ell_{max}(i)\right) \times \left(\sum_{i=1}^{K} \ell_{max}(i)\right)} \subset C^{\left(\sum_{i=1}^{K} \ell_{max}(i)\right) \times \left(\sum_{i=1}^{K} \ell_{max}(i)\right)},$$

it is a negative definite matrix, and the cardinality of the optimal solution set

$$\arg\left(\max_{(Q_1,\cdots,Q_K)\in V^2} \sum_{i=1}^{K} \sum_{\ell=1}^{\ell_{max}(i)} \log\left(1 + (\Lambda_i)_{\ell,\ell} (Q_i)_{\ell,\ell}\right) \right)$$

is one, i.e.,

$$\left(\arg\left(\max_{(Q_1,\cdots,Q_K)\in V^2} \sum_{i=1}^{K} \sum_{\ell=1}^{\ell_{max}(i)} \log\left(1 + (\Lambda_i)_{\ell,\ell} (Q_i)_{\ell,\ell}\right) \right) \right)^{\sharp} = 1.$$

Therefore, it is easily obtained that the cardinality of the optimal solution set of

$$\max_{(Q_1,\cdots,Q_K)\in V} \left\{ \sum_{i=1}^{K} \log\left(\det\left(I_m + \sum_{t=1,t\neq i}^{K} H_t^{\dagger}\overline{Q}_t H_t + H_i^{\dagger} Q_i H_i \right)\right) \middle| \begin{array}{c} \overline{Q}_t \text{ is} \\ \text{given} \end{array} \right\}$$

is equal to one, corresponding to the equivalence (4.15) of the optimization problems and the definitions of V, V^1 and V^2. An important consequence is that the diagonal uniqueness condition of Corollary 4.4.7 is obtained.

The final optimization problem in (4.15) can be easily solved by some optimization algorithms, for instance, as well known, the conventional water-filling algorithm, or the water-filling algorithm with Fibonacci search presented in Chapter 2. According to Corollary 4.4.7, convergence of Algorithm 4.2 is acquired. If $\exists H_i = 0$ (the null matrix), a guarantee of convergence of Algorithm 4.2 to the global optimal solution is obtained from the above derivation and the compactness of the feasible set.

According to the discussion of convergence of Algorithm 4.2, the following corollary is obtained.

Corollary 4.4.9. *For the dual MIMO MAC problem (or the problem in (4.1)) of the MIMO BC model, the iterative water-filling algorithm (with or without Fibonacci search) under the sum power constraint (i.e., Algorithm 4.2) converges to the optimal solution of the dual MIMO MAC problem. In addition, as the number of iterative steps increases, the corresponding objective value approaches to the optimal objective*

value of (4.1).

In other words, the optimal solution of (4.1) can be found by the iterative water-filling algorithm (Algorithm 4.2) with or without Fibonacci search.

Chapter 5

Conclusions and Future Work

In this thesis, three new water-filling algorithms with Fibonacci search are presented for finding optimal input covariances for the single-user MIMO channel, the MIMO multiple access channel (MAC) and the MIMO broadcast channel (BC). Rigorous convergence proofs for the algorithms are also established.

In detail, Chapter 2 of this thesis presents the Fibonacci search method for speeding up the computation of optimal solutions to the single-user problem. It also describes how the optimal input covariance is constructed, and offers a formal proof of optimality for the constructed solutions. Because the algorithms for the MIMO MAC and the MIMO BC are iterative algorithms in which a single-user problem is solved at each step, the Fibonacci search and the proof of optimality together allow us to find solutions to these multiuser problems more conveniently and efficiently.

In Chapter 3 we studied iterative water-filling algorithms for the multiuser MIMO MAC. In particular, we exposed a weakness in previous convergence proof proposed

by others. Then we established a rigorous proof of convergence for the iterative water-filling algorithm based on fixed point theory. This was done via a complex-valued framework that better matches practical communication systems than the real-valued framework used in previous work. Also, we showed how Fibonacci search could be incorporated into the iterative water-filling algorithm for the MIMO MAC to reduce the computational cost.

In Chapter 4 we studied iterative water-filling algorithms for the multiuser MIMO BC. We explained that there is a so-called dual MIMO MAC for the MIMO BC, and that this dual is equivalent from the sum rate perspective, in the sense that the sum rates are equal and the optimal input covariances can be computed from each other, e.g., [20]. The dual MIMO MAC is of significant interest because the sum rate optimization problem is convex, whereas that for the MIMO BC is not. However, the extension of the iterative water-filling algorithm for the MIMO MAC to the dual MIMO MAC of the MIMO BC is not straightforward (e.g., [20]), because the power constraint couples the stages of the iterative water-filling algorithm, whereas those stages are not coupled in the MAC case. In our analysis, we exposed a weakness in the previous convergence proof proposed by others. Then, we established a rigorous proof of convergence of the iterative water-filling algorithm for the case of the MIMO BC, based on an extended fixed point theory. Also, we showed how Fibonacci search could be incorporated into the iterative water-filling algorithms for the MIMO BC to reduce the computational cost.

On a more general level, this thesis sets up a unified theoretical framework, based

on generalized fixed point theory, for the discussion of convergence of iterative algorithms for the optimal input covariances for MIMO systems. Under this framework, a rigorous convergence proof for each of the constructed algorithms was presented.

Although we managed to answer some of the basic algorithmic questions concerning iterative water-filling algorithms, a number of questions and directions remain open for future work, two of which we will highlight below.

The utilization of the iterative water-filling algorithm with Fibonacci search for the optimization of the sum rate of a MIMO relay system (RS) will be proposed and studied. The MIMO RS of interest consists of a single source-destination pair, with multiple half-duplex decode-and-forward MIMO relays and no direct path from the source to the destination. This system is the concatenation of a MIMO BC (source to relays) and a MIMO MAC (relays to destination). In order to find the solution to the problem of optimizing the rate of the MIMO RS, a hybrid iterative water-filling algorithm with Fibonacci search will be presented, which will combine the iterative water-filling algorithm for the MIMO BC with that for the MIMO MAC. Convergence of this hybrid algorithm will be studied and proved based on our discussions in Chapters 2, 3 and 4, under the unified framework established in this thesis.

The reader might recall from previous discussions that the Fibonacci search can speed up the computation of water-filling, which is one step of the iterative water-filling algorithms (IWFAs). In fact, it may also improve the performance of other aspects of IWFAs. For example, we can replace the simple spacer step in the existing

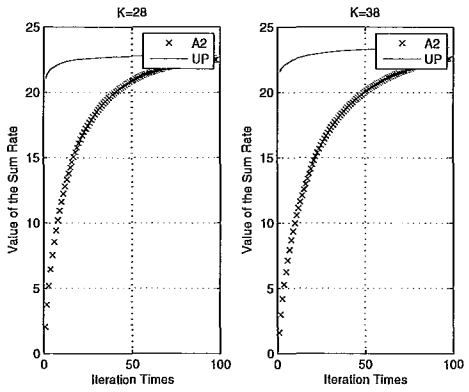


Figure 5.1: Performance of UP Compared with A2, as K=28 and 38

IWFAs for the MIMO BC (cf. [20]) by a more sophisticated spacer step inspired by Fibonacci search. Although the detailed machinery is still being researched, some preliminary numerical experiments, such as that provided below, suggest that this new spacer step can significantly improve the convergence rate of IWFA algorithms for the MIMO BC.

Consider a MIMO BC with $m = 8$ transmitter antennas, $n = 8$ receiver antennas at each of the $K = 28$ and $K = 38$ receivers, and a sum power bound of $P = 2$. The convergence of the sum rate for Algorithm 2 from [20] (Algorithm 4.2 in this thesis), denoted A2, and that for the algorithm with the Fibonacci-search-inspired spacer step, denoted UP, are plotted in Figure 5.1 for the case of a single realization of channel matrices from the standard i.i.d. Rayleigh distribution. As can be seen from this figure, the proposed algorithm converges much more quickly than Algorithm 2 of [20].

In conclusion, we would like to point out that the quest for better algorithms for the sum rate optimization problem in MIMO communications system has led to some

important progress in optimization theory, in addition to significant successes in practice. As this thesis provides ample evidence, the iterative water-filling algorithm can be efficiently performed for a wide variety of MIMO communications systems, and this now raises the hope that the analogous pursuit of constructive capacity-approaching algorithms for some of the other MIMO communications systems discussed above might actually be tractable. The end result of such a pursuit, if it is successful, will have a significant impact in practice, but in addition we believe that there are also more algorithmic techniques to be discovered en route.

Chapter 6

Appendix

6.1 Appendix-I: Complex Gaussian Random Vectors

For any $z \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times m}$, let us define

$$\widehat{z} = \begin{pmatrix} \mathbb{R}e\,(z) \\ \mathbb{I}m(z) \end{pmatrix}$$

and

$$\widehat{A} = \begin{pmatrix} \mathbb{R}e\,(A) & -\mathbb{I}m(A) \\ \mathbb{I}m(A) & \mathbb{R}e\,(A) \end{pmatrix}.$$

A complex random vector $\xi \in \mathbb{C}^n$ is said to be Gaussian if the real random vector

$\widehat{\xi} \in \mathbb{R}^{2n}$ consisting of the real and imaginary parts of ξ,

$$\widehat{\xi} = \begin{pmatrix} \mathbb{R}e\,(\xi) \\ \mathbb{I}m\,(\xi) \end{pmatrix},$$

is Gaussian [35, 39, 40]. In fact, any complex random vector, without restricting to a complex Gaussian random vector, has such a real and expanded random vector. The relationship between these two complex and real vectors is a one to one mapping. Let us recall

$$E\left[\widehat{\xi}\right] \triangleq \int_{\mathbb{R}^{2n}} x dP_{\widehat{\xi}}(x) \in \mathbb{R}^{2n}$$

and

$$E\left[\left(\widehat{\xi} - E\left[\widehat{\xi}\right]\right)\left(\widehat{\xi} - E\left[\widehat{\xi}\right]\right)^T\right] \triangleq \int_{\mathbb{R}^{2n}} \left(x - E\left[\widehat{\xi}\right]\right)\left(x - E\left[\widehat{\xi}\right]\right)^T dP_{\widehat{\xi}}(x)$$

$\in \mathbb{R}^{2n\times 2n}$, the former is called the mean of $\widehat{\xi}$ and the latter is called the covariance of $\widehat{\xi}$. According to standard Lebesgue integration [29] on $\mathbb{R}^{2n}$, the mean and covariance of $\widehat{\xi}$ can be found respectively. Thus, to specify the distribution of a complex Gaussian random vector ξ, it is necessary to specify the mean and covariance of $\widehat{\xi}$, namely,

$$E\left[\widehat{\xi}\right] \text{ and } E\left[\left(\widehat{\xi} - E\left[\widehat{\xi}\right]\right)\left(\widehat{\xi} - E\left[\widehat{\xi}\right]\right)^T\right].$$

The definitions of the mean and covariance are also suitable for the case of any complex random vector.

According to standard Lebesgue integration [29] on $\mathbb{C}^n$, mean μ and covariance Q

of ξ can be defined as follows.

$$\mu = E[\xi], \quad \text{where } E[\xi] \triangleq \int_{\mathbb{C}^n} x dP_\xi(x) \in \mathbb{C}^n.$$

$$Q = E\left[(\xi - \mu)(\xi - \mu)^\dagger\right],$$

where

$$E\left[(\xi - \mu)(\xi - \mu)^\dagger\right] \triangleq \int_{\mathbb{C}^n} (x - \mu)(x - \mu)^\dagger dP_\xi(x) \in \mathbb{C}^{n \times n}.$$

A complex Gaussian random vector ξ is said to be circularly symmetric [27, 26, 35] if the covariance of the corresponding vector $\widehat{\xi}$ has the structure

$$E\left[\left(\widehat{\xi} - E\left[\widehat{\xi}\right]\right)\left(\widehat{\xi} - E\left[\widehat{\xi}\right]\right)^T\right] = \frac{1}{2}\begin{pmatrix} \mathbb{R}e(Q) & -\mathbb{I}m(Q) \\ \mathbb{I}m(Q) & \mathbb{R}e(Q) \end{pmatrix} \tag{6.1}$$

for some Hermitian positive semidefinite matrix $Q \in \mathbb{C}^{n \times n}$. Note that the real part of a Hermitian matrix is symmetric, and the imaginary part of a Hermitian matrix is skew-symmetric. Thus the matrix appearing in (6.1) is real and symmetric. In this case $E\left[(\xi - E[\xi])(\xi - E[\xi])^\dagger\right] = Q$, and thus, a circularly symmetric complex Gaussian random vector ξ is specified by its mean and variance.

Let ξ be a circularly symmetric complex Gaussian random vector. Then the probability density function (with respect to the Radon-Nikodym derivative of the standard Lebesgue measure [29] on $\mathbb{C}^n$) of a circularly symmetric complex Gaussian random vector with mean μ and covariance Q is derived by the following:

Due to the definition of ξ and the relationship between $\widehat{\xi}$ and ξ,

$$f_{\widehat{\xi}}\left(\widehat{x};\widehat{\mu},\widehat{Q}\right) = \frac{1}{(2\pi)^{\frac{2n}{2}}\left[\det\left(\frac{1}{2}\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\widehat{x}-\widehat{\mu}\right)^{\dagger}\left(\frac{1}{2}\widehat{Q}\right)^{-1}\left(\widehat{x}-\widehat{\mu}\right)\right\}$$
$$= \frac{1}{\pi^{n}\left[\det\left(\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\left(\widehat{x}-\widehat{\mu}\right)^{\dagger}\left(\widehat{Q}\right)^{-1}\left(\widehat{x}-\widehat{\mu}\right)\right\}.$$

The next step is to simplify the exponent part from the previous exponent function in order to finally remove the symbol `` $\widehat{\ }$ ". This can be done by utilizing the following algebraic property.

Due to $C = A^{-1}$ being equivalent to $\widehat{C} = \left(\widehat{A}\right)^{-1}$, we have

$$\frac{1}{\pi^{n}\left[\det\left(\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\left(\widehat{x}-\widehat{\mu}\right)^{\dagger}\left(\widehat{Q}\right)^{-1}\left(\widehat{x}-\widehat{\mu}\right)\right\}$$
$$= \frac{1}{\pi^{n}\left[\det\left(\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\left(\widehat{x}-\widehat{\mu}\right)^{\dagger}\widehat{\left(Q^{-1}\right)}\left(\widehat{x}-\widehat{\mu}\right)\right\}.$$

The next step is to remove the symbol `` $\widehat{\ }$ " in the preceding exponent function. Due to $z = x + y$ being equivalent to $\widehat{z} = \widehat{x} + \widehat{y}$, $\mathbb{R}e\left(x^{\dagger}y\right) = \widehat{x}^{\dagger}\widehat{y}$ and $y = Ax$ being equivalent to $\widehat{y} = \widehat{A}\widehat{x}$, we have

$$\frac{1}{\pi^{n}\left[\det\left(\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\left(\widehat{x}-\widehat{\mu}\right)^{\dagger}\widehat{\left(Q^{-1}\right)}\left(\widehat{x}-\widehat{\mu}\right)\right\}$$
$$= \frac{1}{\pi^{n}\left[\det\left(\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\left(x-\mu\right)^{\dagger}Q^{-1}\left(x-\mu\right)\right\}.$$

We may remove the symbol`` $\widehat{\ }$" in the determinant: Due to $\det\left(\widehat{A}\right) = |\det(A)|^{2}$, we have

$$\frac{1}{\pi^n\left[\det\left(\widehat{Q}\right)\right]^{\frac{1}{2}}}\exp\left\{-\left(x-\mu\right)^{\dagger}Q^{-1}\left(x-\mu\right)\right\}$$

$$=\frac{1}{\pi^n\det(Q)}\exp\left\{-\left(x-\mu\right)^{\dagger}Q^{-1}\left(x-\mu\right)\right\}.$$

Therefore,

$$\begin{aligned}f_{\widehat{\xi}}\left(\widehat{x};\widehat{\mu},\widehat{Q}\right)&=\frac{1}{\pi^n\det\left(Q\right)}\exp\left\{-\left(x-\mu\right)^{\dagger}Q^{-1}\left(x-\mu\right)\right\}\\&=\det\left(\pi Q\right)^{-1}\exp\left\{-\left(x-\mu\right)^{\dagger}Q^{-1}\left(x-\mu\right)\right\}.\end{aligned}$$

According to the uniqueness of the Radon-Nikodym derivative [29] and the correspondence between $\mathbb{R}^{2n}$ and $\mathbb{C}^n$, we have

$$f_{\xi}\left(x;\mu,Q\right)=\det\left(\pi Q\right)^{-1}\exp\left\{-\left(x-\mu\right)^{\dagger}Q^{-1}\left(x-\mu\right)\right\},$$

as the probability density function of ξ.

Remark 6.1.1. *For [35], it only uses (4d) and (4h) and thus it only can obtain the probability density distribution of the real random vector of a complex random vector. Only due to the uniqueness of the Radon-Nikodym derivative, mentioned above, in measure theory, may we acquire the probability density distribution of the complex random vector.*

6.2 Appendix-II: Maximum of Entropy

The channel capacity is dependent on the definition of the mutual information. At the same time the mutual information can be computed also by introducing the differential entropy and the conditional entropy. Therefore, the mutual information,

the differential entropy and the conditional entropy are revisited, referring to [10]. If familiarity of these concepts are assumed, we may skip them over to (6.2).

Definition 6.2.1 (Mutual Information). *Assume that $\xi \in \mathbb{C}^n$ and $\eta \in \mathbb{C}^n$ are two complex continuous random vectors, and $p(x)$ and $p(y)$ are the corresponding probability density functions. The mutual information $\mathbb{I}(\xi;\eta)$ between the two random vectors is defined as:*

$$\mathbb{I}(\xi;\eta) \triangleq \int_{\mathbb{C}^n}\int_{\mathbb{C}^n} p(x)\, p(y|x) \log \frac{p(x)\, p(y|x)}{p(x)\, p(y)} dxdy,$$

where $p(y|x)$ denotes the conditional probability density function.

In information theory, the mutual information of two random variables (or vectors) is a quantity that measures the mutual dependence of the two variables (or vectors).

In information theory, the following concept of the differential entropy is measurement for the entropy of a random variable (or vector).

Definition 6.2.2 (Differential Entropy). *Assume that ξ is a complex continuous random vector, and $p(x)$ is the corresponding probability density function. Then*

$$\mathbb{H}(\xi) \triangleq -\int_{\mathbb{C}^n} p(x) \log p(x)\, dx$$

is called the differential entropy of ξ.

In information theory, the conditional entropy quantifies the remaining entropy of a random variable (or vector) ξ given that the value of a second random variable (or

vector) η is known.

Definition 6.2.3 (Conditional Entropy). *Assume that ξ and η are two complex continuous random vectors, $p(x)$ and $p(y)$ are the corresponding probability density functions and $p(x,y)$ is the corresponding joint probability density function. Then the conditional entropy of ξ for given η is defined as:*

$$\mathbb{H}(\xi|\eta) \triangleq \int_{\mathbb{C}^n}\int_{\mathbb{C}^n} p(x,y)\log p(x|y)\,dxdy.$$

The following proposition offers the mathematical relationship among the mutual information, the differential entropy and the conditional entropy.

Proposition 6.2.4. *Assume that ξ and η are two continuous random vectors. Then*

$$\mathbb{I}(\xi;\eta) = \mathbb{H}(\xi) - \mathbb{H}(\xi|\eta)$$

and

$$\mathbb{I}(\xi;\eta) = \mathbb{H}(\eta) - \mathbb{H}(\eta|\xi).$$

The differential entropy of a complex Gaussian variable (or vector) ξ with mean μ and covariance Q is derived as follows. Due to the definition of $\mathbb{H}(\xi;\mu,Q)$, we have

$$\mathbb{H}(\xi;\mu,Q) = E_\xi\left[-\log f_\xi(\xi;\mu,Q)\right], \tag{6.2}$$

where E_ξ is the expectation operator of ξ, i.e., $E_\xi[\xi] \triangleq \int_{\mathbb{C}^n} x p_\xi(x)\,dx$.

Due to the form of the probability density function of the circularly symmetric complex Gaussian random vector ξ, we have

$$E_\xi \left[-\log f_\xi \left(\xi; \mu, Q \right) \right] = \log \det \left(\pi Q \right) + E \left[\left(\xi - \mu \right)^\dagger Q^{-1} \left(\xi - \mu \right) \right].$$

Due to the definition and basic properties of the trace operator, we may write

$$\log \det \left(\pi Q \right) + E \left[\left(\xi - \mu \right)^\dagger Q^{-1} \left(\xi - \mu \right) \right]$$
$$= \log \det \left(\pi Q \right) + E \left[\mathrm{Tr} \left(\left(\xi - \mu \right) \left(\xi - \mu \right)^\dagger Q^{-1} \right) \right].$$

Due to the commutative property for the product of the trace and expectation operators, we have

$$\log \det \left(\pi Q \right) + E \left[\mathrm{Tr} \left(\left(\xi - \mu \right) \left(\xi - \mu \right)^\dagger Q^{-1} \right) \right]$$
$$= \log \det \left(\pi Q \right) + \mathrm{Tr} \left(E \left[\left(\xi - \mu \right) \left(\xi - \mu \right)^\dagger \right] Q^{-1} \right).$$

Then the definition of the covariance of ξ implies that

$$\log \det \left(\pi Q \right) + \mathrm{Tr} \left(E \left[\left(\xi - \mu \right) \left(\xi - \mu \right)^\dagger \right] Q^{-1} \right) = \log \det \left(\pi Q \right) + \mathrm{Tr} \left(Q Q^{-1} \right),$$

and using the definition of the logarithm function and the fact,
$e \triangleq \lim_{m \to \infty} \left(1 + \frac{1}{m} \right)^m$, we have

$$\log \det \left(\pi Q \right) + \mathrm{Tr} \left(Q Q^{-1} \right) = \log \det \left(\pi Q \right) + \log e^n.$$

This can be simplified to

$$\log\det(\pi Q) + \log e^n = \log\det(\pi e Q).$$

The following proposition, which states that a circularly symmetric complex Gaussian variable (or vector) is the entropy maximizer, highlights the importance of circularly symmetric complex Gaussian vectors. [35] also claims this proposition. For proving that a circularly symmetric complex Gaussian variable (or vector) is the entropy maximizer, [35] uses the argument, i.e., $\log\gamma_Q(x)$ is a linear combination of the terms $x_i x_j^*$. But it is incorrect and unnecessary. In addition, the last step of deriving $\mathbb{H}(p) - \mathbb{H}(\gamma_Q) \leq 0$, and $\mathbb{H}(p) - \mathbb{H}(\gamma_Q) = 0$ implying $p = \gamma_Q$ are not proved. Thus, we offer a formal and alternative proof.

Proposition 6.2.5. *Suppose that the complex random vector $\xi \in \mathbb{C}^n$ has zero mean and ξ satisfies $E\left[\xi\xi^\dagger\right] = Q$, i.e., $E\left[\xi_i\xi_j^\dagger\right] = Q_{i,j}, 1 \leq i,j \leq n$. Then the entropy of ξ satisfies $\mathbb{H}(f(\xi;\mu,Q)) \leq \log\det(\pi e Q)$, with equality if and only if ξ is a circularly symmetric complex Gaussian random variable (or vector).*

Our proof is partly based on the proof given by [35].

The following two important facts are needed to complete our proof. *The first important fact is:*

$$E_\eta[\log p_\eta(\eta)] = E_\xi[\log p_\eta(\xi)].$$

The second one is a simple but crucial inequality in our proof. *The second important fact is:*

$$\log x \leq x - 1, \forall x > 0; \ \log x = x - 1, \text{ iff } x = 1.$$

The first one needs a proof that is given as follows.

Lemma 6.2.6. *If assumptions are the same as those of Proposition 6.2.5, the random vector ξ and η satisfy the assumptions and η is a circularly symmetric complex Gaussian random vector, then*

$$E_\eta \left[\log p_\eta (\eta)\right] = E_\xi \left[\log p_\eta (\xi)\right].$$

The Proof of Lemma 6.2.6. Due to the definitions of η and (6.2), which qualify the relationship expression between the differential entropy and the mean, we have

$$E_\eta \left[\log p_\eta (\eta)\right] = \int_{\mathbb{C}^n} \left[-\log\det(\pi Q) - (x-\mu)^\dagger Q^{-1}(x-\mu)\right] p_\eta (x)\, dx.$$

Due to the linearity property of integration, we have

$$\int_{\mathbb{C}^n} \left[-\log\det(\pi Q) - (x-\mu)^\dagger Q^{-1}(x-\mu)\right] p_\eta (x)\, dx$$
$$= -\log\det(\pi Q) - \int_{\mathbb{C}^n} (x-\mu)^\dagger Q^{-1}(x-\mu)\, p_\eta (x)\, dx.$$

Due to the circular invariance property of the trace, $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA)$, we have

$$-\log\det(\pi Q) - \int_{\mathbb{C}^n} (x-\mu)^\dagger Q^{-1}(x-\mu)\, p_\eta (x)\, dx$$
$$= -\log\det(\pi Q) - \int_{\mathbb{C}^n} \mathrm{Tr}\left(Q^{-1}(x-\mu)(x-\mu)^\dagger\right) p_\eta (x)\, dx.$$

Because the order of the trace and integration can be interchanged, one has

$$-\log\det(\pi Q) - \int_{\mathbb{C}^n} \mathrm{Tr}\left(Q^{-1}(x-\mu)(x-\mu)^\dagger\right) p_\eta (x)\, dx$$
$$= -\log\det(\pi Q) - \mathrm{Tr}\left(\int_{\mathbb{C}^n} Q^{-1}(x-\mu)(x-\mu)^\dagger\, p_\eta (x)\, dx\right).$$

Due to the linearity property of the integration, we have

$$-\log\det(\pi Q) - \mathrm{Tr}\left(\int_{\mathbb{C}^n} Q^{-1}(x-\mu)(x-\mu)^{\dagger} p_\eta(x)\,dx\right)$$
$$= -\log\det(\pi Q) - \mathrm{Tr}\left(Q^{-1}\int_{\mathbb{C}^n}(x-\mu)(x-\mu)^{\dagger} p_\eta(x)\,dx\right).$$

The assumption that the variances of ξ and η are the same implies

$$Q = \int_{\mathbb{C}^n}(x-\mu)(x-\mu)^{\dagger} p_\eta(x)\,dx = \int_{\mathbb{C}^n}(x-\mu)(x-\mu)^{\dagger} p_\xi(x)\,dx,$$

$$-\log\det(\pi Q) - \mathrm{Tr}\left(Q^{-1}\int_{\mathbb{C}^n}(x-\mu)(x-\mu)^{\dagger} p_\eta(x)\,dx\right)$$
$$= \int_{\mathbb{C}^n} -\log\det(\pi Q)\, p_\xi(x)\,dx + \mathrm{Tr}\left(Q^{-1}\int_{\mathbb{C}^n} -(x-\mu)(x-\mu)^{\dagger} p_\xi(x)\,dx\right).$$

Because of the basic property, $\log(ab) = \log(a) + \log(b)$, of the logarithm, it is obtained that

$$\int_{\mathbb{C}^n} -\log\det(\pi Q)\, p_\xi(x)\,dx + \mathrm{Tr}\left(Q^{-1}\int_{\mathbb{C}^n} -(x-\mu)(x-\mu)^{\dagger} p_\xi(x)\,dx\right)$$
$$= \int_{\mathbb{C}^n} \log\left(\det(\pi Q)^{-1}\exp\left\{\mathrm{Tr}\left(-Q^{-1}(x-\mu)(x-\mu)^{\dagger}\right)\right\}\right) p_\xi(x)\,dx.$$

The circular invariance property, $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)$, of the trace implies

$$\int_{\mathbb{C}^n} \log\left(\det(\pi Q)^{-1}\exp\left\{\mathrm{Tr}\left(-Q^{-1}(x-\mu)(x-\mu)^{\dagger}\right)\right\}\right) p_\xi(x)\,dx$$
$$= \int_{\mathbb{C}^n} \log\left(\det(\pi Q)^{-1}\exp\left\{-(x-\mu)^{\dagger} Q^{-1}(x-\mu)\right\}\right) p_\xi(x)\,dx.$$

Finally, due to the definition of $E_\xi[\log p_\eta(\xi)]$, i.e.,

$$\int_{\mathbb{C}^n} \log\left(\det(\pi Q)^{-1}\exp\left\{-(x-\mu)^{\dagger} Q^{-1}(x-\mu)\right\}\right) p_\xi(x)\,dx = E_\xi[\log p_\eta(\xi)],$$

we have $E_\eta[\log p_\eta(\eta)] = E_\xi[\log p_\eta(\xi)]$. □

The proof of Proposition 6.2.5 is offered as follows.

Proof. Let $p_\xi : \mathbb{C}^n \longrightarrow \mathbb{R}$ be the probability density function of ξ. According to the assumption of the random vector ξ,

$$E\left(\xi\xi^\dagger\right) = \int_{\mathbb{C}^n} xx^\dagger p_\xi(x)\, dx = Q.$$

Let η be a circularly symmetric complex Gaussian variable (or vector) with zero mean and variance $E\left(\eta\eta^\dagger\right) = Q$. Let the probability density function of η be $p_\eta(x)$.

The definitions of $\mathbb{H}(\xi)$ and $\mathbb{H}(\eta)$ imply

$$\mathbb{H}(\xi) - \mathbb{H}(\eta) = -E_\xi\left[\log p_\xi(\xi)\right] + E_\eta\left[\log p_\eta(\eta)\right]. \tag{6.3}$$

The first important fact holding is followed by

$$-E_\xi\left[\log p_\xi(\xi)\right] + E_\eta\left[\log p_\eta(\eta)\right] = -E_\xi\left[\log p_\xi(\xi)\right] + E_\xi\left[\log p_\eta(\xi)\right].$$

Because of the linearity property of the expectation,

$$-E_\xi\left[\log p_\xi(\xi)\right] + E_\xi\left[\log p_\eta(\xi)\right] = E_\xi\left[\log \frac{p_\eta(\xi)}{p_\xi(\xi)}\right].$$

Due to the second important fact holding and basic properties of Lebesgue integration, it is to see

$$E_\xi\left[\log \frac{p_\eta(\xi)}{p_\xi(\xi)}\right] \leq E_\xi\left[\frac{p_\eta(\xi)}{p_\xi(\xi)} - 1\right]. \tag{6.4}$$

The definition of the expectation implies

$$E_\xi\left[\frac{p_\eta(\xi)}{p_\xi(\xi)}-1\right]=E_\xi\left[\frac{p_\eta(\xi)}{p_\xi(\xi)}\right]-E_\xi[1]=1-1=0.$$

Hence, according to (6.3), $\mathbb{H}(\xi)-\mathbb{H}(\eta)\leq 0$.

Therefore, taking note of the second important fact, the entropy of ξ satisfies

$$\mathbb{H}(\xi)=-E_\xi\left[\log p_\xi(\xi)\right]\leq \log\det(\pi e Q),$$

with equality if and only if ξ is a circularly symmetric complex Gaussian random variable (or vector) under the given mean and variance. □

Remark 6.2.7. *(6.4) is explained as follows. First, define* $u\log u|_{u=0}\triangleq \lim_{u\downarrow 0} u\log u =$

0 *and* $\frac{u}{u}|_{u=0} \triangleq \lim_{u \downarrow 0} \frac{u}{u} = 1$. *Second, (6.4) holds because*

$$\begin{aligned}
E_\xi\left[\log \frac{p_\eta(\xi)}{p_\xi(\xi)}\right] &= \int_{\mathbb{C}^n} \log\left(\frac{p_\eta(x)}{p_\xi(x)}\right) p_\xi(x)\, dx \\
&= \left(\int_{\{x|p_\xi(x)=0\}} + \int_{\{x|p_\xi(x)>0\}}\right) \log\left(\frac{p_\eta(x)}{p_\xi(x)}\right) p_\xi(x)\, dx \\
&= \int_{\{x|p_\xi(x)=0\}} 0 dx + \int_{\{x|p_\xi(x)>0\}} \log\left(\frac{p_\eta(x)}{p_\xi(x)}\right) p_\xi(x)\, dx \\
&\leq \int_{\{x|p_\xi(x)>0\}} \left(\frac{p_\eta(x)}{p_\xi(x)} - 1\right) p_\xi(x)\, dx \\
&\leq \int_{\mathbb{C}^n} \frac{p_\eta(x)}{p_\xi(x)} p_\xi(x)\, dx - \left(\int_{\{x|p_\xi(x)=0\}} + \int_{\{x|p_\xi(x)>0\}}\right) p_\xi(x)\, dx \\
&= \int_{\mathbb{C}^n} \frac{p_\eta(x)}{p_\xi(x)} p_\xi(x)\, dx - \int_{\mathbb{C}^n} p_\xi(x)\, dx \\
&= E_\xi\left[\frac{p_\eta(\xi)}{p_\xi(\xi)}\right] - E_\xi[1] \\
&= E_\xi\left[\frac{p_\eta(\xi)}{p_\xi(\xi)} - 1\right].
\end{aligned}$$

Using the definition of the mutual information, we have that

$$C(H,P) = \max_{p_x} \{\mathbb{H}(y) - \mathbb{H}(y|x) \,|S \succeq 0, \mathrm{Tr}(S) \leq P\},$$

where the model (2.1) implies

$$\max_{p_x} \{\mathbb{H}(y) - \mathbb{H}(y|x) \,|S \succeq 0, \mathrm{Tr}(S) \leq P\}$$

$$= \max_{p_x} \{\mathbb{H}(y) - \mathbb{H}(z) \,|S \succeq 0, \mathrm{Tr}(S) \leq P\}.$$

The assumptions of z in model (2.1) also imply

$$\max_{p_x} \{\mathbb{H}(y) - \mathbb{H}(z) \,|S \succeq 0, \mathrm{Tr}(S) \leq P\}$$

$$= \max_{p_x} \{\mathbb{H}(y) \,|S \succeq 0, \mathrm{Tr}(S) \leq P\} - \mathbb{H}(z).$$

Note that we may assume that x satisfies $E\left(x^\dagger x\right) \leq P$ and is a zero mean random vector. Furthermore for such an x, if x is a zero mean random vector with covariance $E\left(xx^\dagger\right) = S$, then y is a zero mean random vector with covariance $E\left(yy^\dagger\right) = HSH^\dagger + I_r$, which results from the form of model (2.1) and the linearity of the expectation operation, and by Proposition 6.2.5 among such y vectors the entropy is the largest when y is a circularly symmetric complex Gaussian random vector, which is the case when x is a circularly symmetric complex Gaussian random vector by the two facts at the end part of last section. Thus, we can further restrict our attention to the circularly symmetric complex Gaussian random vector x. In this case the mutual information is given by $\log\left(\det\left(I_r + HSH^\dagger\right)\right)$.

The two facts (refer to [27, 26, 35]) are claimed as follows. A linear transformation of a circularly symmetric complex Gaussian random vector is a circularly symmetric complex Gaussian random vector. The set of circularly symmetric complex Gaussian random vectors is closed for addition. They are used for calculating the channel capacity in the following section.

6.3 Appendix-III: Proofs of the Lemmas in Section 2.4

Lemma 6.3.1. *For the channel H, there is a unitary matrix U such that $U^{\dagger}H^{\dagger}HU = diag(\lambda_1, \cdots, \lambda_t)$ (the diagonal matrix) and*

$$\max\left\{\log\left(\det\left(I_r + HSH^{\dagger}\right)\right) | S \succeq 0, Tr(S) \leq P\right\} =$$

$$\max\left\{\log\left(\det\left(I_t + diag\left(\lambda_1, \cdots, \lambda_t\right) S\right)\right) | S \succeq 0, Tr(S) \leq P\right\}$$

and $U^{\dagger}S_l U = S_r$, where S_l and S_r are two optimal solutions of the two optimization problems mentioned above, respectively.

Proof. According to the matrix theory, it is easily known that there is a unitary matrix U such that

$$U^{\dagger}H^{\dagger}HU = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) \tag{6.5}$$

(the diagonal matrix) and the two maximum points exist due to the compactness of the two constraints. Let

$$S_l \in \arg\max\left\{\log\left(\det\left(I_r + HSH^{\dagger}\right)\right) | S \succeq 0, \mathrm{Tr}\left(S\right) \leq P\right\}.$$

Because U is a unitary matrix and $\det\left(I + AB\right) = \det\left(I + BA\right)$ with appropriate dimensions of the matrices, we have

$$\log\left(\det\left(I_r + HS_lH^{\dagger}\right)\right) = \log\left(\det\left(I_t + U^{\dagger}H^{\dagger}HUU^{\dagger}S_lU\right)\right).$$

Due to (6.5),

$$\log\left(\det\left(I_t + U^\dagger H^\dagger H U U^\dagger S_l U\right)\right) = \log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) U^\dagger S_l U\right)\right).$$

Since $S_l \succeq 0$ and $\mathrm{Tr}\left(S_l\right) \leq P$,

$$\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) U^\dagger S_l U\right)\right)$$

$$\leq \max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) U^\dagger S U\right)\right) | U^\dagger S U \succeq 0, \mathrm{Tr}\left(U^\dagger S U\right) \leq P\right\}.$$

As the unitary similarity transformation keeps the semidefinite positiveness and trace,

$$\max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) U^\dagger S U\right)\right) | U^\dagger S U \succeq 0, \mathrm{Tr}\left(U^\dagger S U\right) \leq P\right\}$$

$$\leq \max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) S\right)\right) | S \succeq 0, \mathrm{Tr}\left(S\right) \leq P\right\}.$$

Hence,

$$\max\left\{\log\left(\det\left(I_r + H S H^\dagger\right)\right) | S \succeq 0, \mathrm{Tr}\left(S\right) \leq P\right\} \leq$$

$$\max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) S\right)\right) | S \succeq 0, \mathrm{Tr}\left(S\right) \leq P\right\}.$$

On the other hand,

$$\forall S_r, S_r \in \arg\max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) S\right)\right) | S \succeq 0, \mathrm{Tr}\left(S\right) \leq P\right\}.$$

Because the definition of matrix Λ,

$$\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) S_r\right)\right) \leq \log\left(\det\left(I_t + U^\dagger H^\dagger H U S_r\right)\right).$$

Due to $\det(I+AB) = \det(I+BA)$,

$$\log\left(\det\left(I_t + U^\dagger H^\dagger H U S_r\right)\right) \leq \log\left(\det\left(I_r + H U S_r U^\dagger H^\dagger\right)\right).$$

As the unitary similarity transformation keeps the semidefinite positiveness and trace, we get

$$\log\left(\det\left(I_r + H U S_r U^\dagger H^\dagger\right)\right)$$

$$\leq \max\left\{\log\left(\det\left(I_r + H U S_r U^\dagger H^\dagger\right)\right) | U S_r U^\dagger \succeq 0, \mathrm{Tr}\left(U S_r U^\dagger\right) \leq P\right\}.$$

The same reason implies

$$\max\left\{\log\left(\det\left(I_r + H U S_r U^\dagger H^\dagger\right)\right) | U S_r U^\dagger \succeq 0, \mathrm{Tr}\left(U S_r U^\dagger\right) \leq P\right\}$$

$$\leq \max\left\{\log\left(\det\left(I_r + H S H^\dagger\right)\right) | S \succeq 0, \mathrm{Tr}(S) \leq P\right\}.$$

Thus,

$$\max\left\{\log\left(\det\left(I_t + \mathrm{diag}(\lambda_1, \cdots, \lambda_t)\, S\right)\right) | S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

$$\leq \max\left\{\log\left(\det\left(I_r + H S H^\dagger\right)\right) | S \succeq 0, \mathrm{Tr}(S) \leq P\right\}.$$

Therefore,

$$\max\left\{\log\left(\det\left(I_r + H S H^\dagger\right)\right) | S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

$$= \max\left\{\log\left(\det\left(I_t + \mathrm{diag}(\lambda_1, \cdots, \lambda_t)\, S\right)\right) | S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

and further $U^\dagger S_l U = S_r$, where S_l and S_r are two optimal solutions of the two optimization problems respectively. □

Lemma 6.3.2. *For the channel H, there is a unitary matrix U such that $U^\dagger H^\dagger H U = \Lambda$ (the diagonal matrix) and*

$$\max\left\{\log\left(\det\left(I_r + HSH^{\dagger}\right)\right)|S \succeq 0, Tr(S) \leq P\right\}$$

$$= \max\left\{\log\left(\det\left(I_t + \Lambda^{\frac{1}{2}} S \Lambda^{\frac{1}{2}}\right)\right)|S \succeq 0, Tr(S) \leq P\right\},$$

and $U^{\dagger} S_l U = S_r$, *where* S_l *and* S_r *are two optimal solutions of the two optimization problems mentioned above, respectively.*

Proof. According to **Lemma 6.3.1**, we have

$$\max\left\{\log\left(\det\left(I_r + HSH^{\dagger}\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

$$= \max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) S\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}.$$

According to the definition of the matrix Λ and (6.5), we have

$$\max\left\{\log\left(\det\left(I_t + \mathrm{diag}\left(\lambda_1, \cdots, \lambda_t\right) S\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

$$= \max\left\{\log\left(\det\left(I_t + \Lambda S\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}.$$

Due to the definition of the square root for the matrix Λ, we get

$$\max\left\{\log\left(\det\left(I_t + \Lambda S\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

$$= \max\left\{\log\left(\det\left(I_t + \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} S\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}.$$

For the reason that $\det(I + AB) = \det(I + BA)$, we have

$$\max\left\{\log\left(\det\left(I_t + \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} S\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}$$

$$= \max\left\{\log\left(\det\left(I_t + \Lambda^{\frac{1}{2}} S \Lambda^{\frac{1}{2}}\right)\right)|S \succeq 0, \mathrm{Tr}(S) \leq P\right\}.$$

□

Bibliography

[1] Barry, J. R., Lee, E. A., and Messerschmitt, D. G., *Digital Communication*, 3rd Edition, Springer, 2004.

[2] Bertsekas, D. P., *Nonlinear Programming*, 2nd Edition, Athena Scientific, Belmont, MA, 1999.

[3] Bertsekas, D. P., and Tsitsiklis, J. N., *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, MA, 1997.

[4] Biglieri, E., Calderbank, R., Constantinides, A., Goldsmith, A., Paulraj, A., and Poor, H. V., *MIMO Wireless Communications*, Cambridge University Press, 2007.

[5] Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A., *Numerical Optimization: Theoretical and Practical Aspects*, 2nd Edition, Springer, 2006.

[6] Boyd, S., and Vandenberghe, L., *Convex Optimization*, Cambridge University Press, 2004.

[7] Carleial, A. B., *Interference channels*, IEEE Trans. Inf. Theory, Vol. 24, pp. 60-70, 1978.

[8] Cioffi, J. M., Dudevoir, G. P., Eyuboglu, M. V., and Forney, G. D., *MMSE decision feedback equalizers and coding: Part I and II*, IEEE Trans. Comm., Vol. 43, pp. 2582-2604, 1995.

[9] Cioffi, J. M., and Forney, G. D., *Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise*, in Communications, Computation, Control and Signal Processing: a tribute to T. Kailath, A. Paulraj, V. Roychowdhury, and C. D. Shaper, Kluwer Academic Publishers, Eds. 1997.

[10] Cover, T. M., and Thomas, J. A., *Elements of Information Theory*, New York: Wiley, 1991.

[11] Diggavi, S. N., and Cover, T. M., *Worst additive noise under covariance constraints*, IEEE Trans. Inf. Theory, Vol. 47, pp. 3072-3081, 2001.

[12] Erez, U., Shamai, S., and Zamir, R., *Capacity and lattice strategies for cancelling known interference*, IEEE Trans. Inf. Theory, Vol. 51, pp. 3820 - 3833, 2005.

[13] Eyuboglu, M. V., and Forney Jr., G. D., *Trellis precoding: combined coding, precoding and shaping for intersymbol interference channels*, IEEE Trans. Inf. Theory, Vol. 38(II), pp. 301-314, 1992.

[14] Fan, K., *Minimax theorems*, Proc. Nat. Acad. Sci., Vol. 39, pp. 42-47, 1953.

[15] Forney Jr., G. D., *Trellis shaping*, IEEE Trans. Inf. Theory, Vol. 38(II), pp. 281-300, 1992.

[16] Foschini, G. J., and Miljanic, Z., *A simple distributed automonous power control algorithm and its convergence*, IEEE Trans. Veh. Tech., Vol. 42, pp. 641-646, 1993.

[17] Goldsmith, A., *Wireless Communications*, Cambridge University Press, 2005.

[18] Haykin, S., *Transmit power control techniques for wireless communication systems*, U.S. patent, USPTO Application No.: 2009-0047916, Feb. 2009.

[19] Ihara, S., *On the capacity of channels with additive non-Gaussian noise*, Information and Control, Vol. 37, pp. 34-39, 1978.

[20] Jindal, N., Rhee, W., Vishwanath, S., Jafar, S.A., and Goldsmith, A., *Sum power iterative water-filling for multi-antenna Gaussian broadcast channels*, IEEE Trans. Inf. Theory, Vol. 51, pp. 1570-1580, 2005.

[21] Kailath, T., Sayed, A., and Hassibi, B., *Linear Estimation*, Prentice Hall, 2000.

[22] Kelly, J. L., *General Topology*, 2nd Edition, Springer-Verlag, 1975.

[23] M. Kobayashi and G. Caire, *An Iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC*, IEEE J. Sel. Areas Commun., Vol. 24, pp. 1640-1646, 2006.

[24] Luenberger, D. G., *Optimization by Vector Space Methods*, New York, John Wiley and Sons, Inc., 1969.

[25] Magnus, J. R., and Neudecker, H., *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd Edition, Wiley, 1999.

[26] Neeser, F. D., and Massey, J. L., *Proper complex random processes with applications to information theory*, IEEE Trans. Inf. Theory, Vol. 39, pp. 1293-1302, 1993.

[27] Picinbono, B., *On circularity*, IEEE Trans. Signal Process. Vol. 42. pp. 3473-3482, 1994.

[28] Rockafellar, R. T., *Convex Analysis*, Princeton University Press, 1970.

[29] Rudin, W., *Real and Complex Analysis*, 3rd Edition, McGraw-Hill, 1985.

[30] Sato, H., *The capacity of the Gaussian interference channel under strong interference*, IEEE Trans. Inf. Theory, Vol. 27, pp. 786-788, 1981.

[31] Sato, H., *An outer bound on the capacity region of broadcast channels*, IEEE Trans. Inf. Theory, Vol. 24, pp. 374-377, 1978.

[32] Shannon, C. E., *Two-way communication channels*, in Proc. 4th Berkeley Symp. Math. Stat. Prob., pp. 611-644, University of California Press, 1961.

[33] Shannon, C. E., *A mathematical theory of communication*, Bell Sys. Tech. J., Vol. 27, pp. 379-423, 1948.

[34] Sponsor: IEEE Computer Society, and IEEE Microwave Theory and the Technology Society, *P802.16 Rev 2/D5*, IEEE Standards Activities Department, 2008.

[35] Telatar, E., *Capacity of multi-antenna Gaussian channels*, Europ. Trans. on Telecomm., Vol. 10, pp. 585-596, 1999.

[36] Tomlinson, M., *New automatic equalizer employing modulo arithmetic*, Electr. Let., Vol. 7, pp. 138-139, 1971.

[37] Varanasi, M. K., and Guess, T., *Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Guassian multiple-access channel*, in Proc. Asilomar Conf. Signal System Computers, pp. 1405-1409, 1997.

[38] Vishwanath, S., Jindal, N., and Goldsmith, A., *Duality, achievable rates, and sum-rate capacity of MIMO broadcast channels*, IEEE Trans. Inf. Theory, Vol. 49, pp. 2658-2668, 2003.

[39] Wang, Z. K., and Yang, X. Q., *Birth and Death Processes and Markov Chains*, Springer-Verlag, 1992.

[40] Wang, Z. K., *Stochastic Process*, Science Press of China, 1965.

[41] Witsenhausen, H. S., *A determinant maximization problem occurring in the theory of data communication*, SIAM J. Appl. Math., Vol. 29, pp. 515-522, 1975.

[42] Yu, W., Rhee, W., Boyd, S., and Cioffi, J. M., *Iterative water-filling for Gaussian vector multi-access channels*, IEEE Trans. Inf. Theory, Vol. 50, pp. 145-152, 2004.

[43] Zakovic, S., and Pantelides, C., *An interior point algorithm for computing saddle points of constrained continuous minimax*, Annals of Operations Research, Vol. 99, pp. 59-77, 2000.

[44] Zangwill, W. I., *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.