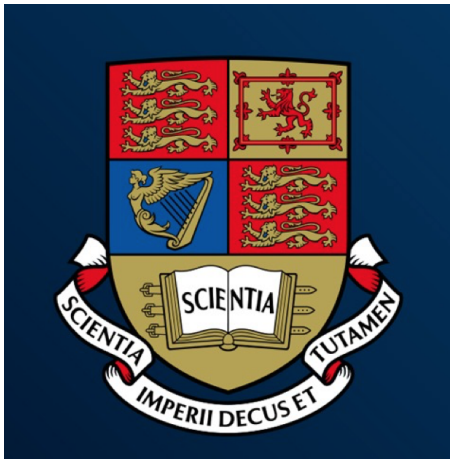
Imperial College London

Department of Electrical and Electronic Engineering

Final Year Project Report 2020



Project Title:	**Design and Evaluation of Beyond 5G Wireless Communication Systems**
Student:	**Xinyuan Xu**
CID:	**01183830**
Course:	**EEE4**
Project Supervisor:	**Dr Bruno Clerckx**
Second Marker:	**Dr Wei Dai**

Plagiarism Statement

I confirm that I have written this report myself.

I have read the EEE Department Plagiarism Committee Basic Principles on Plagiarism in Assessed Coursework.

I am aware that the College views plagiarism/collusion very seriously and that an infringement could lead to expulsion from the College.

I affirm that I have submitted, or will submit, electronic copies of my final year project report to both Blackboard and the EEE coursework submission system.

I affirm that the two copies of the report are identical.

I affirm that I have provided explicit references for all material in my Final Report which is not authored by me and represented as my own work.

Acknowledgements

First of all, I would like to express my gratefulness to my project supervisor Dr Bruno Clerckx, whose patient mentoring always set out a clear goal for me to work towards. Secondly, I would like to say thank you to Dr Yijie (Lina) Mao, who has always helped me a lot when I had questions or faced problems. Finally, I want to thank my family and friends. Their support and encouragement must be appreciated, especially in this difficult 2020 ravaged by the pandemic.

List of Figures

Contents

1 Abstract

This project concerns the design and evaluation of a novel transmission scheme called "Rate-Splitting" (RS),which has the potential to become part of the next generation MIMO communication systems. The principle and theories of RS has been learnt and its performance has been compared with other conventional approaches. The main emphasis of this project is the algorithm for RS precoder design and the optimization method behind. Two-user scenario has been simulated with MATLAB throughout the project. First, the so called "Weighted Sum Rate Maximization using Weighted MMSE algorithm" has been studied, which might also be referred as Alternating Optimization. This algorithm provides an excellent way for Multiple User Linear Precoder calculation and is proved to be capacity achieving when compared to DPC rate region. Second, the ergodic rate region of RS with partial CSIT has been studied, involving techniques like Sample Average Approximation and WMMSE approach. The later is based on the previously mentioned Alternating Optimization. Finally, the feasibility of the above convex optimization algorithms has been experimented, after the addition of new QoS constraints. Solving optimization problem using CVX might not be feasible when the individual users' rate needs to be no smaller than a threshold rate.

2 Introduction

As we all know, the latest technology everyday consumer could enjoy falls under the name of 5G, the fifth generation of wireless communication technology. I personally use a Huawei 5G mobile phone. In UK, there are already around 140 towns or cities with 5G coverage and more constructions are well underway. At the end of 2019, 52 cities in China have been covered by commercially enabled 5G networks and 400 thousand additional 5G base stations were planned to be deployed.

Among 5G and of course 4G, one of the core technologies is our beloved MIMO, multiple input multiple output system. MIMO could offer a lot of technical benefits, like spatial multiplexing gain, diversity gain, just to name a few. These technical benefits brings many advantages which are tangile to users, such as increased data rate, reduction in air-latency, improved energy efficiency and interference suppression etc.

However, the benefits MIMO offer us have some cost. One of the most important one is around channel state knowledge at transmitter (CSIT). The channel estimation may not be sufficiently accurate, and maybe only limited feedback is available, depending on the system design. The CSIT inaccuracy could leads to residual multi-user interference, and this has become one of the major bottlenecks of system performance. So here comes the entry point of Rate-Splitting(RS), instead of keeping the accurate CSIT assumption, RS boldy faces the reality and addresses directly the problem of imperfect CSIT.

Rate Splitting turns out not to be a brand-new idea, it could be dated back to research on two-user interference channel. But here we are applying it to a MU-MIMO system, on downlink, so broadcast channels. In the system we have K users. First, the message of each user would be split into two parts, a part called common message, the other private message. The private message of each user would be coded individually. The common messages of all users could be combined, to get one final common message, and it would be coded with a publicly available codebook. All the messages would be precoded by a linear beamformer before transmission. In the channel the signals are affected by noise and interference. Then at each receiver, the common message would first be decoded. Successive Interference Cancellation (SIC) would be applied before each user decode their private message. Piecing together the user private message and its share of common message we get the original message of each user.

The beauty of RS, or how could RS be powerful, could be demonstrated by its difference with existing architectures. In non-orthogonal multiple access (NOMA), Superposition Coding and Successive Interference Cancellation is used, and it aims at decoding the interference from other users completely. In Space Division Multiple Access, one jargon is Multiple User Linear Precoding (MULP). After separating users in spatial domain, the interference left is treated as noise. Rate Splitting is able to adjust between the two schemes. By controlling the power, or rate or proportion of the common message, the amount of interference to decode is flexible. It has been shown that Rate Splitting Multiple Access (RSMA) generalizes and outperforms NOMA and

SDMA [12].

Organization

The rest of the report is structured as follow:

- Section 3 is Background, which introduces several concepts related to this project. Hopefully, they would help to understand this report. Also, they are record of some more theoretical learning during this project, which is as important as the MATLAB codes and simulations.
- Section 4 studies a crucial algorithm: "Weighted Sum Rate Maximization using Weighted MMSE" [2], which is also called Alternating Optimization (AO), assuming perfect channel knowledge.
- Section 5 studies two key tools used in Rate Splitting under partial CSIT scenario: Sample Average Approximation(SAA) and WMMSE apporach.
- Section 6 reports an experiment on the feasibility of optimization algorithms used in MATLAB simulation of Section 5, after an extra constraint on minimum individual user rate is imposed.
- Section 7 looks back and evaluate this final year project. Section 8 is conclusion and future work.

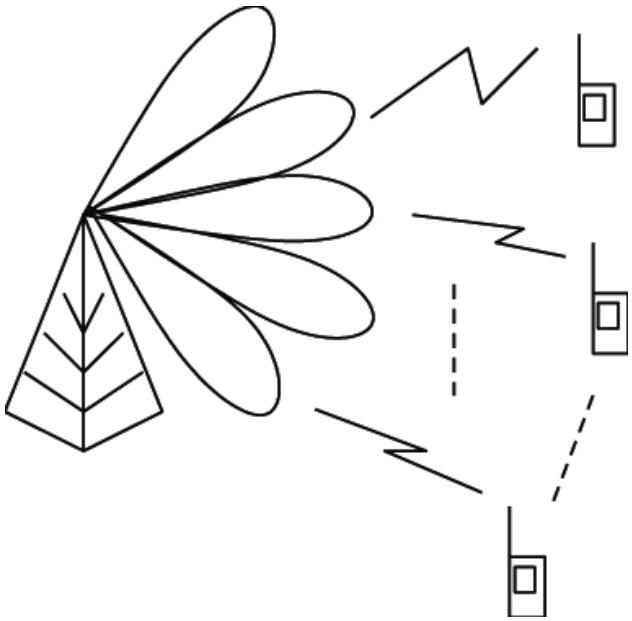


Figure 1: An illustration of SDMA [1]

3 Background

3.1 SDMA and MULP

Nowadays, more and more Access Points(AP) are equiped with multiple antenna, this enables a "new" method of multiple access: Space Division Multiple Access (SDMA). Rather than dividing time period or frequency band and allocate to different users, users are separated on transmitter side in space domain. For instance, users at different directions of arrival from the base station could communicate simultaneously in the same frequency band, if the beamformer successfully separates them in angular domain.

SDMA is often achieved by Multiple User Linear Precoding (MULP). Unlike Dirty Paper Coding, which achieves the capacity region of Gaussiam MIMO Broadcast Channel (BC) [17], MULP has much lower computational complexity and could realize performance close to DPC, dispite it is suboptimal. Therefore SDMA using MULP benefits from the spatial multiplexing gain while keeping the complexity of receiver and transmitter low. [12]

The problem of SDMA using MULP could be demonstrated by using an example in Multiple-Input-Single-Output (MISO) system:

- The base station has N_t transmit antennas;
- There are K users, each equipped with single antenna;
- $\mathbf{x} \in \mathbb{C}^{Nt \times 1}$ represents transmitted signal in a single channel use;
- The transmit power is constrained by $\mathbb{E}\{||\mathbf{x}||\} \leq P_t$;
- $\mathbf{h}_k$ represents the channel experienced between the base station and user-k;
- $n_k \sim \mathcal{CN}(0, \sigma_{n,k}^2)$ represents complex circular symmetric additive white Gaussian noise (AWGN) experienced at user-k;
- $\mathbf{p}_k \in \mathbb{C}^{Nt \times 1}$ is the precoder of user k. $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k]$ is the precoder matrix;
- $\mathbf{s} = [s_1, s_2, ..., s_k]$ represents the collection of data streams of users;
- $\mathbf{x} = \mathbf{Ps}$.

The signal received at user-k is given by:

$$y_k = \mathbf{h}_k^H \mathbf{x} + n_k, \forall k \in K \tag{1}$$

Assume noise power is normalized to unity, the Signal-to-Interference-plus-Noise-Ratio(SINR) experienced at user-k is given by:

$$\gamma_k = \frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{\sum_{j \neq k, j \in K} |\mathbf{h}_k^H \mathbf{p}_j|^2 + 1} \tag{2}$$

The capacity or achievable rate of each user is therefore $R_k = \log_2(1 + \gamma_k)$.

Usually, a weighted sum rate (WSR) optimization problem is considered, as sum rate maximization problem is just a special case of equal weights. With relative weighting of a user represented by u_k, the WSR problem is formulated as:

$$\begin{aligned} R_{MULP} = \quad & \max \sum_{k \in K} u_k R_k \\ \text{s.t.} \quad & tr(PP^H) \leq P_T \end{aligned} \tag{3}$$

This optimisation problem could be solved by an algorithm called "Weighted Sum Rate Maximization using Weighted MMSE" proposed in [2]. This will be further closly studied in Section 4.

3.2 Dirty Paper Coding

For an undergraduate student studying wireless communication, Dirty Paper Coding(DPC) is probably first learnt as the capacity region performance bound of a Gaussian MU-MIMO channel. The name of DPC comes from a beautiful analogy [5]: Communication over a channel with output Y, input X, State S and i.i.d noise Z could be viewed as writing on a piece of dirty paper. What is intended to write is like input x. State S, which might be viewed as multi-user interference, is some careless ink dots, perfectly know to the writer (transmitter). The task of writing a legible message on this dirty paper is analogues to transmission in such a channel.

Professor Costa pointed out in this correspondence than instead of using part of the limited power to cancel S, the optimum encoding scheme is to adapt its signal to state S. Code words should be chosen far apart so that they could be distinguishable on output. In my imprecise understanding, this is like trying to avoid the ink dots on paper when writing a message, rather than using more ink trying to cover it up. Furthermore, it has been proved that all rates with $R < C^{\star} = C(P/N)$ are achievable. Capacity is only limited by signal to noise ratio and state S will not affect transmission provided that it is perfectly known at transmitter.

Developing from the Costa precoding technique, the sum capacity of the vector Gaussian broadcast channel has been computed [15]. The most important idea in this paper is to utilize the duality between up-link (Multiple Access Channel - MAC) and down-link (Broadcast Channel - BC). The noise

covariance matrix $\mathbf{Q}$ could also be interpreted as a cost matrix.

Furthermore, it has been proved in [17] that the DPC rate region equals to the capacity region of the MIMO BC. Instead relying on MAC-BC duality, a new notion of enhanced channel was used. This equality has been shown to hold under many different input constraints.

A practical optimization algorithm which could be used to find the capacity region of DPC has been proposed in [16]. This algorithm is also based on the BC-MAC duality. In this paper, more practical aspects in a cellular wireless network has been considered.

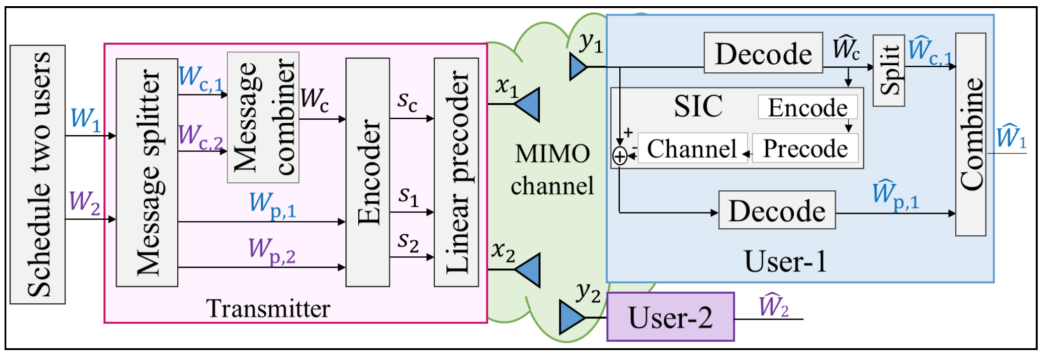


Figure 2: Two-user system architecture with rate splitting [12]

3.3 Basics of Rate Splitting

3.3.1 Two user interference channel

The origin of Rate Splitting could be traced back to early work by Carleiai, Han and Kobayashi. Rate Splitting has been mention in several placed in the book [7]. In particular in Chapter 6 of the book, rate splitting is used to prove Han-Kobayashi Inner Bound, which could be the best-known bound on the capacity region of the Discrete Memory-less Interference Channel. This is a kind of 'there is nothing new on earth', that a probably tiny idea somewhere could be developed into something great later.

3.3.2 Rate Splitting Architecture: a two user example

A simple two user example in MISO BC system could help to understand the principle and operation of rate splitting [4]. As shown in Figure 2, the system could be divided into 10 steps:

1. Each user has its message intended to transmit W_k;

2. For each user's message, it is divided into a private part $W_{p,k}$ and a common part $W_{p,k}$;

3. The common messages of the 2 users are combined into W_c;

4. W_c is encoded into data stream s_c with a codebook available to both users;

5. Each private message $W_{p,k}$ is encoded into data stream s_k, with codebook specific to each user;

6. All the data streams are linearly precoded before transmission;

7. At each receiver, the common message is first decoded;

8. Then the common message is removed from received signal using SIC;

9. Decoding the output of SIC, the private message of each user is obtained;

10. Each user pieces together the receiver common and private message to get the complete received message.

Notations:

- Define data streams $\mathbf{s} = [s_c, s_1, s_2]$ and assume $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}$;
- $\mathbf{p}_c$ is the precoder of the total common message;
- $\mathbf{p}_k$ is the precoder of user-k's private message;
- the transmit power contraint is given by: $|\mathbf{p}_c|^2 + |\mathbf{p}_1|^2 + |\mathbf{p}_2|^2 \leq P$;

- the channel vector experienced by each user is represented by $\mathbf{h}_k$;
- $n_k \sim \mathcal{CN}(0,1)$ represents the circular symmetric complex additive white Gaussian noise (AWGN) experienced at user-k;

The transmitted signal is given by:

$$\mathbf{x} = \mathbf{p}_c s_c + \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2 \tag{4}$$

The received signal at user-k is given by:

$$y_k = \mathbf{h}_k^H \mathbf{x} + n_k, \quad k = 1, 2 \tag{5}$$

The rate of private message of user-k is given by the well known log(1+SINR) equation:

$$R_k = \log_2(1 + \frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{1 + |\mathbf{h}_k^H \mathbf{p}_j|^2}), \quad j \neq k. \tag{6}$$

The potential rate of common message $R_c(k), k = 1, 2$ is given by:

$$R_c(k) = \log_2(1 + \frac{|\mathbf{h}_k^H \mathbf{p}_c|^2}{1 + |\mathbf{h}_1^H \mathbf{p}_1|^2 + |\mathbf{h}_2^H \mathbf{p}_2|^2}) \tag{7}$$

The rate of common message is therefore:

$$R_c = \min(R_c(1), R_c(2)) \tag{8}$$

Then the sum rate can be represented by:

$$R_{sum} = R_1 + R_2 + R_c \tag{9}$$

3.3.3 Degree of Freedom

Not that precisely speaking, Degree of Freedom (DoF) is the number of interference-free streams that can superimpose in time domain while not affecting each other, in a single channel use. In DoF analysis, $O(SNR^{-\alpha})$ represents the speed at which CSIT error decays when SNR increases. It has been found that an extra DoF of $1-\alpha$ can be achieved with rate splitting, because decoding common message at each receiver decodes part of the interference [3].

While the DoF of an optimum non-RS scheme is limited to $\max(1, K\alpha)$,it has been proved in [11] that the DoF of an optimum rate splitting transmission scheme with K users is given by:

$$\lim_{P_t \to \infty} \frac{\mathbb{E}_{\hat{H}}[\mathbb{R}_{RS}(P_t)]}{\log_2(P_t)} = 1 + (K-1)\alpha \tag{10}$$

Therefore for all $\alpha \in (0, 1)$ DoF of rate splitting is strictly greater than that of non-rate-splitting schemes.

3.3.4 CSIT

As mentioned before, the pre-condition of enjoying the benefits of MU-MIMO depends on an accurate channel estimation of all users in a cell and the feedback of these estimations by receivers to transmitter. As the number of user equipment and antenna increases, there is a huge burden of accurate CSIT on MIMO network. One major bottleneck of MIMO wireless network is the multi-user interference problem due to CSIT inaccuracy[3]. Rate splitting is an attempt to directly tackle this problem.

It has been shown that rate splitting could help relax the requirement on CSIT [10]. If a constant sum rate loss is kept, it has been found that rate splitting has a feedback overhead reduction advantage, compared to Zero Forcing Beamforming (ZFBF) with quantized precoding. Then if the feedback overhead is fixed, rate splitting could offer a big SNR gain over conventional scheme.

The following subsections contain topics surveyed when doing background reading but unfortunately not learnt in detail and not worked on in the project These subsections could be treated as summary learning notes of some other related topics during this project.

3.4 Different Architectures of Rates Splitting

The Rate Splitting architecture introduced previously (with only 2 users) is probably the simplest possible. There are definitely more sophisticated architecture variations.

Hierarchical Rate Splitting(HRS) was first proposed in [6] which discusses applying RS in massive MIMO. In [13], a Two-layer HRS has been explained: First the users are assigned into groups and rate splitting is carried out; then each group of users could be treated as a super user and rate splitting is applied again to these superusers. So there would be a super common message for all the user groups, common messages for each user group intend for users only in that group, and private messages of all the users. In this way, only two layers of SIC is required at each receiver. Two-layer HRS is a low complexity strategy in general Rate Splitting framework, which could have many layers of SC-SIC and common messages.

Topological Rate Splitting (TRS) has been introduced in [9]. Previously, there is only one transmitter in the system, whereas TRS is motivated in a K-cell MISO IC with imperfect CSIT. In TRS the message of each user is rate splitted into a private message and several common messages. Each common message is intended to be decoded by only a group of users rather than all the users in the system. The paper [9] showed that TRS yields a very good achievable DoF region. Benefited from the interference management ability, TRS has been shown to outperform conventional schemes, in a realistic scenario example.

3.5 Cell Free Massive MIMO

One possible definition of Massive MIMO is that at least hundreds of antennas are employed at the base station. The potential benefits of Massive MIMO includes huge improvements in throughput and energy efficiency, better link reliability, reduced latency and so on. However, Massive MIMO faces many problems/challenges, including increased hardware complexity, higher energy consumption of signal processing, difficulty in attaining accurate CSIT. Rate Splitting has been applied to massive MIMO with imperfect CSIT in [6]. As mentioned in previous subsection, Hierarchical Rate Splitting has been proposed, which benefits from the two-tier precoder structure. The two layers of common message tackle inter and intra group interference. This could potentially help massive MIMO in its sensitivity to CSIT quality.

The idea of Cell Free Massive MIMO could be found in [14]. Instead of centralizing antenna arrays at base station serving a dedicated area, a much lower number of users are served simultaneously and collectively by a high number of distributed single-antenna Access Points (APs). The traditional concept of cell, originated probably from frequency reuse and coverage area, no longer exists in this concept. The backhull network and a central processing unit enable cooperation between APs. Potential benefits of Cell Free Massive MIMO includes more uniform high quality service for all users, low computational complexity (conjugate beamforming), increased 95%-likely per-user throughput and enhanced immunity to shadow fading spatial correlation.

4 Alternating Optimisation

This section studies the paper [2]. The paper used the relationship between weighted sum rate and weighted MMSE in the MIMO-BC to propose algorithm that solves the weighted sum rate problem of MULP mentioned in previous section. Although this paper does not mentioned rate splitting, this algorithm "weighted sum rate maximization using weighted MMSE" (WSRM-WMMSE), or "Alternating Optimization" (AO), provides the key idea that is adapted to solve the weighted sum rate problem in RS transmission scheme with partial CSIT [11].

In this section, perfect channel state information at transmitter has been assumed. The objective of AO is to find the transmit beamforming design (or precoder) so that the weighted sum rate is maximized. Transmit beamforming design means finding the linear transmit filter, rather than non-linear ones. Non-linear transmission (using DPC) could have better performance but has much higher computational burden. In addition, weighted sum rate (WSR) rather than sum rate (SR) is considered as user weights helps user prioritization thus is closer to real world application.

The algorithm proposed in paper [2] guarantees convergence to local WSR maximum. The paper claims that the WSR problem could be solved as a WMMSE-problem. This algorithm is called "Alternating Optimization" because it iterates between calculating WMMSE transmit filter coefficients, calculating MMSE receiver filter and updating weighting matrix. Since all three steps are formulated by close form expressions, this algorithm has low

complexity, high efficiency and excellent convergence property.

4.1 System Model

- In the MIMO system, there is only 1 transmitter with P transmit antennas;
- There are K users (receivers), each equipped with Q receive antennas;
- $\mathbf{H}_k \in \mathbb{C}^{Q \times P}$ represents the channel matrix between the base station and user-k; the entries of this matrix are complex values representing the gain between transmit and receive antennas; The channel matrix of each user is assumed to stay constant over the transmit window;
- input data vector: $\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k \in \mathbb{C}^{Q \times 1}$;
- beamformers: $\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_k \in \mathbb{C}^{P \times Q}$;
- transmited vector $\mathbf{x} \in \mathbb{C}^{Q \times 1}$: $= \sum_{k=1}^{K} \mathbf{B}_k \mathbf{d}_k$;
- circular symmetric AWGN experienced at user-k: $\mathbf{v}_k \in \mathbb{C}^{Q \times 1}$; It is assume to have identity noise covariance matrix: $\mathbf{R}_{v_k v_k} = \mathbf{I}_Q$;
- received complex signal vector at user-k: $\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{v}_k$;
- total transmission power is limited to E_{tx};
- weight of user-k: a constant u_k;
- effective noise covariance matrix: $\mathbf{R}_{\tilde{v}_k \tilde{v}_k}$;

- MSE-matrix for user k given that the MMSE receive filter is applied: $\mathbf{E}_k$;
- constant (per iteration) weight matrix of user-k: $\mathbf{W}_k \in \mathbb{C}^{Q_k \times Q_k}$

4.2 Objective functions, Gradients and Equivalence

The weighted sum rate maximization problem can be written as a minimization of negatives:

$$\begin{aligned} &\min_{\mathbf{B}_1,\ldots,\mathbf{B}_k} \sum_k -u_k R_k \\ \text{s.t.} \quad &\sum_{k=1}^{K} Tr(\mathbf{B}_k \mathbf{B}_k^H) = E_{tx} \end{aligned} \tag{11}$$

The gradient of this problem could be written as [2]:

$$\begin{aligned} \nabla_{\mathbf{B}_k} f = &-u_k \mathbf{H}_k^H \mathbf{R}_{\tilde{v}_k \tilde{v}_k}^{-1} \mathbf{H}_k \mathbf{B}_k \mathbf{E}_k \\ &+ \Big(\sum_{i=1, i \neq k}^{K} u_i \mathbf{H}_i^H \mathbf{R}_{\tilde{v}_i \tilde{v}_i}^{-1} \mathbf{H}_i \mathbf{B}_i \mathbf{E}_i \mathbf{B}_i^H \mathbf{H}_i^H \mathbf{R}_{\tilde{v}_i \tilde{v}_i}^{-1} \mathbf{H}_i \Big) \mathbf{B}_k + \lambda \mathbf{B}_k \end{aligned} \tag{12}$$

On the other hand, the weighted minimum mean square error (WMMSE) minimization problem is that:

$$\begin{aligned} &\min_{\mathbf{B}_1,\ldots,\mathbf{B}_k} \sum_k Tr(\mathbf{W}_k \mathbf{E}_k) \\ \text{s.t.} \quad &\sum_{k=1}^{K} Tr(\mathbf{B}_k \mathbf{B}_k^H) = E_{tx} \end{aligned} \tag{13}$$

The gradient of the WMMSE transmit filter design problem is obtained as [2]:

$$\begin{aligned}\nabla_{\mathbf{B}_k} g = &-\mathbf{H}_k^H \mathbf{R}_{\tilde{v}_k \tilde{v}_k}^{-1} \mathbf{H}_k \mathbf{B}_k \mathbf{E}_k \mathbf{W}_k \mathbf{E}_k \\ &+ \Big(\sum_{i=1, i \neq k}^{K} \mathbf{H}_i^H \mathbf{R}_{\tilde{v}_i \tilde{v}_i}^{-1} \mathbf{H}_i \mathbf{B}_i \mathbf{E}_i \mathbf{W}_i \mathbf{E}_i \mathbf{B}_i^H \mathbf{H}_i^H \mathbf{R}_{\tilde{v}_i \tilde{v}_i}^{-1} \mathbf{H}_i \Big) \mathbf{B}_k + \lambda \mathbf{B}_k \end{aligned} \tag{14}$$

Comparing the two gradient equations 12 and 14, we can see that the gradient of the two problem can be identical if:

$$u_k = \mathbf{W}_k \mathbf{E}_k \tag{15}$$

Therefore the paper [2] claims that the WMMSE minimization problem could be used to find the transmit filters $\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_k$ that also solve the WSR maximization problem. This "substitution" is the key to how Alternating Optimization works.

4.3 AO: WSRBF-WMMSE algorithm

Necessary Equations:

- The effective noise covariance matrix at user-k:

$$\mathbf{R}_{\tilde{v}_k\tilde{v}_k} = \mathbf{I}_k + \sum_{i=1,i\neq k}^{K} \mathbf{H}_k\mathbf{B}_i\mathbf{B}_i^H\mathbf{H}_k^H \tag{16}$$

- The MMSE receive filter at user-k:

$$\mathbf{A}_k^{MMSE} = \mathbf{B}_k^H\mathbf{H}_k^H(\mathbf{H}_k\mathbf{B}_k\mathbf{B}_k^H\mathbf{H}_k^H + \mathbf{R}_{\tilde{v}_k\tilde{v}_k})^{-1} \tag{17}$$

- The MSE matrix for user-k given that the MMSE recieve filter is applied:

$$\mathbf{E}_k = (\mathbf{I}_k + \mathbf{B}_k^H\mathbf{H}_k^H\mathbf{R}_{\tilde{v}_k\tilde{v}_k}^{-1}\mathbf{H}_k\mathbf{B}_k)^{-1} \tag{18}$$

- WMMSE transmit filter structure:

$$\bar{\mathbf{B}} = (\mathbf{H}^H\mathbf{A}^H\mathbf{W}\mathbf{A}\mathbf{H} + \frac{Tr(\mathbf{W}\mathbf{A}\mathbf{A}^H)\mathbf{I}_P}{E_{tx}})^{-1}\mathbf{H}^H\mathbf{A}^H\mathbf{W} \tag{19}$$

$$\mathbf{W}_{[QK\times QK]} = diag(\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_k)$$
$$\mathbf{A}_{[QK\times QK]} = diag(\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_k)$$
$$\mathbf{H}_{[QK\times P]} = [H_1^T, H_2^T, ..., H_k^T]$$

- scale the signal to satisfy the transmit power constraint:

$$\mathbf{B}^{WMMSE} = \sqrt{\frac{E_{tx}}{Tr(\bar{\mathbf{B}}\bar{\mathbf{B}}^H)}}\ \bar{\mathbf{B}} \tag{20}$$

WSRBF-WMMSE algorithm [2]

1. initialization: set $n = 0$, set $\mathbf{B}_k^n = \mathbf{B}_k^{init} \forall k$

2. loop until convergence:

   (a) increment counter: $n = n + 1$

   (b) compute $\mathbf{A}_k^n$ given $\mathbf{B}_i^{n-1} \forall i$ for all k using (17)

   (c) compute $\mathbf{W}_k^n$ given $\mathbf{B}_i^{n-1} \forall i$ for all k using (18) and (15)

   (d) compute $\mathbf{B}^n$ given $\mathbf{A}^n$ and $\mathbf{W}^n$ using (19) and (20)

3. find rate for user-k using :

$$R_k = \log \det(\mathbf{E}_k^{-1}) \tag{21}$$

The paper [2] also proposes another small variation of the above algorithm for diagonal weighting matrix, in which streams of each user are decorrelated.

4.4 Simulation Results

The most intuitive results of Alternating Optimisation algorithm is probably the two user rate region.

The code used to produce the results could be found at https://github.com/James-Xu-Xinyuan/Final-Year-Project/tree/master/WSR-WMMSE-MIMO-BC .

Throughout this subsection, some parameters are kept constant:

- $P = 4$ transmit antennas, $Q = 2$ receive antennas, and $K = 2$ users;
- The weight of user-1 is always 1; the weight of user-2 is $[-3, -1 : 0.05 : 1, 3]$ so the plot would be the convex hull of 43 points;
- Convergence tolerance is 10^{-6}, which could be relaxed if the calculation time is too long or only a rough estimate is needed;
- Unit of rate is always bits/s/Hz if not specified;

The following figures are reproductions based on Figure 5 of [2], demonstrating the achievable rate region of Alternating Optimization algorithm.

- Figure 3 shows the output of algorithm of 2 random user channels.

- Figure 4 and 5 studies the output with different user angles, that is fixing the user channel to: $\mathbf{H}_1 = [1, 1, 1, 1; 1, 1, 1, 1]$; and $\mathbf{H}_2 = [\exp(i * 0 * \theta), \exp(i * 1 * \theta), \exp(i * 2 * \theta), \exp(i * 3 * \theta); \exp(i * 0 * \theta), \exp(i * 1 * \theta), \exp(i * 2 * \theta), \exp(i * 3 * \theta)]$;;

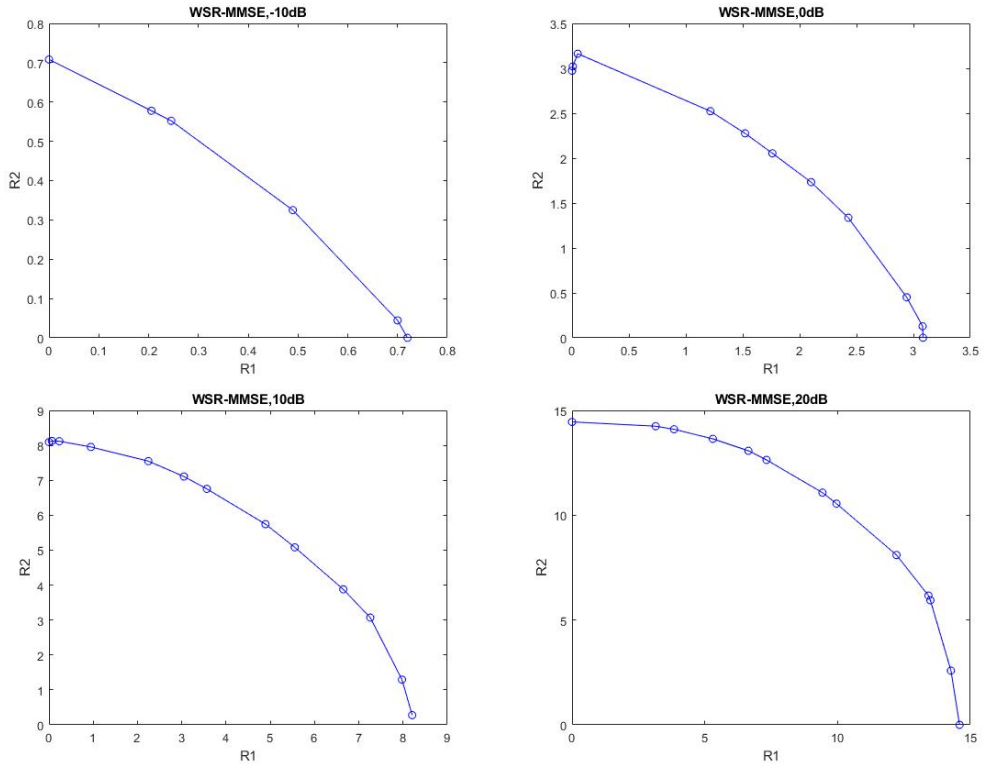


Figure 3: Achievable Rate region for WSRBF-WMMSE algorithm, of 2 random user channels

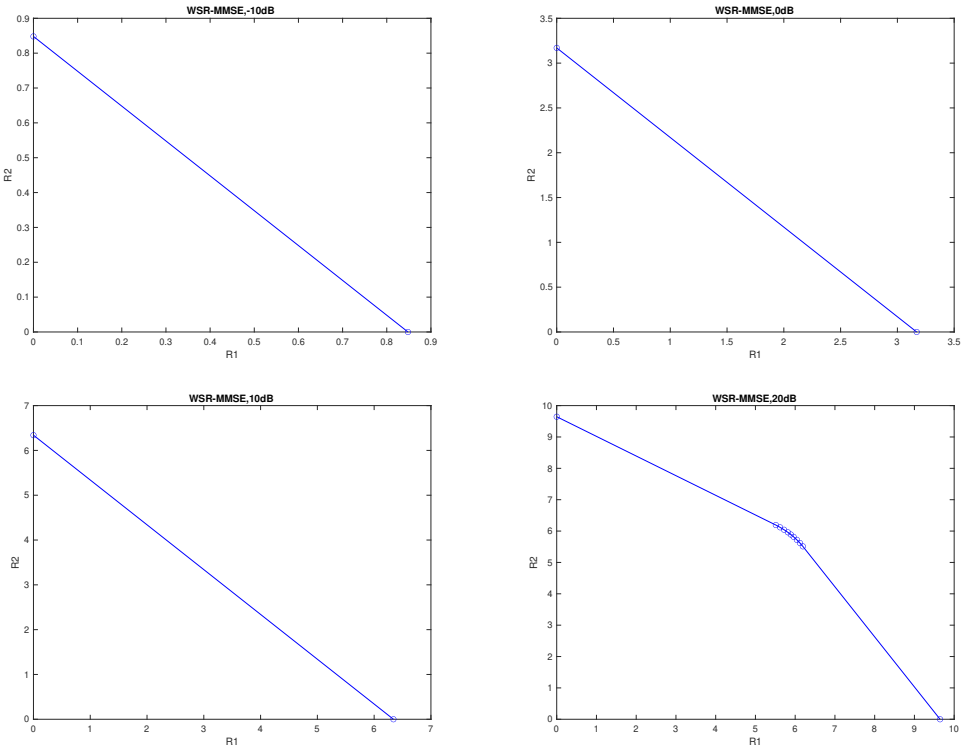


Figure 4: Achievable Rate region, user angle = $\frac{\pi}{9}$

- It can be observed from Figure 4 that when the user angle is small ($\theta = \frac{\pi}{9}$), or user channels are very aligned,the rate region looks smaller than that of the random channel. Also, the shape or rate region is more or less triangles (except 20dB), that increasing the rate of one user by giving him higher weight needs to sacrifice the rate achievable of the other user.

- Figure 5 shows that when the two users are less correlated ($\theta = \frac{\pi}{3}$), the rate region is larger than that in Figure 4; The marginal benefit (rate gained) of one user is larger than the marginal cost (rate sacrificed)

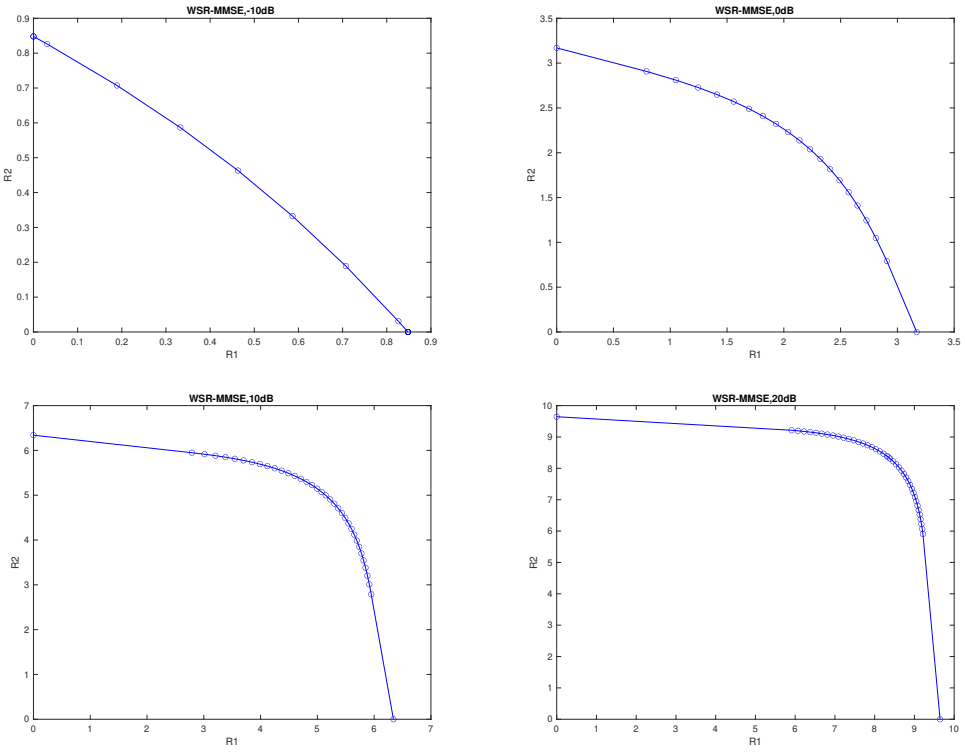


Figure 5: Achievable Rate region, user angle = $\frac{\pi}{3}$

when relative user weights changes, expcept when SNR is too low ($\gamma = -10dB$). The shape of rate region is close to a pentagon, which is ideal for a multi-user MIMO broadcast channel transmission.

- The rate regions at different SNR of $\theta = \frac{\pi}{2}$ looks almost the same as $\theta = \frac{\pi}{3}$, which is Figure 5, so it not included.

- Figure 6 is obtained by applying a bias of $\frac{1}{\sqrt{2}}$ on the channel matrix of user-2. It might be not very obvious, but observing the intersection of rate region with the two axis, it can be observed that rate of user 2 is schrinked proportionally compared to Figure 5.

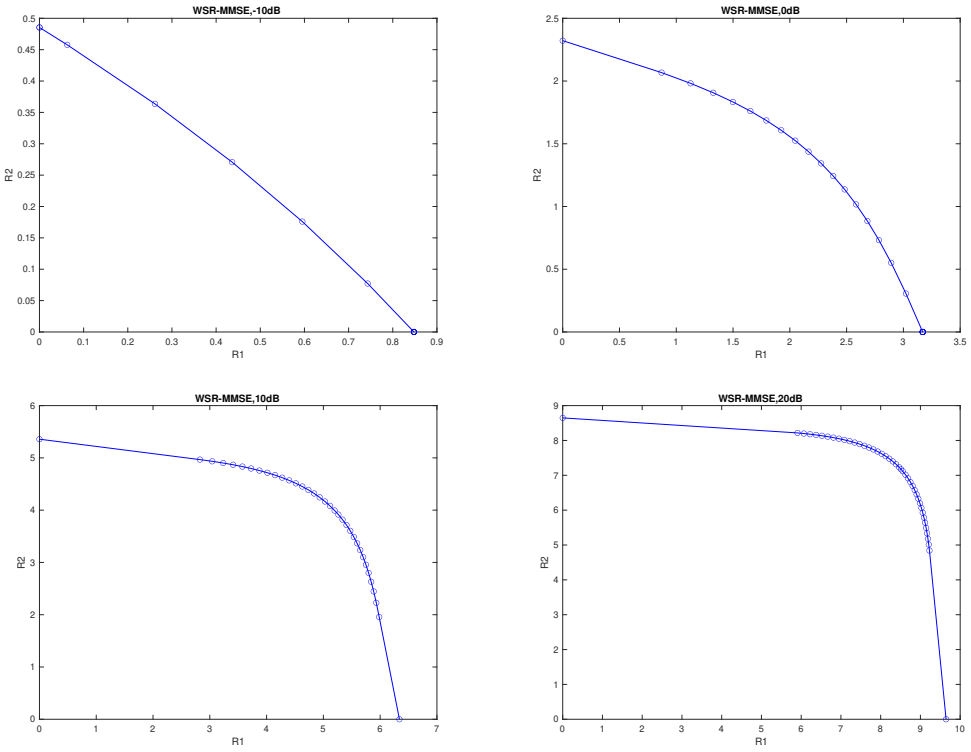


Figure 6: Achievable Rate region, bias = $\frac{1}{\sqrt{2}}$

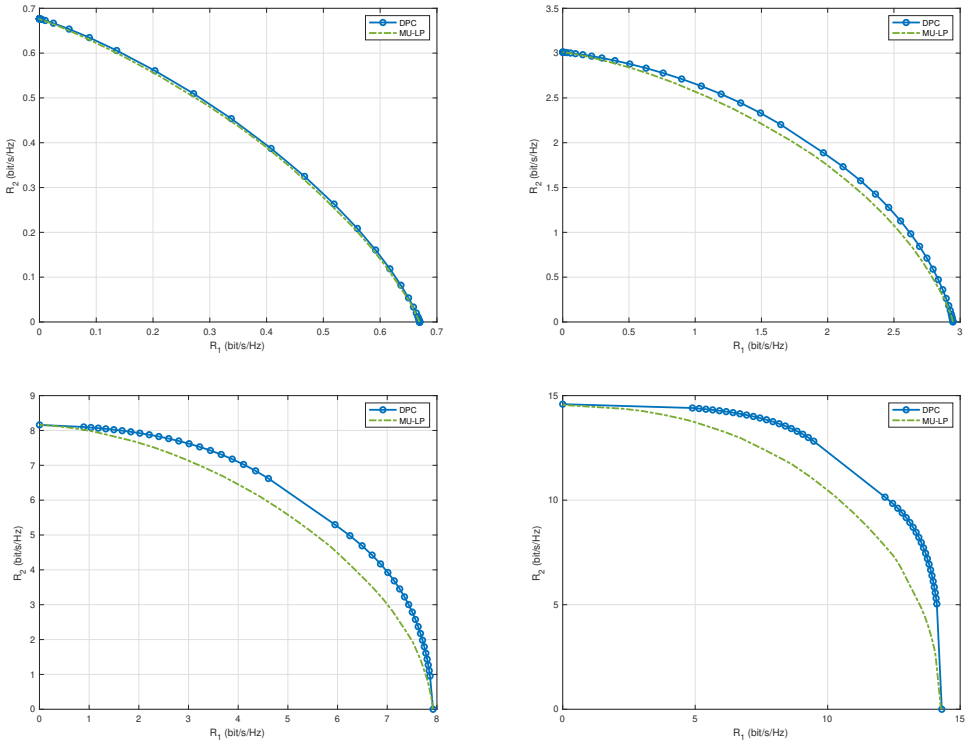


Figure 7: Comparing rate region with DPC

The orginal paper [2] claims that the the rate region achieved by WSRBF-WMMSE algorithm is nearly optimal, when compared to DPC rate region. Paper [12] suggests using the average rate region of 100 random channels to confirm that above claim. Using the DPC rate region code provided in the package, Figure 7 is obtained. The 4 sub-figures still follow the order of $[-10, 0, 10, 20]$dB.

5 Rate Splitting with Partial CSIT

This section studies how rate splitting handles partial CSIT and how maximum sum rate for linear precoding is achieved in a multiuser MISO downlink system, which is explained in paper [11]. As there is no perfect channel knowledge at transmitter, Ergodic Sum Rate (ESR) over a large number of transmission frames has been considered to be the key performance to maximize. Two important approaches has been introduced in [11]: Sample Average Approximation (SAA) and WMMSE approach.

As mention in the Background Section, the high multiplexing gain of MIMO system requires high quality CSIT. Because normally receivers like mobile phones can not cooperate with each other, it is the responsibility of the base station to deal with multi-user interference in the broadcast channel, usually in the form of transmit precoding. However, the precision of CSIT could be limited by many factors, such as finite feedback rate threshold, feedback delays, Doppler effects and so on.

Normally, the sum rate maximization problem could be very challenging because the problem is non-convex. However, the transformation of WSR problem into WMMSE problem proposed in [2] as explained in the previous section convert the problem into block-wise convex and could be solved by Alternating Optimization(AO).

5.1 System Model

- Assume imperfect CSIT is available with no feedback delay and errors dacay as $O(SNR^{-\alpha})$ and $\alpha \in [0, 1]$. More specifically in later simulation, variance of error $\sigma_e^2 = P_t^{-\alpha}$ and $\alpha = 0.6$ unless specified otherwise; α measures CSIT quality so is called 'Quality Scaling Factor'; α approaching infinity suggests perfect CSIT while α approaching zero suggests CSIT quality does not changed to due to change in SNR;

- In the system there are one base station with N_t antennas, and K single antenna users, with $N_t \geq K$;

- $\mathbf{h}_k \in \mathbb{C}^{N_T}$: the complex channel gain vector between transmitter and receiver of user-k;

- $\mathbf{x} \in \mathbb{C}^{N_T}$ is the transmit signal (already after precoding), which is constrained by input power: $\mathbb{E}(\mathbf{x}\mathbf{x}^H) \leq P_t$;

- $n_k \sim \mathcal{CN}(0, 1)$ represents the circular symmetric complex additive white Gaussian noise (AWGN) experienced at user-k, which is assume to be independent and identically distributed (i.i.d) among all users;

- the signal to noise ration at transmitter is defined as $\mathbf{SNR} = \frac{P_t}{\sigma_n^2}$, where the noise variance σ_n^2 is strictly positive and constant;

- $\hat{\mathbf{H}}$ represents an estimate of the channel. $\tilde{\mathbf{H}}$ represents the error of this estimation. So the real channel $\mathbf{H} = \hat{\mathbf{H}} + \tilde{\mathbf{H}}$. and the conditional probability density function $f_{\mathbf{H}|\hat{\mathbf{H}}}$ describes the CSIT error.

- $\mathbf{P} \triangleq [\mathbf{p}_c, \mathbf{p}_1, ..., \mathbf{p}_K]$, represents the procoder matrix containing precoder for the common stream and private stream for all users, in which $\mathbf{p}_i \in \mathbb{C}^{N_t}$.

- $\mathbf{s} \triangleq [s_c, s_1, ..., s_K]^T \in \mathbb{C}^{K+1}$ represents the symbols of common/private streams in a given channel use;

- Therefore, the transmit signal is given by: $\mathbf{x} = \mathbf{Ps} = \sum_{i=1}^{K} \mathbf{p}_i s_i$;

- transmit power constraint: $\text{tr}(\mathbf{PP}^H) \leq P_t$;

- $$S_{c,k} = |\mathbf{h}_k^H \mathbf{p}_c|^2, \quad S_k = |\mathbf{h}_k^H \mathbf{p}_k|^2, \quad I_k = \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{p}_i|^2 + \sigma_n^2 \tag{22}$$

  At a given channel state, the average received power of user-k is given by:

  $$T_{c,k} = S_{c,k} + S_k + I_k, \quad I_{c,k} = T_k = S_k + I_k. \tag{23}$$

- [11] The instantaneous Signal to Interference plus Noise Ratio (SINR) of private and common streams, after Successive Interference Cancellation, at the output of receiver-k:

  $$\gamma_{c,k} \triangleq \frac{S_{c,k}}{I_{c,k}} \quad \gamma_k \triangleq \frac{S_k}{I_k} \tag{24}$$

- Therefore the corresponding instantaneous achievable rate is: (assuming Gaussian Codebook)

  $$R_{c,k} = \log_2(1 + \gamma_{c,k}) \quad R_k = \log_2(1 + \gamma_k) \tag{25}$$

5.2 Average Sum Rate and Sample Average Approximation

Short term power constraint ($\text{tr}(\mathbf{P}\mathbf{P}^H) \leq P_t$) has been adopted because it makes the problem of maximizing ESR more tractable. With the imperfect available CSIT, solving the ESR maximization problem requires 1) the precoding scheme needs to be aware of the MU interference 2)transmission must be reliable or instantaneous / ergodic rate should not be overestimated by transmitter.

As there is no perfect channel knowledge, the base station can not estimate the instantaneous rate. However, it can calculate the Average Rate (AR) : $\bar{R}_k \triangleq \mathbb{E}_{H|\hat{H}}[R_k|\hat{\mathbf{H}}]$. This is not hard to understand - conditional expectation of instantaneous rate given the estimated channel. The paper [11] suggests a crucial like between ER and AR, that ER experience by the user could be expressed as:

$$\mathbb{E}_{(H,\hat{H})}[R_k(H,\hat{H})] = \mathbb{E}_{\hat{H}}[\mathbb{E}_{H|\hat{H}}[R_k(H,\hat{H})|\hat{\mathbf{H}}]] = \mathbb{E}_{\hat{H}}[\bar{R}_k(\hat{H})] \tag{26}$$

This has a nice interpretation [11]: by averaging the ARs over the distribution of $\hat{\mathbf{H}}$, we get the ERs. Applying this to the rate splitting context, we get the ASR maximization problem:

$$\begin{aligned} \max_{\bar{R}_c,\mathbf{P}} \quad & \bar{R}_c + \sum_{k=1}^{K} \bar{R}_k \\ s.t. \quad & \bar{R}_{c,k} \geq \bar{R}_c, \forall k \in \mathbb{K} \\ & tr(\mathbf{P}\mathbf{P}^H) \leq P_t \end{aligned} \tag{27}$$

The first line of inequality is equivalent to finding the minimum of AR for common stream, so that reliable transmission of common message to all users are possible. Note that the problem formulated in Equation 27 is stochastic in nature. In order to transform it into a deterministic nature, Sample Average Approximation (SAA) method is used. The original paper defines Sample Average Functions (SAFs): $\bar{R}_{c,k}^{(M)} \triangleq \frac{1}{M}\sum_{m=1}^{M} R_{c,k}^{(m)}$ and $\bar{R}_{k}^{(M)} \triangleq \frac{1}{M}\sum_{m=1}^{M} R_{k}^{(m)}$. That is the ARs are approximated by averaging rates associated with the m-th realization and each i.i.d realization is a sample drawn from the conditional probability distribution $f_{\mathbf{H}|\hat{\mathbf{H}}}(\mathbf{H}|\hat{\mathbf{H}})$. With the precoder matrix $\mathbf{P}$ fixed over the M realizations, the SAA of Equation 27 is converted into:

$$\begin{aligned} \max_{\bar{R}_c,\mathbf{P}} \quad & \bar{R}_c + \sum_{k=1}^{K} \bar{R}_k^{(M)} \\ s.t. \quad & \bar{R}_{c,k}^{(M)} \geq \bar{R}_c, \forall k \in \mathbb{K} \\ & tr(\mathbf{P}\mathbf{P}^H) \leq P_t \end{aligned} \tag{28}$$

The author of [11] proved that as M goes to infinity, the global optimum solutions of (deterministic) Equation 28 converge to the solutions of stochastic Equation 27.

5.3 WMMSE Algorithm: augmented AWSMSE minimization

The WMSE approach studied in the previous section and proposed in [2] helps to reformulate Equation 28 into an augmented WMSE problem. The word augmented is used to emphasize two differences: 1) Optimization variables includes the weights, 2) logarithms of the weights are incorporated into the cost function. Based on them, this subsection explains briefly the process to derive Rate-WMMSE relationship and the transformation to deterministic augmented AWSMSE minimiation problem [11]. This explanation is a bit brief but it covers all the necessary equations used in my MATLAB simulation codes.

1. the Optimum Minimum MSE equalizers:

$$g_{c,k}^{MMSE} = \mathbf{p}_c^H \mathbf{h}_k T_{c,k}^{-1} \quad and \quad g_k^{MMSE} = \mathbf{p}_k^H \mathbf{h}_k T_k^{-1} \tag{29}$$

2. the MMSEs:

$$\varepsilon_{c,k}^{MMSE} \triangleq \min_{g_{c,k}} \varepsilon_{c,k} = T_{c,k}^{-1} I_{c,k} \quad and \quad \varepsilon_k^{MMSE} \triangleq \min_{g_k} \varepsilon_k = T_k^{-1} I_k \tag{30}$$

3. the augemented WMSE:

$$\xi_{c,k} = u_{c,k}\varepsilon_{c,k} - \log_2(u_{c,k}) \quad and \quad \xi_k = u_k\varepsilon_k - \log_2(u_k) \tag{31}$$

4. Connecting Rate and WMMSE:

$$\xi_{c,k}^{MMSE} \triangleq \min_{u_{c,k},g_{c,k}} \xi_{c,k} = 1 - R_{c,k} \quad and \quad \xi_k^{MMSE} \triangleq \min_{u_k,g_k} \xi_k = 1 - R_k \tag{32}$$

5. the optimum MMSE weights

$$u^{\star}_{c,k} = u^{MMSE}_{c,k} \triangleq (\varepsilon^{MMSE}_{c,k})^{-1} \quad and \quad u^{\star}_{k} = u^{MMSE}_{k} \triangleq (\varepsilon^{MMSE}_{k})^{-1} \tag{33}$$

6. AR-AWMMSE is obtained by taking expectation over the conditional distribution $f_{\mathbf{H}|\hat{\mathbf{H}}}(\mathbf{H}|\hat{\mathbf{H}})$. This is an average version of Equation 32:

$$\begin{aligned} \bar{\xi}^{MMSE}_{c,k} &\triangleq \mathbb{E}_{H|\hat{H}}[\min_{u_{c,k},g_{c,k}} \xi_{c,k}|\hat{H}] = 1 - \bar{R}_{c,k} \\ \bar{\xi}^{MMSE}_{k} &\triangleq \mathbb{E}_{H|\hat{H}}[\min_{u_{k},g_{k}} \xi_{k}|\hat{H}] = 1 - \bar{R}_{k} \end{aligned} \tag{34}$$

7. following the same process of Sample Average Approximation, that is $\bar{\xi}^{(M)}_{c,k} \triangleq \frac{1}{M}\sum_{m=1}^{M} \bar{\xi}^{(m)}_{c,k}$ and $\bar{\xi}^{(M)}_{k} \triangleq \frac{1}{M}\sum_{m=1}^{M} \bar{\xi}^{(m)}_{k}$, the following relationship is obtained:

$$\begin{aligned} \bar{\xi}^{MMSE(M)}_{c,k} &\triangleq \min_{u_{c,k},g_{c,k}} \bar{\xi}^{(M)}_{c,k} = 1 - \bar{R}_{c,k} \\ \bar{\xi}^{MMSE(M)}_{k} &\triangleq \min_{u_{k},g_{k}} \bar{\xi}^{(M)}_{k} = 1 - \bar{R}_{k} \end{aligned} \tag{35}$$

With the help of all the equations above detailed in [11], the SAA problem is converted to an augmented AWSMSE minimization problem:

$$\begin{aligned} \min_{\bar{\xi}_c,\mathbf{P},\mathbf{U},\mathbf{G}} \quad & \bar{\xi}_c + \sum_{k=1}^{K} \bar{\xi}^{(M)}_{k} \\ s.t. \quad & \bar{\xi}^{(M)}_{c,k} \leq \bar{\xi}_c, \forall k \in \mathbb{K} \\ & tr(\mathbf{P}\mathbf{P}^{H}) \leq P_t \end{aligned} \tag{36}$$

Where $\mathbf{U}, \mathbf{G}$ contains the set of optimum MMSE equalizers and optimum MMSE weights of both common and private part, associated with each conditional realization. $\bar{\xi}_c$ is the AWMSE of common stream and $\bar{\xi}_k$ is the AWMSE of private stream of user-k.

5.4 Alternating Optimization Algorithm

Steps: [11]

- Initialization:
  - $n = 0$;
  - augmented AWSMSE $\mathbb{A}_{RS}^{[n]} = 0$;
    (for code in practice, a small non-zero value is used if tolerance is checked against a relative change of $\mathbb{A}_{RS}^{[n]}$)
  - $\mathbf{P}$: MRC-SVD initialization is suggested because of its good overall performance;
- Loop until $\mathbb{A}_{RS}^{[n]}$ converges;
  1. $n = n + 1$, $\mathbf{P}^{[n-1]} = \mathbf{P}^{[n]}$
  2. update $(\mathbf{G}, \mathbf{U})$ for a given $\mathbf{P}$
  3. update SAFs for all $k \in \mathbb{K}$:
     $\bar{\psi}_{c,k}$, $\bar{\psi}_k$, $\bar{\mathbf{f}}_{c,k}$, $\bar{\mathbf{f}}_k$, $\bar{t}_{c,k}$, $\bar{t}_k$, $\bar{u}_{c,k}$, $\bar{u}_k$, $\bar{v}_{c,k}$, $\bar{v}_k$;
  4. update the precoder: $\mathbf{P} = \arg\min \mathbb{A}_{RS}^{[n]}(P_t)$

Sample Average Functions:

Equalizers and weights of each realization are updated according to equations in the previous subsection and ensemble averages over the M realizations are obtained in each n-th iteration. The rest follows the same idea of Sample Average Approximation, with each realization given by:

$$
\begin{aligned}
t_{c,k}^{(m)} &= u_{c,k}^{(m)} |g_{c,k}^{(m)}|^2 & t_k^{(m)} &= u_k^{(m)} |g_k^{(m)}|^2 \\
\psi_{c,k}^{(m)} &= t_{c,k}^{(m)} \mathbf{h}_k^{(m)} \mathbf{h}_k^{(m)H} & \psi_k^{(m)} &= t_k^{(m)} \mathbf{h}_k^{(m)} \mathbf{h}_k^{(m)H} \\
\mathbf{f}_{c,k}^{(m)} &= u_{c,k}^{(m)} \mathbf{h}_k^{(m)} g_{c,k}^{(m)H} & \mathbf{f}_k^{(m)} &= u_k^{(m)} \mathbf{h}_k^{(m)} g_k^{(m)H} \\
v_{c,k}^{(m)} &= \log_2(u_{c,k}^{(m)}) & v_k^{(m)} &= \log_2(u_k^{(m)})
\end{aligned}
$$

In each iteration the new precoders are found by putting the augmented AWSMSE minimization problem into CVX optimisation ([8]):

$$
\begin{aligned}
\mathbb{A}_{RS}^{[n]}(P_t) = \min_{\bar{\xi}_c, \mathbf{P}} \quad & \bar{\xi}_c + \sum_{k=1}^{K} (\sum_{i=1}^{K} \mathbf{p}_i^H \bar{\psi}_k \mathbf{p}_i + \sigma_n^2 \bar{t}_k - 2\mathbb{R}(\bar{\mathbf{f}}_k^H \mathbf{p}_k) + \bar{u}_k - \bar{v}_k) \\
s.t. \quad & \mathbf{p}_c^H \bar{\psi}_{c,k} \mathbf{p}_c + \sum_{i=1}^{K} \mathbf{p}_i^H \bar{\psi}_{c,k} \mathbf{p}_i + \sigma_n^2 \bar{t}_{c,k} - 2\mathbb{R}(\bar{\mathbf{f}}_{c,k}^H \mathbf{p}_c) + \bar{u}_{c,k} - \bar{v}_{c,k} \leq \bar{\xi}_c, \forall k \in \mathbb{K} \\
& ||\mathbf{p}_c||^2 + \sum_{k=1}^{K} ||\mathbf{p}_k||^2 \leq P_t
\end{aligned}
\tag{37}
$$

5.5 Simulation Results

The simulation target of this section is 2-user ER region, so the results are based on reproduction of Figure 8 in paper [11].

The code used to produce the results could be found at https://github.com/James-Xu-Xinyuan/Final-Year-Project/tree/master/Partial%20CSIT .

Throughout this subsection, some parameters are kept constant:

- $Nt = 4$ transmit antennas, $Nr = 2$ receive antenna per user, and $K = 2$ users;
- The weight of user-1 is always 1; the weights of user-2 is $[-3, -1 : 0.05 : 1, 3]$ so each line onthe plot would be the convex hull of 43 points;
- Convergence tolerance is 10^{-4}, in relative sense, so that when the value changes less than 0.1% after a new iteration, it is considered as convergence. This number is taken because practically it is enough to produce a smooth output figure.
- Unit of rate is always bits/s/Hz if not specified;
- Except Figure 10, there is no bias between the 2 user channels. That is the norm of channel matrix are always equal. In Figure 10, a bias of $\frac{1}{\sqrt{2}}$ is applied to channel estimate (estimation error, and real channel matrix) of user 2.

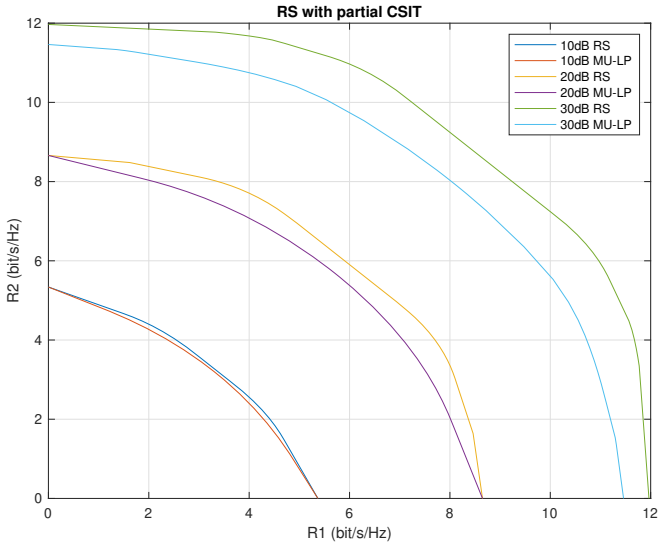


Figure 8: 2-user ER region for $\sigma_e^2 = P_t^{-0.6}$,$\theta = \frac{\pi}{3}$

Figure 8 to 10 study the output with different user angles, that is fixing the estimate of user channels to: $\mathbf{H}_{k=1,est} = [1, 1, 1, 1]$; and $\mathbf{H}_{k=2,est} = [\exp(i * 0 * \theta), \exp(i * 1 * \theta), \exp(i * 2 * \theta), \exp(i * 3 * \theta)]$;

Figure 8 clearly shows that with partial CSIT, Rate Splitting has larger rate region than non-RS (MULP). The performance difference is visually more manifest at higher SNR.

Figure 9 shows that the gap between rate splitting and NoRS is even larger when $\theta = \frac{\pi}{9}$ is small. Rate Splitting has greater advantage when the channel of two users are more aligned/dependent.

Figure 10 shows that when an extra bias is applied to user-2, the statement of Rate Splitting outperforms NoRS with 2-user ER region still holds,

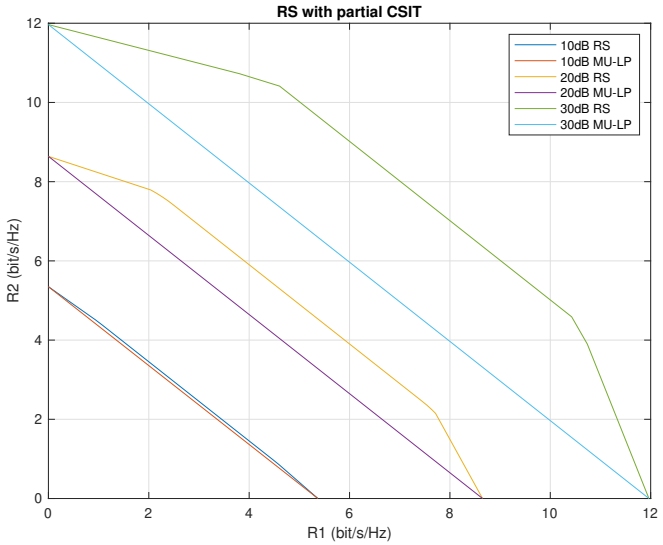


Figure 9: 2-user ER region for $\sigma_e^2 = P_t^{-0.6}$,$\theta = \frac{\pi}{9}$

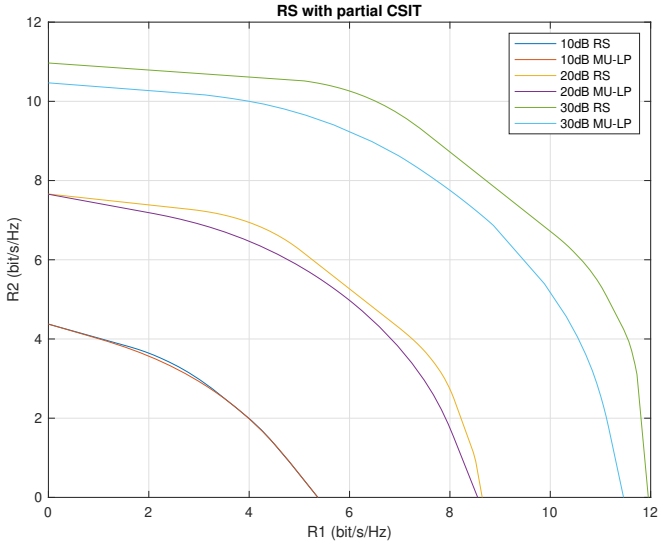
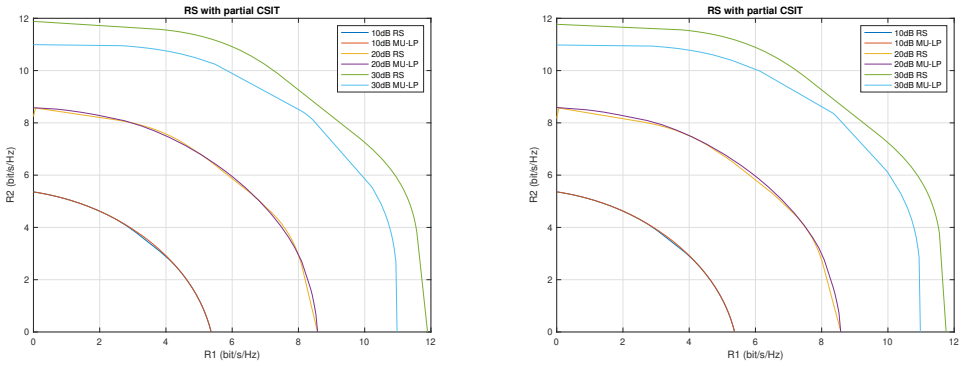


Figure 10: 2-user ER region for $\sigma_e^2 = P_t^{-0.6}$,$\theta = \frac{\pi}{3}$, bias $= \frac{1}{\sqrt{2}}$



(a) random estimated user channels (b) $\theta = \frac{\pi}{1.999}$

Figure 11: 2-user ER region for $\sigma_e^2 = P_t^{-0.6}$

although the achievable rates of user-2 has been squeezed down.

Figure 11a and 11b are not distinguishable with human eye. Figure 11a is produced by generating the user channel estimate using randn function. Figure 11b is generated with fixed user estimate channel with angle difference of $\frac{\pi}{1.999}$. This shows that when the user channels (real or estimate) are almost orthogonal, the increase of achievable rate provided by rate splitting is more limited, as compared to previous figures.

In Figure 12, the quality scaling factor α has been reduced from 0.6 for all previous figures to 0.06. With almost fixed quality w.r.t SNR, Rate Splitting could not offer much rate gain as expected. There is almost no additional performance on achievable rate at 10dB and 20dB SNR. The performance gap is still very small at SNR = 30dB (about 0.5 bits/sec/Hz). In addition, the shape of ER region looks a bit like rate region of TDMA, that increasing

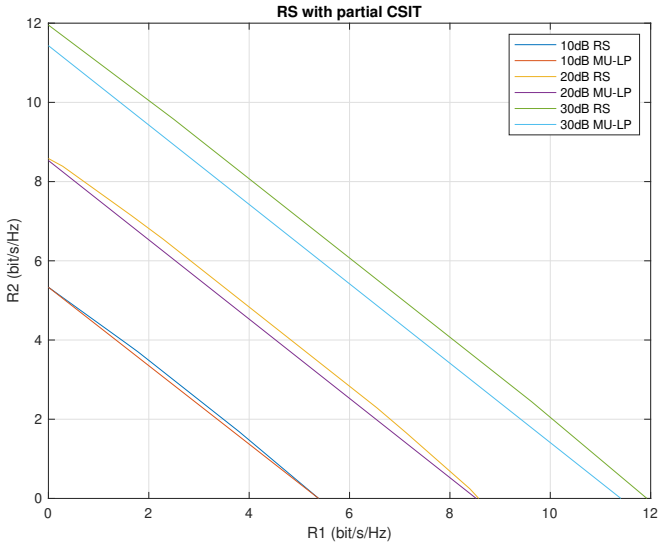


Figure 12: 2-user ER region for $\sigma_e^2 = P_t^{-0.06}$,$\theta = \frac{\pi}{3}$

the rate of one user needs to sacrifice the same amount from the other user.

6 Experiments on the feasibility of optimization algorithm

This section is about an experiment on the feasibility of optimization algorithm used in the previous section, after additional Quality of Service (QoS) constraints are added.

QoS certainly contains many performance metrics. The QoS requirement added in this section is the minimum achievable rate of each user (also referred to as rate threshold R_{th}) that needs to be guaranteed. Therefore, in addition to code used in the simulation of previous section (based on Equation 37), two extra lines of constraints are added: $R_k \geq R_{th}$, $k = 1, 2$. This constraint has been applied to MULP rate region optimization code as well. Previously, the constraint on minimum user rate could be thought as $R_k \geq 0$ but is always satisfied as $\log(1 + x), x \geq 0$ would never return something negative.

The introduction of extra constraints naturally changes the nature of the problem, and this could potentially makes the optimization infeasible. In CVX in MATLAB [8], one of the output parameters of an optimization problem is "cvx_status", with value either "Solved" or "Infeasible". For example, when R_{th} is outside the ER region, the QoS requirement could be satisfied with 0 probability; therefore the outcome should be "Infeasible" for all the input estimated channel matrix. Rate-WMMSE relationship from [11] is used to help construct the new optimization constraints : $\bar{\xi}_k^{MMSE(M)} = 1 - \bar{R}_k^{(M)}$.

6.1 Experiment 1: Equal Weight

Experiment Setup:

- Some basic parameters are kept the same as what is used in simulating 2-user ER region: (relative) tolerance = 10^{-3}, $N_t = 4$; $N_r = 1$, CSIT quality scaling: $P_e = P_t^{(} - 0.6)$; the number of realization in Sample Average Approximation is the same: $M = 1000$;
- SNRdB=[10 20 30]; feasibility is tested at 3 different SNR;
- 45 points per SNRdB has been tested : $R\_thresholds = [1, 1.5, 2 : 0.1 : 6, 6.5, 7]$;
- Specifically in this experiment, the weight pair of the two users is kept at equal weight: $u1 = u2 = 1$;
- For each tuple of (SNR, R_{th}), 100 randomly generated channel matrix are used. These channel matrix are inputs to the estimated channel in the algorithm; Feasibility is calculated as the percentage of of channel matrices that optimization is possible in MATLAB-CVX simulation;
- With Rate Splitting scheme, the two decoding orders are both tested. The channel is counted as feasible if at least one decoding order returns feasible.
- With Rate Splitting scheme, there are 4 different candidate initialization methods [11]: the common part precoder could be initialized with SVD or just random recoder and the private part precoder could use

MRC or ZF. The channel (at that decoding order) is counted as feasible if at least one initialization method returns feasible;

- With both RS and MULP, regardless of decoding order and precoder initialization methods, the Alternating Optimization should converge if the 1st iteration returns feasible and precoders. Even so, code has been designed to count it as infeasible if a later iteration returns "Infeasible". However, it has been observed during manually tracing and step in simulation that "infeasible" channel always fails immediately at the 1st iteration. The student has never observed "infeasible" in later iteration output if the 1st iteration is "Solved", but this could be limited by the number of observations;

- Another option of detecting infeasible optimization is to put "try-catch" statement around the CVX code. However, this is a bit inconvenient that the infeasbile optimization would still return 2 private and 1 common precoders but filled with "NaN". Error would only be catched in the next iteration due to this value. And "try-catch" could potentially cover up other types of error. So in the end "cvx_status" has been used as the primary indicator.

- For optimization in MULP, there is no decoding order and there are only 2 different precoder initialization methods because there is no precoder for common part.

- This experiment has colossal computational burden: The computation burden of 1 SNR value and the array of R_{th} is at similar level compared with finding the rate region of 1 SNR value and the array of user weight

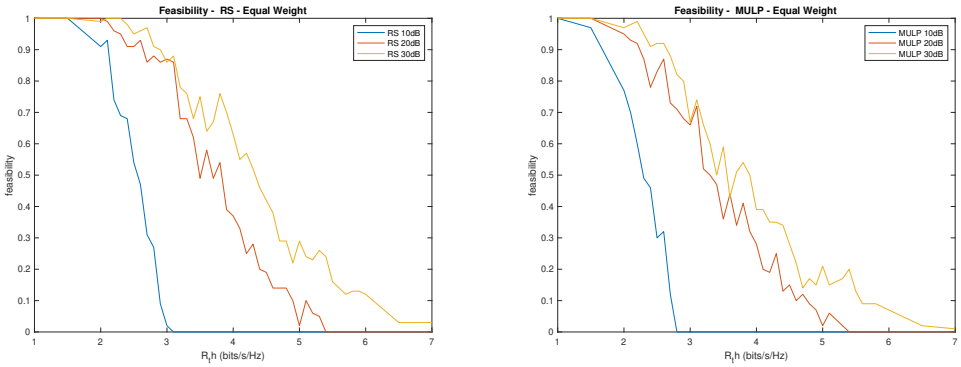


Figure 13: Feasibility Experiment on RS and MULP optimization algorithm, with equal user weight and three SNRdB

pairs, which always took more than 24 hours on my personal computer. But in this experiment, each tuple of (SNR, R_{th}) needs to test 100 random channels (ideally larger number is needed because the output lines were still not very smooth in the end). This simulation would never be successfully completely on my old MacBook. Thanks to my dear Supervisor who granted access to Imperial HPC and thanks to Lina who clarified the trick of batch processing, the timely simulations were enabled remotely and in parallel on the super computer.

Figure 13 shows the simulation results of this experiment:

1. Sanity check:
   - optimization should always be possible when R_{th} is very small;
   - optimization should never be feasible when R_{th} is outside the rate region;

- the higher the R_{th}, the lower the feasiblity should be.

2. It seems that on the optimization feasibility side, MULP and RS has no difference in nature. The student believes that the difference of curve position with equal SNR is just a result of Rate Region difference shown in the previous section.

3. The points in figures with feasibility (y-value) between 0 and 1 suggests that although with R_{th} inside the rate region, optimization maybe infeasible for some channels. After setting up breakpoints and temporarily remove rate threshold constraints to monitor optimization progress, an interesting observation has been made by the student. Although Alternating Optimization has been shown with very good convergence property [2] [11], the convergence of WMMSE would still take at least a few iterations. In the 1st iteration, precoders for common and private part together may use only fraction of total available power, thus rates calculated with these precoders are below threshold. This could be one possible reason for infeasible optimization of a particular channel after introducing the QoS rate constraint. Also, this phenomenon is easier to spot out at high SNR (30dB).

6.2 Experiment 2: different user weight pairs and fixed SNR

Experiment 2 is a variation of Experiment 1. Changes in experiment setups include:

- instead of SNRdB = 10, 20 and 30, in Experiment 2a SNR is fixed to 10dB and in Experiment 2b SNR is fixed to 20dB.
- the weight pair between the 2 users are no longer constant: in Experiment 2a u2 = [3,1,1/3] and in Experiment 2b u2 = [10,3,1,1/3,0.1]. u1 has always been fixed to a vector of 1s.
- In Experiment 2a R_thresholds = [1,1.5,2:0.1:6,6.5,7]. In Experiment 2b, it is changed to R_thresholds = [1,1.5:0.1:5.5,6] in the hope that the region of interest could be better tested.
- Experiment 2b was carried out intended to verify the observation made in Experiment 2a.

From Figure 14, it seems that changing the relative weight between the two users doesn't impact the feasiblity of optimization. Figure 15 confirms this.

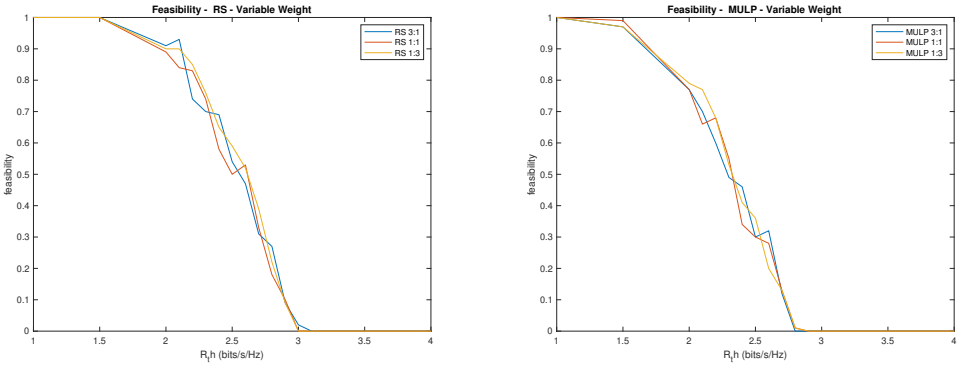


Figure 14: Experiment 2a

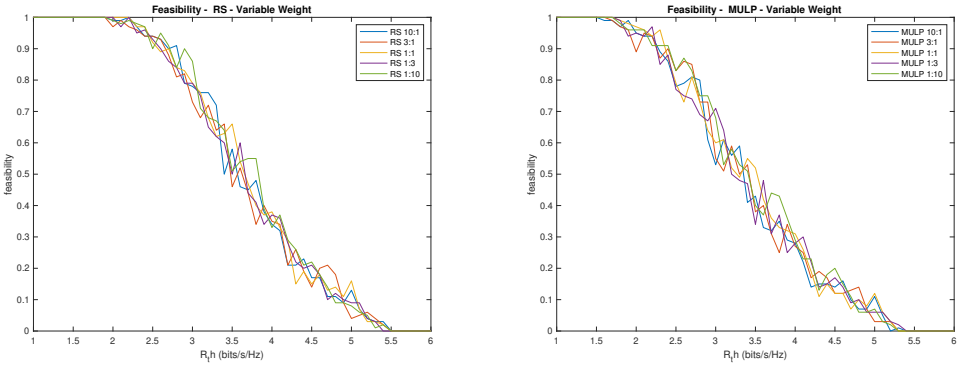


Figure 15: Experiment 2b

7 Evaluation

This final year project is my first time doing a research oriented project. Albeit good or bad, it has been a significant part of my final year and it gives me a glimpse into research in wireless communication.

Personally, one way to evaluate the project is simply just looking back. In the Autumn term, there were so many words that I don't have any idea, in the papers that have been brought up in the first project meeting. Although I have been reminded that this project is very demanding, pressure still stroke me hard. However, after many chapters of textbook been self-studied and numerous pages of note got recorded, I started to understand what is going on in those academic literature, bit by bit. Knowledge was absorbed like water into sponges. I feel that self-study, with some objectives (the project) in mind, has been very enjoyable and productive. This process also served as a good preview of Wireless Communication course in the spring term. Then in the spring term, a practice presentation has been scheduled by my supervisor to remind me and other students the project, for which I am really grateful. Then the pandemic hit UK, lock-down and all other related things impacted heavily on learning in this central London University, including progree in this project. In the summer term, only 4 weeks were left after I finished my last examination. But I am glad that in 2 weeks I was able to learn the new tool of HPC (college super computer) and batch processing, so that I could carry out the feasibility experiments. It always fells good that something could be done before the end, even though it might be just little. Half way writing this report, my laptop screen was suddenly broken and I

panicked. But luckily I was allowed to use the TV in the common room to keep on my writing job.

However, a lot could be done better. First, as reflected on the length of this report, the amount of work done in this final year project is a bit deficient. It is one thing to plan what to do, it is another to actually balance project work with lessons, coursework and exams. More time should have been spent on the project and lessons have been learnt in project time management, planning and adjusting when incidents come at surprise. Second, a lot of time has been spent on writing my own code, a large part of which was of course debugging. Finding out what has gone wrong in the code has been very time consuming and frustrating. But successfully dealing the problems has given my a bit more understanding to the content. The last but not the least, I need to be more resilient when facing difficulty. Throughout the project I have been very nervous and sometimes I don't think I can do it subconsciously. If I had more determination and deliberation, maybe I could do this project better.

8 Conclusion and Future Work

To sum up, the major work of this final year project could be divided into 3 parts. The first part is to learn the WSRBF-WMMSE or Alternating Optimization algorithm from [2]. The second part is to learn now Rate Splitting deal with partial CSIT and its ER region from [11]. For each of these two parts, reproducing a figure in the original paper has been the key learning outcome. The last part is doing the feasibility experiment on the optimization algorithm with extra QoS rate constraints. The backbone of this project has been optimization methods.

All the code files, data files and figures could be found my personal GitHub repository: https://github.com/James-Xu-Xinyuan/Final-Year-Project .

The first option for future works is naturally what has been planned but is not achieved. Based on coursework 3 of Wireless Communication module, Rate Splitting could be applied to simulation with user deployment with shadow fading and long term SINR, proportional fair scheduling, interference between cells, users and streams, variation of channel space and time correlation, and so on. Another good possibility is to apply Rate Splitting to Cell Free Massive MIMO mentioned in the background section.

PS: Although this project has been pretty hard for me, I still like what I have experienced and a lot has been learnt. Furthermore, I want to express again my greatest gratitude to my supervisor Dr Bruno Clerckx and Dr Yijie (Lina) Mao, whose support is vital for the completion of this project.

Bibliography

[1] Acampora Anthony and Haipeng Jin. "Bounds on the Outage-Constrained Capacity Region of Space-Division Multiple-Access Radio Systems". In: *EURASIP Journal on Advances in Signal Processing* 2004 (Aug. 2004). DOI: 10.1155/S1110865704402297.

[2] S. S. Christensen et al. "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design". In: *IEEE Transactions on Wireless Communications* 7.12 (Dec. 2008), pp. 4792–4799. ISSN: 1558-2248. DOI: 10.1109/T-WC.2008.070851.

[3] B. Clerckx et al. "Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution". In: *IEEE Communications Magazine* 54.5 (May 2016), pp. 98–105. ISSN: 1558-1896. DOI: 10.1109/MCOM.2016.7470942.

[4] B. Clerckx et al. "Rate-Splitting Unifying SDMA, OMA, NOMA, and Multicasting in MISO Broadcast Channel: A Simple Two-User Rate Analysis". In: *IEEE Wireless Communications Letters* (June 2019), pp. 1–1. ISSN: 2162-2345. DOI: 10.1109/LWC.2019.2954518.

[5] M. Costa. "Writing on dirty paper (Corresp.)" In: *IEEE Transactions on Information Theory* 29.3 (May 1983), pp. 439–441. ISSN: 1557-9654. DOI: 10.1109/TIT.1983.1056659.

[6] M. Dai et al. "A Rate Splitting Strategy for Massive MIMO With Imperfect CSIT". In: *IEEE Transactions on Wireless Communications* 15.7 (2016), pp. 4611–4624.

[7] A.E. Gamal and Y.H. Kim. *Network Information Theory*. Cambridge University Press, 2011. ISBN: 9781139503143. URL: https://books.google.co.uk/books?id=Ack2AAAAQBAJ.

[8] Michael Grant and Stephen Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. http://cvxr.com/cvx. Mar. 2014.

[9] C. Hao and B. Clerckx. "MISO Networks With Imperfect CSIT: A Topological Rate-Splitting Approach". In: *IEEE Transactions on Communications* 65.5 (2017), pp. 2164–2179.

[10] C. Hao, Y. Wu, and B. Clerckx. "Rate Analysis of Two-Receiver MISO Broadcast Channel With Finite Rate Feedback: A Rate-Splitting Approach". In: *IEEE Transactions on Communications* 63.9 (Sept. 2015), pp. 3232–3246. ISSN: 1558-0857. DOI: 10.1109/TCOMM.2015.2453270.

[11] H. Joudeh and B. Clerckx. "Sum-Rate Maximization for Linearly Precoded Downlink Multiuser MISO Systems With Partial CSIT: A Rate-Splitting Approach". In: *IEEE Transactions on Communications* 64.11 (Nov. 2016), pp. 4847–4861. ISSN: 1558-0857. DOI: 10.1109/TCOMM.2016.2603991.

[12] Yijie Mao, Bruno Clerckx, and Victor O.K. Li. "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA". In: *EURASIP Journal on Wireless Communications and Networking* 2018.1 (May 2018), p. 133. ISSN: 1687-1499. DOI: 10.1186/s13638-018-1104-7. URL: https://doi.org/10.1186/s13638-018-1104-7.

[13] Yijie Mao, Bruno Clerckx, and Victor O.K. Li. "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA". In: *EURASIP Journal on Wireless Communications and Networking* 2018.1 (May 2018), p. 133. ISSN: 1687-1499. DOI: 10.1186/s13638-018-1104-7. URL: https://doi.org/10.1186/s13638-018-1104-7.

[14] H. Q. Ngo et al. "Cell-Free Massive MIMO Versus Small Cells". In: *IEEE Transactions on Wireless Communications* 16.3 (Mar. 2017), pp. 1834–1850. ISSN: 1558-2248. DOI: 10.1109/TWC.2017.2655515.

[15] P. Viswanath and D. N. C. Tse. "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality". In: *IEEE Transactions on Information Theory* 49.8 (Aug. 2003), pp. 1912–1921. ISSN: 1557-9654. DOI: 10.1109/TIT.2003.814483.

[16] H. Viswanathan, S. Venkatesan, and H. Huang. "Downlink capacity evaluation of cellular networks with known-interference cancellation". In: *IEEE Journal on Selected Areas in Communications* 21.5 (2003), pp. 802–811.

[17] H. Weingarten, Y. Steinberg, and S. S. Shamai. "The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel". In: *IEEE Transactions on Information Theory* 52.9 (2006), pp. 3936–3964.