Imperial College London
Department of Electrical and Electronic Engineering

Novel Transmission and Beamforming Strategies for Multiuser MIMO with Various CSIT Types

Mingbo Dai

November 2016

Supervised by Dr. Bruno Clerckx

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Electrical and Electronic Engineering of Imperial College London
and the Diploma of Imperial College London

Statement of Originality

I hereby certify that I am the sole author of this thesis and this thesis is the result of my own work. I certify that, to the best of my knowledge, any ideas/techniques or other material from the work of others included in my thesis have been fully acknowledged in accordance with the standard referencing practices. The material of this thesis has not been and will not be submitted for another degree at any other university or institution.

Copyright Declaration

To my family and girlfriend

Abstract

In multiuser multi-antenna wireless systems, the transmission and beamforming strategies that achieve the sum rate capacity depend critically on the acquisition of perfect Channel State Information at the Transmitter (CSIT). Accordingly, a high-rate low-latency feedback link between the receiver and the transmitter is required to keep the latter accurately and instantaneously informed about the CSI. In realistic wireless systems, however, only imperfect CSIT is achievable due to pilot contamination, estimation error, limited feedback and delay, etc. As an intermediate solution, this thesis investigates novel transmission strategies suitable for various imperfect CSIT scenarios and the associated beamforming techniques to optimise the rate performance.

First, we consider a two-user Multiple-Input-Single-Output (MISO) Broadcast Channel (BC) under statistical and delayed CSIT. We mainly focus on linear beamforming and power allocation designs for ergodic sum rate maximisation. The proposed designs enable higher sum rate than the conventional designs. Interestingly, we propose a novel transmission framework which makes better use of statistical and delayed CSIT and smoothly bridges between statistical CSIT-based strategies and delayed CSIT-based strategies.

Second, we consider a multiuser massive MIMO system under partial and statistical CSIT. In order to tackle multiuser interference incurred by partial CSIT, a Rate-Splitting (RS) transmission strategy has been proposed recently. We generalise the idea of RS into the large-scale array. By further exploiting statistical CSIT, we propose a novel framework Hierarchical-Rate-Splitting that is particularly suited to massive MIMO systems.

Third, we consider a multiuser Millimetre Wave (mmWave) system with hybrid analog/digital precoding under statistical and quantised CSIT. We leverage statistical CSIT to design digital precoder for interference mitigation while all feedback overhead is reserved for precise analog beamforming. For very limited feedback and/or very sparse channels, the proposed precoding scheme yields higher sum rate than the conventional precoding schemes under a fixed total feedback constraint. Moreover, a RS transmission strategy is introduced to further tackle the multiuser interference, enabling remarkable saving in feedback overhead compared with conventional transmission strategies.

Finally, we investigate the downlink hybrid precoding for physical layer multicasting with a limited number of RF chains. We propose a low complexity algorithm to compute the analog precoder that achieves near-optimal max-min performance. Moreover, we derive a simple condition under which the hybrid precoding driven by a limited number of RF chains incurs no loss of optimality with respect to the fully digital precoding case.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Bruno Clerckx for his invaluable guidance throughout my PhD study. I greatly appreciate his insightful suggestions on research directions and continuous assistance in paper revision. I do learn quite a bit from his comprehensive and profound understanding of both science and industry, and his dedication to research. It's a privilege to have worked with him to make this 'imperfect' world better.

I would like to thank Professor K.K. Leung at Imperial College London and Professor Zhiguo Ding at Lancaster University for their helpful comments and advices on this thesis.

I would like to extend my sincere gratitude to my friends and colleagues from the Communications and Signal Processing group at Imperial College: Chenxi for fruitful discussion on various technical problems, Yang for accompanying me in the lab for countless (150) weekends, Hamdi and John for encouraging me and memorable Gym/Basketball time, Enrico and He for all the useful and useless conversations.

Special heartfelt thanks go to my wonderful family for their understanding and support of every decision I made. Their boundless love takes me all the way to explore the world. And last, but by no means least, I am truly thankful to my girlfriend Jane (QingQing) for her love and faith in me. Our dreams about future have been and will always be a source of motivation for me to move ahead.

Mingbo Dai
London, November 2016.

Contents

1. Introduction **17**
1.1. Multiuser MIMO Systems . . . 17
1.2. Performance Metrics . . . 18
1.3. Channel Models . . . 18
1.3.1. I.I.D. Rayleigh Channels . . . 19
1.3.2. Correlated Rayleigh Channels . . . 19
1.3.3. Finite Scatterer Channels . . . 20
1.4. Precoding Techniques . . . 20
1.5. CSIT Types . . . 22
1.5.1. Perfect CSIT . . . 22
1.5.2. Statistical CSIT . . . 22
1.5.3. Delayed CSIT . . . 23
1.5.4. Partial CSIT . . . 23
1.6. Mathematical Tool . . . 24
1.6.1. Convex Optimization . . . 24
1.6.2. Deterministic Equivalent . . . 24
1.7. Literature Review . . . 25
1.7.1. Transmission Strategies with Statistical CSIT Only . . . 25
1.7.2. Transmission Strategies with Delayed CSIT Only . . . 26
1.7.3. Transmission Strategies with Statistical and Delayed CSIT . . . 27
1.7.4. Transmission Strategies with Partial CSIT Only . . . 27
1.7.5. Transmission Strategies with Partial and Statistical CSIT . . . 30
1.7.6. Beamforming Strategies with Quantised CSIT Only . . . 31
1.7.7. Beamforming Strategies for Multicasting with Perfect CSIT . . . 32
1.8. Summary of Contributions . . . 32
1.9. Publications . . . 35
1.10. Notations . . . 36

2. A Novel Transmission Strategy for Two-User MISO Broadcast Channel with Statistical and Delayed CSIT **37**
2.1. System Model . . . 37
2.2. SBF Strategy . . . 38

2.3. AMAT Strategy 40
2.3.1. Rate Approximation 40
2.3.2. Precoder Design 41
2.4. SAMAT Strategy 44
2.5. SAMAT Discussion 49
2.6. Performance Evaluation 50
2.6.1. Precoders Comparison for SBF 51
2.6.2. Precoders Comparison for AMAT 51
2.6.3. Performance of SAMAT 53
2.7. Summary and Conclusion 56

3. A Novel Transmission Strategy for Multiuser Massive MIMO System with Partial and Statistical CSIT 57
3.1. System Model 57
3.2. Rate-Splitting 58
3.2.1. RS Transmission 59
3.2.2. Precoder Design 60
3.2.3. Asymptotic Rate Analysis 61
3.2.4. Power Allocation 62
3.3. Hierarchical-Rate-Splitting 63
3.3.1. HRS Transmission 63
3.3.2. Precoder Design 65
3.3.3. Asymptotic Rate Analysis 66
3.3.4. Power Allocation 67
3.4. Performance Evaluation 69
3.4.1. RS 70
3.4.2. HRS 73
3.4.3. RS vs. HRS 77
3.5. Summary and Conclusion 78

4. A Novel Multiuser Millimeter Wave Beamforming Strategy with Quantised and Statistical CSIT 79
4.1. System Model 80
4.2. Multiuser Hybrid Precoding 81
4.2.1. One-Stage Feedback Scheme 82
4.2.2. Two-Stage Feedback Scheme 85
4.2.3. One-stage vs. Two-stage Feedback Scheme 87
4.2.4. Signalling and Feedback Protocol 88
4.3. Rate Splitting with Hybrid Precoding 89
4.3.1. Precoder design 90

4.3.2. Power allocation design 91
4.4. Performance Evaluation 93
4.4.1. Multiuser Hybrid Precoding 93
4.4.2. Rate Splitting with Hybrid Precoding 98
4.5. Summary and Conclusion 99

5. A Novel Hybrid Precoding Strategy for Physical Layer Multicasting with Perfect CSIT 101
5.1. Problem Formulation 101
5.1.1. A Sufficient Number of RF Chains 102
5.1.2. A Limited Number of RF chains 102
5.2. Proposed Solutions 103
5.2.1. A Low Complexity Algorithm 103
5.2.2. Achieving Optimum with a Reduced Number of RF Chains . . 104
5.3. Performance Evaluation 105
5.4. Summary and Conclusion 108

6. Conclusion and Future work 109
6.1. Conclusion 109
6.2. Future Work 110

Bibliography 113

A. Proofs for Chapter 2 123
A.1. Lemma 123
A.2. Lemma 123
A.3. Proof of Theorem 2.1 124
A.4. Proof of Proposition 2.1 125
A.5. Proof of Proposition 2.2 126
A.6. Proof of Proposition 2.3 126
A.7. Proof of Theorem 2.2 128

B. Proofs for Chapter 3 129
B.1. Proof of Proposition 3.1 129
B.2. Proof of Proposition 3.2 129
B.3. Proof of Proposition 3.3 130
B.4. Proof of Corollary 3.1 132

C. Proofs for Chapter 4 133
C.1. Proof of Proposition 4.1 133
C.2. Proof of Proposition 4.3 133

List of Figures

1.1. Diagram of multiuser MISO downlink system. 18
1.2. Diagram of hybrid precoding method. 21

2.1. Convergence of iterative algorithm 1&2. 43
2.2. Block diagram of the proposed SAMAT scheme. 44
2.3. Ergodic sum rate of SBF with WE and with GE precoder in spatially correlated channel. 51
2.4. Ergodic sum rate of AMAT with various precoders in a setup with $M = 4$. 52
2.5. Ergodic sum rate of AMAT with various precoders in a setup with $M = 2$. 52
2.6. Ergodic sum rate comparison between the proposed SAMAT scheme and various baselines, in a setup with SNR = 20 dB, $|\phi_A - \phi_B| \leq \frac{\pi}{2}, M = 2$. 54
2.7. Ergodic sum rate comparison between the proposed SAMAT scheme and various baselines, in a setup with $|\phi_A - \phi_B| \leq \frac{\pi}{2}, |t_A| = 0.95, |t_B| = 0.9, M = 2$. 54
2.8. Comparison of the ergodic sum rate vs. SNR between SAMAT and baselines, $|t_A| = 0.95$, $|t_B| = 0.9$, $\phi_A, \phi_B \in U(0, 2\pi)$, $M = 4$. 55

3.1. Sum rate comparison: RS asymptote (AS) approximation vs. RS Monte Carlo (MC) simulation for various CSIT qualities and SNRs. 70
3.2. Sum rate comparison: RS vs. conventional No-RS and TDMA, in a setup with $\tau^2 = 0.4$. 71
3.3. Sum rate gain of RS over conventional No-RS for various antennas/users ratio ($\eta = M/K$). 71
3.4. Sum rate gain of MBF-precoded RS over conventional RZF-precoded No-RS, in a setup with $K = 5$. 72
3.5. Sum rate comparison: HRS asymptote (AS) approximation vs. HRS Monte-Carlo (MC) simulation in disjoint eigen-subspace scenario. . . . 74
3.6. Sum rate comparison: HRS asymptote (AS) approximation vs. HRS Monte-Carlo (MC) simulation in overlapping eigen-subspace scenario. 74
3.7. Sum rate comparison: HRS vs. various baselines under partial CSIT in disjoint eigen-subspace scenario. 76
3.8. Sum rate comparison: HRS vs. various baselines under partial CSIT in overlapping eigen-subspace scenario. 76

3.9. Sum rate comparison: HRS vs. RS and No-RS with one/two-tier precoding under various CSIT qualities, in a setup with SNR = 30 dB. . 77

4.1. Block diagram of multiuser mmWave downlink system model with hybrid precoding and limited feedback. Solid lines indicates the feedback requirement of the hybrid precoding scheme in Section 4.2.1. Dash lines represents the additional requirement of the hybrid precoding scheme in Section 4.2.2. 80
4.2. Illustration of signalling and feedback procedure for various schemes. . 88
4.3. Comparison among various hybrid precoding schemes, in a setup with $B = 6, L = 2$. 94
4.4. Comparison among various hybrid precoding schemes, in a setup with $B = 6, L = 15$. 94
4.5. Comparison among various hybrid precoding schemes, in a setup with $B = 10, L = 2$. 95
4.6. Comparison among various hybrid precoding schemes, in a setup with $B = 10, L = 15$. 96
4.7. Comparison among the hybrid precoding schemes with various M, in a setup with $L = 4, B = 8$, SNR = -5 dB. 97
4.8. Comparison among the hybrid precoding schemes with various B, in a setup with $L = 4, M = 32$, SNR = 5 dB. 97
4.9. Comparison between No-RS with extra feedback and RS, in a setup with $M = 32, K = 4, L = 8, B_{RF} = 4$ and ZF digital precoder. 98
4.10. Comparison between No-RS with extra feedback and RS, in a setup with $M = 32, K = 4, L = 8, B_{RF} = 4$ and SLNR-based digital precoder. 99

5.1. Max-min rate comparison between the proposed hybrid precoding design and various baselines, in a setup with $M = 6, K = 2$. 106
5.2. Max-min rate CDF comparison between the proposed hybrid precoding design and various baselines, in a setup with $M = 8, K = 3$. 106
5.3. Max-min rate comparison between the proposed hybrid precoding design and various baselines, in a setup with $M = 10, K = 3$. 108

List of Tables

1.1. Feedback rate and latency requirements for various CSIT types. 22

2.1. Precoder optimisation algorithms for AMAT. 42

4.1. Hybrid precoding schemes 83
4.2. SCA algorithm 91
4.3. Overall RS transmission with hybrid precoding 92

5.1. Precoding design algorithm 104

1. Introduction

Wireless network continues to transform the way people communicate and access information. The currently deployed 3G/4G wireless communication technologies has delivered the ubiquitous high-speed mobile broadband services. Further developing 5G technologies that enable highly efficient and reliable networks will come to refine the transmission strategies along with the corresponding beamforming designs.

Multiple-Input-Multiple-Output (MIMO) technology has become an essential component in 3G/4G/5G wireless communication standards. Compared to Single-Input-Single-Output (SISO), MIMO provides spatial multiplexing and diversity by employing multiple antennas at the transmitter and receiver. In addition, multiple antennas allow the base-station to simultaneously serve multiple users on the same time-frequency resource in both the uplink and the downlink. To realise multiuser transmission specifically in the downlink, perfect (i.e., accurate and up-to-date) Channel State Information at the Transmitter (CSIT) is required to deal with multiuser interference as the receivers (i.e., mobile users) are usually scattered over the service area and cannot cooperate. Yet, obtaining perfect CSIT is challenging in practical systems and has become progressively difficult due to the increasing number of antennas and access points in 5G.

So far, most techniques have been designed on the assumption of perfect CSIT and applied to practical scenarios with imperfect CSIT. Hence, the system performance would be degraded to different extents due to various CSIT imperfectness. There is a clear demand to design wireless networks accounting for imperfect CSIT and the consequent multiuser interference. Specifically, considerable efforts have been made to devise dedicated transmission and beamforming strategies for various imperfect CSIT scenarios.

1.1. Multiuser MIMO Systems

Consider multiuser Multiple-Input-Single-Output (MISO) downlink systems, which is illustrated in Figure 1.1. A Base Station (BS) equipped with multiple antennas serves a group of single-antenna users. These users can be served by Time-Division-Multiple-Access (TDMA) or Frequency-Division-Multiple-Access (FDMA) techniques with time or frequency sharing among users. However, single user transmission strategy (e.g., TDMA/FDMA) is limited by low spectral efficiency. There exist a number of technologies that the BS (i.e., transmitter) uses to simultaneously serve those users within a time/frequency slot. To name just a few, for instance, Broadcast Channel (BC) refers

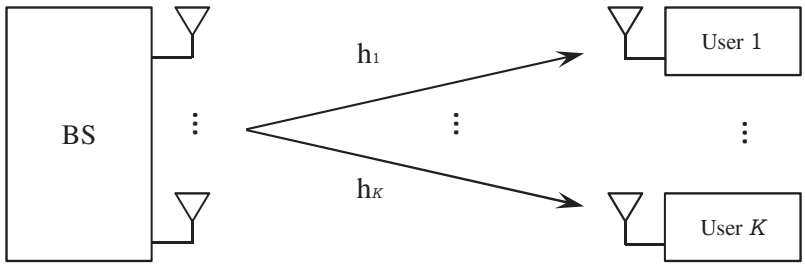

Figure 1.1.: Diagram of multiuser MISO downlink system.

to that the BS simultaneously sends dedicated (i.e., private) messages to each user [1]. Interfering broadcast channel indicates that multiple BS in a cellular network simultaneously transmit messages to a group of users in their own cells while creating interference to each other [2]. Multicasting refers to that the BS transmits a common message which contains the information that all users need [3]. Multi-group multicasting indicates that the BS transmits multiple common messages each containing the information required by a certain group of users [4]. In this thesis, we mainly focus on broadcast channel in Chapter 2, 3, 4, and on multicasting in Chapter 5.

1.2. Performance Metrics

To evaluate the system performance, a list of throughput metrics are of considerable interest, including but not limited to Degree-of-Freedom (DoF), sum rate, weighted sum rate and minimum rate. Specifically, DoF is defined as the number of interference-free streams that the BS can simultaneously transmit at high Signal-to-Noise-Ratio (SNR). Sum rate refers to the sum of individual data rate that each user achieves, which measures the system performance as a whole. Weighted sum rate indicates the sum of a weighted individual data rate of each user, which maintains various user priorities or fairness. Minimum rate refers to the minimum individual rate out of all users which measures the worst data rate that users can experience. We shall clarify here that the proposed transmission strategies can be evaluated by any of aforementioned metrics. In this thesis, we mainly focus on the sum rate metric in Chapter 2, 3, 4, and on minimum rate in Chapter 5.

1.3. Channel Models

Channel model is a simple representation of the communication channel between the BS and the user. The channels vary randomly, or fade, according to a statistical distribution [5]. Fast fading refers to that the channel impulse response changes rapidly within the symbol duration due to reflections of local objects and the motion of the objects relative

to those objects. By contrast, slow fading refers to that the channel impulse response remains within the symbol duration.

Let us focus on fast fading channels and briefly introduce independently and identically distributed (i.i.d.) Rayleigh channel, correlated Rayleigh channel and finite scatterer (or Ray-based geometric) channel. In this thesis, we mainly consider correlated Rayleigh channel in Chapter 2, 3, and finite scatterer channel in Chapter 4, 5.

1.3.1. I.I.D. Rayleigh Channels

Considering a MISO downlink channel, we denote the channel vector of a given user by $\mathbf{h} \in \mathbb{C}^M$, where M is the number of transmit antennas. The i.i.d. Rayleigh fading channel indicates that each entry of the channel vector $\mathbf{h}$ is i.i.d. zero-mean unit-variance complex Gaussian distributed, i.e., $h_m \sim \mathcal{CN}(0,1)$. It is also known as uncorrelated Rayleigh channel and occurred when the antenna spacings and/or the angular spreading of the energy at both sides of the wireless link are sufficiently large such that the various channel correlation becomes very small and negligible [5].

1.3.2. Correlated Rayleigh Channels

When the aforementioned conditions do not hold, channel correlation can be observed. According to [6] and [7], the Rayleigh channel correlation model can be respectively expressed as

$$\mathbf{h} = \mathbf{R}^{1/2}\,\mathbf{h}_w \tag{1.1}$$

$$\mathbf{h} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\,\mathbf{h}_w \tag{1.2}$$

where $\mathbf{h}_w$ is the i.i.d. channel vector. $\mathbf{R} = \mathbb{E}\big(\mathbf{h}\mathbf{h}^H\big)$ denotes the channel covariance matrix of $\mathbf{h}$ and is also known as spatial correlation matrix. By operating eigen-decomposition as $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, $\mathbf{U}$ consists of eigenvectors and $\mathbf{\Lambda}$ contains the associated eigenvalues. The i.i.d. Rayleigh channel model can be viewed as a special case of the correlated Rayleigh channel model when $\mathbf{R} = \mathbf{I}$.

Let us first consider a single parameter exponential correlation model [5]

$$\mathbf{R} = \begin{bmatrix} 1 & t & \dots & t^{M-1} \\ t^H & 1 & \dots & t^{M-2} \\ \vdots & & \ddots & \\ (t^H)^{M-1} & \dots & t^H & 1 \end{bmatrix}, \tag{1.3}$$

where t denotes the transmit correlation coefficient $t = |t|\,e^{j\phi}$, $\phi \in [0,\, 2\pi]$. In this exponential model, large (small) $|t|$ corresponds to high (low) correlation. This single parameter correlation model is a special case of virtual channel representation (VCR) which is valid for uniform linear arrays with half-wavelength spacing.

Some works in the literature also consider a geometrical one-ring scattering model by assuming a diffuse two-dimensional field of isotropic scatterers around the users [7]. The correlation between the channel coefficients of antennas $1 \leq i, j \leq M$ is given by

$$[\mathbf{R}]_{i,j} = \frac{1}{2\Delta} \int_{\theta-\Delta}^{\theta+\Delta} e^{-j\frac{2\pi}{\lambda}\Psi(\alpha)(\mathbf{r}_i - \mathbf{r}_j)} d\alpha, \tag{1.4}$$

where θ is the azimuth angle of a given user with respect to the orientation perpendicular to the array axis. Δ indicates the angular spread of departure to that user. $\Psi(\alpha) = [\cos(\alpha), \sin(\alpha)]$ is the wave vector for a planar wave impinging with the angle of α, λ is the wavelength and $\mathbf{r}_i = [x_i, y_i]^T$ is the position vector of the i-th antenna.

1.3.3. Finite Scatterer Channels

The physical finite scattering channel model has been widely used in Massive/Millimetre-Wave (mmWave) MIMO systems [8,9]. For instance, the channels in the mmWave bands tend to be sparse and with limited paths [10,11]. Under this model, the channel vector from a given user is defined as

$$\mathbf{h} = \sqrt{\frac{M}{L}} \sum_{l=1}^{L} g_l \, \mathbf{a}(\theta_l) = \sqrt{\frac{M}{L}} \mathbf{A}\mathbf{g}, \tag{1.5}$$

where the path gain vector $\mathbf{g}_k = [g_1, \cdots, g_L]$ has i.i.d. $\mathcal{CN}(0,1)$ entries and varies independently across different time slots. $\mathbf{A} = [\mathbf{a}(\theta_1), \cdots, \mathbf{a}(\theta_L)] \in \mathbb{C}^{M \times L}$ contains L steering vectors and $\theta_l \in [0, \pi]$ are the angle-of-departure (AoD) of lth path. The steering vector is closely related to the antenna array structure, operating frequency and antenna spacing, etc. Take uniform linear array (ULA) as an example. Under the plane wave and balanced narrowband array assumptions, the array steering vector can be written as

$$\mathbf{a}(\theta_l) = \frac{1}{\sqrt{M}} [1, e^{j2\pi\frac{d}{\lambda}\cos(\theta_l)}, \cdots, e^{j2\pi\frac{(M-1)d}{\lambda}\cos(\theta_l)}]^T, \tag{1.6}$$

where λ is the wavelength and $d = \frac{\lambda}{2}$ is the antenna spacing.

1.4. Precoding Techniques

Let us consider multiuser MISO broadcast channel and momentarily assume perfect channel state information at the transmitter (CSIT). Precoding is a processing technique that exploits CSIT to deal with multiple streams transmission.

From the perspective of design, the precoding techniques can be categorised into non-linear and linear methods. The non-linear precoding technique, Dirty Paper Coding (D-PC), achieves the sum rate capacity of broadcast channel [12]. Other non-linear precoding techniques such as Tomlinson-Harashima Precoding (THP) [13] and Vector Purturbation

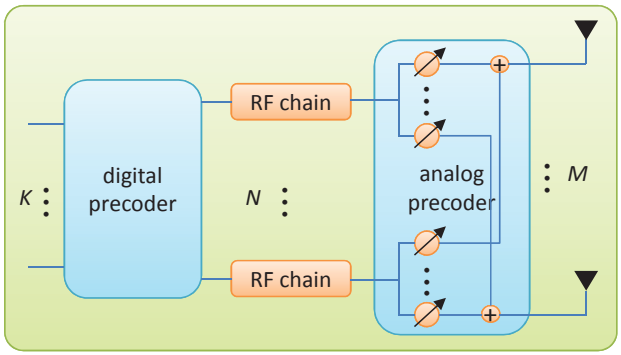

Figure 1.2.: Diagram of hybrid precoding method.

(VP) [14] offer very good sum rate performance.

On the other side, typical linear precoding techniques include matched precoding, zero-forcing (ZF) and regularised-zero-forcing (RZF). Specifically, matched precoding designs the precoder along with the channel direction, maximizing the desired signal power. ZF projects each users signal in the nullspace of all other users' channels to minimise the multiuser interference. RZF strikes a balance between matched precoding and ZF, i.e., desired signal enhancement and interference elimination[1]. Matched precoding, ZF and RZF will be detailed in Chapter 3 and 4. Generally speaking, non-linear precoding outperforms linear precoding in terms of the performance metrics introduced in Section 1.2. Nevertheless, linear precoding methods are of more interest due to their implementation simplicity in practical systems. In this thesis, we mainly focus on linear precoding methods. We also develop novel precoding strategies to enable better performance.

From the perspective of implementation, the precoding techniques can be categorised into digital precoding, analog precoding and hybrid precoding methods. Digital precoding is realised when the BS has a dedicated radio-frequency (RF) chain for each antenna element. The digital precoding vector (or matrix) can be arbitrarily constructed in the complex domain $\mathbb{C}^M$. When the BS has multiple transmit antenna but only one RF chain, analog precoding as in [10, 15, 16] can be implemented by analog phase shifters. However, the analog phase shifter has only control in its phase and is subject to a constant modulus constraint, leading to reduced precoding flexibility. When the BS has less RF chains than the number of transmit antennas, a hybrid precoding is implemented by a high-dimensional analog precoder using cost-efficient analog phase shifters network, cascaded with a reduced-dimensional digital precoder [17]. The illustration of the hybrid precoding structure is shown in Figure 1.2.

In this thesis, we mainly consider digital precoding in Chapter 2, 3, and hybrid precoding in Chapter 4, 5.

[1] Matched precoding, ZF, RZF, THP and VP are somehow the analog at the transmitter side of the detection based on MRC, ZF, MMSE, SIC and the sphere decoder, respectively. For more information, please refer to [5].

CSIT type	Feedback Rate	Latency
Perfect CSIT	High	Low
Statistical CSIT	Low	Low
Delayed CSIT	High	High
Partial/Quantised CSIT	Low	Low

Table 1.1.: Feedback rate and latency requirements for various CSIT types.

1.5. CSIT Types

So far, we have introduced a number of multiuser MIMO system (1.1) where the BS serves a group of users by precoding techniques (1.4) in order to achieve certain performance 1.2. The performance of these systems and corresponding precoding techniques heavily depends on the acquisition of CSIT. Table 1.1 shows various CSIT types and the associated requirements on the feedback link in frequency division duplexing (FDD) systems[2]. In this thesis, we mainly consider statistical and delayed CSIT in Chapter 2, partial and statistical CSIT in Chapter 3, quantised and statistical CSIT in Chapter 4 and perfect CSIT in Chapter 5.

1.5.1. Perfect CSIT

In FDD systems, the downlink channel is estimated by each user during the downlink training phase and then fed back to the BS via uplink signalling. This approach requires a *high-rate low-latency* feedback link between the user and the BS to keep the latter continuously and instantaneously informed about the Channel State Information (CSI). In time division duplexing (TDD) systems, the BS acquires CSI by uplink training and channel reciprocity. Perfect CSIT requires accurate channel estimation and ideal channel reciprocity.

However, obtaining high quality instantaneous CSIT represents a challenge [3], especially in fast fading channels. In FDD systems, the feedback link suffers from error, delay, etc. As an intermediate solution, imperfect CSIT imposes less stringent restrictions on feedback rate and latency.

1.5.2. Statistical CSIT

Statistical CSIT refers to the second-order channel statistics, i.e., the spatial correlation matrix $\mathbf{R} = \mathbb{E}\big(\mathbf{h}\mathbf{h}^H\big)$, where $\mathbf{h}$ denotes the channel vector. The channel statistics can be viewed as a sort of imperfect knowledge of the instantaneous channel. Such information is closely related to the scattering environment and the array structure. Due to the slowly-

[2]While some results developed in this work can be applied to various imperfect CSIT scenarios either in TDD or in FDD mode, we mainly focus on FDD system for clarity.

[3]In TDD systems, the uplink training and estimating approach is subject to antenna miscalibration and pilot contamination, leading to an imperfect CSIT.

varying nature of these related factors, the long-term channel statistics can be accurately obtained via *low-rate low-latency* feedback and therefore easily known to the BS.

When $\mathbf{R}$ is rank deficient, it can be used to completely eliminate multiuser interference by projecting the signal of each user in the nullspace of other users' channel covariance matrices. When $\mathbf{R}$ is full rank, this cannot be achieved. Nevertheless, such statistical CSIT still works effectively in mitigating multiuser interference at finite/practical SNR when the channels are highly correlated (but still full rank) [18].

1.5.3. Delayed CSIT

In FDD systems, the channel state has to be measured by the user and fed back to the BS. High quality CSI can be obtained at the user side whereas the BS may acquire this information with a delay larger than the coherence time of the channel such that the reported CSI known at the BS is uncorrelated with the current CSI. This occurs when the coherence time of the channel becomes short due to high mobility for example. Nevertheless, the accurate but completely delayed/outdated CSIT has been proven very useful in terms of high-SNR multiplexing gain [19] as well as finite-SNR data rate [20]. The scheme proposed in [19] that makes use of delayed CSIT will be revisited in Chapter 2. The accurate delayed CSIT still requires *high-rate* feedback but relieves the restriction on the latency of feedback link, i.e., *high-latency*.

1.5.4. Partial CSIT

In FDD systems, when the channel information is quantised and then reported via a rate-limited feedback link within the coherence time, only a partial CSI can be acquired by the BS. Nevertheless, partial CSIT only requires a *low-rate low-latency* feedback link[4]. In TDD systems, the BS obtains CSI by channel reciprocity. Still, only a partial current CSIT is attainable due to antenna miscalibration and pilot contamination.

When the CSIT error variance τ^2 decays with SNR (P) as $O(P^{-\alpha})$ for some constant $0 \leq \alpha \leq 1$, conventional multiuser transmission strategies that the BS transmits private messages to each user using linear precoding (e.g., ZF) achieve the sum Degree-of-Freedom (DoF) 2α in a two-user MISO broadcast channel [21]. Such a sum DoF performance reveals the obstacle of a family of linear precoding schemes relying on imperfect CSIT as $\alpha \to 0$. For example, the sum DoF is worse than TDMA for $\alpha < 0.5$ (i.e., a DoF of 1 is guaranteed by TDMA even with no CSIT) and becomes interference-limited for $\alpha = 0$.

To address this issue and tackle the residual interference, a Rate-Splitting transmission strategy was recently proposed in to achieve better sum DoF than conventional single-user/multiuser transmission strategies [21]. More technical details of RS will be given in Section 1.7.4.

[4]Partial delayed CSIT has the least restriction on feedback rate and latency, i.e., *low-rate high-latency*. Such CSIT scenario is not covered in this thesis.

1.6. Mathematical Tool

To solve the problems formulated in this thesis, a number of mathematical tools are required, including but not limited to matrix analysis, probability integrals, optimization techniques and deterministic equivalent of large dimensional random matrix. We here briefly describe convex optimization technique and deterministic equivalent. For more information, please refer to [22], [7, 23] and references therein.

1.6.1. Convex Optimization

Convex optimization is a special class of mathematical optimization problems which typically includes least-squares and linear programming problems. Typically, the convex optimization problem is written as [22]

$$\text{minimise} \quad f_0(x) \tag{1.7}$$

$$\text{subject to} \quad f_i(x) \leq b_i, \quad , i = 1, \cdots, m \tag{1.8}$$

where the functions $f_0, \cdots, f_m$ are convex, i.e., satisfy

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y), \quad \forall \lambda \in [0, 1]. \tag{1.9}$$

After interior-point methods was developed to efficiently and optimally solve convex optimization problem, the convex optimization technique can be applied to a large class of problems such as semidefinite programming (SDP), second-order cone programming (SOCP), quadratically constrained quadratic programming (QCQP), etc.

For non-convex problems, obtaining the optimal solution is a NP-hard problem. Some of these problems can be approximated in a relaxed or conservative manner as a convex problem. Then, the optimal solution of the approximated/relaxed problem is suboptimal but may be a good candidate to the original problem. The semidefinite relaxation (SDR) [3] and successive convex approximation (SCA) [24] are two examples of such relaxation/approximation techniques.

1.6.2. Deterministic Equivalent

In the well-established field of large dimensional random matrix theory, deterministic equivalent (DE) approach is a very powerful tool to provide accurate deterministic approximations of random quantities. DE approach plays a significant role in 5G wireless networks, where the number of antennas and access points increases dramatically. DE can be used to compute some key system performance metrics, for example, the Signal-to-Interference-and-Noise-Ratio (SINR) of large dimensional antenna array MISO systems [7].

The DE of a sequence of random quantities is defined as [25]

Definition 1. *The deterministic equivalent of a sequence of random complex values* $(X_M)_{M\geq 1}$ *is a deterministic sequence* $(\overline{X}_M)_{M\geq 1}$*, which approximates* X_M *such that*

$$X_M - \overline{X}_M \overset{M\to\infty}{\longrightarrow} 0, \tag{1.10}$$

almost surely.

Let us first review two basic results on large-dimensional random vectors that will be useful afterwards.

Lemma 1.1. *[26, Lemma 4] Let* $\mathbf{A} \in \mathbb{C}^{M\times M}$ *be uniformly bounded spectral norm w.r.t.* M *and* $\mathbf{x}, \mathbf{y} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ *be mutually independent of* $\mathbf{A}$*. Then, we have almost surely that*

$$\frac{1}{M}\mathbf{x}^H\mathbf{A}\mathbf{x} - \frac{1}{M}tr(\mathbf{A}) \overset{M\to\infty}{\longrightarrow} 0, \tag{1.11}$$

$$\frac{1}{M}\mathbf{x}^H\mathbf{A}\mathbf{y} \overset{M\to\infty}{\longrightarrow} 0. \tag{1.12}$$

For more results on deterministic equivalent, please refer to [7, Theorem 1, Lemma 6, Lemma 7], [23] and references therein.

1.7. Literature Review

In multiuser MISO BC, schemes that achieve the sum rate capacity and the capacity region require perfect (i.e., accurate and instantaneous) CSIT [27–30]. However, perfect CSIT is hardly achievable in practice due to feedback delay, limited feedback overhead [31] and channel estimation errors [5]. CSIT imperfection is therefore a major obstacle for MIMO systems to realise its spectral and energy efficiency benefits. There have been plentiful efforts in designing transmission strategies and beamforming techniques to overcome the detrimental effects of various imperfect CSIT. In what follows, we revisit what have been done in this direction and analyse their strengths and weaknesses.

1.7.1. Transmission Strategies with Statistical CSIT Only

The transmit beamforming scheme driven by statistical CSIT is denoted as Statistical Beamforming (SBF). Recently, SBF precoded by Generalised Eigenvector (GE) has been shown to maximise the ergodic sum rate at high SNR for M-user M-transmit-antenna MISO BC when $M = 2$ [6] or $M \to \infty$ [32]. Nevertheless, the optimal precoder for the general $M > 2$ case is still unknown. With only statistical CSIT, there has been no investigation on maximising the ergodic sum rate for two-user, arbitrary M-transmit-antenna case.

1.7.2. Transmission Strategies with Delayed CSIT Only

When the CSIT is completely outdated, it was proved to still benefit the multiplexing gain (or DoF). In K-user K-transmit-antenna MISO BC, it was shown by M. Maddah-Ali and D. Tse in [19] that one can achieve a total DoF of

$$\frac{K}{\sum_{k=1}^{K} \frac{1}{k}} \tag{1.13}$$

per second per Hz in this channel. In other words, we can achieve a sum rate that scales with SNR as

$$\frac{K}{\sum_{k=1}^{K} \frac{1}{k}} \log_2 \mathrm{SNR} + O(1) \quad \text{bits/s/Hz}. \tag{1.14}$$

Take two-user ($K = 2$) as an example. The sum DoF of $\frac{4}{3}$ can be achieved by the transmission scheme proposed in [19] (known as MAT scheme). Let us revisit MAT in the two-user MISO BC to understand how the delayed CSIT is useful. Denote the channel between the BS and user A in time slot t as $\mathbf{h}_t$ and similarly $\mathbf{g}_t$ for user B. The MAT transmission scheme is expanded into three slots. The transmitted signals $\mathbf{x}_t \in \mathbb{C}^K$ in time slot t are given by

$$\mathbf{x}_1 = \mathbf{s}_A \tag{1.15}$$

$$\mathbf{x}_2 = \mathbf{s}_B \tag{1.16}$$

$$\mathbf{x}_3 = \left[\mathbf{g}_1^H \mathbf{s}_A + \mathbf{h}_2^H \mathbf{s}_B,\ 0\right]^T. \tag{1.17}$$

where $\mathbf{s}_k = [s_{k1}, s_{k2}]^T$ represents the two private symbols intended to user $k = A, B$. $\mathbf{h}_2^H \mathbf{s}_B$ and $\mathbf{g}_1^H \mathbf{s}_A$ are the interference overheard by user A in slot 1 and user B in slot 2, respectively. At time slot 1 (2), the BS transmits two private symbols intended to user A (B), which creates interference to each other. At time slot 3, the BS has the channel knowledge $\mathbf{g}_1, \mathbf{h}_2$. Thus, the BS can reconstruct and retransmit the overheard interference using single antenna. The received signal at user A is described as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{h}_1^H \\ \mathbf{0} \\ h_{31}^* \mathbf{g}_1^H \end{bmatrix}}_{\text{rank}=2} \mathbf{s}_A + \underbrace{\begin{bmatrix} \mathbf{0} \\ \mathbf{h}_2^H \\ h_{31}^* \mathbf{h}_2^H \end{bmatrix}}_{\text{rank}=1} \mathbf{s}_B + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}, \tag{1.18}$$

where n_t is Gaussian additive noise. Eq. (1.18) can be further arranged as

$$\begin{bmatrix} y_1 \\ y_3 \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^H \\ h_{31}^* \mathbf{g}_1^H \end{bmatrix} \mathbf{s}_A + \begin{bmatrix} n_1 \\ n_3 - h_{31}^* n_2 \end{bmatrix}, \tag{1.19}$$

which is a two-dimensional independent interference-free observation of $\mathbf{s}_A$ obtained at user A. It implies that we can successfully decode two independent messages for user A

within three time slots and likewise two independent messages for user B. Therefore, the total/sum DoF of $\frac{4}{3}$ is enabled by MAT scheme.

In a nutshell, the main idea of MAT scheme is to exploit delayed CSIT to align the overheard interference into a one-dimensional subspace. By doing so, the interference is eliminated without affecting the dimension of the useful signals. For $K > 2$ case, please refer to [19] for more details.

Interestingly, the interference alignment can be operated in a different manner. An alternative MAT (AMAT) transmission scheme was proposed in [33] where the transmitted signals $\mathbf{x}_t$ are given as

$$\mathbf{x}_1 = \mathbf{s}_A + \mathbf{s}_B \tag{1.20}$$

$$\mathbf{x}_2 = \left[\mathbf{h}_1^H \mathbf{s}_B,\ 0\right]^T \tag{1.21}$$

$$\mathbf{x}_3 = \left[\mathbf{g}_1^H \mathbf{s}_A,\ 0\right]^T. \tag{1.22}$$

In contrast to MAT, AMAT transmits two private symbols $\mathbf{s}_A$ to user A and two private symbols $\mathbf{s}_B$ to user B in a superimposed manner. Then, the overheard interference of user A and B at time slot 1 will be reconstructed and retransmitted in time slot 2 and 3, respectively. It is easy to verify that this variant of MAT alignment can achieve the same sum DoF as MAT. Please refer to [33] for further details. When it comes to finite SNR rate, the work [34] generalised the MAT as GMAT and achieved a higher data rate by constructing precoders which strike a balance between desired signal enhancement and interference alignment.

1.7.3. Transmission Strategies with Statistical and Delayed CSIT

Consider a two-user MISO BC with multiple antennas at the BS. With both statistical and delayed CSIT at hand, authors in [20] developed an enhanced MAT strategy, denoted as VMAT, yielding a higher sum rate than the original MAT at finite SNR. However, in highly-correlated channel, the rate performance of VMAT is still inferior to SBF which exploits only statistical CSIT [20]. In brief, statistical channel information is not fully exploited. In this thesis, we fill the gap between what has been achieved and what can be achieved with both statistical and delayed CSIT.

1.7.4. Transmission Strategies with Partial CSIT Only

In realistic wireless communication systems, only partial CSIT is acquired by the BS due to limited feedback[5] or channel estimation error. The downlink channel of user k can be mathematically modelled as

$$\mathbf{h}_k = \widehat{\mathbf{h}}_k + \widetilde{\mathbf{h}}_k, \tag{1.23}$$

[5] The up-to-date results in [35] show that the downlink channel can be effectively inferred from the uplink channel (i.e., channel reciprocity in FDD mode). Then, the proposed techniques in [35] can potentially eliminate the channel feedback. Yet, the CSIT inference is still subject to channel estimation error.

where $\widehat{\mathbf{h}}_k$ is the imperfect estimate of the true channel $\mathbf{h}_k$ at the BS and $\widetilde{\mathbf{h}}_k$ denotes the channel estimation error. The imperfect channel estimate can partially eliminate multiuser interference and enable partial DoF, as explained below. We denote transmit power by P and assume unit noise power. Without loss of generality, we suppose that the CSIT error $\|\widetilde{\mathbf{h}}_k\|^2$ scales as $O(P^{-\alpha})$ for $\alpha \in [0,\, 1]$. It is worth highlighting that α relies on various practical interpretations, e.g., the number of feedback bits in FDD systems [36,37] and Doppler process of the fading in time correlated channels [33, 38].

Consider a multiuser MISO BC with the conventional multiuser transmission strategy, the BS transmits dedicated/private messages to each user. The transmitted signal vector $\mathbf{x}$ is expressed as the superposition of statistically independent signals $\mathbf{x}_k \in \mathbb{C}^M$ destined to K users, i.e.,

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{x}_k, \tag{1.24}$$

Based on the channel estimate, the BS performs the linear precoding such as ZF

$$\mathbf{x}_k \in \mathrm{Null}\big([\widehat{\mathbf{h}}_1, \ldots, \widehat{\mathbf{h}}_{k-1}, \widehat{\mathbf{h}}_{k+1}, \ldots, \widehat{\mathbf{h}}_K]^H\big) \tag{1.25}$$

and uniform power allocation across users, yielding a received signal given by

$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{x}_k}^{O(P^1)} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{j \neq k} \mathbf{x}_j}^{O(P^{1-\alpha})} + \overbrace{n_k}^{O(P^0)} \tag{1.26}$$

which contains residual multiuser interference terms $\widetilde{\mathbf{h}}_k^H \mathbf{x}_j$, $\forall j \neq k$. The interference power scales as $O(P^{1-\alpha})$ by following the CSIT error scaling. The additive noise is independent of signal strength and thus the noise power does not scale with P, i.e., $O(P^0)$. Then, the SINR scales as $O(P^\alpha)$ and it follows that the achievable rate of user k at high SNR can be written as $R_k = \alpha \log_2(\mathrm{SNR}) + O(1)$. Intuitively, only a fraction α of signalling dimension is accessible by user k due to the residual multiuser interference. Moreover, $\alpha = 1$ implies perfect CSIT in the sense of DoF, as the multiuser interference that scales as $O(P^0)$ can be drawn by the noise. By contrast, $\alpha = 0$ implies no CSIT in DoF sense as the desired signal is drawn by the residual interference.

Then, the sum DoF achieved by ZF based on partial CSIT is given by $K\alpha$. Take a two-user ($K = 2$) MISO broadcast channel as an example, the sum DoF 2α is worse than single-user transmission strategy (e.g., TDMA) for $\alpha < 0.5$. It is worth noting that the latter enables a DoF of 1 even with no CSIT. Moreover, we note that in order to access the fraction α of signalling dimension, partial transmit power P^α enables the same DoF of α, i.e.,

$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{x}_k}^{O(P^\alpha)} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{j \neq k} \mathbf{x}_j}^{O(P^0)} + \overbrace{n_k}^{O(P^0)}. \tag{1.27}$$

Naturally, there must exist a general transmission strategy that makes better use of partial CSIT and full transmit power to achieve higher sum DoF (i.e., at least a sum DoF of 1 for arbitrary CSIT quality).

To this end, a Rate-Splitting (RS) transmission strategy was recently proposed[6] [33, Lemma 2], [21]. Specifically, we can split one selected user's signal (e.g., user k) into a common part ($\mathbf{x}_c$) and a private part ($\mathbf{x}_k$). Then, the private signals ($\mathbf{x}_{j|j\neq k}$) intended to other users and the private signal $\mathbf{x}_k$ are designed according to (1.25) by using partial transmit power P^α as (1.27). The common signal $\mathbf{x}_c$ is broadcasted to all users using the residual transmit power $P - P^\alpha$. Following this design, the transmitted signal is constructed as

$$\mathbf{x} = \mathbf{x}_c + \sum_{k=1}^{K} \mathbf{x}_k, \tag{1.28}$$

and the received signal of user k writes as

$$y_k = \overbrace{\mathbf{h}_k^H \mathbf{x}_c}^{O(P^1)} + \overbrace{\mathbf{h}_k^H \mathbf{x}_k}^{O(P^\alpha)} + \overbrace{\widetilde{\mathbf{h}}_k^H \sum_{j\neq k} \mathbf{x}_j}^{O(P^0)} + \overbrace{n_k}^{O(P^0)}. \tag{1.29}$$

At the user side, the common signal $\mathbf{x}_c$ is decoded by treating all private signals as noise. The received signal power of $\mathbf{x}_c$ scales as $O(P^1)$ due to $P - P^\alpha \sim P$ at high SNR for $\alpha < 1$. Then, $\mathbf{x}_c$ can be decoded by all users with a SINR that scales as $O(P^{1-\alpha})$, resulting in a DoF of $1 - \alpha$. While $\mathbf{x}_c$ is not required by all users, it has to be decoded by all users such that the users can decode their own private signals after removing $\mathbf{x}_c$ from the received signal. The SINR of $\mathbf{x}_k$ is akin to (1.27) and scales as $O(P^\alpha)$, leading to a DoF of α. For convenience, the conventional transmission strategy shown in (1.24) is referred to as No-RS strategy. Thus, the DoF from private signals transmission in RS with partial power remains the same ($K\alpha$) as the conventional No-RS strategy. The RS transmission strategy achieves a DoF gain of $1-\alpha$, enabling a sum DoF of $1 + (K-1)\alpha$.

The benefits of RS can be extended from high-SNR DoF region to finite-SNR rate region. When the CSIT error variance is fixed (i.e., $\alpha = 0$), linear precoding techniques with uniform power allocation lead to multiuser interference, which ultimately create a rate ceiling at high SNR [36]. To circumvent this problem, one can adaptively tune the power allocation parameters among users as a function of SNR hence obtaining a single user transmission at extremely high SNR. Such an adaptive per-user power allocation bridges in a continuous manner the single-user mode and the multiuser mode.

By contrast, RS can provide a rate performance beyond just operating the adaptive per-user power allocation. By optimising the transmit beamformer and power allocation

[6] In [33], the authors characterised the optimal DoF region of a two-user MISO broadcast channel with a mixture of imperfect current CSIT and perfect delayed CSIT. The corner points $(1, \alpha)$ and $(\alpha, 1)$ of the DoF region can be achieved with the Rate-Splitting approach, which does not exploit delayed CSIT and is applicable to the scenarios with only imperfect current CSIT.

parameters for both RS and conventional multiuser transmission strategy, RS shows significant sum rate gain over the conventional baseline at finite SNR [39]. In the context of a two-user MISO broadcast channel with quantised CSIT, [37] has also validated the rate benefits provided by RS over conventional multiuser transmission strategy. However, both the optimisation method proposed in [39] and the analysis in [37] can hardly be extended to massive MIMO systems.

1.7.5. Transmission Strategies with Partial and Statistical CSIT

As discussed before, the feedback overhead to obtain high quality CSIT is unaffordable in massive MIMO systems. When it comes to designing precoders on the basis of reduced CSI feedback, a two-tier precoder relying on both short- and long-term CSIT has been proposed by several authors [40–47]. The dimensionality reduction offered by the two-tier precoder was shown to be very beneficial to multiuser MIMO in deployments with spatially correlated fading [40, 41]. This precoder structure also made its way to realistic systems as IEEE 802.16m and LTE-A [42, 43].

Consider a multiuser massive MIMO system. When users are clustered into groups according to the similarity of their channel covariance matrices, [44] proposed a two-tier precoding approach to achieve massive-MIMO-like gains with highly reduced-dimensional CSIT. More precisely, the outer precoder controls inter-group interference based on long-term CSIT (the channel covariance matrices) while the inner precoder controls intra-group interference based on short-term effective channel (the channel concatenated with the outer precoder) with a reduced-dimension. This precoding scheme is referred to as Joint Spatial Division and Multiplexing (JSDM). The finding of [44] has been generalised into multi-polarised system [45], where antenna polarisation can be regarded as long-term CSI and used to further reduce the signalling overhead for CSIT acquisition. The work [46] proposed a SLNR-based outer precoder design and [47] developed a low complexity iterative algorithm to compute the outer precoder.

However, the system performance of the aforementioned two-tier precoding schemes is highly degraded by two limiting factors: inter-group and intra-group interference. When the eigen-subspaces spanned by the dominant eigenvectors of groups' spatial correlation matrices severely overlap, the outer precoder design may leak power (inter-group interference) to unintended groups. A typical example of overlapping eigen-subspace is that of users in different groups sharing common scatters. Furthermore, randomly located users are not naturally partitioned into groups with the same covariance matrix. When user grouping techniques (e.g., K-mean clustering) are applied, the inter-group interference cannot be completely eliminated by the outer precoder [48]. In addition, the reduced-dimensional effective CSIT might be imperfect due to limited feedback, which leads to intra-group interference.

There has been no investigation on how to deal with the problem of imperfect user

grouping and imperfect CSIT and to further enhance the system performance. Inspired by the fact that RS can effectively tackle multiuser interference, we can get some insights and develop novel transmission strategy to cope with the problem of inter-group and intra-group interference.

1.7.6. Beamforming Strategies with Quantised CSIT Only

Millimetre wave communication has been recognised as a promising technology in 5G cellular network for its large transmission bandwidth [49, 50]. To compensate the severe pathloss of mmWave link, large-scale antenna array is needed to provide high precoding gains [51]. However, the prohibitive cost and power consumption of RF chains (i.e., mixed-signal components) at mmWave bands makes the fully digital precoding infeasible. To tackle this RF hardware constraint, a hybrid precoding transceiver architecture has been recently proposed, where the large-scale antenna array is driven by a small number of RF chains [8]. The hybrid precoder is implemented by a high-dimensional RF beamformer using cost-efficient analog phase shifters, cascaded with a reduced-dimensional digital precoder [52].

Considering single user MIMO systems, the hybrid precoder is designed to approach the performance of fully digital precoder by solving a matrix factorisation problem [53]. When it comes to multiuser MIMO systems, [54] maximised the sum rate by iteratively optimising the analog and digital precoder until convergence. The work [55] analysed the rate performance in the large array regime for a given hybrid precoder design. Furthermore, [56] considered a partially-connected phase shifter networks. All these works determined the hybrid precoder assuming perfect full dimensional CSIT.

In practical mmWave systems, only partial CSIT is attainable through channel estimation [57, 58] and quantisation. With limited feedback, [17] proposed a low complexity hybrid precoding approach based on a two-stage feedback scheme. The analog precoder (or equivalently, RF beamformer) is designed to maximise the desired signal power of each user by beam search and feedback. Then, the digital precoder depends on the random vector quantisation (RVQ) and feedback of the effective channel (the channel concatenated with the RF beamformer). This hybrid precoding method relying on two-stage feedback requires a complicated signalling and feedback procedure. So far, there has been no investigation on how to simplify the signalling and feedback procedure while maintaining the rate performance. Moreover, the multiuser interference is still a limiting factor for mmWave systems and leads to a performance degradation. To alleviate this issue, a rate splitting transmission strategy was recently proposed in [21]. The benefit of RS in the context of multiuser mmWave systems with hybrid precoding has never been investigated.

1.7.7. Beamforming Strategies for Multicasting with Perfect CSIT

Wireless multicasting has been emerging as a key enabling technology to efficiently address the overwhelming traffic demands (e.g., popular video delivery to a number of mobile devices) in 5G cellular networks. The optimisation problem of transmit beamforming for quality-of-service (QoS) requirements and for max-min fairness was proved to be NP-hard [3]. This NP-hard optimisation problem can be approximated by a convex SDP problem using a SDR relaxation [3]. However, the solution of the relaxed problem is not always feasible for the original problem. To address this issue, several iterative algorithms were proposed in [59–61]. In order to facilitate applications for real-time systems, [62] proposed a non-iterative and simple-yet-effective linear precoding strategy. Furthermore, the problem of multicasting was extended to multiple co-channel groups [63], per-antenna power constraint [64, 65] and massive MIMO deployments [66], respectively.

With perfect CSIT, all the aforementioned works assume digital precoding which requires a dedicated RF chain for each antenna element at the BS. Unfortunately, such a requirement is very costly and therefore unrealistic for massive MIMO systems (corresponding to a large number of RF chains) and mmWave MIMO systems (due to expensive mmWave mixed-signal components). To address this issue, hybrid analog/digital precoder is typically used. A very recent work [67] applied hybrid precoder into physical layer multicasting. However, the authors computed the hybrid precoder as a sparse weighted combination of predefined vectors and minimised the ℓ_1-norm of the weights, rather than solving the original problem of minimising the transmission power under the QoS constraints. So far, even with perfect CSIT, there has been no investigation on the max-min fairness of hybrid precoding multicasting.

1.8. Summary of Contributions

We depart from the analysis in literature review and propose novel transmission and beamforming strategies to enhance the system performance. A summary of main contributions of each chapter is given as follows.

- **Chapter 2: A Novel Transmission Strategy for Two-User MISO Broadcast Channel with Statistical and Delayed CSIT**

  We focus on linear beamforming design and power allocation scheme for ergodic rate maximisation in a two-user MISO system with statistical and delayed CSIT. Firstly, with statistical CSIT only, we focus on statistical beamforming (SBF) design that maximises a lower bound on the ergodic sum rate. Secondly, relying on both statistical and delayed CSIT, an iterative algorithm is proposed to compute the precoding vectors of Alternative MAT (AMAT), which maximises an approximation of the ergodic sum rate with equal power allocation. Finally, we propose a transmission strategy, denoted as Statistical Alternative MAT (SAMAT), which

exploits both channel statistics and delayed CSIT. Via proper power allocation, the SAMAT framework is proposed to softly bridge between SBF and AMAT at any SNR in arbitrary spatial correlation condition.

- **Chapter 3: A Novel Transmission Strategy for Multiuser Massive MIMO System with Partial and Statistical CSIT**

  To tackle the detrimental effects of partial CSIT and resultant multiuser interference, a Rate-Splitting (RS) approach has been proposed recently, which splits one selected user's message into a common and a private part, and superimposes the common message on top of the private messages. The common message is drawn from a public codebook and decoded by all users. We generalise the idea of RS into the large-scale array regime. By further exploiting the channel second-order statistics, we propose a novel and general framework Hierarchical-Rate-Splitting (HRS) that is particularly suited to massive MIMO systems. HRS simultaneously transmits private messages and two kinds of common messages that are decoded by all users and by a subset of users, respectively. We analyse the asymptotic sum rate of RS/HRS and optimise the precoders of the common messages. A closed-form power allocation is derived which provides insights into the effects of various system parameters.

- **Chapter 4: A Novel Multiuser Millimeter Wave Beamforming Strategy with Quantised and Statistical CSIT**

  In multiuser mmWave systems, hybrid analog/digital precoding is typically employed. To compute the hybrid precoder, the conventional two-stage feedback scheme determines the analog beamformer by beam search and feedback to maximise the desired signal power of each user. The digital precoder is designed based on quantisation and feedback of effective channel to mitigate multiuser interference. Alternatively, we propose a one-stage feedback scheme which effectively reduces the complexity of the signalling and feedback procedure. Specifically, the second-order channel statistics are leveraged to design digital precoder for interference mitigation while all feedback overhead is reserved for precise analog beamforming. Under a fixed total feedback constraint, we investigate the conditions under which the one-stage feedback scheme outperforms the conventional two-stage counterpart. Moreover, a rate splitting (RS) transmission strategy is introduced to further tackle the multiuser interference and enhance the rate performance. Consider (1) RS precoded by the one-stage feedback scheme and (2) conventional transmission strategy precoded by the two-stage scheme with the same first-stage feedback as (1) and also certain amount of extra second-stage feedback. We show that (1) can achieve a sum rate comparable to that of (2). Hence, RS enables remarkable saving in the second-stage training and feedback overhead.

- **Chapter 5: A Novel Hybrid Precoding Strategy for Physical Layer Multicasting with Perfect CSIT**

  We investigate the problem of hybrid analog/digital precoding for multicasting with a limited number of RF chains. Considering a total transmit power constraint over the RF chains, the goal is to maximise the minimum (max-min) received SNR among all users. We propose a low complexity algorithm to compute the RF precoder that achieves near-optimal max-min performance. Moreover, we derive a simple condition under which the hybrid precoding driven by a limited number of RF chains incurs no loss of optimality with respect to the fully digital precoding case.

- **Chapter 6: Conclusion and Future Work**

  We summarise the thesis, discuss its findings and contributions and also outline a number of directions for future research.

1.9. Publications

The works in this thesis have resulted in a few papers that have been published or submitted for publication.

Journals and Magazine

- **M. Dai**, B. Clerckx, D. Gesbert and G. Caire, "A Rate Splitting Strategy for Massive MIMO with Imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611-4624, July 2016.
- B. Clerckx, H. Joudeh, C. Hao, **M. Dai**, and B. Rassouli, "Rate Splitting for MIMO Wireless Networks: A Promising PHY-Layer Strategy for LTE Evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98-105, May 2016.
- **M. Dai** and B. Clerckx, "Hybrid Precoding for Physical Layer Multicasting," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 228-231, Feb. 2016.
- **M. Dai** and B. Clerckx, "Transmit Beamforming for MISO Broadcast Channels with Statistical and Delayed CSIT," *IEEE Trans. Commun.*, vol. 63, no. 4, pp. 1202-1215, April 2015.
- **M. Dai** and B. Clerckx, "Multiuser Millimeter Wave Beamforming Strategies with Quantized and Statistical CSIT", *submitted to IEEE Trans. Wireless Commun.*, Nov. 2016.

Conference

- **M. Dai**, B. Clerckx, D. Gesbert and G. Caire, "A Hierarchical Rate Splitting Strategy for FDD Massive MIMO under Imperfect CSIT," in *Proc. IEEE Int. Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Sept. 2015, pp. 80-84.

Related patents not covered in this thesis

- B. Clerckx, **M. Dai** and J. Park, "Transmit Antenna Selection in Massive MIMO Systems with Limited Feedback," PCT/CN2014/093699, 2014.
- J. Park, B. Clerckx and **M. Dai**, "Transceiver Architecture (Precoding, Channel Estimation and Feedback) for Distributed/Polarized Multi-user MIMO Systems under Imperfect CSIT," PCT/CN2014/093614, 2014.

1.10. Notations

The following notations are used throughout the thesis. Bold lower case and upper case letters denote vectors and matrices, respectively. The notations $[\mathbf{X}]_i$,$[\mathbf{X}]_{i,j}$, $\mathbf{X}^T$, $\mathbf{X}^H$, $\mathrm{tr}(\mathbf{X})$, $\det(\mathbf{X})$ denote the i-th column, the entry in the i-th row and j-th column, the transpose, conjugate transpose, trace and determinant of a matrix $\mathbf{X}$. We use $\mathrm{Span}(\mathbf{X})$ and $\mathrm{Null}(\mathbf{X})$ to denote the column space and null space of $\mathbf{X}$. $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ indicate the largest and smallest eigenvalues of a matrix and their corresponding eigenvectors are denoted by $\mathbf{u}_{\max}(\mathbf{X})$ and $\mathbf{u}_{\min}(\mathbf{X})$, respectively. Let $\Phi_{PD} = \{\mathbf{X} \in \mathbb{C}^{M\times M} \ |\mathbf{X} \text{ is positive definite}\}$. $\mathbf{I}$ is the identity matrix and $\mathbf{1}_i$ is the i-th column of $\mathbf{I}$. The notation $\mathrm{diag}(\cdot)$ stands for a diagonal matrix whereas $\mathbb{E}(\cdot)$ is the expectation operator. The l_0-norm and l_2-norm of $\mathbf{x}$ are denoted by $\|\mathbf{x}\|_0$ and $\|\mathbf{x}\|$, respectively. We denote $\mathrm{Exp}(c)$ as the exponential distribution with parameter c and $U(a,b)$ as the uniform distribution.

2. A Novel Transmission Strategy for Two-User MISO Broadcast Channel with Statistical and Delayed CSIT

In this chapter, we consider a two-user MISO BC with statistical and delayed CSIT. We propose a novel and general transmission framework, denoted as SAMAT, which smoothly bridges the gap between sum rate achieved by statistical CSIT strategies and by delayed CSIT strategies. Specifically, we focus on precoder design and power allocation scheme to enhance the ergodic sum rate in a two-user MISO system.

The transmission strategies with statistical CSIT, denoted as statistical beamforming (SBF), have been reviewed in Section 1.7.1. The transmission strategies with delayed CSIT, denoted as MAT/AMAT, have been reviewed in Section 1.7.2. When both statistical and delayed CSIT are attained at the BS, the transmission strategy (VMAT) has been briefly discussed in 1.7.3. It is worth noticing that the rate performance of VMAT is still inferior to SBF which exploits only statistical CSIT [20] in highly-correlated channels. In other words, the benefits of statistical CSIT is not fully extracted by [20]. In this chapter, we fill the gap between what has been achieved and what can be achieved with both statistical and delayed CSIT.

2.1. System Model

Consider a MISO BC where the transmitter equipped with M antennas ($M \geq 2$) wishes to send private messages to two users each with a single antenna. Perfect CSI is instantaneously available at the user side whereas the BS acquires an outdated version of this information. Statistical CSIT characterised by the spatial correlation matrix is assumed at the BS. This is a reasonable assumption because long-term channel statistics are more related to the scattering environment and independent of the transmission period.

Rayleigh fading channel model is considered, which implies that the spatial statistics can be completely depicted by the second-order moments of the channel [6]. Specifically, we denote the channel between the transmitter and user A in time slot j as $\mathbf{h}_j$ and similarly $\mathbf{g}_j$ for user B

$$\mathbf{h}_j = \mathbf{R}_A^{1/2}\,\mathbf{h}_{w,j} \tag{2.1}$$

$$\mathbf{g}_j = \mathbf{R}_B^{1/2}\,\mathbf{g}_{w,j}, \tag{2.2}$$

where $\mathbf{h}_{w,j}$ and $\mathbf{g}_{w,j}$ are $M\times 1$ vectors with i.i.d. $\mathcal{CN}(0,1)$ entries. They are assumed constant within one time slot and varying independently across time slots. $\mathbf{R}_A$ and $\mathbf{R}_B$ are full rank positive definite covariance matrices[1] for user A and B respectively, which can be decomposed as $\mathbf{R}_k = \mathbf{V}_k\mathbf{\Lambda}_k\mathbf{V}_k^H$, $k = A, B$. $\mathbf{V}_k \in \mathbb{C}^{M\times M}$ is a unitary matrix whose columns are eigenvectors of $\mathbf{R}_k$, while the diagonal $\mathbf{\Lambda}_k$ that contains the eigenvalues of $\mathbf{R}_k$ is normalised as $\mathrm{tr}(\mathbf{\Lambda}_k) = M$. $\mathbf{\Lambda}_k = \mathbf{I}$ indicates the k-th channel is spatially uncorrelated while $\mathrm{rank}(\mathbf{\Lambda}_k) = 1$ implies it is fully correlated [68].

Let us first investigate the statistical CSIT only transmission strategy (i.e., SBF) in Section 2.2. Then, we discuss the delayed CSIT transmission strategy (i.e., AMAT) in Section 2.3. Last, we proceed to elaborate the new transmission strategy SAMAT in Section 2.4.

2.2. SBF Strategy

In this section, we address the statistical precoding problem of a two-user M-transmit-antenna system by maximising the ergodic sum rate. Recently, SBF with Generalised Eigenvector (SGEBF) has been shown to maximise the ergodic sum rate at high SNR for M-user M-transmit-antenna MISO BC when $M = 2$ [6] or $M \to \infty$ [32]. Nevertheless, the optimal precoder for the general $M > 2$ case is still unknown because of a lack of closed-form ergodic sum rate expression. In [18], the Generalised Eigenvector (GE) solution is arrived based on ergodic Signal-to-Leakage-and-Noise Ratio (SLNR), which leverages independence between the numerator and the denominator of SLNR. We will focus on addressing the original ergodic sum rate problem.

The transmitted signal of SBF writes as

$$\mathbf{x} = \sqrt{\rho}\,\mathbf{w}\,s_A + \sqrt{\rho}\,\mathbf{q}\,s_B, \tag{2.3}$$

where $\rho = \frac{P}{2}$, $\mathbf{w}$ and $\mathbf{q}$ are the unit norm precoding vectors of user A and B, respectively. For simplicity, we will look at the rate performance of user A only and a similar derivation can be easily extended to user B. The received signal at the receiver side is given as $y = \sqrt{\rho}\,\mathbf{h}^H\mathbf{w}\,s_A + \sqrt{\rho}\,\mathbf{h}^H\mathbf{q}\,s_B + n_A$, where $\mathbf{h} \in \mathbb{C}^M$ is the channel vector and $n_A \sim \mathcal{CN}(0,1)$ is the standard complex additive white Gaussian noise (AWGN). The achievable ergodic rate of user A is given by

$$R_A = \mathbb{E}\left[\log_2\left(1 + \mathrm{SINR}_A\right)\right], \tag{2.4}$$

where $\mathrm{SINR}_A = \frac{\rho|\mathbf{h}^H\mathbf{w}|^2}{1+\rho|\mathbf{h}^H\mathbf{q}|^2}$ is the instantaneous signal-to-interference-and-noise ratio (SINR) by treating the multi-user interference as noise. Consequently, the ergodic sum rate of the system with linear beamforming is $R_{\mathrm{sum}} \triangleq R_A + R_B$.

[1] For rank deficient case, the symbol intended to user i is simply precoded by a column vector in $\mathrm{Null}(\mathbf{R}_j)$. By doing this, the overheard interference of each symbol can be completely removed. Thus, we can transmit two symbols at one time instant, achieving a sum DoF of 2 as if we have perfect CSIT.

Theorem 2.1. *For any M, the ergodic sum rate of SBF in a two-user MISO broadcast channel at high SNR can be lower bounded by*

$$R_{sum} \geq \log_2 \left(\frac{\mathbf{w}^H \mathbf{R}_A \mathbf{w}}{\mathbf{w}^H \mathbf{R}_B \mathbf{w}} \frac{\mathbf{q}^H \mathbf{R}_B \mathbf{q}}{\mathbf{q}^H \mathbf{R}_A \mathbf{q}} \right), \tag{2.5}$$

and the precoders that maximise the lower bound in (2.5) *are generalised eigenvectors (GE) given by*

$$\mathbf{w}_{GE} = \mathbf{u}_{max}(\mathbf{R}_B^{-1}\mathbf{R}_A), \quad \mathbf{q}_{GE} = \mathbf{u}_{max}(\mathbf{R}_A^{-1}\mathbf{R}_B). \tag{2.6}$$

The corresponding lower bound of the ergodic sum rate is

$$R_{sum,lb} = \log_2 \left(\chi\big(\mathbf{R}_B^{-1}\mathbf{R}_A\big) \right) = \log_2 \left(\chi\big(\mathbf{R}_A^{-1}\mathbf{R}_B\big) \right), \tag{2.7}$$

where $\chi(\cdot) = \frac{\lambda_{max}(\cdot)}{\lambda_{min}(\cdot)}$ is the condition number.

Proof. A detailed proof is relegated in Appendix A.3. □

A special case of (2.7) (when the two users share the same set of statistical eigenmodes but orthogonal dominant eigenvectors and $M = 2$) is confirmed by [18, Corollary 2]. In the low SNR regime where the interference can be completely ignored, the optimal choice is to send along the dominant statistical eigen-mode of the user's own channel [69]. At intermediate SNR, however, [6] has shown the difficulty of finding a closed-form expression of the optimal precoders even for $M = 2$ case. Instead, it is solved by an exhaustive search operated upon a linearly combined high- and low-SNR solution. In the general $M > 2$ case, we compute only a high-SNR solution and avoid the line search method. The simulation results show that it works well at practical SNR.

Remark 2.1. *The closed-form precoders that maximise the ergodic sum rate of SBF is difficult to compute due to the coupled nature in the SINR expression. To solve this problem, we can use an alternative SLNR metric, which is defined as $SLNR_A = \frac{\rho|\mathbf{h}^H\mathbf{w}|^2}{1+\rho|\mathbf{g}^H\mathbf{w}|^2}$. At high SNR, the maximisation of a lower bound on $\mathbb{E}[SLNR_A]$ also leads to the solution* (2.6)*. Similarly, the effectiveness of the SLNR metric in designing multi-user transmit beamforming vectors has been examined in [6, 18, 70].*

In contrast with statistical beamforming with generalised eigenvectors (SGEBF), the statistical beamforming with weakest eigenvectors (SWEBF) computes the precoding vectors as $\mathbf{w}_{WE} = \mathbf{u}_{min}(\mathbf{R}_B)$, $\mathbf{q}_{WE} = \mathbf{u}_{min}(\mathbf{R}_A)$. However, the rate performance of SWEBF is unfavourable in the scenario where both channels of user A and B have similar weakest eigen-direction (e.g., co-located users). Specifically, the precoding vectors which are designed to remove the interference also cancel out the intended signal. In contrast, the GE beamforming approach obtains a balance between interference cancellation and desired signal enhancement. In other words, SGEBF exhibits robustness with respect to different channels compared to SWEBF.

2.3. AMAT Strategy

The AMAT transmission strategy that makes use of delayed CSIT is reviewed in Section 1.7.2. Under equal power allocation, the ergodic rate performance of AMAT can be further enhanced by precoding design with the help of additional statistical CSIT. In this section, we develop an iterative algorithm to compute the precoding vectors of AMAT, which maximises an approximation of the ergodic sum rate.

Our idea is inspired by [20], where the authors [20] has shown that additional channel statistics enable a higher achievable sum rate compared with the original MAT. However, there are big differences between this work and [20]. Firstly, they released the power constraint in the interference retransmission phase (stage II), which leads to a variation of the total transmit power. We here control the power consumption by using a long-term power constraint. Secondly, an efficient iterative algorithm is developed to compute the statistical precoders to maximise an approximation of the ergodic sum rate. Particularly, monotonic convergence of the algorithm is proved.

2.3.1. Rate Approximation

Recall that the AMAT strategy transmits four symbols at the first time slot, each two intended to user A and B respectively. Then, the overheard interferences are reconstructed and retransmitted in the second and third slots with delayed CSIT of the first slot. The transmitted signals of AMAT with equal power distribution, denoted as ρ, can be expressed as

$$\begin{aligned} \mathbf{x}_1 &= \sqrt{\rho}\,\mathbf{W}\,\mathbf{s}_A \;+\; \sqrt{\rho}\,\mathbf{Q}\,\mathbf{s}_B \\ \mathbf{x}_2 &= \sqrt{\rho}\left[\mathbf{h}_1^H\mathbf{Q}\,\mathbf{s}_B,\; 0\right]^T \\ \mathbf{x}_3 &= \sqrt{\rho}\left[\mathbf{g}_1^H\mathbf{W}\,\mathbf{s}_A,\; 0\right]^T \end{aligned} \tag{2.8}$$

where $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2], \mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2]$ are the precoding matrices of user A and B, respectively. Moreover, $\mathbf{w}_1, \mathbf{w}_2, \mathbf{q}_1, \mathbf{q}_2$ are unit norm vectors and will be optimised to maximise the ergodic sum rate of AMAT. For simplicity of exposition, we focus on the performance of user A and similar results can be symmetrically applied to user B. The signal vector received by user A is given by

$$\mathbf{y}_A = \sqrt{\rho}\begin{bmatrix} \mathbf{h}_1^H\mathbf{W} \\ \mathbf{0} \\ h_{31}^*\mathbf{g}_1^H\mathbf{W} \end{bmatrix}\mathbf{s}_A + \sqrt{\rho}\begin{bmatrix} \mathbf{h}_1^H\mathbf{Q} \\ h_{21}^*\mathbf{h}_1^H\mathbf{Q} \\ \mathbf{0} \end{bmatrix}\mathbf{s}_B + \begin{bmatrix} n_{A1} \\ n_{A2} \\ n_{A3} \end{bmatrix}, \tag{2.9}$$

where $\mathbf{y}_A \triangleq [y_{A1}, y_{A2}, y_{A3}]^T$ denotes the received signals over three time slots and h_{jm} denotes the channel coefficient between m-th transmit antenna and user A in time slot j. $n_{Aj} \sim \mathcal{CN}(0,1)$ is the normalised complex AWGN. After further interference elimination,

the received signal vector becomes

$$\widetilde{\mathbf{y}}_A = \sqrt{\rho}\,\widetilde{\mathbf{H}}\,\mathbf{s}_A + \begin{bmatrix} h_{21}^* n_{A1} - n_{A2} \\ n_{A3} \end{bmatrix}, \tag{2.10}$$

where $\widetilde{\mathbf{H}} = [(h_{21}^* \mathbf{h}_1^H \mathbf{W})^T, (h_{31}^* \mathbf{g}_1^H \mathbf{W})^T]^T$. By utilising the Minimum Mean Square Error (MMSE) receiver with Successive Interference Cancellation (SIC), the ergodic rate achieved per slot by user A is written as

$$R_A = \frac{1}{3}\,\mathbb{E}\left[\log_2 \det\left(\mathbf{I} + \rho \widetilde{\mathbf{H}}^H \mathbf{K}^{-1} \widetilde{\mathbf{H}}\right)\right], \tag{2.11}$$

where $\mathbf{K}$ is the covariance matrix of the noise vector in (2.10) and given by $\mathbf{K} = \mathrm{diag}(1 + |h_{21}|^2, 1)$. It is challenging to obtain the closed-form expression of the ergodic rate, especially for $M > 2$ case. Hence, we optimise the linear beamforming vectors based on the following analytical approximation of R_A.

Proposition 2.1. *In spatially correlated Rayleigh fading channel, the ergodic rate of user A for AMAT can be approximated as*

$$R_A \approx \frac{2}{3}\log_2\left(1 + \rho\sqrt{e^a \Theta_A}\right), \tag{2.12}$$

where

$$\Theta_A = tr(\mathbf{W}^H \mathbf{R}_A \mathbf{W}) tr(\mathbf{W}^H \mathbf{R}_B \mathbf{W}) - tr(\mathbf{W}^H \mathbf{R}_A \mathbf{W} \mathbf{W}^H \mathbf{R}_B \mathbf{W}) \tag{2.13}$$

and $a = e\,Ei(-1) - 2\gamma$, $Ei(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt$ *is the exponential integral.*

Proof. See Appendix A.4. □

Then, we can obtain that $R_{\text{sum}} = R_A + R_B \approx \frac{2}{3}\log_2(1+\rho\sqrt{e^a\Theta_A}) + \frac{2}{3}\log_2(1+\rho\sqrt{e^a\Theta_B})$, where $\Theta_B = \mathrm{tr}(\mathbf{Q}^H \mathbf{R}_A \mathbf{Q})\mathrm{tr}(\mathbf{Q}^H \mathbf{R}_B \mathbf{Q}) - \mathrm{tr}(\mathbf{Q}^H \mathbf{R}_A \mathbf{Q}\mathbf{Q}^H \mathbf{R}_B \mathbf{Q})$. It shows that the AMAT strategy exploiting delayed CSIT enables a DoF of $\frac{4}{3}$ at high SNR, while the beamforming based on statistical CSIT makes no contribution to the DoF gain. However, the ergodic rate performance at practical SNR benefits from spatial correlation. In (2.12) and (2.13), we observe that the ergodic rate relies on the precoders and the spatial correlation matrices. Since the latter terms are invariable, the ergodic rate R_A and R_B only depend on Θ_A ($\mathbf{W}$) and Θ_B ($\mathbf{Q}$), respectively. To maximise the ergodic sum rate, the precoders $\mathbf{W}$ and $\mathbf{Q}$ can be independently designed. Let us focus on Θ_A only and optimise $\mathbf{W}$. Likewise, we can obtain the optimal $\mathbf{Q}$ that maximises Θ_B.

2.3.2. Precoder Design

1) *Multi-antenna case* ($M > 2$): It is analytically challenging to obtain a closed-form expression of the beamforming vectors that maximise (2.13) and further (2.12). For such a problem where joint optimisation is difficult but the objective function is convex in each

1: **Initialise**: Set iteration index $m = 0$, and randomly generate $\mathbf{w}_1^{(0)}$, $\mathbf{w}_2^{(0)}$
2: **Repeat**
3: $m \leftarrow m + 1$
4: Update $\mathbf{w}_1^{(m)}$ with GradAct [**Algorithm 1**], or Max-Eig [**Algorithm 2**]
5: Update $\mathbf{w}_2^{(m)}$ with GradAct or with Max-Eig
6: **Until** $|\Theta_A^{(m)} - \Theta_A^{(m-1)}| \leq \epsilon$

Table 2.1.: Precoder optimisation algorithms for AMAT.

of the optimisation variables $\mathbf{w}_1$ and $\mathbf{w}_2$, an alternating algorithm, also known as Block Coordinate Descent, has been widely used in [71, 72]. More specifically, we maximise (2.13) by sequentially fixing one vector and updating the other. Fix $\mathbf{w}_2$ and focus on $\mathbf{w}_1$ (vice versa, the following derivations still hold). We can reformulate the subproblem as

$$\begin{aligned} \max_{\|\mathbf{w}_1\|=1} \Theta_A(\mathbf{w}_1) \quad = \quad & \mathbf{w}_1^H \mathbf{R}_A \mathbf{w}_1 \mathbf{w}_2^H \mathbf{R}_B \mathbf{w}_2 + \mathbf{w}_1^H \mathbf{R}_B \mathbf{w}_1 \mathbf{w}_2^H \mathbf{R}_A \mathbf{w}_2 \\ & - \mathbf{w}_1^H \mathbf{R}_A \mathbf{w}_2 \mathbf{w}_2^H \mathbf{R}_B \mathbf{w}_1 - \mathbf{w}_1^H \mathbf{R}_B \mathbf{w}_2 \mathbf{w}_2^H \mathbf{R}_A \mathbf{w}_1. \end{aligned} \tag{2.14}$$

Since Θ_A is convex in $\mathbf{w}_1$[2], the classical gradient ascent (GradAct) method can be used to determine the optimal solution (step 4 of Table 2.1). Once the optimal $\mathbf{w}_1$ is obtained in terms of certain $\mathbf{w}_2$, the process is repeated the other way around (step 5), leading to an iterative algorithm. Since the steepest ascent direction acts as the best direction to increase the objective function, a proper step size can be computed for a non-decreasing objective value, i.e., $\Theta_A^{(m,4)} \leq \Theta_A^{(m,5)} \leq \Theta_A^{(m+1,4)}$ where $\Theta_A^{(m,4)}$ refers to the objective value at step 4 in the m-th iteration in Table 2.1 (Algorithm 1). The convergence of Algorithm 1 is ensured, since Θ_A is monotonically increased (non-decreased) after each iteration and upper bounded. Even though the optimal solution is obtained for each subproblem, the iterative algorithm cannot guarantee the global optimal beamforming vectors.

Alternatively, (2.14) is quadratic in $\mathbf{w}_1$ and the optimal solution can be obtained by eigen-decomposition. Rewrite (2.14) as $\Theta_A(\mathbf{w}_1) = \mathbf{w}_1^H \mathbf{M}(\mathbf{w}_2) \mathbf{w}_1$, where $\mathbf{M}(\mathbf{w}_2) = \mathbf{w}_2^H \mathbf{R}_B \mathbf{w}_2 \mathbf{R}_A + \mathbf{w}_2^H \mathbf{R}_A \mathbf{w}_2 \mathbf{R}_B - \mathbf{R}_A \mathbf{w}_2 \mathbf{w}_2^H \mathbf{R}_B - \mathbf{R}_B \mathbf{w}_2 \mathbf{w}_2^H \mathbf{R}_A$. The closed-form solution is the maximum eigenvector (Max-Eig),

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}_1\|=1} \Theta_A(\mathbf{w}_1) = \mathbf{u}_{\max}\left(\mathbf{M}\left(\mathbf{w_2}\right)\right). \tag{2.15}$$

With (2.15) at hand, we can easily compute the optimal precoders by the proposed iterative alogithm. In Fig.2.1, we show by two cases ($M = 4, 8$) that the iterative algorithm converges very fast, where the covariance matrices are randomly generated.

2) *Two-antenna case* ($M = 2$): A special case of considerable interest is the two transmit antenna scenario. The optimal precoders can be easily obtained as follows.

[2]The convexity can be easily proved with the second order condition, which is omitted here for conciseness.

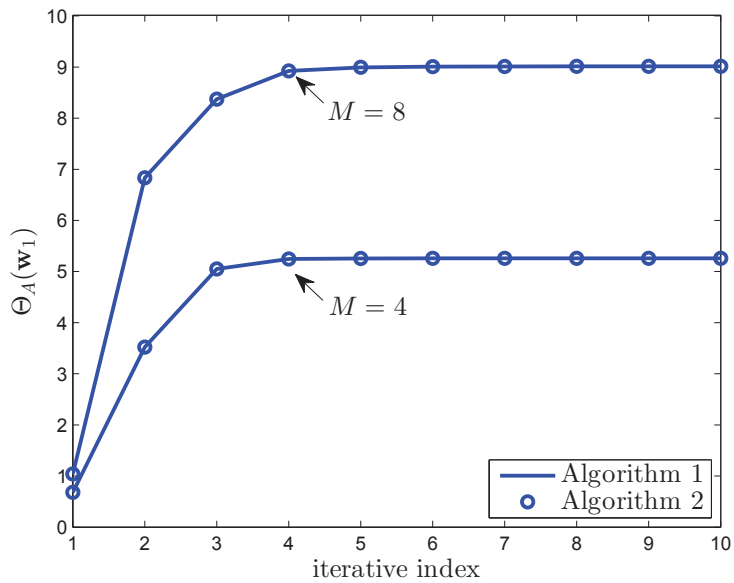

Figure 2.1.: Convergence of iterative algorithm 1&2.

Proposition 2.2. *For two-user MISO BC with $M = 2$ and spatially correlated Rayleigh fading, any unitary beamforming matrix is optimal to maximise* (2.13) *and further* (2.12).

Proof. A detailed proof is provided in Appendix A.5. □

This proposition reveals that any orthogonal beamforming vectors with equal power allocation achieve the same ergodic sum rate. This observation can be verified by the Fig. 2.5 in Section 2.4. Then, let us compute the equal power allocation ρ.

For (A)MAT-based schemes, the transmit power in stage II is inherently dependent on the channel realisation that changes rapidly. The power consumption in each transmission period hardly keeps constant. A less restrictive metric is the long-term average power constraint. Accordingly, the long-term average power consumption for AMAT is represented by

$$\begin{aligned}\bar{P}_{\mathrm{c}} &= 4\rho + \rho\,\mathrm{tr}(\mathbf{Q}^H\mathbf{R}_B\mathbf{Q}) + \rho\,\mathrm{tr}(\mathbf{W}^H\mathbf{R}_A\mathbf{W}) \\ &= \rho\,(4 + \mathbf{q}_1^H\mathbf{R}_B\mathbf{q}_1 + \mathbf{q}_2^H\mathbf{R}_B\mathbf{q}_2 + \mathbf{w}_1^H\mathbf{R}_A\mathbf{w}_1 + \mathbf{w}_2^H\mathbf{R}_A\mathbf{w}_2) \\ &\leq \rho\,(4 + 2M)\,. \end{aligned} \tag{2.16}$$

where the inequality (2.16) is obtained by using $\mathbf{x}_1^H\mathbf{R}\mathbf{x}_1 + \mathbf{x}_2^H\mathbf{R}\mathbf{x}_2 \leq \lambda_1(\mathbf{R}) + \lambda_2(\mathbf{R}) \leq \mathrm{tr}\,(\mathbf{R})$, where unit-norm $\mathbf{x}_i$ are mutually orthogonal and $\lambda_i(\mathbf{R})$ corresponds to the i-th largest eigenvalue [73], [74]. In order to maintain the power constraint, equal power allocation is calculated as $\rho = \frac{3P}{4+2M}$ (e.g., $\rho = \frac{3P}{8}$ for two transmit antennas). Equality in (2.16) holds for $M = 2$ case, which also justifies proposition 2.2 in the sense that orthonormal precoders (optimally) use up the entire power budget.

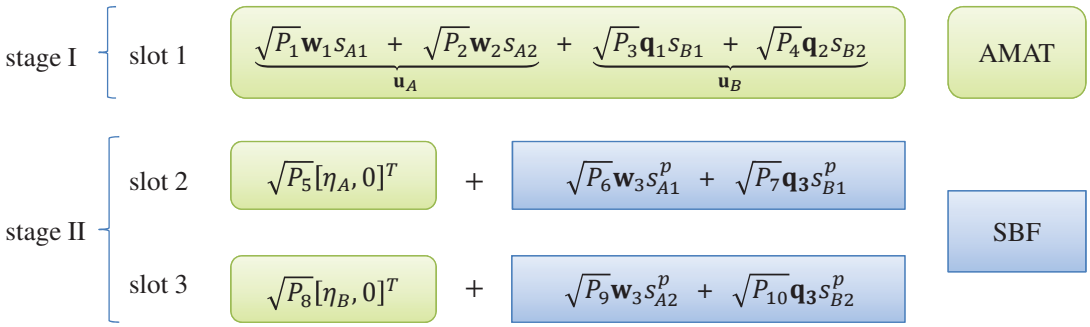

Figure 2.2.: Block diagram of the proposed SAMAT scheme.

2.4. SAMAT Strategy

In this section, we elaborate the SAMAT framework which bridges the ergodic rate performance gap between the SBF and AMAT presented in last two sections. This can be done by first deriving a tractable approximation of ergodic sum rate of SAMAT and then optimising the power allocation based on the approximation.

The proposed SAMAT framework is indeed a superposition of AMAT (expanded into three time slots) and SBF (in second and third time slots), shown in Fig. 2.2. In slot 1, the transmitter superposes four private symbols $s_{A1}, s_{A2}, s_{B1}, s_{B2}$ and sends them to both users. Denote $\mathbf{u}_A = \mathbf{W}\mathbf{P}_A^{1/2}\mathbf{s}_A$ and $\mathbf{u}_B = \mathbf{Q}\mathbf{P}_B^{1/2}\mathbf{s}_B$ as the encoded symbols with statistical beamformer and power allocation, where $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2], \mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2], \mathbf{P}_A = \mathrm{diag}(P_1,\ P_2), \mathbf{P}_B = \mathrm{diag}(P_3,\ P_4)$. $\mathbf{s}_k = [s_{k1}, s_{k2}]^T$ represents the symbols intended to user k and $\mathbb{E}\{\mathbf{s}_k\mathbf{s}_k^H\} = \mathbf{I}$.

At the end of first slot, each user receives its desired signal as well as the overheard interference due to the superposed transmission. Denote $\eta_A = \mathbf{h}_1^H\mathbf{u}_B$ and $\eta_B = \mathbf{g}_1^H\mathbf{u}_A$ as the interference overheard by user A and B, respectively. In slot 2 and 3, the transmitter has access to $\mathbf{h}_1$ and $\mathbf{g}_1$. Then, η_A and η_B can be reconstructed and broadcast via a single antenna. This step helps both users eliminate the overheard interference and reinforce the desired signals. In addition, new private messages $s_{A1}^p, s_{A2}^p, s_{B1}^p, s_{B2}^p$ are sent to both users in a superposed fashion and this extra transmission makes use of statistical CSIT only. $P_k \geq 0, k = 1, \ldots, 10$ indicate the power allocated to each symbol. $\mathbf{w}_k$ and $\mathbf{q}_k, k = 1, \ldots, 3$ denote $M \times 1$ unit-norm precoders which depend only on statistical CSIT.

The proposed SAMAT scheme facilitates a smart use of statistical and/or delayed CSIT. With statistical CSIT only, SAMAT with $P_1 = P_3 = P_5 = P_8 = 0$ boils down to SBF in each time slot. With delayed CSIT only, SAMAT with $P_6 = P_7 = P_9 = P_{10} = 0$ becomes AMAT and enables a sum DoF of $\frac{4}{3}$ at high SNR. If the transmitter has both statistical and delayed CSIT, we will show that proper power allocation and statistical precoding can make room for extra symbols transmission. In this case, the proposed SAMAT framework

allows for the parallel transmission of SBF on top of AMAT while outperforming AMAT and SBF at any SNR.

Based on Fig. 2.2, the transmitted signals are written as

$$\begin{aligned}
\mathbf{x}_1 &= \mathbf{u}_A + \mathbf{u}_B && (2.17)\\
\mathbf{x}_2 &= \sqrt{P_5}\left[\eta_A,\ 0\right]^T + \sqrt{P_6}\,\mathbf{w}_3\, s_{A1}^p + \sqrt{P_7}\,\mathbf{q}_3\, s_{B1}^p && (2.18)\\
\mathbf{x}_3 &= \sqrt{P_8}\left[\eta_B,\ 0\right]^T + \sqrt{P_9}\,\mathbf{w}_3\, s_{A2}^p + \sqrt{P_{10}}\,\mathbf{q}_3\, s_{B2}^p. && (2.19)
\end{aligned}$$

The long-term average power constraint is considered

$$\begin{aligned}
\bar{P}_{\mathrm{c}} &= \mathbb{E}[\mathrm{tr}(\mathbf{x}_1\mathbf{x}_1^H)] + \mathbb{E}[\mathrm{tr}(\mathbf{x}_2\mathbf{x}_2^H)] + \mathbb{E}[\mathrm{tr}(\mathbf{x}_3\mathbf{x}_3^H)]\\
&= \sum_{i=1}^{4} P_i + P_6 + P_7 + P_9 + P_{10} + P_5(\lambda_{A1}P_3 + \lambda_{A2}P_4) + P_8(\lambda_{B1}P_1 + \lambda_{B2}P_2)\\
&\leq 3P, && (2.20)
\end{aligned}$$

where $\lambda_{A1} = \mathbf{q}_1^H\mathbf{R}_A\mathbf{q}_1$, $\lambda_{A2} = \mathbf{q}_2^H\mathbf{R}_A\mathbf{q}_2$, $\lambda_{B1} = \mathbf{w}_1^H\mathbf{R}_B\mathbf{w}_1$, $\lambda_{B2} = \mathbf{w}_2^H\mathbf{R}_B\mathbf{w}_2$. The expectation is taken over the input signals and the channels and P denotes the average power budget of the transmitter for each time slot.

Next, let us analyse the ergodic sum rate of SAMAT. Hereafter, we focus on user A and similar results can be derived for user B. The received signal of user A can be written as

$$\mathbf{y}_A = \mathbf{H}_1\mathbf{P}_A^{1/2}\mathbf{s}_A + \mathbf{H}_2\mathbf{P}_B^{1/2}\mathbf{s}_B + \mathbf{H}_3\,\mathbf{s}_A^p + \mathbf{H}_4\,\mathbf{s}_B^p + \mathbf{n}_A, \tag{2.21}$$

where

$$\mathbf{H}_1 \triangleq \begin{bmatrix} \mathbf{h}_1^H\mathbf{W} \\ \mathbf{0} \\ \sqrt{P_8}\,h_{31}^*\mathbf{g}_1^H\mathbf{W} \end{bmatrix} \quad \mathbf{H}_3 \triangleq \begin{bmatrix} 0 & 0 \\ \sqrt{P_6}\mathbf{h}_2^H\mathbf{w}_3 & 0 \\ 0 & \sqrt{P_9}\,\mathbf{h}_3^H\mathbf{w}_3 \end{bmatrix}, \tag{2.22}$$

$$\mathbf{H}_2 \triangleq \begin{bmatrix} \mathbf{h}_1^H\mathbf{Q} \\ \sqrt{P_5}\,h_{21}^*\mathbf{h}_1^H\mathbf{Q} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{H}_4 \triangleq \begin{bmatrix} 0 & 0 \\ \sqrt{P_7}\,\mathbf{h}_2^H\mathbf{q}_3 & 0 \\ 0 & \sqrt{P_{10}}\,\mathbf{h}_3^H\mathbf{q}_3 \end{bmatrix}, \tag{2.23}$$

and $\mathbf{y}_A \triangleq [y_{A1}, y_{A2}, y_{A3}]^T$, $\mathbf{s}_A^p \triangleq [s_{A3}, s_{A4}]^T$, $\mathbf{s}_B^p \triangleq [s_{B3}, s_{B4}]^T$. $\mathbf{n}_A \triangleq [n_{A1}, n_{A2}, n_{A3}]^T$ with $n_{Aj} \sim \mathcal{CN}(0,1)$. The decoding procedure that mainly uses interference alignment and cancellation (similar to [19]) is described as follows. Denote $\widetilde{\mathbf{y}}_A$ as the received signal after subtracting $\sqrt{P_5}\,h_{21}^* \cdot y_{A1}$ from y_{A2} and retaining y_{A1}[3]. First, decode the private

[3] A(MAT)-based schemes use one observation to completely remove the overheard interference and two independent observations remain to resolve two symbols. By contrast, we cancel the overheard interference while we maintain all three observations. The reasons are explained as follows: 1) in some cases (e.g., highly correlated channel) where SAMAT boils down to SBF in each time slot, symbols need to be decoded slot by slot. However, conventional decoding method causes rate loss because one observation is dropped; 2) the ergodic rate of the proposed decoding method is slightly better than the conventional one, due to one more observation even with strong interference.

symbols ($\mathbf{s}_A$) by regarding the extra symbols ($\mathbf{s}_A^p, \mathbf{s}_B^p$) as interference

$$\widetilde{\mathbf{y}}_A = \widetilde{\mathbf{H}}_1 \mathbf{P}_A^{1/2} \mathbf{s}_A + \underbrace{\widetilde{\mathbf{H}}_2 \mathbf{P}_B^{1/2} \mathbf{s}_B + \mathbf{H}_3\, \mathbf{s}_A^p + \mathbf{H}_4\, \mathbf{s}_B^p + \widetilde{\mathbf{n}}_A}_{\mathbf{z}}, \tag{2.24}$$

where $\widetilde{\mathbf{H}}_1 = [(\mathbf{h}_1^H \mathbf{W})^T, -(\sqrt{P_5}\, h_{21}^* \mathbf{h}_1^H \mathbf{W})^T, (\sqrt{P_8}\, h_{31}^* \mathbf{g}_1^H \mathbf{W})^T]^T$, $\widetilde{\mathbf{H}}_2 = [(\mathbf{h}_1^H \mathbf{Q})^T,\ \mathbf{0}^T,\ \mathbf{0}^T]^T$ and $\widetilde{\mathbf{n}}_A = [n_{A1}, n_{A2} - \sqrt{P_5}\, h_{21}^* n_{A1}, n_{A3}]^T$. $\mathbf{K}$ is the covariance matrix of the interference plus noise vector $\mathbf{z}$ and given by $\mathbf{K} = \mathrm{diag}(k_1, k_2, k_3)$, where $k_1 = 1 + |\mathbf{h}_1^H \mathbf{Q} \mathbf{P}_B|^2, k_2 = 1 + P_5 |h_{21}|^2 + P_6 |\mathbf{h}_2^H \mathbf{w}_3|^2 + P_7 |\mathbf{h}_2^H \mathbf{q}_3|^2, k_3 = 1 + P_9 |\mathbf{h}_3^H \mathbf{w}_3|^2 + P_{10} |\mathbf{h}_3^H \mathbf{q}_3|^2$. To resolve $\mathbf{s}_A$, MMSE-SIC receiver is applied on (2.24) and the ergodic sum rate of $\mathbf{s}_A$ can be written as

$$R_{\mathbf{s}_A} = \mathbb{E}\left[\log_2 \det\left(\mathbf{I} + \mathbf{P}_A \widetilde{\mathbf{H}}_1^H \mathbf{K}^{-1} \widetilde{\mathbf{H}}_1 \mathbf{P}_A\right)\right]. \tag{2.25}$$

Once $\mathbf{s}_A$ is obtained, we subtract it from $\widetilde{\mathbf{y}}_A$. Then, we decode the extra symbols by taking the second and third entries of $\widetilde{\mathbf{y}}_A$ as $\widehat{\mathbf{y}}_A$

$$\widehat{\mathbf{y}}_A = \widehat{\mathbf{H}}_3\, \mathbf{s}_A^p + \widehat{\mathbf{H}}_4\, \mathbf{s}_B^p + \widehat{\mathbf{n}}_A, \tag{2.26}$$

where $\widehat{\mathbf{H}}_3 = \mathrm{diag}(\sqrt{P_6}\, \mathbf{h}_2^H \mathbf{w}_3,\ \sqrt{P_9}\, \mathbf{h}_3^H \mathbf{w}_3)$, $\widehat{\mathbf{H}}_4 = \mathrm{diag}(\sqrt{P_7}\, \mathbf{h}_2^H \mathbf{q}_3,\ \sqrt{P_{10}}\, \mathbf{h}_3^H \mathbf{q}_3)$ and $\widehat{\mathbf{n}}_A = [n_{A2} - \sqrt{P_5}\, h_{21}^* n_{A1}, n_{A3}]^T$. The covariance matrix of $\widehat{\mathbf{n}}_A$ is given by $\mathbf{N} = \mathrm{diag}(1 + P_5 |h_{21}|^2,\ 1)$. The ergodic sum rate of $\mathbf{s}_A^p$ is given by

$$R_{\mathbf{s}_A}^p = \mathbb{E}\left[\log_2 \det\left(\mathbf{I} + \widehat{\mathbf{H}}_3^H (\mathbf{N} + \widehat{\mathbf{H}}_4 \widehat{\mathbf{H}}_4^H)^{-1} \widehat{\mathbf{H}}_3\right)\right]. \tag{2.27}$$

It is challenging to obtain the closed-form expression for the ergodic rate, we rather derive a tractable approximation.

Proposition 2.3. *The achievable ergodic sum rate per slot at user A with linear beamforming can be approximated as* $R_A \triangleq \frac{1}{3}(R_{\mathbf{s}_A} + R_{\mathbf{s}_A^p})$ *where*

$$R_{\mathbf{s}_A} \approx \log_2\left(1 + \delta_{A1}\left(\tau_{A1} P_1 + \tau_{A2} P_2\right) + \delta_{A2}\left(\lambda_{B1} P_1 + \lambda_{B2} P_2\right) + \delta_{A1}\delta_{A2}\Theta_A P_1 P_2\right) \tag{2.28}$$

$$R_{\mathbf{s}_A^p} \approx \log_2\left(1 + \frac{\tau_{A3} P_6}{1 + P_5 + \lambda_{A3} P_7}\right) + \log_2\left(1 + \frac{\tau_{A3} P_9}{1 + \lambda_{A3} P_{10}}\right) \tag{2.29}$$

Similarly, we have $R_B \triangleq \frac{1}{3}(R_{\mathbf{s}_B} + R_{\mathbf{s}_B^p})$ *and*

$$R_{\mathbf{s}_B} \approx \log_2\left(1 + \delta_{B1}\left(\tau_{B1} P_3 + \tau_{B2} P_4\right) + \delta_{B2}\left(\lambda_{A1} P_3 + \lambda_{A2} P_4\right) + \delta_{B1}\delta_{B2}\Theta_B P_3 P_4\right) \tag{2.30}$$

$$R_{\mathbf{s}_B^p} \approx \log_2\left(1 + \frac{\tau_{B3} P_7}{1 + \lambda_{B3} P_6}\right) + \log_2\left(1 + \frac{\tau_{B3} P_{10}}{1 + P_8 + \lambda_{B3} P_9}\right) \tag{2.31}$$

where

$$\begin{aligned}
\delta_{A1} &= \frac{1}{1+\lambda_{A1}P_3+\lambda_{A2}P_4}+\frac{P_5}{1+P_5+\tau_{A3}P_6+\lambda_{A3}P_7}\\
\delta_{B1} &= \frac{1}{1+\lambda_{B1}P_1+\lambda_{B2}P_2}+\frac{P_8}{1+P_8+\lambda_{B3}P_9+\tau_{B3}P_{10}}\\
\delta_{A2} &= \frac{P_8}{1+\tau_{A3}P_9+\lambda_{A3}P_{10}}, \quad \delta_{B2}=\frac{P_5}{1+\lambda_{B3}P_6+\tau_{B3}P_7}
\end{aligned} \tag{2.32}$$

$$\begin{aligned}
\lambda_{A1} &= \mathbf{q}_1^H\mathbf{R}_A\mathbf{q}_1, \quad \lambda_{A2}=\mathbf{q}_2^H\mathbf{R}_A\mathbf{q}_2, \quad \lambda_{B1}=\mathbf{w}_1^H\mathbf{R}_B\mathbf{w}_1\\
\lambda_{B2} &= \mathbf{w}_2^H\mathbf{R}_B\mathbf{w}_2, \quad \tau_{A1}=\mathbf{w}_1^H\mathbf{R}_A\mathbf{w}_1, \quad \tau_{A2}=\mathbf{w}_2^H\mathbf{R}_A\mathbf{w}_2\\
\tau_{B1} &= \mathbf{q}_1^H\mathbf{R}_B\mathbf{q}_1, \quad \tau_{B2}=\mathbf{q}_2^H\mathbf{R}_B\mathbf{q}_2, \quad \lambda_{A3}=\mathbf{q}_3^H\mathbf{R}_A\mathbf{q}_3\\
\lambda_{B3} &= \mathbf{w}_3^H\mathbf{R}_B\mathbf{w}_3, \quad \tau_{A3}=\mathbf{w}_3^H\mathbf{R}_A\mathbf{w}_3, \quad \tau_{B3}=\mathbf{q}_3^H\mathbf{R}_B\mathbf{q}_3
\end{aligned} \tag{2.33}$$

$$\begin{aligned}
\Theta_A &= tr(\mathbf{W}^H\mathbf{R}_A\mathbf{W})\, tr(\mathbf{W}^H\mathbf{R}_B\mathbf{W}) - tr(\mathbf{W}^H\mathbf{R}_A\mathbf{W}\mathbf{W}^H\mathbf{R}_B\mathbf{W})\\
\Theta_B &= tr(\mathbf{Q}^H\mathbf{R}_A\mathbf{Q})\, tr(\mathbf{Q}^H\mathbf{R}_B\mathbf{Q}) - tr(\mathbf{Q}^H\mathbf{R}_A\mathbf{Q}\mathbf{Q}^H\mathbf{R}_B\mathbf{Q}).
\end{aligned} \tag{2.34}$$

Proof. Refer to Appendix A.6 for proof. □

Remark 2.2. *Compared with the interference quantisation approach in [33], the analog transmission induces a noise enhancement. Namely, interference alignment cancels the overheard interference while scaling up the noise by P_5 (P_8). This noise enhancement can be observed in from* (2.24) *to* (2.31)*. At low SNR, the proposed SAMAT strategy behaves as SBF in each time slot. The scaling factors are small and therefore the effect of noise enhancement is negligible. The gain over AMAT mainly comes from extra symbol transmission and statistical precoding. At high SNR, the proposed SAMAT strategy behaves as AMAT, achieving a DoF of $\frac{4}{3}$. In this case, the ergodic rates of extra symbols can be eliminated by noise enhancement. Namely, we have little benefit by transmitting extra symbols. However, the proposed SAMAT strategy still achieves significant gain over SBF and AMAT by power allocation optimisation and statistical precoding.*

With predefined beamforming vectors, the proposed SAMAT strategy softly bridges between SBF and AMAT by power control. Let us concentrate on two cases.

case 1: bridge between SWEBF and AMAT, $\mathbf{w}_1=\mathbf{u}_{\max}(\mathbf{R}_B)$, $\mathbf{q}_1=\mathbf{u}_{\max}(\mathbf{R}_A)$, $\mathbf{w}_2=\mathbf{w}_3=\mathbf{w}_{WE}=\mathbf{u}_{\min}(\mathbf{R}_B)$, $\mathbf{q}_2=\mathbf{q}_3=\mathbf{q}_{WE}=\mathbf{u}_{\min}(\mathbf{R}_A)$;

case 2: bridge between SGEBF and AMAT, $\mathbf{w}_1=\mathbf{u}_{\min}(\mathbf{R}_B^{-1}\mathbf{R}_A)$, $\mathbf{q}_1=\mathbf{u}_{\min}(\mathbf{R}_A^{-1}\mathbf{R}_B)$, $\mathbf{w}_2=\mathbf{w}_3=\mathbf{w}_{GE}=\mathbf{u}_{\max}(\mathbf{R}_B^{-1}\mathbf{R}_A)$, $\mathbf{q}_2=\mathbf{q}_3=\mathbf{q}_{GE}=\mathbf{u}_{\max}(\mathbf{R}_A^{-1}\mathbf{R}_B)$.

case 1 is used to show the efficacy of the power allocation optimisation technique by which the proposed SAMAT strategy can softly bridge between SWEBF and AMAT. Beyond this, ***case 2*** makes better use of statistical CSIT in the sense that SGEBF exhibits better robustness compared with SWEBF. Instead of using the optimised AMAT

precoders in transmission stage I ($\mathbf{w}_1, \mathbf{w}_2, \mathbf{q}_1, \mathbf{q}_2$ as developed in Section 2.3), we use the precoders above ($\mathbf{w}_1, \mathbf{w}_2, \mathbf{q}_1, \mathbf{q}_2$ as WE/GE precoder) and the motivations are explained as follows. Firstly, the optimal precoders in Section 2.3 that maximise the ergodic sum rate of AMAT under equal power allocation are not necessarily optimal for SAMAT with power control. Secondly, SAMAT boils down to SBF at low to intermediate SNR in highly correlated channel, where the optimised AMAT precoders may cause a poorer rate performance compared to the WE/GE precoders. In order to softly bridge between SBF and AMAT, we adopt the precoder design as above.

Accordingly, $R_{\text{sum}} \triangleq R_A + R_B$ and the ergodic sum rate optimisation problem is formulated as

$$\max_{\{P_i\}} R_{\text{sum}} \quad \text{s.t.} \quad \bar{P}_{\text{c}} = 3P, \; P_i \geq 0 \quad i = 1, \ldots, 10. \tag{2.35}$$

It was shown that the sum rate optimisation problem is generally NP hard [75]. Thus, an algorithm achieving global optimum cannot be expected. Nevertheless, Sequential Quadratic Programming (SQP) algorithm provides an efficient way to solve non-linear constrained optimisation problem. An overview on SQP is provided in [76–78]. Briefly, a quadratic approximation of the Lagrangian function is made by applying quasi-Newton updating method. The consequent QP subproblem can be optimally solved and then the solution is used as a search direction. With proper line search, an estimate of the solution is computed for the next iteration. This SQP algorithm can guarantee a super-linear convergence to a local minimum.

In order to get insights into the optimal power allocation, a necessary condition for optimality of the constrained problem (2.35) is identified from the Karush-Kuhn-Tucker (KKT) conditions. Accordingly, the optimum power allocation that depends on the precoders and the channel covariance matrices is stated as follows.

Theorem 2.2. *At high SNR, the optimal power allocation that maximises R_{sum} in* (2.35) *should satisfy*

$$\frac{P_1}{P_2} = \frac{1 + \lambda_{B2} P_8}{1 + \lambda_{B1} P_8}, \qquad \frac{P_3}{P_4} = \frac{1 + \lambda_{A2} P_5}{1 + \lambda_{A1} P_5} \tag{2.36}$$

where $\lambda_{A1}, \lambda_{A2}, \lambda_{B1}, \lambda_{B2}$ are defined in (2.33).

Proof. The full proof is relegated in Appendix A.7. □

Remark 2.3. *As can be seen from* (2.36)*, the power allocation depends on the spatial correlation, the precoder design as well as SNR. Take* **case 1** *as an example, $\lambda_{B2} = \mathbf{w}_2^H \mathbf{R}_B \mathbf{w}_2 = \lambda_{min}(\mathbf{R}_B)$ while $\lambda_{B1} = \lambda_{max}(\mathbf{R}_B)$. Then, $\frac{1+\lambda_{B2}P_8}{1+\lambda_{B1}P_8} \leq 1$ implies that $P_1 \leq P_2$ and likewise $P_3 \leq P_4$. This implies that more power needs to be allocated on the weaker eigen-mode ($\mathbf{w}_2$) to constrain the interference imposed to the other desired symbol of the same user. As mentioned before, the power consumption in time slot 2 and 3 relies on the beamformer design. Such power allocation method enables to compress the interference and makes room for delivering two more private symbols. Moreover, consider*

i.i.d Rayleigh fading channels where $\mathbf{R}_A = \mathbf{R}_B = \mathbf{I}$. $\lambda_{B2} = \mathbf{w}_2^H \mathbf{R}_B \mathbf{w}_2 = 1$ *and likewise* $\lambda_{B1} = \lambda_{A1} = \lambda_{A2} = 1$. *Based on* (2.36) *and symmetry, the optimal power allocation satisfies* $P_1 = P_2 = P_3 = P_4$ *and therefore SAMAT boils down to AMAT. It makes sense because the channels are uncorrelated and there is no room in the spatial domain to suppress the interference. In this case, equal power allocation is the optimal choice.*

To operate the proposed SAMAT transmission protocol, the signalling and feedback procedure is described as follows. Using LTE-A framework [79], channel state information reference signals (CSI-RS) are transmitted to enable the receiver to measure the short-term CSI and the long-term CSI (channel covariance matrix), which are then fed back to the transmitter via a delayed but assumed perfect feedback link. The long-term CSI only varies at a very slow pace and is therefore not affected by the delay. However by the time the transmitter has acquired the short term CSI, the channel has changed and the transmitter only has knowledge of a completely stale short-term CSIT. Based on the long-term and the short-term CSIT, the transmitter computes the precoders and the power allocation and constructs the transmitted signals that are then transmitted using demodulation reference signals (DM-RS) [43].

2.5. SAMAT Discussion

The proposed SAMAT transmission strategy is inspired by [33]. The authors in [33] developed a superimposition scheme of AMAT and ZF-based new privates transmission, which exploits optimally both the delayed CSIT and partial CSIT. This new scheme smoothly bridges between MAT and PZFBF (i.e., multiuser transmission with ZF beamforming based on perfect CSIT) scheme in terms of sum DoF.

In our work, we aim to bridge between AMAT and SBF in terms of ergodic sum rate based on statistical and delayed CSIT. Although our framework is analog to [33], there are essential distinctions between them. A first distinction lies in the channel model. We exploit spatial correlation to compress the multiuser interference and make room for extra symbols transmission while they make use of time correlation. The power allocation in [33] depends on SNR and quality of current CSIT while our power allocation strategy relies on SNR, precoder design as well as spatial correlation.

The secondary distinction lies in the encoding/decoding strategy (and hence the transmission protocol). More specifically, interference quantisation is crucial for [33], where the overheard interference symbol with a reduced power is transmitted with full power in order to save channel resources. Interference quantisation is proposed to solve the consequent problem of power mismatch (which scales with transmit power). By decoding the interference symbols first, [33] equivalently obtains one AMAT transmission plus two ZF transmissions. DoF gain at high SNR can be obtained over the original AMAT scheme.

In contrast with [33], the overheard interference is multicast by analog transmission in our scenario and the reason is threefold. Firstly, we retransmit the interference symbols

after scaling them by constant (i.e., not scaling with the transmit power) factors P_5 and P_8. P_1, P_2, P_3 and P_4 in the main AMAT transmission scale with the transmit power at high SNR to achieve the DoF of $\frac{4}{3}$. However, to guarantee the power constraint, the multiplication terms $P_5(\lambda_{A1}P_3 + \lambda_{A2}P_4)$ and $P_8(\lambda_{B1}P_1 + \lambda_{B2}P_2)$ in eq. (2.20) limit P_5 and P_8 to some constants. Secondly, interference quantisation would prevent the proposed SAMAT strategy from bridging SBF at low SNR. More specifically, in slot 2 and 3, SAMAT should behave as SBF at low SNR and should therefore allocate most of the transmit power to the extra symbols. Only very little power is left to transmit the overheard interference symbol. Following [33], if interference quantisation is applied, the digitised interference should be decoded first by treating the extra symbols as noise. In this case, however, the decoding would fail because the noise power would overwhelm the desired signal power.

Finally, due to the inherent properties of the channel model (full-rank channel covariance matrix), a sum DoF strictly larger than $\frac{4}{3}$ cannot be achieved in our case (contrary to [33]). Hence, the SAMAT transmission and reception strategies are not motivated by DoF maximisation. With SAMAT, a sum DoF of $\frac{4}{3}$ is achieved where the extra private symbols are not used to increase the DoF at high SNR (contrary to [33]) but to significantly boost the sum rate at low/finite SNR. This implies that the retransmitted overheard interference does not have to be decoded first in SAMAT (contrary to [33]) but can simply be aligned and cancelled so as to decode the private and extra symbols. Recall again that [33] relies on interference quantisation to decode first the overheard interference and then the private messages in order to increase the sum DoF beyond $\frac{4}{3}$.

2.6. Performance Evaluation

In this section, we provide numerical results to show the effectiveness of the proposed precoder design and power allocation scheme. A single parameter exponential correlation model [80] is considered as

$$\mathbf{R}_k = \begin{bmatrix} 1 & t_k & \dots & t_k^{M-1} \\ t_k^H & 1 & \dots & t_k^{M-2} \\ \vdots & & \ddots & \\ (t_k^H)^{M-1} & \dots & t_k^H & 1 \end{bmatrix}, \tag{2.37}$$

where t_k denotes the transmit correlation coefficient $t_k = |t_k|\, e^{j\phi_k}$, $\phi_k \in [0,\, 2\pi]$, $k = A, B$. We use high (low) correlation to indicate large (small) condition number of the spatial correlation matrix, which corresponds to large (small) $|t_k|$ in the exponential model. A large family of spatial correlation has been tested to verify our analysis. With the aid of the optimisation tool in Matlab, 'fmincon' is used to implement the SQP algorithm.

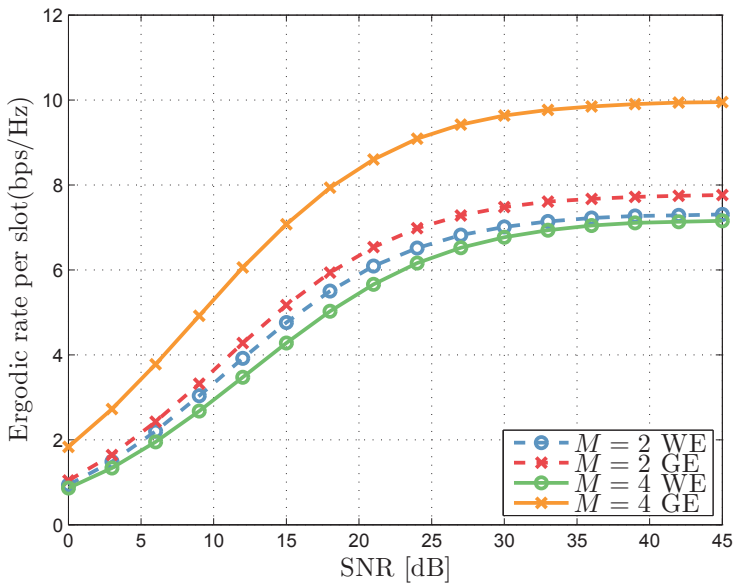

Figure 2.3.: Ergodic sum rate of SBF with WE and with GE precoder in spatially correlated channel.

2.6.1. Precoders Comparison for SBF

In Fig. 2.3, we plot the ergodic sum rate of SBF with WE/GE precoders, averaged over the randomness in the channel realisations and in the correlation coefficient phase ϕ_k. The amplitudes of channel correlation coefficients of both users are given by $|t_A| = 0.95$, $|t_B| = 0.9$ and the superiority of SGEBF over SWEBF is illustrated by two cases ($M = 2, 4$). In Fig. 2.3, GE beamforming vector shows robustness for large M as well as varying scattering environment (i.e., ϕ_k). Interestingly, SWEBF performs even worse for larger M, which is inherently caused by the idea of zero forcing. The WE precoder is designed to reduce the interference to the unintended user, but may cancel out the desired signal of the intended user. In other words, as M increases, the $M \times 1$ WE precoder $\mathbf{w} = \mathbf{u}_{\min}(\mathbf{R}_B)$ may fall into the $(M-1)$ dimensional Null($\mathbf{h}$) with higher probability.

2.6.2. Precoders Comparison for AMAT

In Fig. 2.4, we compare the ergodic sum rate performance of AMAT with different precoding methods. t_k is randomly generated by $|t_k| \in U(0,1), \phi_k \in U(0, 2\pi)$. ORG denotes original AMAT that the transmitter sends symbols simply using 2 out of M antennas. WE and GE are statistical precoders defined in Section 2.2. The optimal precoders (OPT) is computed by the proposed iterative algorithm in Table 2.1. We can observe that OPT achieves a better ergodic sum rate than other baselines.

Meanwhile, Fig. 2.5 confirms the validity of Proposition 2.2. It can be seen that any orthogonal beamforming vectors constituting a unitary matrix are optimal for $M = 2$

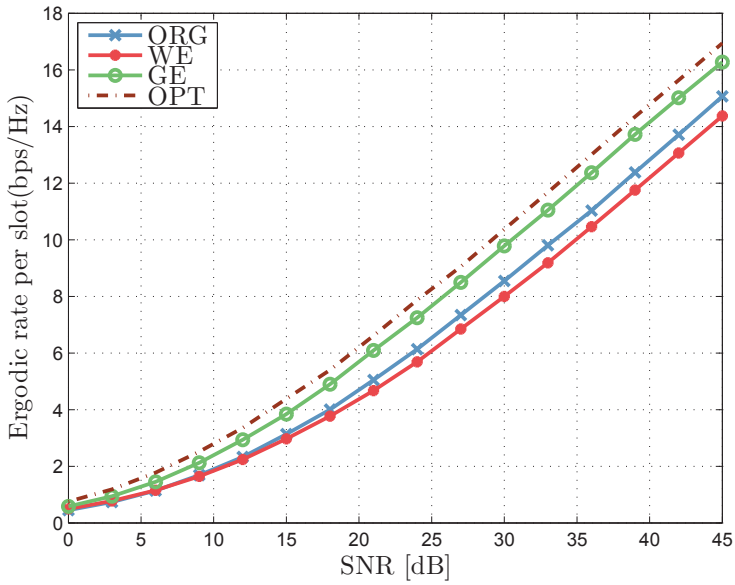

Figure 2.4.: Ergodic sum rate of AMAT with various precoders in a setup with $M = 4$.

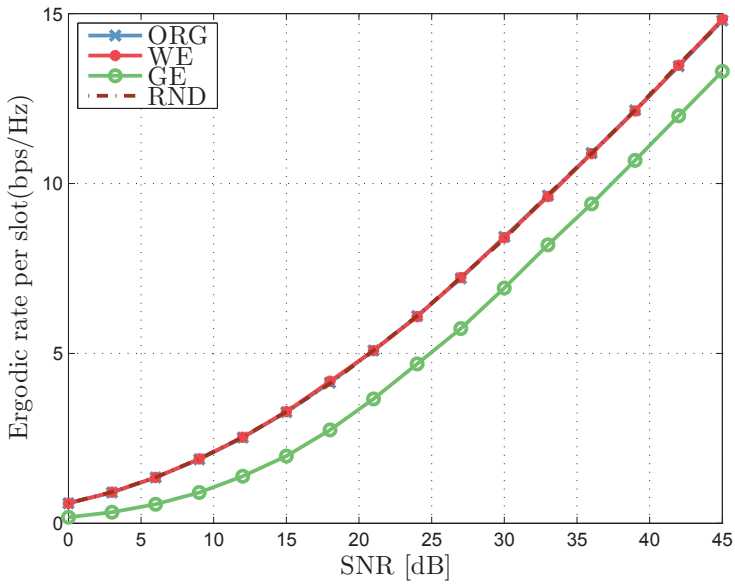

Figure 2.5.: Ergodic sum rate of AMAT with various precoders in a setup with $M = 2$.

case. More specifically, RND indicates that $\mathbf{W}$ and $\mathbf{Q}$ are randomly generated unitary matrices. WE precoders, corresponding to $\mathbf{u}_{\max}(\mathbf{R}_k)$ and $\mathbf{u}_{\min}(\mathbf{R}_k)$ $k = A, B$, also form unitary matrices. ORG becomes an 2×2 identity matrix. All these precoders show optimality in terms of the ergodic sum rate whereas GE does not, because GE precoders fail to form a unitary matrix (due to the fact that neither $\mathbf{R}_A^{-1}\mathbf{R}_B$ nor $\mathbf{R}_B^{-1}\mathbf{R}_A$ is a normal matrix).

2.6.3. Performance of SAMAT

Fig. 2.6 depicts the achievable ergodic sum rates of various schemes with two transmit antennas ($M = 2$): original AMAT, SBF with WE precoders and the proposed SAMAT (***case 1***). We set $|t_A| = |t_B| = |t|$ that varies between 0 and 1, i.e., from uncorrelated to highly correlated channels. Furthermore, ϕ_A, ϕ_B are randomly generated with $|\phi_A - \phi_B| \geq \frac{\pi}{2}$ and SNR = 20 dB. As $|t|$ increases, the sum rate of SBF gradually goes up while a sharp rise occurs at very high correlation level. Because in highly correlated channels, linear beamforming based on statistical information restrains the remaining interference small enough. A special case is the fully correlated channel (i.e., $|t| = 1$), where the overheard interference can be completely eliminated.

Moreover, the rate performance of original AMAT also depends on the transmit correlation of the channel. In $M = 2$ case, ORG precoders for original AMAT become $\mathbf{W} = \mathbf{Q} = \mathbf{I}_{2\times 2}$. Observe in (2.12) and (2.13) that the ergodic rate is a function of $\mathbf{R}_A$ and $\mathbf{R}_B$. Specifically, $\Theta_A = \Theta_B = \mathrm{tr}(\mathbf{R}_A)\mathrm{tr}(\mathbf{R}_B) - \mathrm{tr}(\mathbf{R}_A\mathbf{R}_B)$. With the correlation model in (2.37) and the specific phases ϕ_A, ϕ_B, a positive/negative impact of transmit correlation amplitude $|t|$ can be easily computed: $\Theta_A = \Theta_B = 2\big(1 - |t|^2| \cdot \cos(|\phi_A - \phi_B|)\big)$. As $|t|$ increases, the transmit correlation is beneficial for $|\phi_A - \phi_B| > \frac{\pi}{2}$ while it is detrimental for $|\phi_A - \phi_B| < \frac{\pi}{2}$. When $|\phi_A - \phi_B| = \frac{\pi}{2}$, the ergodic rate keeps constant irrespective of $|t|$.

The cross point between SWEBF and AMAT is determined by the spatial correlation level and SNR. Fig. 2.6 reveals that the proposed SAMAT strategy obtains strictly higher rate than SWEBF and AMAT by exploiting both statistical and delayed CSIT. Based on (2.1), we can regard the long-term channel statistics as a kind of imperfect current CSI. Fig. 2.6 coincides with Fig. 1 in [33] in the sense that the proposed strategies softly bridge between SWEBF (PZFBF in [33]) and AMAT (MAT in [33]) in terms of the ergodic sum rate (DoF).

In addition, for given channel covariance matrices, the ergodic sum rate of these strategies can be plotted versus SNR. It can be observed from Fig. 2.7 that SAMAT achieves higher rate than SWEBF as well as AMAT along the entire SNR region. It acts as SWEBF at low SNR while it utilises the DoF capability of AMAT in the high SNR regime. As a comparison, PZFBF with perfect CSIT enables a sum DoF of 2 at high SNR. Briefly, when $|t_A|, |t_B| \longrightarrow 0$, the SAMAT transmission protocol boils down to AMAT s-

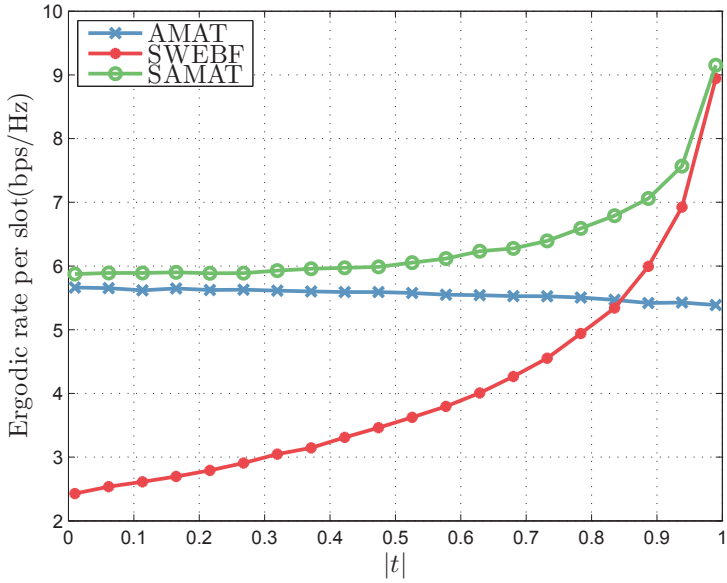

Figure 2.6.: Ergodic sum rate comparison between the proposed SAMAT scheme and various baselines, in a setup with SNR = 20 dB, $|\phi_A - \phi_B| \leq \frac{\pi}{2}, M = 2$.

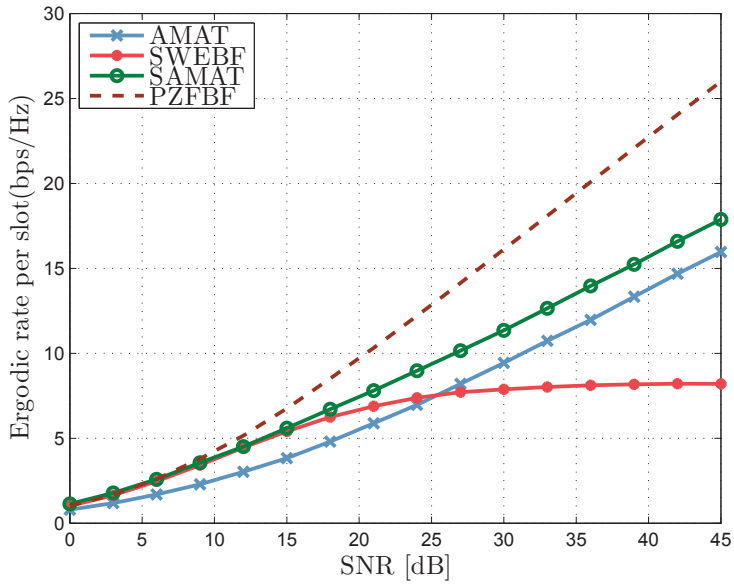

Figure 2.7.: Ergodic sum rate comparison between the proposed SAMAT scheme and various baselines, in a setup with $|\phi_A - \phi_B| \leq \frac{\pi}{2}, |t_A| = 0.95, |t_B| = 0.9, M = 2$.

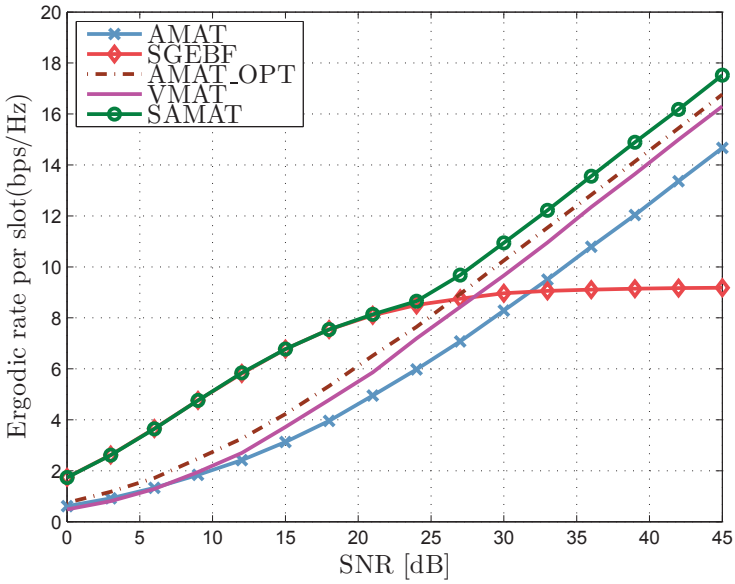

Figure 2.8.: Comparison of the ergodic sum rate vs. SNR between SAMAT and baselines, $|t_A| = 0.95$, $|t_B| = 0.9$, $\phi_A, \phi_B \in U(0, 2\pi)$, $M = 4$.

ince no correlation can be exploited to enhance the rate performance. Consider the other extreme $|t_A|, |t_B| \longrightarrow 1$ but $|\phi_A - \phi_B| \longrightarrow 0$, it indicates highly correlated channels but their weakest eigen-modes lie in the similar direction. The rate performance of SWEBF is unfavourable and therefore SAMAT also behaves as AMAT.

In Fig. 2.7, the power allocation of SAMAT from the SQP algorithm at SNR = 30 dB is given as $P_1 = 525.2, P_2 = 81.6, P_3 = 579.3, P_4 = 59.4, P_5 = 5.9, P_8 = 4.3$, $\lambda_{A1} = 0.05$, $\lambda_{A2} = 1.95, \lambda_{B1} = 0.1, \lambda_{B2} = 1.9$. It can be easily verified that Theorem 2.2 (i.e., Equation (2.36)) holds.

Fig. 2.8 illustrates the benefits of the proposed strategy with the power allocation optimisation. The transmitter antennas $M = 4$ and robust GE precoders are considered. ϕ_A and ϕ_B are randomly generated. Specifically, AMAT indicates the original AMAT with equal power allocation only exploiting delayed CSIT. AMAT_OPT denotes AMAT with the optimal beamforming vectors developed in Section 2.3. SGEBF denotes the statistical beamforming with GE precoders. Moreover, we compare the proposed SAMAT (***case 2***) strategy with VMAT [20]. As mentioned before, the power constraint of VMAT in stage II was released. To make a fair comparison, we also apply the long-term power constraint for VMAT and scale it down to $3P$.

In Fig. 2.8, we observe that AMAT_OPT enables around 5 dB enhancement over original AMAT at high SNR. VMAT achieves almost the same ergodic sum rate as AMAT_OPT, since both schemes exploit statistical CSIT under equal power allocation. However, the SBF scheme still outperforms both of them in a certain range of low to intermediate SNR. The proposed SAMAT framework which is precoded by GE with closed-form

power allocation outperforms all these strategies. Meanwhile, with the optimised power allocation computed by the SQP algorithm, the SAMAT strategy maximises the ergodic sum rate (further 2 dB over AMAT_OPT). The enhancement over VAMT/AMAT_OPT mainly comes from power allocation optimisation.

To sum up, SAMAT boils down to AMAT in low-correlated/uncorrelated channels while for highly correlated scenario where SBF outperforms AMAT, it behaves as SBF in the low to mediate SNR regime and as AMAT at high SNR. In addition, the optimised power values satisfy Theorem 2.2.

2.7. Summary and Conclusion

This chapter aimed to maximise the ergodic sum rate with both statistical and delayed CSIT in a two-user MISO broadcast channel. We considered the robust design of statistical beamforming vectors for arbitrary transmit antennas, showing the optimality of dominant generalised eigenvectors in maximising a lower bound of the ergodic sum rate. Moreover, the optimal precoders were designed to maximise the rate approximation of AMAT under equal power allocation. An iterative algorithm was developed to compute these precoders.

The SBF and AMAT strategies show superiority to each other in different regime of spatial correlation and SNR. The SAMAT transmission protocol was proposed to integrate SBF and AMAT for a wide range of SNR and spatial correlation condition with an arbitrary number of transmit antennas. In low correlated channel, the SAMAT strategy boils down to AMAT because limited spatial correlation can be exploited to enhance the ergodic sum rate. For highly correlated scenario, it employs the advantage of SBF in the low to intermediate SNR region and the DoF capability of AMAT at high SNR. To sum up, the proposed SAMAT strategy yields a significant ergodic sum rate enhancement over both SBF and AMAT. At low SNR, the gain mostly comes from extra symbols transmission. At high SNR, it is achieved by power allocation optimisation and statistical precoding. Numerical results were provided to confirm the analysis and designs.

For multiuser ($K > 2$) case, direct implementation of A(MAT) schemes involves the explosion of time slots. For instance, $K = 3$ MAT scheme requires 11 time slots to complete the whole transmission [19]. As [20] suggests, a user scheduling/pairing can be considered. The proposed SAMAT scheme can serve a pair of users at a time while all pairs are scheduled in a round-robin manner.

3. A Novel Transmission Strategy for Multiuser Massive MIMO System with Partial and Statistical CSIT

We shift the focus to design transmission strategy with partial and statistical CSIT. In FDD systems, downlink channel estimation and uplink feedback of instantaneous CSI is widely used in MIMO networks. In TDD systems, downlink channel can be instantaneously measured by uplink training and channel reciprocity. Rather than delayed CSIT considered in previous chapter, it is also worth an effort to look into the partial CSIT. As introduced in Section 1.7.4, a recently proposed Rate-Splitting (RS) transmission strategy can deal with partial CSIT.

In this chapter, we generalise the idea of RS into the large-scale array regime. The rationale behind is explained as below. It is well known that massive MIMO is a promising candidate for improving the spectrum and energy efficiency for 5G cellular networks. However, for a large number of transmit antennas, the downlink training represents a significant obstacle and the corresponding feedback overhead is typically too large to afford. With limited feedback for massive MIMO system, it is very likely that only partial CSIT is available at the BS. There has been no investigation on the performance of RS in massive MIMO systems. We will fill the gap by analysing its rate gain over the conventional transmission strategy.

Moreover, statistical CSIT only requires low-rate feedback link and can be easily and accurately obtained by the BS, as discussed in Section 1.5.2. It is of great interest to investigate how statistical CSIT helps RS to further enhance the system performance. We indeed incorporate the idea of RS into the conventional transmission strategy with partial and statistical CSIT as discussed in 1.7.5. More specifically, we propose a novel and general framework Hierarchical-Rate-Splitting (HRS) that is particularly suited to massive MIMO system. Deterministic equivalent sum rate expressions are derived and power allocation schemes are developed for both RS and HRS.

3.1. System Model

Consider a single cell downlink system where the BS equipped with M antennas transmits messages to K single-antenna users over a spatially correlated Rayleigh fading channel.

Consider a geometrical one-ring scattering model [5], the correlation between the channel coefficients of antennas $1 \leq i, j \leq M$ is given by

$$[\mathbf{R}_k]_{i,j} = \frac{1}{2\Delta_k} \int_{\theta_k - \Delta_k}^{\theta_k + \Delta_k} e^{-j\frac{2\pi}{\lambda}\Psi(\alpha)(\mathbf{r}_i - \mathbf{r}_j)} d\alpha, \tag{3.1}$$

where θ_k is the azimuth angle of user k with respect to the orientation perpendicular to the array axis. Δ_k indicates the angular spread of departure to user k. $\Psi(\alpha) = [\cos(\alpha), \sin(\alpha)]$ is the wave vector for a planar wave impinging with the angle of α, λ is the wavelength and $\mathbf{r}_i = [x_i, y_i]^T$ is the position vector of the i-th antenna. With the Karhunen-Loeve model, the downlink channel of user k $\mathbf{h}_k \in \mathbb{C}^M$ is expressed as

$$\mathbf{h}_k = \mathbf{U}_k \mathbf{\Lambda}_k^{\frac{1}{2}} \mathbf{g}_k, \tag{3.2}$$

where $\mathbf{\Lambda}_k \in \mathbb{C}^{r_k \times r_k}$ is a diagonal matrix containing the non-zero eigenvalues of the spatial correlation matrix $\mathbf{R}_k$, and $\mathbf{U}_k \in \mathbb{C}^{M \times r_k}$ consists of the associated eigenvectors. The slowly-varying channel statistics $\mathbf{R}_k$ can be accurately obtained via a rate-limited feedback link or via uplink-downlink reciprocity and is assumed perfectly known to both BS and users. We also assume block fading channel and $\mathbf{g}_k \in \mathbb{C}^{r_k}$ has i.i.d. $\mathcal{CN}(0,1)$ entries. For each channel use, linear precoding is employed at the BS to support simultaneous downlink transmissions to K users. The received signals can be expressed as

$$\mathbf{y} = \mathbf{H}^H \mathbf{x} + \mathbf{n}, \tag{3.3}$$

where $\mathbf{x} \in \mathbb{C}^M$ is the linearly precoded signal vector subject to the transmit power constraint $\mathbb{E}[||\mathbf{x}||^2] \leq P$, $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K]$ is the downlink channel matrix, $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ is the AWGN vector and $\mathbf{y} \in \mathbb{C}^K$ is the received signal vector of K users.

Due to limited feedback (e.g., quantised feedback with a fixed number of quantisation bits) or imperfect channel estimation, only an imperfect channel estimate $\widehat{\mathbf{h}}_k$ is available at the BS and modelled as [81]

$$\widehat{\mathbf{h}}_k = \mathbf{U}_k \mathbf{\Lambda}_k^{\frac{1}{2}} \widehat{\mathbf{g}}_k = \mathbf{U}_k \mathbf{\Lambda}_k^{\frac{1}{2}} \left(\sqrt{1 - \tau_k^2}\, \mathbf{g}_k + \tau_k \mathbf{z}_k \right), \tag{3.4}$$

where $\mathbf{z}_k$ has i.i.d. $\mathcal{CN}(0,1)$ entries independent of $\mathbf{g}_k$. $\tau_k \in [0,1]$ indicates the quality of instantaneous CSIT for user k, i.e., $\tau_k = 0$ implies perfect CSIT whereas for $\tau_k = 1$ the CSIT estimate is completely uncorrelated with the true channel.

3.2. Rate-Splitting

In this section, we introduce the rate-splitting (RS) transmission strategy and elaborate on the precoder design, asymptotic rate performance as well as power allocation. The sum rate gain of RS over conventional strategy is quantified.

3.2.1. RS Transmission

Let us firstly consider conventional multiuser strategy (referred to as No-RS) with one-tier precoder. The transmitted signal and the corresponding received signal of user k can be written as

$$\begin{aligned} \mathbf{x} &= \mathbf{W}\mathbf{s} = \sum_{k=1}^{K} \sqrt{P_k}\mathbf{w}_k\, s_k, \\ y_k &= \sqrt{P_k}\,\mathbf{h}_k^H \mathbf{w}_k\, s_k + \underbrace{\sum_{j\neq k}^{K} \sqrt{P_j}\,\mathbf{h}_k^H \mathbf{w}_j\, s_j}_{\text{multiuser interference}} + n_k, \end{aligned} \tag{3.5}$$

where $\mathbf{s} = [s_1, \cdots, s_K]^T \in \mathbb{C}^K$ is the data vector intended for the K users. Based on the imperfect channel estimate $\widehat{\mathbf{H}} = [\widehat{\mathbf{h}}_1, \cdots, \widehat{\mathbf{h}}_K]$, the $M \times K$ precoder $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K]$ is designed as ZF or Regularised-ZF (RZF). In the presence of partial CSIT with fixed error variance, the sum rate of No-RS with uniform power allocation is multiuser interference-limited at high SNR. In order to tackle the interference, one can adaptively schedule a smaller number of users to transmit as the SNR increases, resulting in TDMA at extremely high SNR. However, such an adaptive scheduling is computationally heavy for a large number of users.

To tackle this issue, we introduce the RS transmission strategy that is well described in Section 1.7.4. In the RS strategy, the transmitted signal can be written as

$$\mathbf{x} = \sqrt{P_c}\,\mathbf{w}_c\, s_c + \textstyle\sum_{k=1}^{K} \sqrt{P_k}\,\mathbf{w}_k\, s_k, \tag{3.6}$$

where $\mathbf{w}_c$ is the unit-norm precoding vector of the common message. We here perform a uniform[1] power allocation for the private messages. The interest is on how to balance the power allocation between the common and private messages. Hence, the powers allocated to each message are given by $P_c = P(1-t)$ and $P_k = Pt/K$, where $t \in (0, 1]$ denotes the fraction of the total power that is allocated to the private messages. The decoding procedure is performed as follows. Each user decodes first the common message s_c by treating all private messages as noise. After eliminating the decoded common message by SIC, each user decodes its own private messages. By plugging (3.6) into (3.3), the SINRs of the common message and the private message experienced by user k are written as

$$\mathrm{SINR}_k^c = \frac{P_c\,|\mathbf{h}_k^H \mathbf{w}_c|^2}{\sum_{j=1}^{K} P_j\,|\mathbf{h}_k^H \mathbf{w}_j|^2 + 1}, \quad \mathrm{SINR}_k^p = \frac{P_k\,|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j\neq k} P_j\,|\mathbf{h}_k^H \mathbf{w}_j|^2 + 1}. \tag{3.7}$$

The achievable rate of the common message is given as $R_c^{RS} = \log_2(1 + \mathrm{SINR}^c)$, where

[1] A per-user power allocation optimisation can further enhance the rate performance of RS, as has been done in [39]. We here consider a uniform power allocation to the private messages for both RS and conventional No-RS.

$\mathrm{SINR}^c = \min_k \{\mathrm{SINR}_k^c\}$ ensuring that the common message can be successfully decoded by all users. The sum rate of the private messages is given as $R_p^{RS} = \sum_{k=1}^K R_k^{RS} = \sum_{k=1}^K \log_2(1+\mathrm{SINR}_k^p)$. Then, the sum rate of RS is $R_{\mathrm{sum}}^{RS} = R_c^{RS} + R_p^{RS}$.

3.2.2. Precoder Design

The precoder ($\mathbf{w}_k$) of the private message for multiuser MISO broadcast channel has been investigated in [82] assuming perfect CSIT. The structure of the optimal $\mathbf{w}_k$ is a generalisation of RZF precoding. In the presence of partial CSIT, the optimal precoders of the private messages are still unknown in simple closed-form expression and have to be optimised numerically following e.g. [39]. The optimisation is particularly complex in large scale antenna array systems. Nevertheless, building upon [7, 37, 44], RZF based on the channel estimates $\widehat{\mathbf{H}}$ would be a suitable strategy for the precoders of the private messages. Hence,

$$\mathbf{W} = \xi\, \widehat{\mathbf{M}}\widehat{\mathbf{H}}, \tag{3.8}$$

where $\widehat{\mathbf{M}} = (\widehat{\mathbf{H}}\widehat{\mathbf{H}}^H + M\,\varepsilon\,\mathbf{I}_M)^{-1}$ and $\varepsilon = K/(MP)$. To satisfy the transmit power constraint, the normalisation scalar is set as $\xi^2 = K/\mathrm{tr}(\widehat{\mathbf{H}}^H\widehat{\mathbf{M}}^H\widehat{\mathbf{M}}\widehat{\mathbf{H}})$.

The precoder $\mathbf{w}_c$ is designed to maximise the achievable rate of the common message, i.e., $\log_2(1+\mathrm{SINR}^c)$ with $\mathrm{SINR}^c = \min_k \{\mathrm{SINR}_k^c\}$. From (1.11), different channel estimates become asymptotically orthogonal in the large-scale array regime [66]. Thus, we are able to design the precoder $\mathbf{w}_c$ in the subspace $\mathcal{S} = \mathrm{Span}(\widehat{\mathbf{H}})$, i.e.,

$$\mathbf{w}_c = \sum_k a_k \widehat{\mathbf{h}}_k, \tag{3.9}$$

which can be interpreted as a weighted matched beamforming (MBF). The corresponding optimisation problem is formulated as

$$\mathcal{P}1: \max_{\mathbf{w}_c \in \mathcal{S},\, \|\mathbf{w}_c\|^2=1} \ \min_k \ \pi_k \cdot |\mathbf{h}_k^H \mathbf{w}_c|^2, \tag{3.10}$$

where $\pi_k = \frac{P_c}{\sum_{k=1}^K P_k\, |\mathbf{h}_k^H \mathbf{w}_k|^2 + 1}$ and the optimal solution $\{a_k^\star\}$ is shown as below.

Proposition 3.1. *As $M \to \infty$, the asymptotically optimal solution of problem $\mathcal{P}1$ is given by* (3.9) *with*

$$a_k^\star = \frac{1}{\sqrt{M \cdot \sum_{j=1}^K \frac{\pi_k\,(1-\tau_k^2)}{\pi_j\,(1-\tau_j^2)}}}, \ \forall k. \tag{3.11}$$

Proof. See Appendix B.1. □

Remark 3.1. *The precoder $\mathbf{w}_c^\star$ is optimised such that all the K users experience the same SINR* (3.7) *w.r.t. the common message. Specially, the equally weighted MBF with $a_k^\star = 1/\sqrt{MK}$ is optimal when the condition $\pi_k\,(1-\tau_k^2) = \pi_j\,(1-\tau_j^2)$, $\forall k \neq j$ is satisfied.*

Nevertheless, we employ $a_k^\star = 1/\sqrt{MK}$ rather than (3.11) for arbitrary cases, in order to obtain a more insightful and tractable asymptotic sum rate expression in the sequel.

3.2.3. Asymptotic Rate Analysis

We shall omit the proof of the following asymptotic SINRs of RS, which are straightforwardly established based on the approach of [7].

Theorem 3.1. *As $M, K \to \infty$ with a fixed ratio $\eta = \frac{M}{K}$, the SINRs (3.7) of RS asymptotically converge as*

$$SINR_k^c - SINR_k^{c,\circ} \overset{M\to\infty}{\longrightarrow} 0, \quad SINR_k^p - SINR_k^{p,\circ} \overset{M\to\infty}{\longrightarrow} 0, \tag{3.12}$$

almost surely, where

$$SINR_k^{c,\circ} = \frac{P(1-t)(1-\tau_k^2)\eta}{\frac{Pt}{K}\left(\xi^\circ\right)^2\left(\Upsilon_k^\circ\Omega_k + \Phi_k\right) + 1}, \quad SINR_k^{p,\circ} = \frac{\frac{Pt}{K}\left(\xi^\circ\right)^2\Phi_k}{\frac{Pt}{K}\left(\xi^\circ\right)^2\Upsilon_k^\circ\Omega_k + 1}, \tag{3.13}$$

and

$$(\xi^\circ)^2 = \frac{K}{\Psi^\circ}, \quad \Psi^\circ = \frac{1}{M}\sum_{j=1}^{K}\frac{m_j'}{(1+m_j^\circ)^2}, \quad \Phi_k = \frac{(1-\tau_k^2)(m_k^\circ)^2}{(1+m_k^\circ)^2} \tag{3.14}$$

$$\Upsilon_k^\circ = \frac{1}{M}\sum_{j\neq k}\frac{m_{j,k}'}{(1+m_j^\circ)^2}, \quad \Omega_k = \frac{1-\tau_k^2(1-(1+m_k^\circ)^2)}{(1+m_k^\circ)^2} \tag{3.15}$$

with $\mathbf{m}' = [m_1', \cdots, m_K']^T$ and $\mathbf{m}_k' = [m_{1,k}', \cdots, m_{K,k}']^T$ defined by

$$\mathbf{m}' = (\mathbf{I}_K - \mathbf{J})^{-1}\mathbf{v}, \qquad \mathbf{m}_k' = (\mathbf{I}_K - \mathbf{J})^{-1}\mathbf{v}_k, \tag{3.16}$$

where $\mathbf{J}, \mathbf{v}$ and $\mathbf{v}_k$ are given as

$$[\mathbf{J}]_{i,j} = \frac{\frac{1}{M}tr(\mathbf{R}_i\mathbf{T}\mathbf{R}_j\mathbf{T})}{M(1+m_j^\circ)^2}, \tag{3.17}$$

$$\mathbf{v}_k = \left[\frac{1}{M}tr(\mathbf{R}_1\mathbf{T}\mathbf{R}_k\mathbf{T}), \cdots, \frac{1}{M}tr(\mathbf{R}_K\mathbf{T}\mathbf{R}_k\mathbf{T})\right]^T, \tag{3.18}$$

$$\mathbf{v} = \left[\frac{1}{M}tr(\mathbf{R}_1\mathbf{T}^2), \cdots, \frac{1}{M}tr(\mathbf{R}_K\mathbf{T}^2)\right]^T, \tag{3.19}$$

and with m_k° and $\mathbf{T}_g$ the unique solutions of

$$m_k^\circ = \frac{1}{M}tr(\mathbf{R}_k\mathbf{T}), \quad \mathbf{T} = \left(\frac{1}{M}\sum_{j=1}^{K}\frac{\mathbf{R}_j}{1+m_j^\circ} + \varepsilon\,\mathbf{I}_M\right)^{-1}. \tag{3.20}$$

By applying the continuous mapping theorem [83], it follows from (3.12) that $R_k^{RS} - R_k^{RS,\circ} \overset{M\to\infty}{\longrightarrow} 0$, where $R_k^{RS,\circ} = \log_2(1 + \mathrm{SINR}_k^{p,\circ})$ and that $R_c^{RS} - R_c^{RS,\circ} \overset{M\to\infty}{\longrightarrow} 0$, where

$R_c^{RS,\circ} = \log_2(1 + \mathrm{SINR}^{c,\circ})$ with $\mathrm{SINR}^{c,\circ} = \min_k \{\mathrm{SINR}_k^{c,\circ}\}$. The asymptotic sum rate of the private messages is $R_p^{RS,\circ} = \sum_{k=1}^K \log_2(1 + \mathrm{SINR}_k^{p,\circ})$ and it follows that $\frac{1}{K}(R_p^{RS} - R_p^{RS,\circ}) \overset{M\to\infty}{\longrightarrow} 0$. Then, an approximation $R_{\mathrm{sum}}^{RS,\circ}$ of the sum rate of RS is obtained as

$$R_{\mathrm{sum}}^{RS,\circ} = R_c^{RS,\circ} + R_p^{RS,\circ}. \tag{3.21}$$

According to random matrix theory and following [7,44,45], the asymptotic SINR/rate approximations become more accurate for increasing number of transmit antennas. The asymptotic approximations are also tight for large but finite M, e.g., 64 antennas implemented in the typical prototype of massive MIMO [84]. Moreover, simulation results suggest that the asymptotes remain effective even for small system dimension, e.g., $M = 16$ [7].

3.2.4. Power Allocation

The optimal power splitting ratio t that maximises (3.21) can be obtained by line search. Nevertheless, we compute a suboptimal but effective and insightful power allocation by which RS considerably outperforms conventional No-RS strategies. Denote the asymptotic sum rate of the No-RS strategy in (3.5) with RZF and uniform power allocation by $R_{\mathrm{sum}}^{RZF,\circ} = \sum_{k=1}^K \log_2(1 + \mathrm{SINR}_k^{RZF,\circ})$ with $\mathrm{SINR}_k^{RZF,\circ} = \mathrm{SINR}_k^{p,\circ}|_{t=1}$. We can write

$$\mathrm{SINR}_k^{p,\circ} = \frac{\frac{Pt}{K}\left(\xi^\circ\right)^2 \Phi_k}{\frac{Pt}{K}\left(\xi^\circ\right)^2 \Upsilon_k^\circ \Omega_k + 1} \leq \mathrm{SINR}_k^{RZF,\circ} = \frac{\frac{P}{K}\left(\xi^\circ\right)^2 \Phi_k}{\frac{P}{K}\left(\xi^\circ\right)^2 \Upsilon_k^\circ \Omega_k + 1}, \tag{3.22}$$

for $\forall t \in (0,\, 1]$. The basic idea is to allocate a fraction (t) of the total power to transmit the private messages of RS and achieve approximately the same sum rate as conventional No-RS with full power. Then, using the remaining power to transmit the common message of RS enhances the sum rate. The sum rate gain of RS over No-RS with RZF is quantified by

$$\Delta R^{RS,\circ} = R_c^{RS,\circ} + \sum_k \big(\log_2(1 + \mathrm{SINR}_k^{p,\circ}) - \log_2(1 + \mathrm{SINR}_k^{RZF,\circ})\big). \tag{3.23}$$

Proposition 3.2. *The equality of* (3.22) *holds when the power splitting ratio t is given by*

$$t = min\left\{\frac{K}{P\Gamma},\, 1\right\}, \tag{3.24}$$

where $\Gamma = \min_k \left\{\Upsilon_k^\circ \tau_k^2/\Psi^\circ\right\}$ and the sum rate gain $\Delta R^{RS,\circ}$ at high SNR is lower bounded as

$$\Delta R^{RS,\circ} \geq \log_2(1 + SINR^{c,\circ}) - \log_2(e). \tag{3.25}$$

Proof. See Appendix B.2. □

Remark 3.2. *We can get useful insights into the effects of system parameters on the power allocation and sum rate gain. For example, t decreases as τ^2 increases, i.e., the*

power allocated to the private messages is reduced as the channel quality becomes worse ($\tau^2 \to 1$). *Moreover, the power allocated to the private messages* Pt *is fixed at high SNR, in order to place the sum rate of the private messages back to the non-interference-limited regime. By assigning the remaining power* $P - Pt$ *to a common message, the sum rate increases with the available transmit power. As* P *increases,* $t \to 0$ *but* $t \neq 0$, *i.e., RS will never boil down to common message transmission and always exploits the benefits of private messages at low to medium SNR. In addition, from* (3.25), *the rate loss* $\left(R_{sum}^{RZF,\circ} - R_p^{RS,\circ}\right)$ *based on power allocation* (3.24) *is upper bounded by* $\log_2(e) \approx 1.44$ *bps/Hz. Last but not least, as* K *becomes larger,* $t(\leq 1)$ *increases which deteriorates* $SINR_k^{c,\circ} = \frac{M}{K}P(1-t)c$ *for some constant* $c > 0$ *and further the sum rate gain* (3.25).

3.3. Hierarchical-Rate-Splitting

As discussed before, a large number of users (K) degrades the rate benefits of RS. Moreover, the BS requires channel estimate $\widehat{\mathbf{H}}$ with a large dimension to perform conventional No-RS strategies (e.g., ZF/RZF) or RS. In addition, the RS framework can be applied to spatially uncorrelated or correlated channels. Yet in its current form, RS does not explicitly make use of the channel second-order statistics, if known to the transmitter. Motivated by these considerations, we propose a novel and general framework, denoted as Hierarchical-Rate-Splitting, that exploits the knowledge of spatial correlation matrices and two kinds of common messages to enhance the sum rate and alleviate the CSIT requirement as well as the effect of large K.

3.3.1. HRS Transmission

Recently, a multiuser transmission strategy with a two-tier precoder for FDD massive MIMO systems have been proposed to lessen CSIT requirement by exploiting the knowledge of spatial correlation matrix at the transmitter [44–47]. Since the human activity is usually confined in a small region, locations of users tend to be spatially clustered. We make the same assumption as [44] that K users are partitioned into G groups (e.g., via K-mean clustering) and that users in each group share the same spatial correlation matrix $\mathbf{R}_g = \mathbf{U}_g\mathbf{\Lambda}_g\mathbf{U}_g^H$ with rank r_g. We let K_g denote the number of users in group g such that $\sum_{g=1}^{G} K_g = K$. The downlink channel of the g-th group is expressed as $\mathbf{H}_g = [\mathbf{h}_{g1}, \cdots, \mathbf{h}_{gK_g}] = \mathbf{U}_g\mathbf{\Lambda}_g^{1/2}\mathbf{G}_g$, where the elements of $\mathbf{G}_g$ are distributed with $\mathcal{CN}(0,1)$. Then, the transmitted signal of conventional two-tier precoded (TTP) No-RS scheme is expressed as

$$\mathbf{x} = \sum_{g=1}^{G} \mathbf{B}_g\mathbf{W}_g\mathbf{P}_g\,\mathbf{s}_g, \tag{3.26}$$

where $\mathbf{s}_g \in \mathbb{C}^{K_g}$ represents the data streams for the g-th group users. The outer precoder $\mathbf{B}_g \in \mathbb{C}^{M\times b_g}$ is based on long-term CSIT while the inner precoder $\mathbf{W}_g \in \mathbb{C}^{b_g\times K_g}$ depends

on short-term effective channel $\bar{\mathbf{H}}_g = \mathbf{B}_g^H \mathbf{H}_g$. $\mathbf{P}_g \in \mathbb{C}^{K_g \times K_g}$ is the diagonal power allocation matrix with $\mathbf{P}_g = \sqrt{P/K} \cdot \mathbf{I}$. The received signal of the k-th user in g-th group is given by

$$y_{gk} = \sqrt{P_{gk}} \mathbf{h}_{gk}^H \mathbf{B}_g \mathbf{w}_{gk} s_{gk} + \underbrace{\sum_{j \neq k}^{K_g} \sqrt{P_{gj}} \mathbf{h}_{gk}^H \mathbf{B}_g \mathbf{w}_{gj} s_{gj}}_{\text{intra-group interference}} + \underbrace{\sum_{l \neq g}^{G} \mathbf{h}_{gk}^H \mathbf{B}_l \mathbf{W}_l \mathbf{P}_l \, \mathbf{s}_l}_{\text{inter-group interference}} + n_{gk}, \quad (3.27)$$

where $\mathbf{w}_{gk} = [\mathbf{W}_g]_k$. To eliminate inter-group interference, the outer precoder is designed in the nullspace of the eigen-subspace spanned by the dominant eigenvectors of the other groups' spatial correlation matrices. However, the power attached to the weak eigenmodes may leak out to other groups and incur inter-group interference. Furthermore, the intra-group interference cannot be completely removed due to partial CSIT (e.g., limited feedback). To eliminate the interference-limited behaviour at high SNR, one can optimise the groups and the users in each group as a function of the total transmit power and CSIT quality. In general, such an optimisation problem is quite complex.

By generalising the philosophy of RS in Section 3.2, we propose a novel and general Hierarchical RS transmission protocol that consists of an outer RS and an inner RS. By treating each group as a single user, an outer RS tackles the inter-group interference by packing part of one user's message into a common codeword that can be decoded by all users. Likewise, an inner RS copes with the intra-group interference by packing part of one user's message into a common codeword that can be decoded by multiple users in that group. The common messages are superimposed over the private messages and the transmitted signal of HRS can be written as

$$\mathbf{x} = \sqrt{P_{oc}} \mathbf{w}_{oc} s_{oc} + \sum_{g=1}^{G} \mathbf{B}_g \left(\sqrt{P_{ic,g}} \mathbf{w}_{ic,g} s_{ic,g} + \sqrt{P_{gk}} \mathbf{W}_g \mathbf{s}_g \right), \quad (3.28)$$

where $s_{ic,g}$ denotes the inner common message for g-th group while s_{oc} denotes the outer common message for all users. $\mathbf{w}_{ic,g}$ and $\mathbf{w}_{oc}$ are the corresponding unit norm precoding vectors. Similarly to RS, a uniform power allocation is performed on the private messages and we mainly focus on how to balance power allocation between the common and private messages. Hence, let $\beta \in (0, 1]$ represent the fraction of the total power that is allocated to the group (inner common and private) messages. Within each group, $\alpha \in (0, 1]$ denotes the fraction of power given to the private messages. Then, the power allocated to each message is jointly determined by α and β, i.e., $P_{oc} = P(1-\beta), P_{ic,g} = \frac{P\beta}{G}(1-\alpha), P_{gk} = \frac{P\beta}{K}\alpha$.

The decoding procedure is performed as follows. Each user sequentially decodes s_{oc} and $s_{ic,g}$, then remove them from the received signal by SIC. The private message intended to each user can be independently decoded by treating all other private messages as noise. By plugging (3.28) into (3.3), the SINRs of the common messages and the private message

experienced at user k are written as

$$\mathrm{SINR}_{gk}^{oc} = \frac{P_{oc}\,|\mathbf{h}_{gk}^H\mathbf{w}_{oc}|^2}{IN_{gk}}, \quad \mathrm{SINR}_{gk}^{ic} = \frac{P_{ic,g}\,|\mathbf{h}_{gk}^H\mathbf{B}_g\mathbf{w}_{ic,g}|^2}{IN_{gk} - P_{ic,g}\,|\mathbf{h}_{gk}^H\mathbf{B}_g\mathbf{w}_{ic,g}|^2} \tag{3.29}$$

$$\mathrm{SINR}_{gk}^{p} = \frac{P_{gk}|\mathbf{h}_{gk}^H\mathbf{B}_g\mathbf{w}_{gk}|^2}{IN_{gk} - P_{ic,g}|\mathbf{h}_{gk}^H\mathbf{B}_g\mathbf{w}_{ic,g}|^2 - P_{gk}|\mathbf{h}_{gk}^H\mathbf{B}_g\mathbf{w}_{gk}|^2} \tag{3.30}$$

where

$$IN_{gk} = \sum_l P_{ic,l}|\mathbf{h}_{gk}^H\mathbf{B}_l\mathbf{w}_{ic,l}|^2 + \sum_l\sum_j P_{lj}|\mathbf{h}_{gk}^H\mathbf{B}_l\mathbf{w}_{lj}|^2 + 1. \tag{3.31}$$

The achievable rate of the outer common message is given by $R_{oc}^{HRS} = \log_2(1+\mathrm{SINR}^{oc})$ with $\mathrm{SINR}^{oc} = \min\limits_{g,k}\{\mathrm{SINR}_{gk}^{oc}\}$. The sum rate of the inner common messages is given by $R_{ic}^{HRS} = \sum_{g=1}^{G} R_{ic,g}^{HRS} = \sum_{g=1}^{G}\log_2(1+\mathrm{SINR}_g^{ic})$ with $\mathrm{SINR}_g^{ic} = \min\limits_{k}\{\mathrm{SINR}_{gk}^{ic}\}$. The sum rate of the private messages is given as $R_p^{HRS} = \sum_{g=1}^{G}\sum_{k=1}^{K_g} R_{gk}^{HRS} = \sum_{g=1}^{G}\sum_{k=1}^{K_g}\log_2(1+\mathrm{SINR}_{gk}^{p})$. Then, the sum rate of HRS is written as $R_{\mathrm{sum}}^{HRS} = R_{oc}^{HRS} + R_{ic}^{HRS} + R_p^{HRS}$.

3.3.2. Precoder Design

In contrast to RS, HRS has only access to the channel covariance matrices and the effective channel estimates $\widehat{\widetilde{\mathbf{H}}}_g = \mathbf{B}_g^H\widehat{\mathbf{H}}_g$ of dimension $b_g \times K_g$, where $\widehat{\mathbf{H}}_g = \mathbf{U}_g\mathbf{\Lambda}_g^{\frac{1}{2}}\,\widehat{\mathbf{G}}_g = \mathbf{U}_g\mathbf{\Lambda}_g^{1/2}(\sqrt{1-\tau_g^2}\,\mathbf{G}_g + \tau_g\mathbf{Z}_g)$ has dimension of $M\times K_g$. Based on long-term CSIT, the outer precoder $\mathbf{B}_g$ is designed to eliminate the leakage to other groups. Denoting the number of dominant (most significant) eigenvalues of $\mathbf{R}_g$ by r_g^d and collecting the associated eigenvectors as $\mathbf{U}_g^d \in \mathbb{C}^{M\times r_g^d}$, we define $\mathbf{U}_{-g} = [\mathbf{U}_1^d, \cdots, \mathbf{U}_{g-1}^d, \mathbf{U}_{g+1}^d, \cdots, \mathbf{U}_G^d] \in \mathbb{C}^{M\times\sum_{l\neq g} r_l^d}$. According to the singular value decomposition (SVD), we denote by $\mathbf{E}_{-g}^{(0)}$ the left eigenvectors of $\mathbf{U}_{-g}$ corresponding to the $(M - \sum_{l\neq g} r_l^d)$ vanishing singular values. To reduce the inter-group interference while enhancing the desired signal power, $\mathbf{B}_g$ is designed by concatenating $\mathbf{E}_{-g}^{(0)}$ with the dominant eigenmodes of the covariance matrix of the projected channel $\widetilde{\mathbf{H}}_g = (\mathbf{E}_{-g}^{(0)})^H\mathbf{H}_g$. The covariance matrix is decomposed as $\widetilde{\mathbf{R}}_g = (\mathbf{E}_{-g}^{(0)})^H\mathbf{U}_g\mathbf{\Lambda}_g\mathbf{U}_g^H\mathbf{E}_{-g}^{(0)} = \mathbf{F}_g\widetilde{\mathbf{\Lambda}}_g\mathbf{F}_g^H$, where $\mathbf{F}_g$ includes the eigenvectors of $\widetilde{\mathbf{R}}_g$. Denote $\mathbf{F}_g^{(1)}$ as the dominant b_g eigenmodes and then $\mathbf{B}_g$ is given by

$$\mathbf{B}_g = \mathbf{E}_{-g}^{(0)}\mathbf{F}_g^{(1)}. \tag{3.32}$$

The outer precoder $\mathbf{B}_g$ can be interpreted as being the b_g dominant eigenmodes that are orthogonal to the subspace spanned by the dominant eigen-space of groups $l \neq g$. b_g determines the dimension of the effective channel and should satisfy $K_g \leq b_g \leq M - \sum_{l\neq g} r_l^d$ and $b_g \leq r_g^d$. $r_g^d(\leq r_g)$ is a design parameter with a sum rank constraint $\sum_{g=1}^{G} r_g^d \leq M$. The inner precoder $\mathbf{W}_g$ can be designed as RZF, i.e.,

$$\mathbf{W}_g = \xi_g\,\widehat{\overline{\mathbf{M}}}_g\widehat{\overline{\mathbf{H}}}_g, \tag{3.33}$$

where $\widehat{\bar{\mathbf{M}}}_g = (\hat{\bar{\mathbf{H}}}_g \hat{\bar{\mathbf{H}}}_g^H + b_g\, \varepsilon\, \mathbf{I}_{b_g})^{-1}$. By following [7,44,45], the regularisation parameter is set as $\varepsilon = K/bP$ which is equivalent to the MMSE linear filter. b is given by $b = \sum_{g=1}^{G} b_g$. Then, the power normalisation factor is $\xi_g^2 = K_g/\mathrm{tr}(\hat{\bar{\mathbf{H}}}_g^H \widehat{\bar{\mathbf{M}}}_g^H \mathbf{B}_g^H \mathbf{B}_g \widehat{\bar{\mathbf{M}}}_g \hat{\bar{\mathbf{H}}}_g)$.

The precoder $\mathbf{w}_{oc} \in \mathbb{C}^M$ aims to maximise the achievable rate of the outer common message $\log_2(1+\mathrm{SINR}^{oc})$ based on the reduced-dimensional channel estimate $\hat{\bar{\mathbf{H}}}_g \in \mathbb{C}^{b_g \times K_g}$, $\forall g$. However, there exists a dimension mismatch between $\mathbf{w}_{oc}$ and $\hat{\bar{\mathbf{H}}}_g$. To address this problem, we first construct $\check{\mathbf{H}}_g = \mathbf{B}_g \hat{\bar{\mathbf{H}}}_g \in \mathbb{C}^{M \times K_g}$ and $\check{\mathbf{H}} = [\check{\mathbf{H}}_1, \cdots, \check{\mathbf{H}}_G] \in \mathbb{C}^{M \times K}$. From (1.11), the columns of $\check{\mathbf{H}}$ become orthogonal as $M \to \infty$ and we are able to design the precoder $\mathbf{w}_{oc}$ in the subspace $\mathcal{S} = \mathrm{Span}(\check{\mathbf{H}})$. Following Proposition 3.1 and Remark 3.1 in RS, we design the precoder $\mathbf{w}_{oc}$ as an equally weighted MBF, i.e., $\mathbf{w}_{oc} = \xi_{oc} \sum_k \frac{\check{\mathbf{h}}_k}{\sqrt{M}}$, where $\check{\mathbf{h}}_k = [\check{\mathbf{H}}]_k$ and ξ_{oc} normalises $\mathbf{w}_{oc}$ to unit norm.

On the other hand, transmit precoding optimisation in multiuser setup is generally a NP-hard problem. Thus, the optimal precoder of the inner common message $\mathbf{w}_{ic,g}$ that maximises R_{ic}^{HRS} cannot be obtained efficiently. However, when the outer precoder fully eliminates the inter-group interference, $\mathbf{w}_{ic,g}$ can be equivalently designed to maximise $R_{ic,g}^{HRS}$ within each group. Following Proposition 3.1 and Remark 3.1 in RS, we here design $\mathbf{w}_{ic,g}$ as an equally weighted MBF of the effective channel $\hat{\bar{\mathbf{H}}}_g$. Under further assumption that $K \to \infty$, we note that $\widehat{\bar{\mathbf{M}}}_g$ of the inner precoder $\mathbf{W}_g (= \xi_g \widehat{\bar{\mathbf{M}}}_g \hat{\bar{\mathbf{H}}}_g)$ can be approximated by an identity matrix. Hence, $\mathbf{w}_{ic,g}$ can be equivalently designed as an equally weighted MBF of $\mathbf{W}_g$, i.e.,$\mathbf{w}_{ic,g} = \zeta_{ic,g} \hat{\bar{\mathbf{q}}}_g$, where $\hat{\bar{\mathbf{q}}}_g = \frac{1}{K_g} \sum_{k=1}^{K_g} \mathbf{w}_{gk}$ and $\zeta_{ic,g}^2 = 1/(\hat{\bar{\mathbf{q}}}_g^H \mathbf{B}_g^H \mathbf{B}_g \hat{\bar{\mathbf{q}}}_g)$.

3.3.3. Asymptotic Rate Analysis

We shall omit the proof of the asymptotic SINRs of HRS, which is directly established based on the approach of [7].

Theorem 3.2. *As $M, K, b \to \infty$ with fixed ratios $\frac{K}{M}$ and $\frac{b}{M}$, the SINRs of HRS in* (3.29) *and* (3.30) *asymptotically converge as*

$$SINR_{gk}^{oc} - SINR_g^{oc,\circ} \to 0, \quad SINR_{gk}^{ic} - SINR_g^{ic,\circ} \to 0, \quad SINR_{gk}^{p} - SINR_g^{p,\circ} \to 0, \tag{3.34}$$

almost surely, where

$$SINR_g^{oc,\circ} = \frac{\kappa_g P(1-\beta)(1-\tau_g^2)}{\beta\big(\sum_{l\neq g}(\xi_l^\circ)^2 \Upsilon_{gl}^\circ + \big(\xi_g^\circ\big)^2 \Upsilon_{gg}^\circ \Omega_g + \frac{P}{K}\big(\xi_g^\circ\big)^2 \Phi_g\big) + 1}, \tag{3.35}$$

$$SINR_g^{ic,\circ} = \frac{\beta(1-\alpha)\big(\xi_g^\circ\big)^2 \left(\Upsilon_{gg}^\circ \Omega_g + \frac{P}{K}\Phi_g\right)}{\beta \sum_{l\neq g} \left(\xi_l^\circ\right)^2 \Upsilon_{gl}^\circ + \beta\alpha \left(\xi_g^\circ\right)^2 (\Upsilon_{gg}^\circ \Omega_g + \frac{P}{K}\Phi_g) + 1}, \tag{3.36}$$

$$SINR_g^{p,\circ} = \frac{\beta\alpha \frac{P}{K}\big(\xi_g^\circ\big)^2 \Phi_g}{\beta \sum_{l\neq g} \left(\xi_l^\circ\right)^2 \Upsilon_{gl}^\circ + \beta\alpha\big(\xi_g^\circ\big)^2 \Upsilon_{gg}^\circ \Omega_g + 1}, \tag{3.37}$$

with

$$\left(\xi_g^\circ\right)^2 = \frac{K_g}{\Psi_g^\circ}, \quad \Psi_g^\circ = \frac{K_g}{b_g}\frac{m_g'}{(1+m_g^\circ)^2}, \quad \Phi_g = \frac{(1-\tau_g^2)(m_g^\circ)^2}{(1+m_g^\circ)^2} \tag{3.38}$$

$$\Upsilon_{gl}^\circ = \frac{P}{K}\frac{K_g}{b_g}\frac{m_{gl}'}{(1+m_l^\circ)^2}, \quad \Omega_g = \frac{K_g-1}{K_g}\frac{(1-\tau_g^2(1-(1+m_g^\circ)^2))}{(1+m_g^\circ)^2} \tag{3.39}$$

$$m_g' = \frac{\frac{1}{b_g}tr(\bar{\mathbf{R}}_{gg}\mathbf{T}_g\mathbf{B}_g^H\mathbf{B}_g\mathbf{T}_g)}{1-\frac{\frac{K_g}{b_g}tr(\bar{\mathbf{R}}_{gg}\mathbf{T}_g\bar{\mathbf{R}}_{gg}\mathbf{T}_g)}{b_g\,(1+m_g^\circ)^2}}, \quad m_{gl}' = \frac{\frac{1}{b_g}tr(\bar{\mathbf{R}}_{ll}\mathbf{T}_l\bar{\mathbf{R}}_{gl}\mathbf{T}_l)}{1-\frac{\frac{K_g}{b_g}tr(\bar{\mathbf{R}}_{ll}\mathbf{T}_l\bar{\mathbf{R}}_{ll}\mathbf{T}_l)}{b_g\,(1+m_l^\circ)^2}} \tag{3.40}$$

$$\kappa_g = \frac{tr(\bar{\mathbf{R}}_{gg})^2}{\sum_{l=1}^G K_g tr(\bar{\mathbf{R}}_{ll})}, \quad \bar{\mathbf{R}}_{gl} = \mathbf{B}_l^H\mathbf{R}_g\mathbf{B}_l,\ \forall g,l \tag{3.41}$$

and m_g° *and* $\mathbf{T}_g$ *the unique solutions of*

$$m_g^\circ = \frac{1}{b_g}tr\left(\bar{\mathbf{R}}_{gg}\mathbf{T}_g\right), \quad \mathbf{T}_g = \left(\frac{K_g}{b_g}\frac{\bar{\mathbf{R}}_{gg}}{1+m_g^\circ} + \varepsilon\,\mathbf{I}_{b_g}\right)^{-1}. \tag{3.42}$$

It follows from (3.34) $\frac{1}{K}(R_p^{HRS} - R_p^{HRS,\circ}) \overset{M\to\infty}{\longrightarrow} 0$ where $R_p^{HRS,\circ} = \sum_{g=1}^G K_g \log_2(1+\mathrm{SINR}_g^{p,\circ})$, $\frac{1}{G}(R_{ic}^{HRS} - R_{ic}^{HRS,\circ}) \overset{M\to\infty}{\longrightarrow} 0$ where $R_{ic}^{HRS,\circ} = \sum_{g=1}^G \log_2(1+\mathrm{SINR}_g^{ic,\circ})$, and that $R_{oc}^{HRS} - R_{oc}^{HRS,\circ} \overset{M\to\infty}{\longrightarrow} 0$ where $R_{oc}^{HRS,\circ} = \log_2(1+\mathrm{SINR}^{oc,\circ})$ with $\mathrm{SINR}^{oc,\circ} = \min_g \{\mathrm{SINR}_g^{oc,\circ}\}$. Then, an approximation $R_{\mathrm{sum}}^{HRS,\circ}$ of the sum rate of HRS is obtained as

$$R_{\mathrm{sum}}^{HRS,\circ} = R_{oc}^{HRS,\circ} + R_{ic}^{HRS,\circ} + R_p^{HRS,\circ}. \tag{3.43}$$

Likewise, the asymptotic sum rate of conventional TTP scheme in (3.26) converges as $(R_{\mathrm{sum}}^{TTP} - R_{\mathrm{sum}}^{TTP,\circ})/K \overset{M\to\infty}{\longrightarrow} 0$, where $R_{\mathrm{sum}}^{TTP,\circ} = \sum_{g=1}^G K_g \log_2(1+\mathrm{SINR}_g^{TTP,\circ})$ and

$$\mathrm{SINR}_g^{TTP,\circ} = \frac{\frac{P}{K}\left(\xi_g^\circ\right)^2\Phi_g}{\sum_{l\neq g}\left(\xi_l^\circ\right)^2\Upsilon_{gl}^\circ + \left(\xi_g^\circ\right)^2\Upsilon_{gg}^\circ\Omega_g + 1}, \tag{3.44}$$

and the first term in the denominator of (3.44) containing $\Upsilon_{gl}^\circ(\bar{\mathbf{R}}_{gl})$ denotes inter-group interference while the second term with $\Omega_g(\tau^2)$ refers to intra-group interference. The sum rate gain of HRS over conventional two-tier precoding No-RS is quantified by

$$\Delta R^{HRS,\circ} = \sum_g K_g\big(\log_2(1+\mathrm{SINR}_g^{p,\circ}) - \log_2(1+\mathrm{SINR}_g^{TTP,\circ})\big) + R_{oc}^{HRS,\circ} + R_{ic}^{HRS,\circ}. \tag{3.45}$$

3.3.4. Power Allocation

Since α and β are coupled in the SINR expressions (3.35) $\sim$ (3.37), a closed-form and optimal solution that maximises $R_{\mathrm{sum}}^{HRS,\circ}$ cannot be obtained. Following a similar philosophy as in Section 3.2.4, we compute a closed-form suboptimal but effective power allocation method, by which the private messages of HRS are transmitted with a fraction of the

total power and achieve nearly the same sum rate as conventional No-RS strategy with full power, i.e., $R_{\text{sum}}^{TTP,\circ} \approx R_p^{HRS,\circ}$. Then, the remaining power is utilised to transmit the common messages and enhance the sum rate. We can write

$$\text{SINR}_g^{p,\circ} \leq \text{SINR}_g^{TTP,\circ}, \quad \forall g, \tag{3.46}$$

for $\forall \alpha, \beta \in (0, 1]$. Consider two extreme cases: weak and strong inter-group interference. Based on (3.44), the notation of 'weak' implies that the inter-group interference is sufficiently small and thus negligible, i.e., $\Upsilon_{gl}^{\circ} \to 0, \forall g \neq l$. The sum rate $R_{\text{sum}}^{TTP,\circ}$ is limited by the intra-group interference. On the contrary, the notation of 'strong' means that the inter-group interference dominates the rate performance, i.e., $\sum_{l \neq g} (\xi_l^{\circ})^2 \Upsilon_{gl}^{\circ} > (\xi_g^{\circ})^2 \Upsilon_{gg}^{\circ}$.

Proposition 3.3. *The equality of* (3.46) *holds when the power splitting ratios α, β are given as*

$$\beta = 1, \quad \alpha = min \left\{ \frac{K_g}{P \cdot \Gamma_{IG}}, \, 1 \right\} \tag{3.47}$$

in the weak inter-group interference regime, and as

$$\beta = min \left\{ \frac{K}{P \cdot \Gamma_{OG} + K_g}, \, 1 \right\}, \quad \alpha = 1 \tag{3.48}$$

in the strong inter-group interference regime, where

$$\Gamma_{OG} = \underset{g}{min} \left\{ \sum_{l \neq g} \frac{K_g}{K} \frac{tr(\bar{\mathbf{R}}_{gl} \bar{\mathbf{R}}_{ll}^{-1})}{tr(\bar{\mathbf{R}}_{ll}^{-1})} \right\}, \quad \Gamma_{IG} = \underset{g}{min} \left\{ \frac{\tau_g^2}{K} \frac{b_g(K_g - 1)}{tr(\bar{\mathbf{R}}_{gg}^{-1})} \right\}. \tag{3.49}$$

Proof. See Appendix B.3. □

Remark 3.3. *When the inter-group interference is negligible, HRS becomes a set of parallel RS in G groups, i.e., the outer common message is unnecessary. By contrast, when the inter-group interference is the dominant degrading factor, the inner common message and private messages transmission is limited and HRS boils down to RS.For the general inter-group interference case, finding simple closed-form α, β that guarantees a sum rate gain of HRS over conventional No-RS is challenging. Nevertheless, motivated by the design philosophy of power allocation in the extreme cases, we induce a threshold μ by which $SINR_g^{p,\circ} = \mu \cdot SINR_g^{TTP,\circ}$ (e.g., $\mu = 0.9$). By following the proof of Proposition 3.3, we compute the power allocation factors as follows: $\beta = min \left\{ \frac{K}{P \cdot (\Gamma_{OG} + \alpha \Gamma_{IG})}, \, 1 \right\}$, $\alpha = min \left\{ \frac{\mu(P\Gamma_{OG}+1)}{P \cdot (\Gamma_{OG} + (1-\mu)\Gamma_{IG}) + 1}, \, 1 \right\}$. The threshold μ should be carefully designed for certain system setting.*

From (3.47) and (3.48), we have $\alpha = \beta = 1$ at low SNR and HRS becomes conventional two-tier precoding No-RS strategy. Namely, the effect of imperfect CSIT/overlapping eigen-subspaces on the private messages transmission is negligible and common messages

are not needed. Otherwise, the rate performance of conventional two-tier precoding No-RS strategy saturates at high SNR while HRS exploits a fraction of the total power ($\alpha < 1$ or $\beta < 1$) to transmit the common message(s) and enhance the sum rate.

Corollary 3.1. *With power allocation of Proposition 3.3, the sum rate gain $\Delta R^{HRS,\circ}$ at high SNR is lower bounded as*

$$\Delta R^{HRS,\circ} \geq \sum_g \big(\log_2(1 + SINR_g^{ic,\circ}) - \log_2(e)\big), \tag{3.50}$$

in the weak inter-group interference regime, and as

$$\Delta R^{HRS,\circ} \geq \log_2(1 + SINR^{oc,\circ}) - \log_2(e), \tag{3.51}$$

in the strong inter-group interference regime.

Proof. See Appendix B.4. □

Remark 3.4. *Here are some interpretations of Proposition 3.3 and Corollary 3.1.*

- ***Power allocation to the private and common messages**: The intra-group power splitting ratio (α) decreases as τ^2 increases. Namely, in order to alleviate intra-group interference, we should allocate less power to the private messages as the CSIT quality gets worse ($\tau^2 \to 1$). Similarly, the inter-group power splitting ratio (β) drops as the inter-group interference $\Upsilon_{gl}^{\circ}, g \neq l$ becomes larger. From (3.47) $\sim$ (3.48), the power given to the privates messages is an invariant of P at high SNR*

$$\sum_g \sum_k P_{gk} = P\alpha\beta = \begin{cases} \frac{K}{\Gamma_{OG}}, & \text{if } \beta < 1 \\ \frac{K_g}{\Gamma_{IG}}, & \text{otherwise} \end{cases} \tag{3.52}$$

  which places private message decoding back into the non-interference-limited regime. Meanwhile, the power allocated to the common messages linearly increases with P.

- ***Sum rate gain**: HRS exploits the extra power beyond saturation of conventional No-RS strategies to transmit the common message(s), leading to a sum rate that increases with the available transmit power. In the weak inter-group interference regime, HRS becomes a set of parallel inner RS. Based on (3.50), the sum rate gain $\Delta R^{HRS,\circ}$ increases by G bps/Hz for each 3 dB power increment at high SNR. By contrast, HRS boils down to RS in the strong inter-group interference regime and $\Delta R^{HRS,\circ}$ increases by 1 bps/Hz for each 3 dB power increment at high SNR.*

3.4. Performance Evaluation

Numerical results are provided to validate the effectiveness of RS/HRS. Uniform circular array with $M = 100$ antennas are equipped at the BS. Consider the transmit correlation

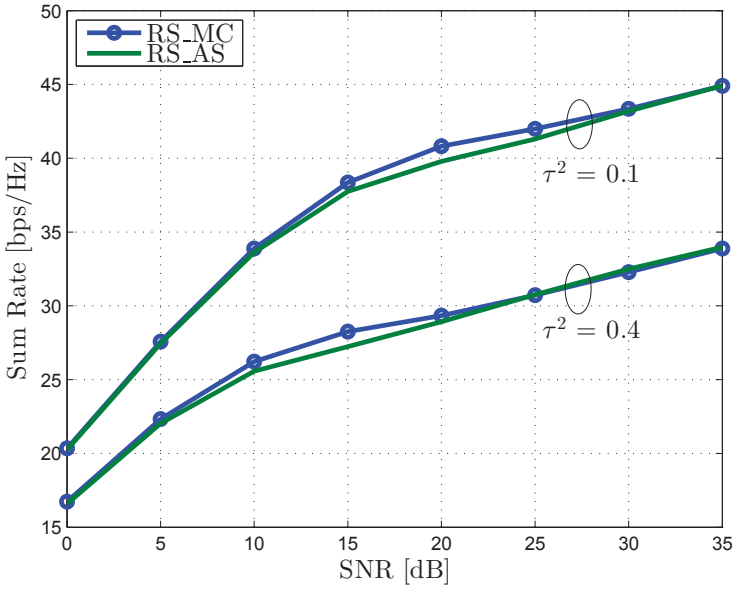

Figure 3.1.: Sum rate comparison: RS asymptote (AS) approximation vs. RS Monte Carlo (MC) simulation for various CSIT qualities and SNRs.

model in (3.1), the antenna elements are equally spaced on a circle of radius λD, for $D = \frac{0.5}{\sqrt{(1-\cos(2\pi/M))^2+\sin(2\pi/M)^2}}$, leading to a minimum distance $\lambda/2$ between any two antennas.

3.4.1. RS

$K = 5$ users are assumed to be distributed uniformly at an azimuth angle $\theta_k = 2\pi k/K$ and angular spread $\Delta_k = \pi/6$. We compare RS with RZF-precoded conventional No-RS (No-RS_RZF) [7] and TDMA. Two types of RS are investigated: exhaustive search (RS_EXS) and closed-form (RS_CLF). Specifically, RS_EXS performs a simulation-based exhaustive search with step 0.01 for the best power splitting ratio t. RS_CLF allocates power by following Proposition 3.2.

In Fig. 3.1, we can observe that the asymptotic approximation $R_{\text{sum}}^{RS,\circ}$ derived in Theorem 3.1 effectively characterises the sum rate of RS for a variety of CSIT qualities and SNRs. Fig. 3.2 shows that RS_CLF achieves almost the same sum rate as RS_EXS. This verifies the efficacy of the proposed power allocation scheme (3.24). Moreover, the multiplexing gain of RS is approaching 1 at high SNR. RS behaves as No-RS_RZF in the low to medium SNR regime. At high SNR, the sum rate of RS linearly increases with the transmit power (dB) as TDMA. By contrast, No-RS_RZF suffers from a rate ceiling effect due to partial CSIT. 'RS_p' denotes the sum rate of the private messages in RS transmission, which confirms the power allocation scheme in Section 3.2.4. Namely, RS allocates a fraction of power to transmit the private messages so as to achieve approx-

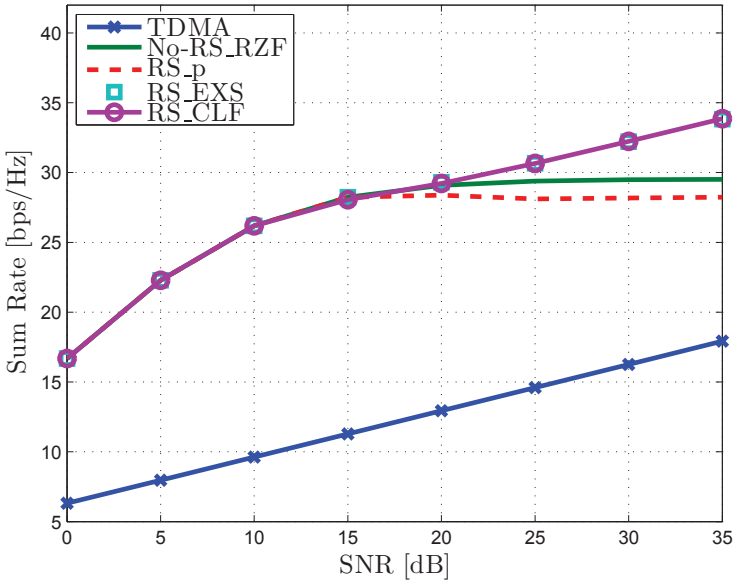

Figure 3.2.: Sum rate comparison: RS vs. conventional No-RS and TDMA, in a setup with $\tau^2 = 0.4$.

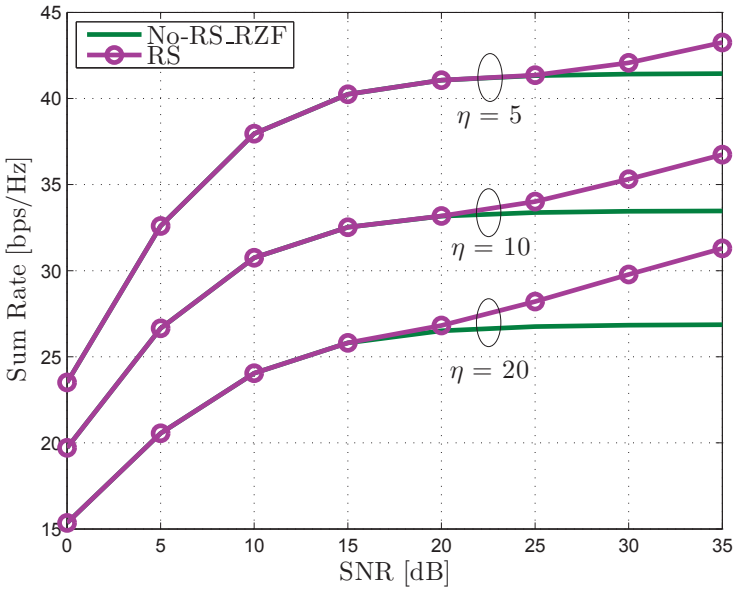

Figure 3.3.: Sum rate gain of RS over conventional No-RS for various antennas/users ratio ($\eta = M/K$).

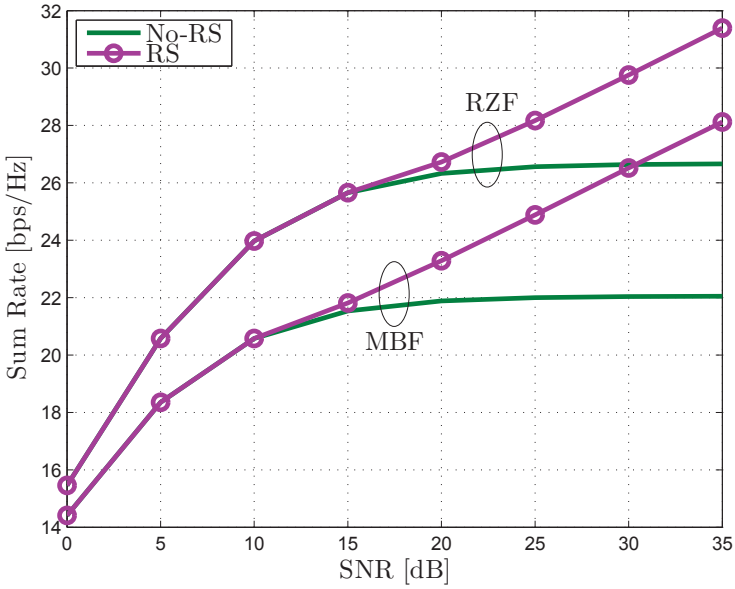

Figure 3.4.: Sum rate gain of MBF-precoded RS over conventional RZF-precoded No-RS, in a setup with $K = 5$.

imately the same sum rate as No-RS_RZF with full power. Then, using the remaining power to transmit the common message enhances the sum rate. In Fig. 3.2, the rate loss $R_{\text{sum}}^{RZF} - R_p^{RS}$ is 1.3 bps/Hz at high SNR, which is lower than $\log_2(e)$ and agrees with Remark 3.2.

In Fig. 3.3, the CSIT quality is set as $\tau^2 = 0.5$. The sum rate gain of RS over No-RS_RZF degrades with larger K (smaller η), which confirms the discussion in Remark 3.2. Namely, the sum rate gain as well as the achievable rate of the common message becomes smaller as the number of users increases.

In general, a low-complexity precoder is desirable for massive MIMO systems [85]. MBF enjoys the lowest precoding complexity while ZF/RZF that achieves a much better performance than MBF involves complicated matrix inversion. Interestingly, RS enables to meet a certain sum rate requirement with a highly-reduced precoding complexity compared with No-RS_RZF. To identify the computational benefits of RS, Fig. 3.4 compares RS (MBF-precoded private messages) with RZF-precoded No-RS. The power splitting ratio t of RS is computed via an exhaustive search[2]. Recall that we assume a predefined set of K user is scheduled, we observe that RS with MBF reaches the same rate performance as No-RS with RZF at SNR = 30 dB. In fact, RS simplifies the precoding design and decreases the computational complexity at the cost of an increased encoding and decoding complexity.

[2]By calculating the asymptotic SINR of RS with MBF-precoded private messages, the closed-form t can be obtained via (3.24).

3.4.2. HRS

For simplicity, we assume $\tau_g = \tau, K_g = \bar{K}, b_g = \bar{b}, \forall g$. Consider $K = 12$ users equally clustered into $G = 4$ groups. We compare the proposed HRS strategy with the following baselines: **Baseline 1 (No-RS with two-tier precoder [44])**, **Baseline 2 (Baseline 1 with user scheduling at the group level)**: Within each group, a single user with the largest effective channel gain is selected and the precoder of the private message intended to each user is MBF. **Baseline 3 (Baseline 1 with user scheduling at the system level)**: User scheduling is performed at the system level such that the best user among all is selected. Two types of HRS are investigated: exhaustive search (HRS_EXS) and closed-form (HRS_CLF). Specifically, HRS_EXS performs a simulation-based exhaustive search with step 0.01 for the best power splitting ratios α and β. HRS_CLF allocates power by following the closed-form solution in Proposition 3.3.

1) Validation of the Asymptotic Rate Analysis: We compare the asymptotic sum rate (3.43) with simulations. Various CSIT qualities have been simulated and $\tau^2 = \{0, 0.4\}$ are taken as examples. For the outer precoder design, we set $\bar{b} = 15$ such that $\bar{K} \leq \bar{b} \leq M - (G-1)r^d$ and $\bar{b} \leq r^d$, where $r^d = 20$[3] includes the dominant eigenvalues of $\mathbf{R}_g, \forall g$. To verify the effectiveness of the asymptotic sum rate approximation of HRS, we consider two scenarios with disjoint and overlapping eigen-subspaces, respectively. As an example, we set $\theta_g = -\frac{\pi}{2} + \frac{\pi}{3}(g-1)$ and $\Delta_g = \Delta = \frac{\pi}{8}, \forall g$ corresponding to disjoint eigen-subspaces ($[\theta_g - \Delta_g, \theta_g + \Delta_g] \bigcap [\theta_l - \Delta_l, \theta_l + \Delta_l] = \emptyset, \forall l \neq g$) while $\Delta_g = \Delta = \frac{\pi}{3}, \forall g$ leading to eigen-subspaces overlap.

In the scenario with disjoint eigen-subspaces, the inter-group interference is negligible and thereby the outer common message is unnecessary. When we further have perfect CSIT ($\tau^2 = 0$), there is no intra-group interference and the inner common messages are not needed. In this case, HRS boils down to two-tier precoding No-RS. Fig. 3.5 shows that the asymptotic sum rate $R_{\text{sum}}^{HRS,\circ}$ matches exactly the simulation result. As the CSIT quality decreases (increased τ^2), inner common messages are exploited to mitigate intra-group interference. The asymptotic approximation (3.43) becomes less accurate but still valid to capture the asymptotic sum rate of HRS.

By contrast, larger angular spread leads to larger rank of the spatial correlation matrix. With outer precoder design as (3.32), a large fraction of power included in the weak eigenmodes is leaked into other groups, leading to strong inter-group interference. Fig. 3.6 indicates that the sum rate of HRS can be approximately captured by (3.43). In the low to medium SNR regime, HRS behaves as two-tier precoded No-RS where the rate gap between the asymptotic approximation and the simulation $\frac{1}{K}(R_p^{HRS,\circ} - R_p^{HRS}) = \frac{1}{K}(R_{\text{sum}}^{TTP,\circ} - R_{\text{sum}}^{TTP})$ is within 0.2 bps/Hz. A similar behaviour can be observed as well in

[3] In this configuration, r^d can be chosen as large as 25 due to $M = 100, G = 4$. Here, we set $r^d = 20$ since a relatively smaller r^d enables to select stronger eigenmodes $\mathbf{F}_g^{(1)}$ thanks to the larger dimension of orthogonal subspace $M - \sum_{l \neq g} r^d$. In fact, simulation results revealed that HRS with $r^d = 20$ can achieve higher rate than that achievable with $r^d = 25$. We omit the simulation results for conciseness.

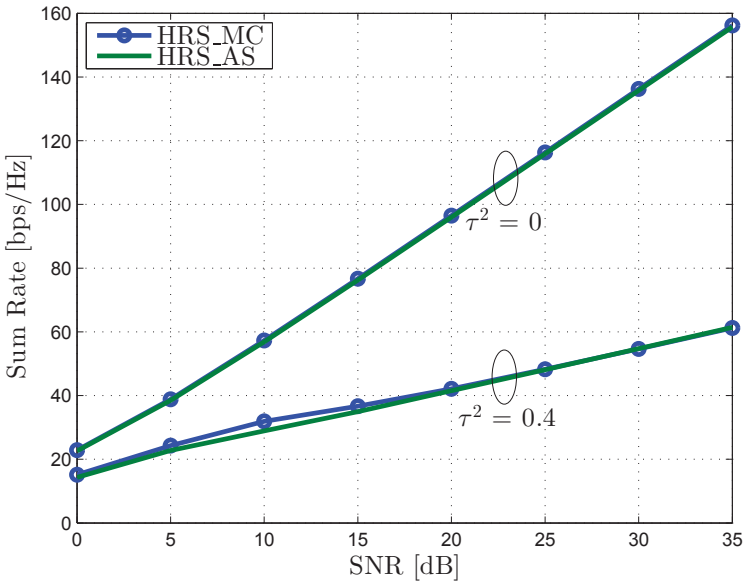

Figure 3.5.: Sum rate comparison: HRS asymptote (AS) approximation vs. HRS Monte-Carlo (MC) simulation in disjoint eigen-subspace scenario.

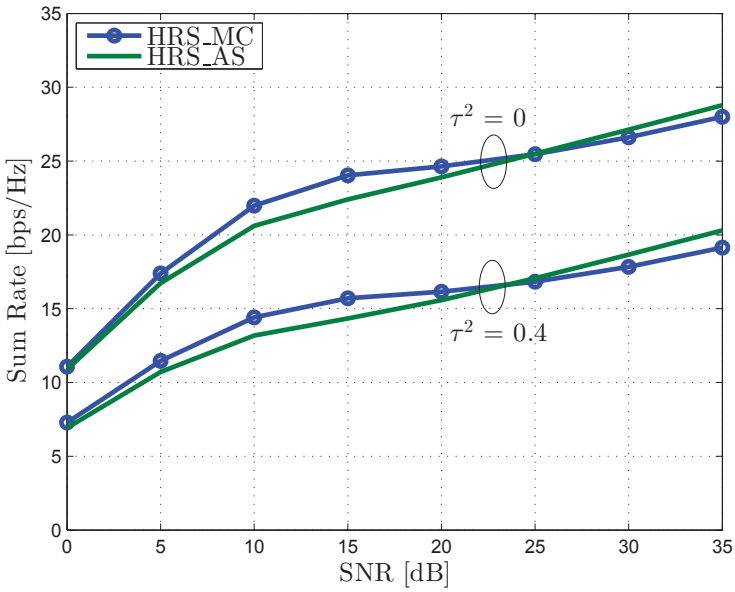

Figure 3.6.: Sum rate comparison: HRS asymptote (AS) approximation vs. HRS Monte-Carlo (MC) simulation in overlapping eigen-subspace scenario.

[7,45]. At high SNR, the simulation R_p^{HRS} is around 1 bps/Hz lower than the asymptotic $R_p^{HRS,\circ}$, because the SINR of the outer common message $\text{SINR}^{oc} = \min_{g,k} \{\text{SINR}_{gk}^{oc}\}$ is upper bounded by its asymptotic $\text{SINR}^{oc,\circ}$ for large but finite M. For example, when the precoder of the outer common message is designed such that all users experience the same $\text{SINR}^{oc,\circ}$, the asymptotic approximation is given by $\text{SINR}^{oc,\circ}$ while the simulated SINR^{oc} is the minimum rate among all. It can be verified that this effect is mitigated when the channel hardens as the number of transmit antennas increases.

2) Rate Performance Comparison: Fig. 3.7 and 3.8 evaluate the benefits of HRS under partial CSIT ($\tau^2 = 0.4$) with the same system configuration as Fig. 3.5 and Fig. 3.6. With disjoint eigen-subspaces (negligible inter-group interference), Fig. 3.7 shows that conventional No-RS strategy with two-tier precoder (Baseline 1) saturates at high SNR due to intra-group interference while user scheduling enables a multiplexing gain of 4 (Baseline 2) and 1 (Baseline 3), respectively. According to Proposition 3.3, HRS becomes a set of parallel inner RS. We observe that the proposed HRS strategy exhibits substantial rate gain over various baselines. For instance, the sum rate gain of HRS ΔR^{HRS} over two-tier precoding No-RS at SNR = 30 dB is 15.5 bps/Hz. Based on (3.50) in Corollary 3.1, the asymptotic sum rate gain $\Delta R^{HRS,\circ}$ is 19.5 bps/Hz. The 4 bps/Hz rate gap between ΔR^{HRS} and $\Delta R^{HRS,\circ}$ is explained as follows. The simulated rate loss $R_{\text{sum}}^{TTP} - R_p^{HRS}$ is indeed upper bounded by $G \log_2(e)$ as (B.16). However, the simulated sum rate of the common messages R_{ic}^{HRS} is around 4 bps/Hz lower than $R_{ic}^{HRS,\circ}$.

With severely overlapping eigen-subspaces (strong inter-group interference), HRS boils down to RS (with reduced-dimensional CSIT) at the system level according to Proposition 3.3, i.e., inner common messages are not transmitted. Fig. 3.8 reveals that HRS outperforms two-tier precoding No-RS with/without user scheduling. The sum rate enhancement of HRS over two-tier precoding No-RS at SNR = 30 dB is 1.5 bps/Hz.

Interestingly, in both settings of Fig. 3.7 and 3.8, the closed-form power allocation achieves almost the same sum rate as that of a simulation-based exhaustive search. This verifies the effectiveness of the power allocation strategy in Proposition 3.3. In Fig. 3.7 and Fig. 3.8, respectively, we observe that the sum rate gain ΔR^{HRS} of HRS over two-tier precoding No-RS increases by nearly G bps/Hz and 1 bps/Hz for any 3 dB increment of power at high SNR, which verifies the discussion of Remark 3.4.

In a nutshell, HRS exhibits robustness w.r.t. CSIT error and eigen-subspaces overlap. HRS behaves as two-tier precoding No-RS at low SNR, where the effect of imperfect CSIT/overlapping eigen-subspaces on the sum rate of transmitting private messages is insignificant. At high SNR, by transmitting common message(s), the asymptotic multiplexing gain of HRS amounts to that of two-tier precoding No-RS with perfect user scheduling. Meanwhile, HRS exploits the rate benefits of two-tier precoding No-RS by transmitting the private messages with a fraction of the total power.

Comparing baseline 1 with baseline 2, we see the importance of user scheduling for No-RS strategy. User scheduling would obviously be useful to HRS as well, but HRS

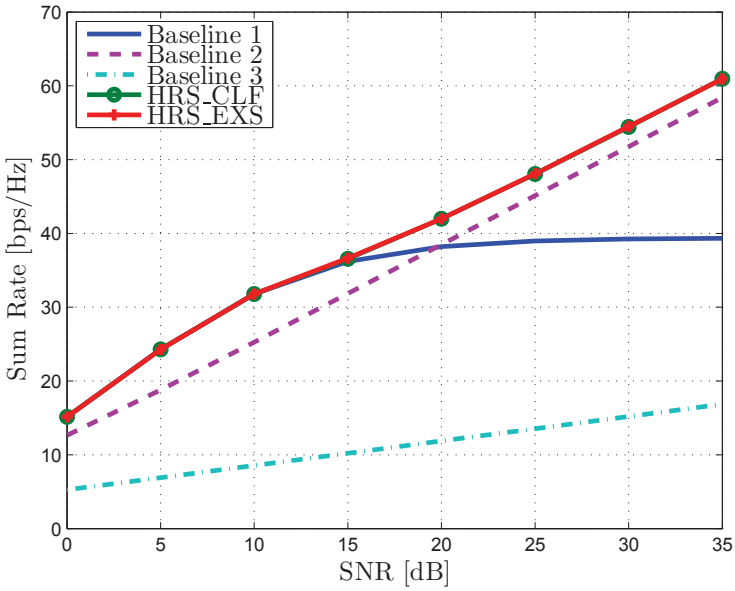

Figure 3.7.: Sum rate comparison: HRS vs. various baselines under partial CSIT in disjoint eigen-subspace scenario.

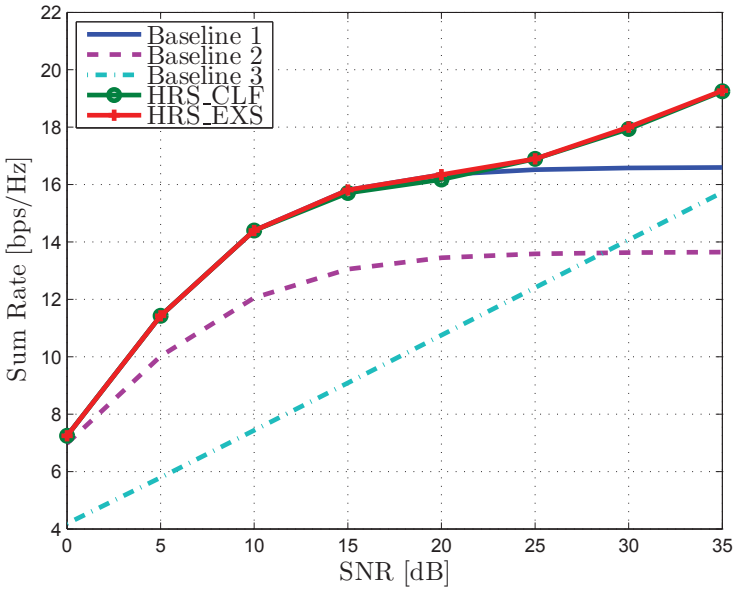

Figure 3.8.: Sum rate comparison: HRS vs. various baselines under partial CSIT in overlapping eigen-subspace scenario.

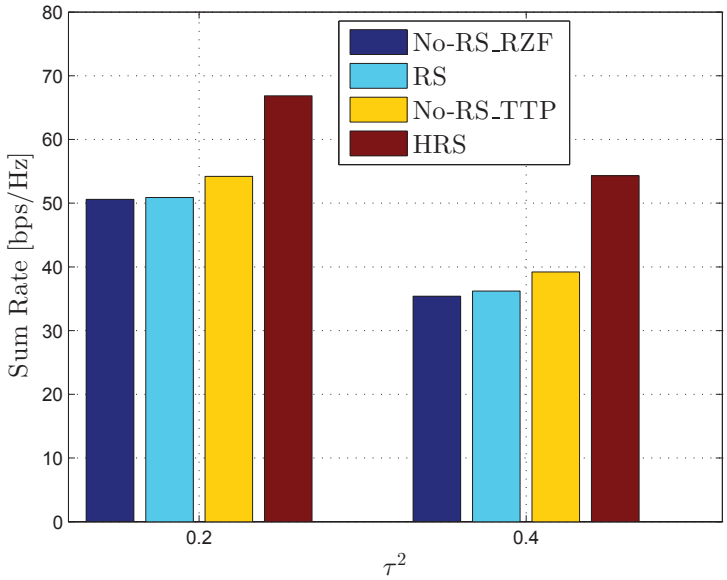

Figure 3.9.: Sum rate comparison: HRS vs. RS and No-RS with one/two-tier precoding under various CSIT qualities, in a setup with SNR = 30 dB.

shows a very competitive performance even without user scheduling. This is particularly attractive in massive MIMO where the number of users (K) can potentially be large. Hence, HRS would decrease the burden on the scheduler and the precoder design but increases the complexity of the encoding and decoding schemes.

3.4.3. RS vs. HRS

To examine the suitability of HRS in spatially correlated massive MIMO, we compare HRS to No-RS_RZF/RS with full-dimensional CSIT in Section 3.2 and two-tier precoding No-RS with reduced-dimensional CSIT for various CSIT qualities. We assume the same system configuration as Fig. 3.7. Overall, Fig. 3.9 shows that a lower CSIT quality (i.e., larger τ^2) degrades the rate performance of these schemes.

In Fig. 3.9, we take $\tau^2 = 0.4$[4] as an example. The sum rate of RS slightly outperforms No-RS_RZF around 1 bps/Hz. Recall that the achievable rate of the common message in RS is the minimum rate among all users ($K = 12$). Thus, the sum rate gain from transmitting a common message appears small at SNR = 30 dB. By contrast, the sum rate gain of HRS over No-RS_TTP is 15.5 bps/Hz. The large gain of HRS is enabled by multiple inner common messages while the achievable rate of each inner common message is the minimum rate among a smaller number of users ($\bar{K} = 3$). This observation

[4]When the CSIT imperfectness is incurred by quantisation error, the number of feedback bits required to achieve a certain CSIT quality is proportional to the dimension of the quantised channel. Then, the CSIT quality of two-tier precoding No-RS/HRS is better than that of one-tier precoding No-RS/RS for a given number of feedback bits. The rate gap between one-tier and two-tier precoding strategies can be even larger. Nevertheless, we assume the same CSIT quality for simplicity of exposition.

confirms that if the channel second-order statistics is available at the BS, the proposed HRS transmission strategy is better suited for massive MIMO deployments than a simple RS transmission strategy.

With reduced-dimensional CSIT, No-RS with two-tier precoder (No-RS_TTP) and HRS achieve much higher rates than No-RS_RZF and RS. This is because the outer precoder exploiting long-term CSIT partitions users into a set of non-interfering groups. Each user experiences interference from fewer users (i.e., only users in the same group) compared with No-RS_RZF and RS.

3.5. Summary and Conclusion

Due to partial CSIT, the rate performance of conventional No-RS schemes is severely degraded by multiuser interference. To tackle this issue, a Rate-Splitting approach has been proposed. RS packs part of one selected user's message into a common message that can be decoded by all users and superimposes the common message on top of the private messages. We generalised the RS strategy into the large-scale array regime. By further exploiting the channel second-order statistics and a two-tier precoding structure, we proposed a novel Hierarchical-Rate-Splitting strategy. Particularly, on top of the private messages, HRS transmits an outer common message and multiple inner common messages that can be decoded by all users and a subset of users, respectively. The outer common message tackles the inter-group interference due to overlapping eigen-subspaces while the inner common messages help with mitigating the intra-group interference due to partial CSIT.

For RS and HRS, we derived the precoder design, asymptotic rate performance and power allocation. Interestingly, to meet a certain sum rate requirement, RS highly decreases the complexity of precoder design and scheduling at the expense of an increase in complexity of the encoding and decoding strategy. Moreover, simulation results showed that the rate performance of conventional No-RS strategies saturates at high SNR due to partial CSIT and the sum rate gain of RS over No-RS with RZF increases with the available transmit power. When users in different groups have disjoint eigen-subspaces, the sum rate gain of HRS over No-RS with two-tier precoder is much larger than the gain achievable with RS. In a nutshell, RS and HRS exhibit robustness w.r.t. CSIT error and/or eigen-subspaces overlaps while HRS is a general and particularly suited framework for the massive MIMO deployments.

We believe that the idea of RS and HRS has a huge potential in a wide range of wireless systems including massive MIMO and single-cell/multi-cell multiuser MIMO broadcast channel, operated at microwave/mmWave with FDD/TDD mode, equipped with sufficient/limited RF chains at the BS. In practice, CSIT is imperfect due to imperfect channel estimation, limited feedback, antenna miscalibration, pilot contamination, etc. The effectiveness of RS and HRS has been demonstrated to mitigate the resultant interference.

4. A Novel Multiuser Millimeter Wave Beamforming Strategy with Quantised and Statistical CSIT

In addition to the massive MIMO considered in prior section, it is well known that Millimeter Wave (mmWave) MIMO is also a promising candidate for improving the spectrum and energy efficiency for 5G cellular networks. In this chapter, we turn to investigate the multiuser mmWave systems with partial (specifically quantised) and statistical CSIT.

When the mmWave system is considered, a hybrid analog and digital precoding is typically employed to tackle the RF hardware constraint. The conventional beamforming strategy determines the hybrid precoder relying on a two-stage feedback scheme [17], as introduced in Section 1.7.6. Specifically, the analog precoder is designed to maximise the desired signal power of each user by beam search and feedback. Then, the digital precoder depends on the random vector quantisation (RVQ) and feedback of the effective channel (the channel concatenated with the analog precoder).

In presence of statistical CSIT, we propose a hybrid precoding design using a one-stage feedback scheme which can effectively reduce the complexity of signalling and feedback overhead. Specifically, we make use of all feedback overhead for the first stage to enable precise design of beamforming directions and take advantage of the second-order channel statistics to mitigate multiuser interference. Hereinafter, this is referred to as One-Stage Feedback plus Statistical CSIT ('OSF + Stat')-based hybrid precoding scheme. To make a fair comparison, we consider an enhanced design of [17] by employing a second-order channel statistics-based quantization codebook in the second-stage feedback. Hereinafter, this is referred to as the Two-Stage Feedback plus Adaptive Codebook ('TSF + Adp CB')-based hybrid precoding scheme. With a fixed total feedback constraint, we mainly investigate the conditions under which the one-stage feedback scheme outperforms the conventional two-stage counterpart.

Moreover, the multiuser interference is still a limiting factor for mmWave systems due to imperfect CSIT and leads to a performance degradation. A rate-splitting (RS) transmission strategy is introduced to tackle this issue. The benefit of RS is discussed in the context of multiuser mmWave systems.

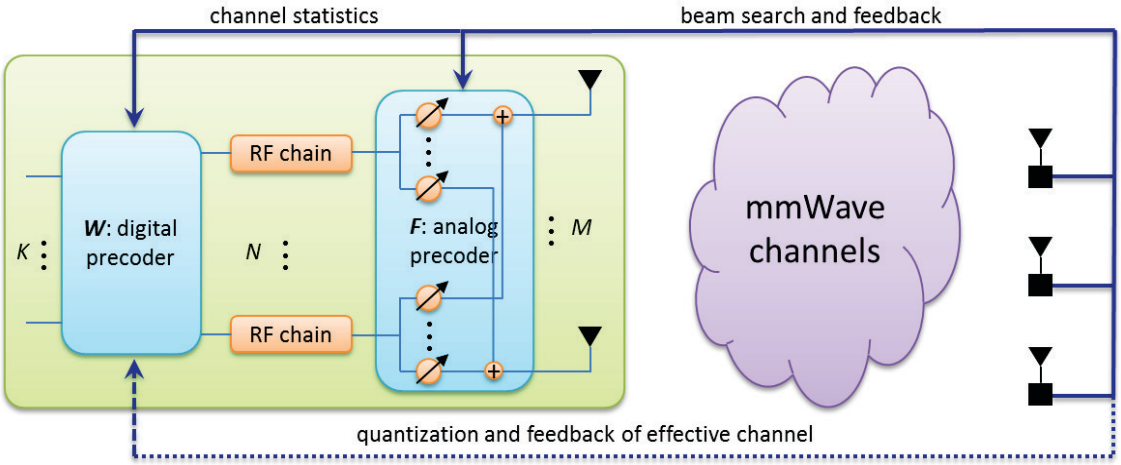

Figure 4.1.: Block diagram of multiuser mmWave downlink system model with hybrid precoding and limited feedback. Solid lines indicates the feedback requirement of the hybrid precoding scheme in Section 4.2.1. Dash lines represents the additional requirement of the hybrid precoding scheme in Section 4.2.2.

4.1. System Model

Consider a multiuser downlink system where the BS equipped with M antennas and N RF chains serves $K(\leq N \leq M)$ single-antenna users over mmWave channels. For simplicity, we assume that the BS only uses K out of N RF chains, which provides a lower bound on the rate performance. A hybrid RF beamformer $\mathbf{F} \in \mathbb{C}^{M\times K}$ and digital precoder $\mathbf{W} \in \mathbb{C}^{K\times K}$ structure is employed at the BS as depicted in Fig. 4.1. Since the RF beamformer is implemented using phase shifting networks, a constant modulus constraint is imposed on its entries. Without loss of generality, we assume that $[\mathbf{F}]_{m,n} = \frac{1}{\sqrt{M}} e^{j\varphi_{m,n}}$.

The channels in the mmWave bands tend to be sparse and we assume a ray-based geometric channel model with limited paths [10, 11]. Under this model, the channel vector from user k is defined as

$$\mathbf{h}_k = \sqrt{\frac{M}{L_k}} \sum_{l=1}^{L_k} g_{k,l}\, \mathbf{a}(\theta_{k,l}) = \sqrt{\frac{M}{L_k}} \mathbf{A}_k \mathbf{g}_k, \tag{4.1}$$

where the path gain vector $\mathbf{g}_k = [g_{k,1}, \cdots, g_{k,L_k}]$ has independent and identical distributed (i.i.d.) $\mathcal{CN}(0,1)$ entries and varies independently across different time slots. $\mathbf{A}_k = [\mathbf{a}(\theta_{k,1}), \cdots, \mathbf{a}(\theta_{k,L_k})] \in \mathbb{C}^{M\times L_k}$ contains L_k steering vectors and $\theta_{k,l} \in [0, \pi]$ are the angle-of-departure (AoD) of lth path. Accordingly, the long-term channel covariance matrix of $\mathbf{h}_k$ can be computed as

$$\mathbf{R}_k = \mathbb{E}\left\{\mathbf{h}_k \mathbf{h}_k^H\right\} = \frac{M}{L_k} \mathbf{A}_k \mathbf{A}_k^H, \tag{4.2}$$

which mainly depends on the AoDs. It is well known that the fading channel statis-

tics (e.g., AoDs) are wide-sense stationary (WSS) due to its scattering-dependency. In particular, the AoDs remain invariant over the entire duration of transmission in typical mmWave scenarios [86]. Hence, prior to the transmission, the AoD information can be efficiently extracted by channel estimation techniques such as [57]. For ease of exposition, $\mathbf{R}_k$ is assumed to be perfectly known at the BS.

While the design in this Section is applicable to arbitrary antenna arrays, we consider a uniform linear array (ULA) for ease of exposition. Under the plane wave and balanced narrowband array assumptions, the array steering vector can be written as

$$\mathbf{a}(\theta_{k,l}) = \frac{1}{\sqrt{M}}[1, e^{j2\pi\frac{d}{\lambda}\cos(\theta_{k,l})}, \cdots, e^{j2\pi\frac{(M-1)d}{\lambda}\cos(\theta_{k,l})}]^T, \tag{4.3}$$

where λ is the wavelength and $d = \frac{\lambda}{2}$ is the antenna spacing.

4.2. Multiuser Hybrid Precoding

In this section, we consider the conventional multiuser transmission strategy with two hybrid precoding schemes, 'OSF + Stat' and 'TSF + Adp CB', based on one-stage and two-stage feedback, respectively. With a fixed total feedback constraint, we explore the conditions under which the 'OSF + Stat' scheme outperforms the 'TSF + Adp CB' counterpart. Moreover, the rate performance of the 'OSF + Stat' scheme is analysed both in single-path channel and in the large-scale array regime with multiple paths.

In the conventional multiuser transmission strategy, the transmitted signal and the received signal of user k can be written as

$$\mathbf{x} = \mathbf{FWPs} = \sum_{k=1}^{K} \sqrt{P_k}\mathbf{F}\mathbf{w}_k\, s_k, \tag{4.4}$$

$$y_k = \sqrt{P_k}\,\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k\, s_k + \underbrace{\sqrt{P_j}\mathbf{h}_k^H \sum_{j\neq k}^{K} \mathbf{F}\mathbf{w}_j\, s_j}_{\text{multiuser interference}} + n_k, \tag{4.5}$$

where $\mathbf{s} = [s_1, \cdots, s_K]^T \in \mathbb{C}^K$ is the data vector intended for the K users. The transmit power is uniformly allocated among users, i.e., $\mathbf{P} = \sqrt{P/K} \cdot \mathbf{I}_K$ and $\|\mathbf{F}\mathbf{w}_k\|^2 = 1$. $n_k \sim \mathcal{CN}(0, 1)$ is the additive white Gaussian noise. Then, the SINR of user k is computed as

$$\text{SINR}_k = \frac{\rho\,|\mathbf{h}_{k,\text{eff}}\mathbf{w}_k|^2}{1 + \rho\sum_{j\neq k}|\mathbf{h}_{k,\text{eff}}\mathbf{w}_j|^2}, \tag{4.6}$$

where we define the effective channel by $\mathbf{h}_{k,\text{eff}} = \mathbf{F}^H\mathbf{h}_k \in \mathbb{C}^{K\times 1}$ and $\rho = P/K$. The sum rate can be written as $R_{\text{sum}} = \sum_{k=1}^{K}\log_2(1 + \text{SINR}_k)$. To facilitate a low complexity design, $\mathbf{F}$ and $\mathbf{W}$ are determined in a decoupled manner [17].

4.2.1. One-Stage Feedback Scheme

The 'OSF + Stat' hybrid precoding scheme uses all feedbacks to precisely design the analog beamsteering while exploiting the statistical CSIT to mitigate multiuser interference. Let $\mathcal{F}$ represent the RF beamsteering codebook, where $\mathcal{F} = \left\{\mathbf{a}(\theta_q)|\theta_q = \frac{\pi q}{Q}, q \in [1, Q]\right\}$ with cardinality $|\mathcal{F}| = Q = 2^B$. For each channel realisation, the BS searches beams in the codebook and user k feeds back the index of the codeword that gives the maximum received power. Efficient beam search algorithm can be found in [87] and references therein. Then, the BS sets the selected codeword as $\mathbf{f}_k$, i.e.,

$$\{\mathbf{f}_k\} = \arg\max_{\mathbf{f}_k \in \mathcal{F}} \ |\mathbf{h}_k^H \mathbf{f}_k|^2. \tag{4.7}$$

We assume full rank $\mathbf{F}$ (i.e., users have different dominant paths and the BS has distinct beamforming direction for each user). The probability of this event goes to one for independently and randomly distributed AoDs, large AoDs space and feedback overhead $B = \log_2(M)$[1].

As $\mathbf{F}$ and $\mathbf{R}_k$ are known to the BS, the BS can equivalently compute the covariance matrix of the effective channel $\mathbf{h}_{k,\text{eff}} = \mathbf{F}^H \mathbf{h}_k \in \mathbb{C}^{K\times 1}$ as $\mathbf{R}_{k,\text{eff}} = \mathbf{F}^H \mathbf{R}_k \mathbf{F} \in \mathbb{C}^{K\times K}$. From (4.2), the sparse nature of mmWave channel leads to low rank $\mathbf{R}_k$ such that $\mathbf{R}_{k,\text{eff}}$ can be used to effectively mitigate the multiuser interference. Noting that $\mathbf{W} \in \mathbb{C}^{K\times K}$, a straightforward design of the digital precoder $\mathbf{w}_k$ of user k lies in the nullspace of $\mathcal{S}_k$, i.e.,

$$\mathbf{w}_k = \text{Null}\{\mathcal{S}_k\}, \tag{4.8}$$

where $\mathcal{S}_k = \text{Span}(\{\mathbf{u}_{\max}(\mathbf{R}_{j,\text{eff}}) : j \neq k\})$ is defined as the space spanned by the dominant eigenvectors of $K \times K$ channel covariance matrices $\mathbf{R}_{j,\text{eff}}$ of all other users $j \neq k$. This design intends to minimise the multiuser interference in a statistical sense but overlooks the desired signal power. To overcome this problem, we adopt a SLNR metric which strikes a balance between the desired signal power and the interference imposed to other users

$$\text{SLNR}_k = \frac{\rho\, |\mathbf{h}_{k,\text{eff}}^H \mathbf{w}_k|^2}{1 + \rho \sum_{j\neq k} |\mathbf{h}_{j,\text{eff}}^H \mathbf{w}_k|^2}. \tag{4.9}$$

The SLNR metric has been widely used for designing multiuser transmit beamforming [6, 70]. Then, $\mathbf{W}$ is designed by maximising a lower bound on the average SLNR

$$\mathbb{E}(\text{SLNR}_k) \geq \text{SLNR}_k^{\text{LB}} = \frac{\rho\, \mathbf{w}_k^H \mathbf{R}_{k,\text{eff}} \mathbf{w}_k}{1 + \rho \sum_{j\neq k} \mathbf{w}_k^H \mathbf{R}_{j,\text{eff}} \mathbf{w}_k}, \tag{4.10}$$

which is obtained by the convexity of $f(x) = 1/x$. The optimal unit norm $\mathbf{w}_k$ that

[1] The total feedback overhead $B = 6 \sim 8$ can support a large antenna array with dimension $64 \sim 256$ and therefore this assumption is valid in practice.

1:	**Input**: 'OSF': RF codebook $\mathcal{F}$ of size $\lvert\mathcal{F}\rvert = 2^B$ or 'TSF': RF codebook $\mathcal{F}$ of size $2^{B_{RF}}$ and statistical CSIT-based digistal codebook $\mathcal{C}_k$ of size $\lvert\mathcal{C}_k\rvert = 2^{B_{BB}}$
2:	**First stage**: Single-user RF beamforming ($\mathbf{F}$)
	Downlink beam search and uplink feedback with (4.7)
3:	**Second stage**: Multiuser digital precoding ($\mathbf{W}$)
4:	'OSF': statistical CSIT ZF (4.8) or SLNR-based (4.11) or
5:	'TSF': channel quantisation and feedback with (4.20) and instantaneous CSIT ZF (4.22) or SLNR-based (4.24)

Table 4.1.: Hybrid precoding schemes

maximises the lower bound in (4.10) is the generalised eigenvector given by [88]

$$\mathbf{w}_k = \mathbf{u}_{\max}\Big(\big(\frac{1}{\rho}\mathbf{I} + \sum_{j\neq k}\mathbf{R}_{j,\text{eff}}\big)^{-1}\mathbf{R}_{k,\text{eff}}\Big). \tag{4.11}$$

Subject to the total transmit power constraint, the digital precoder is then normalised as $\mathbf{w}_k = \frac{\mathbf{w}_k}{\|\mathbf{F}\mathbf{w}_k\|}$. The 'OSF + Stat' hybrid precoding scheme is summarised in Algorithm 4.1. Next, we analyse the achievable sum rate in single-path channels and in the large array regime respectively. The analysis in these special cases gives insights into the rate performance of more general settings.

1) Single-path channels: When $L_k = 1, \forall k$, the channel covariance matrix $\mathbf{R}_k = M\mathbf{a}_k\mathbf{a}_k^H$ is rank one. Equivalently, the direction of the instantaneous channel is perfectly known at the BS. The composite channel matrix can be written as $\mathbf{H} = \mathbf{A} \cdot M\text{diag}\{g_1, \cdots, g_K\}$ with $\mathbf{A} \triangleq [\mathbf{a}_1, \cdots, \mathbf{a}_K]$. To maximise the desired signal power, the BS steers beams to each user via matched beamforming, i.e., $\mathbf{F} = \mathbf{A}$. Then, the effective channel matrix becomes $\bar{\mathbf{H}}_{\text{eff}} = \mathbf{A}^H\mathbf{A}$ and the digital precoder is designed as ZF, i.e., $\mathbf{W} = \bar{\mathbf{H}}_{\text{eff}}^{-1} = (\mathbf{A}^{\mathbf{H}}\mathbf{A})^{-1}$. The achievable rate of user k is given by

$$\begin{aligned} R_k &= \log_2\Big(1 + \frac{\rho M|g_k|^2 \cdot |\mathbf{a}_k^H\mathbf{F}\mathbf{w}_k|^2}{1 + \rho M|g_k|^2 \cdot \sum_{j\neq k}|\mathbf{a}_k^H\mathbf{F}\mathbf{w}_j|^2}\Big) && (4.12)\\ &= \log_2\Big(1 + \frac{\rho M|g_k|^2}{(\mathbf{A}^{\mathbf{H}}\mathbf{A})^{-1}_{k,k}}\Big), && (4.13)\end{aligned}$$

which coincides with [17, Theorem 1]. [17] assumes continuous angles of the RF beam-steering vectors and perfect effective channel knowledge, i.e., both the RF codebooks and the effective channel quantisation codebooks are assumed with infinite resolution. By contrast, we assume perfect statistical CSIT. It implies that in single-path mmWave channels, second-order channel statistics are sufficient to achieve the same rate as that achievable by infinite resolution codebooks.

2) Large array regime: For tractability, the rate analysis of multi-path channel is based on the assumption of a ULA with a large scale transmit antennas. In this case, the channel model (4.1) can be well approximated by virtual channel representation [5, Ch. 3]

$$\mathbf{h}_k = \sqrt{\frac{M}{L}} \sum_{l=1}^{L} g_{k,l}\, \mathbf{e}_{k,l} = \sqrt{\frac{M}{L}} \mathbf{E}_k \mathbf{g}_k, \tag{4.14}$$

where $L_k = L, \forall k$ is assumed for simplicity. The steering vectors $\mathbf{E}_k$ contains L columns of the DFT matrix $\mathbf{E}$. Without loss of generality, we assume the path gains in descending order $|g_{k,1}| \geq ... \geq |g_{k,L}|$. The corresponding channel covariance matrix is expressed as

$$\mathbf{R}_k = \frac{M}{L} \mathbf{E}_k \mathbf{E}_k^H = \frac{M}{L} \mathbf{E} \mathbf{D}_k \mathbf{E}^H, \tag{4.15}$$

where $\mathbf{D}_k = \text{diag}\{\mathbf{d}_k\}$, $\mathbf{d}_k \in \{0, 1\}$ and $\|\mathbf{d}_k\|_0 = L$ (i.e., $\text{rank}(\mathbf{R}_k) = L$). Below, we analyse the achievable rate of the proposed 'OSF + Stat' scheme in two special cases.

Non-overlapped: the AoD spread and user locations are well separated such that users exhibit mutually non-overlapped multi-paths.

Fully overlapped: users are confined into an area such that all users share the same AoDs and thus the same channel covariance matrix $\mathbf{R}$.

Proposition 4.1. *Based on* (4.11)*, the achievable rate of user k of the proposed 'OSF + Stat' hybrid precoding scheme in the large array regime is respectively given as*

$$R_k = \log_2 \left(1 + \rho \frac{M}{L} |g_{k,1}|^2\right) \tag{4.16}$$

in the non-overlapped scenario, and as

$$R_k \geq \log_2 \left(1 + \frac{\rho \frac{M}{L} |g_{k,1}|^2}{1 + \rho \frac{M}{L} \sum_{l=2}^{L} |g_{k,l}|^2}\right) \tag{4.17}$$

in the fully overlapped scenario.

Proof. See Appendix C.1. □

Remark 4.1. *When the AoDs of each user are non-overlapped,* (4.16) *shows that statistical CSIT is able to completely remove the multiuser interference.* (4.16) *also serves as the interference-free per-user rate R_k^s with $\rho = P/K$. When the channels are very sparse (i.e., small L) or have dominant path (i.e., $|g_1| \gg |g_l|, \forall l \geq 2$), we observe in* (4.17) *that the multiuser interference term $\sum_{l=2}^{L} |g_{k,l}|^2$ can be neglected at practical SNR. It implies that the long-term channel statistics enables efficient interference nulling in above scenarios. By contrast, the closed-form achievable rate in partially overlapped AoDs scenario cannot be easily attained due to the complicated structure of $\mathbf{w}_k$ in* (4.11)*. Nevertheless, we note that R_k is upper and lower bounded by* (4.16) *and* (4.17)*, respectively. To get insights into the rate performance of 'OSF + Stat' in partially overlapped AoDs scenario, we quantify the rate gap between* (4.16) *and* (4.17) *as follows.*

Corollary 4.1. *For the proposed 'OSF + Stat' hybrid precoding scheme in the large array regime, the average rate gap ΔR_k between the worst-case (fully overlapped AoDs) and best-case (non-overlapped AoDs) is upper bounded by*

$$\Delta R_k \leq \log_2\Big(1 + \rho M \frac{L-1}{L}\Big). \tag{4.18}$$

Proof. According to (4.16) and (4.17), we have

$$\begin{aligned}\Delta R_k &\leq \mathbb{E}\Big[\log_2\big(1+\rho\frac{M}{L}|g_{k,1}|^2\big) + \log_2\Big(1+\rho\frac{M}{L}\sum_{l=2}^{L}|g_{k,l}|^2\Big) - \log_2\Big(1+\rho\frac{M}{L}\sum_{l}|g_{k,l}|^2\Big)\Big] \\ &\leq \mathbb{E}\Big[\log_2\Big(1+\rho\frac{M}{L}\sum_{l=2}^{L}|g_{k,l}|^2\Big)\Big]. \end{aligned} \tag{4.19}$$

By further using Jensen's inequality, (4.18) is obtained. □

Corollary 4.1 shows the upper bound on the rate gap between arbitrary overlapped scenarios and interference-free single-user transmission. This rate gap upper bound increases with ρ, which suggests that the statistical CSIT can barely mitigate multiuser interference in the fully overlapped AoDs case. Nevertheless, the AoDs of users in mmWave systems are sparse and randomly distributed. Thus, AoDs are unlikely to be heavily overlapped. The proposed 'OSF + Stat' hybrid precoding scheme can still effectively mitigate the multiuser interference, as shown in Section 4.4.1.

4.2.2. Two-Stage Feedback Scheme

Under a fixed total feedback overhead constraint, the 'TSF + Adp CB' hybrid precoding requires to partition the feedback into two stages. The second-stage feedback (B_{BB}) is used to effectively quantise the channel $\mathbf{h}_{k,\text{eff}}$. Rather than using RVQ[2] codebook as in [17], we employ a statistical CSIT adaptive codebook which has been intensively discussed in the literature [32, 41]. Then, the digital precoder is computed based on the effective channel quantisation and thus more capable to mitigate multiuser interference than statistical CSIT-based digital precoder in 'OSF + Stat' scheme. However, the residual feedback ($B_{RF} = B - B_{BB}$) can only enable coarse analog beamforming and may lead to an undesired rate performance.

The RF beamforming follows (4.7) with $|\mathcal{F}| = 2^{B_{RF}}$ and the effective channel is given as $\mathbf{h}_{k,\text{eff}} = \mathbf{F}^H\mathbf{h}_k$. The channel quality information (CQI) (i.e., the effective channel gain $\|\mathbf{h}_{k,\text{eff}}\|$) is assumed perfectly known to the BS while the channel direction information (CDI) $\bar{\mathbf{h}}_{k,\text{eff}} = \mathbf{h}_{k,\text{eff}}/\|\mathbf{h}_{k,\text{eff}}\|$ is quantised, denoted as $\hat{\bar{\mathbf{h}}}_{k,\text{eff}}$, and then fed back to the BS. A classical channel matching metric is adopted for the effective channel quantisation, where

[2]RVQ is known as the asymptotically optimal codebook in the large array regime when used for beamforming in an i.i.d. Rayleigh fading channel [89]. However, the effective channels at the mmWave bands are unlikely to be i.i.d. due to the highly directional link.

each user selects the best codeword that maximises the inner-product with its channel direction, i.e.,

$$\widehat{\bar{\mathbf{h}}}_{k,\text{eff}} = \arg\max_{\mathbf{c}_k \in \mathcal{C}_k} \ |\bar{\mathbf{h}}_{k,\text{eff}}^H \mathbf{c}_k|^2. \tag{4.20}$$

According to (4.2), the geometric multi-path channel model (4.1) can be written as spatially correlated model $\mathbf{h}_k = \mathbf{R}_k^{1/2}\mathbf{g}_k$. Then, the effective channel can be equivalently treated as $\mathbf{h}_{k,\text{eff}} = \mathbf{R}_{k,\text{eff}}^{1/2}\,\mathbf{g}_k$ with positive semi-definite (PSD) $\mathbf{R}_{k,\text{eff}} = \mathbf{F}^H\mathbf{R}_k\mathbf{F}$. In this setting, the channel space is no longer a hypersphere in $\mathbb{C}^K$ but a hyperellipse stretched by the eigenvalues of $\mathbf{R}_{k,\text{eff}}$. Hence, a skewed codebook is preferable which multiplies $\mathcal{C}_{\text{iid}} = \{\mathbf{c}_1, \cdots, \mathbf{c}_{2^{B_{BB}}}\}$ (that is specified to i.i.d. channels) to $\mathbf{R}_{k,\text{eff}}^{1/2}$ [90], i.e.,

$$\mathcal{C}_k = \left\{ \frac{\mathbf{R}_{k,\text{eff}}^{1/2}\,\mathbf{c}_i}{\|\mathbf{R}_{k,\text{eff}}^{1/2}\,\mathbf{c}_i\|} \right\}. \tag{4.21}$$

With the knowledge of $\widehat{\bar{\mathbf{h}}}_{k,\text{eff}}$, $\mathbf{W}$ can be designed as the ZF precoding, i.e.,

$$\mathbf{W} = \widehat{\bar{\mathbf{H}}}_{\text{eff}}\,(\widehat{\bar{\mathbf{H}}}_{\text{eff}}^H\,\widehat{\bar{\mathbf{H}}}_{\text{eff}})^{-1}, \tag{4.22}$$

where $\widehat{\bar{\mathbf{H}}}_{\text{eff}} = [\widehat{\bar{\mathbf{h}}}_{1,\text{eff}}, \cdots, \widehat{\bar{\mathbf{h}}}_{K,\text{eff}}]$. Moreover, inspired by the statistical CSIT SLNR-based design of $\mathbf{W}$ in Section 4.2.1, we here derive an imperfect CSIT SLNR-based design for a fair comparison between the proposed two schemes. Briefly, the SLNR based on the quantised CSIT $\widehat{\mathbf{h}}_{k,\text{eff}} = \|\mathbf{h}_{k,\text{eff}}\| \cdot \widehat{\bar{\mathbf{h}}}_{k,\text{eff}}$ writes as

$$\text{SLNR}_k = \frac{\rho\,|\widehat{\mathbf{h}}_{k,\text{eff}}^H \mathbf{w}_k|^2}{1 + \rho \sum_{j\neq k} |\widehat{\mathbf{h}}_{j,\text{eff}}^H \mathbf{w}_k|^2}, \tag{4.23}$$

and the optimal unit norm $\mathbf{w}_k$ that maximises (4.23) is given by

$$\begin{aligned} \mathbf{w}_k &= \mathbf{u}_{\max}\Big(\big(\frac{1}{\rho}\mathbf{I} + \sum_{j\neq k} \widehat{\mathbf{h}}_{j,\text{eff}}\widehat{\mathbf{h}}_{j,\text{eff}}^H\big)^{-1} \widehat{\mathbf{h}}_{k,\text{eff}}\widehat{\mathbf{h}}_{k,\text{eff}}^H \Big) \\ &= \Big(\frac{1}{\rho}\mathbf{I} + \sum_{j\neq k} \widehat{\mathbf{h}}_{j,\text{eff}}\widehat{\mathbf{h}}_{j,\text{eff}}^H \Big)^{-1} \widehat{\mathbf{h}}_{k,\text{eff}}. \end{aligned} \tag{4.24}$$

Then, we normalise it as $\mathbf{w}_k = \frac{\mathbf{w}_k}{\|\mathbf{F}\mathbf{w}_k\|}$. The 'TSF + Adp CB' hybrid precoding scheme is summarised in Algorithm 4.1. Next, we analyse the effect of B_{BB} on the sum DoF (also referred to as multiplexing gain). Based on the design of (4.20) $\sim$ (4.22), when B_{BB} is small and independent of SNR, the achievable rate of the system eventually saturates (i.e., DoF = 0). In order to achieve a non-zero DoF, B_{BB} should scale at a rate linearly increasing with SNR as follows.

Proposition 4.2. *If B_{BB} is scaled as $B_{BB} = O(\alpha \log_2 P)$ for $\alpha \leq r-1$, the 'TSF + Adp CB' hybrid precoding scheme achieves at least a sum DoF of $K \cdot \frac{\alpha}{r-1}$.*

Proof. As the analog beamformer is determined by (4.7), the effective channel can be equivalently treated as spatially correlated channel $\mathbf{h}_{k,\text{eff}} = \mathbf{R}_{k,\text{eff}}^{1/2}\,\mathbf{g}_k$. Then, the 'TSF + Adp CB' hybrid precoding scheme boils down to the conventional digital precoding scheme. According to [91, Lemma 1], [41], the expected quantisation error of the effective channel can be upper bounded by $2^{-B_{BB}/(r-1)}$, where r is the rank of the $K \times K$ covariance matrix of the effective channel $\mathbf{R}_{k,\text{eff}}$ and $r \leq K$. By straightforwardly following the proof of [36, Theorem 4], we can lower bound the DoF of each user by $\frac{\alpha}{r-1}$ for $B_{BB} = O(\alpha \log_2 P)$. □

Intuitively, in the asymptotical SNR regime $P \to \infty$, the signal power grows linearly with P while the interference power scales with the product of P and the quantisation error. Since the quantisation error is of the order $2^{-\frac{B_{BB}}{r-1}} = P^{-\frac{\alpha}{r-1}}$, the interference power scales as $P^{(1-\frac{\alpha}{r-1})}$ which gives a SINR that scales as $P^{\frac{\alpha}{r-1}}$. Thus, the resulting DoF of each user is $\frac{\alpha}{r-1}$. In order to obtain a sum DoF of $m (m \leq K)$, the number of second-stage feedback bits should scale with SNR

$$B_{BB} = O\Big(m \frac{r-1}{K} \frac{P_{dB}}{3}\Big). \tag{4.25}$$

4.2.3. One-stage vs. Two-stage Feedback Scheme

An analytical comparison between one-stage and two-stage feedback schemes is intractable due to intractability of a closed-form sum rate expression for the 'TSF + Adp CB' scheme. Nevertheless, we can get some insights based on the analysis and discussion in Sections 4.2.1 and 4.2.2. Specifically, we consider the following scenarios: 1. AoDs of each user are well separated; 2. very sparse channel (i.e., $L \to 1$); 3. dominant path (i.e., $|g_1| \gg |g_l|, \forall l \geq 2$); 4. very limited feedback overhead B.

From Proposition 4.1 and Remark 4.1, the multiuser interference can be effectively mitigated by the long-term channel statistics-based digital precoder in scenarios 1-3. The per-user rate of the 'OSF + Stat' scheme approaches the per-user interference-free rate. Thus, the 'OSF + Stat' scheme with $B_{RF} = B$ achieves higher rate than the 'TSF + Adp CB' scheme with $B_{RF} = B - B_{BB}$. In scenario 4, the splitting operation in 'TSF + Adp CB' scheme results in very small B_{RF} and B_{BB}. From (4.7), small B_{RF} leads to coarse RF beamsteering, i.e., mismatch between the RF beamforming and the channel direction. Additionally, small B_{BB} incurs an inaccurate feedback of the effective channel knowledge because the quantisation error is of the order $2^{-\frac{B_{BB}}{r-1}}$. Thus, the 'TSF + Adp CB' scheme with very limited B yields an unfavourable rate performance. In a nutshell, the 'OSF + Stat' scheme outperforms the 'TSF + Adp CB' counterpart in scenarios 1-4 in term of rate performance. Meanwhile, the 'OSF + Stat' scheme highly reduces the complexity of the signalling and feedback procedure by eliminating the second-stage feedback.

On the contrary, the 'TSF + Adp CB' hybrid precoding scheme exceeds the 'OSF + Stat' scheme in the regime of large B, where the 'TSF + Adp CB' scheme with par-

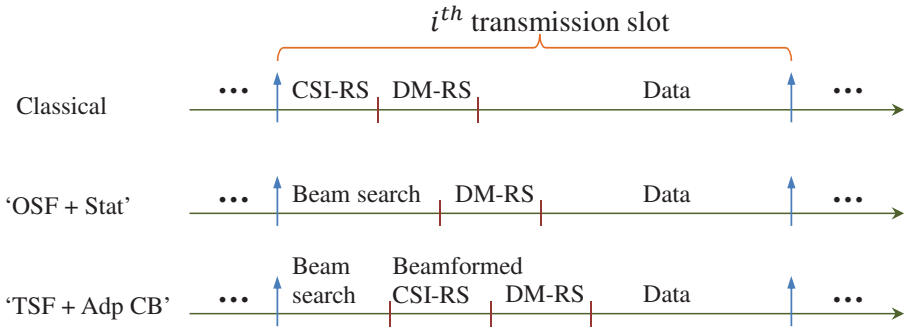

Figure 4.2.: Illustration of signalling and feedback procedure for various schemes.

tial resources B_{RF} is sufficient to provide precise first-stage beamforming. The residual resources B_{BB} gives an accurate channel quantisation (i.e., $2^{-\frac{B_{BB}}{r-1}} \to 0$), resulting in efficient interference elimination. By contrast, the 'OSF + Stat' scheme using all feedback overhead B for the first-stage RF beamsteering can only achieve a marginal RF beamforming gain over the 'TSF + Adp CB' scheme using B_{RF}. Meanwhile, the 'OSF + Stat' scheme based on statistical CSIT-based digital precoder is less capable of mitigating interference.

In Section 4.4, a simulated comparison is provided to show the efficacy of the 'OSF + Stat' scheme over 'TSF + Adp CB' in various aforementioned scenarios. In order to fairly compare them, the optimal feedback allocation between the two stages is numerically computed for the 'TSF + Adp CB' scheme[3].

Moreover, we note that the beam search in the 'OSF + Stat' scheme with $B_{RF} = B$ takes longer than the 'TSF + Adp CB' scheme with $B_{RF} = B - B_{BB}$. Nevertheless, there is not a big gap between them when the total feedback is limited and the feedback allocation of the 'TSF + Adp CB' scheme is optimized, as confirmed by the simulations in Section 4.4.

4.2.4. Signalling and Feedback Protocol

To operate the proposed hybrid precoding schemes, the signalling and feedback procedure is illustrated in Fig. 4.2 and described as follows.

Consider the classical microwave systems with M transmit antennas and $N = M$ RF chains at the BS. Using LTE-A framework [79], CSI-RS are transmitted to enable the users to measure the instantaneous CSI, which is then fed back to the BS. With this channel knowledge, the BS computes the precoders. Then, the BS constructs the transmitted signals that are transmitted along DM-RS [43] to enable the users to detect the desired signal.

In the mmWave systems with one-stage feedback scheme, we first search the beam that

[3]In general, the analytical computation of the optimum feedback allocation between two stages given certain total feedback overhead is not trivial and is beyond the scope of the present work.

maximises the desired signal of each user to overcome the severe pathloss of mmWave link. Based on the analog beamformer and the channel covariance matrix, the BS designs the digital precoder and transmits the DM-RS.

In the mmWave systems with two-stage feedback scheme, beam search is first operated and the beam that maximises the desired signal of each user is fed back to the BS. Then, the BS transmits the beamformed CSI-RS to the users. Each user estimates and reports the effective CSI to the BS. Based on the analog beamformer and the quantised CSIT, the BS determines the digital precoder and then transmits the DM-RS.

4.3. Rate Splitting with Hybrid Precoding

In previous section, the achievable rate of the conventional transmission strategy is highly degraded by multiuser interference due to either AoDs overlap in 'OSF + Stat' scheme or limited feedback in 'TSF + Adp CB' scheme. In the context of multiuser mmWave systems, we introduce a rate splitting (RS) transmission strategy to tackle the residual interference. RS enhances the rate performance and can be applied with both aforementioned hybrid precoding schemes.

With B_{RF} for the analog beamforming, we consider the conventional (i.e., No-RS) transmission strategy with the 'OSF + Stat' hybrid precoding as a baseline. On one side, RS with the 'OSF + Stat' hybrid precoding achieves rate gain over the baseline owing to the benefits of RS. On the other side, by using the same B_{RF} for the analog beamforming and extra feedback B_{BB} for the second-stage channel quantisation, the conventional No-RS with the 'TSF + Adp CB' hybrid precoding also enables rate gain over the baseline. In this section, we mainly investigate how much second-stage feedback can be saved by implementing RS.

The RS transmission strategy superposes a common message on top of all users' private messages. Thus, the conventional No-RS strategy is a sub-scheme of RS. Compared with (4.4), the transmitted signal of RS can be written as

$$\mathbf{x} = \sqrt{P_c}\,\mathbf{F}\,\mathbf{w}_c s_c + \sum_{k=1}^{K} \sqrt{P_k}\,\mathbf{F}\,\mathbf{w}_k\, s_k, \tag{4.26}$$

where $\mathbf{w}_c$ is the precoding vector of the common message s_c. In (4.5), the private message transmissions are interference-limited at high SNR. The basic idea of RS is to transmit the private messages with a fraction of the total power such that the private messages are decoded in the non-interference-limited SNR regime. In addition, a common message is transmitted using the remaining power which gives rise to a rate enhancement [92].

In line with the conventional No-RS strategy, the RF beamformer $\mathbf{F}$ and the digital precoder $\mathbf{w}_k$ are designed as in Section 4.2.1 while uniform power allocation is performed on the private messages. We mainly focus on the power splitting between the common and

private messages and the precoder design of the common message. A fraction $t \in (0, 1]$ of the total power is uniformly allocated to the private messages while the remaining power is given to the common message, i.e., $P_k = Pt/K$ and $P_c = P(1-t)$. At the user side, each user decodes first the common message by treating all private messages as noise. After removing the decoded common message by SIC, each user decodes its own private message. Thus, the SINRs of the common message and the private message experienced by user k are written as

$$\mathrm{SINR}_k^c = \frac{P_c\,|\mathbf{h}_k^H \mathbf{F}\mathbf{w}_c|^2}{1+\sum_{j=1}^{K} P_j\,|\mathbf{h}_k^H \mathbf{F}\mathbf{w}_j|^2}, \tag{4.27}$$

$$\mathrm{SINR}_k^p = \frac{P_k\,|\mathbf{h}_k^H \mathbf{F}\mathbf{w}_k|^2}{1+\sum_{j\neq k} P_j\,|\mathbf{h}_k^H \mathbf{F}\mathbf{w}_j|^2}. \tag{4.28}$$

The achievable rate of the common message is given as

$$R^c = \min_k \{R_k^c\} = \min_k \big\{\log_2(1+\mathrm{SINR}_k^c)\big\}, \tag{4.29}$$

which guarantees that the common message can be successfully decoded by all users. The sum rate of the private messages is given as $R^p = \sum_{k=1}^{K} R_k^p = \sum_{k=1}^{K} \log_2(1+\mathrm{SINR}_k^p)$. Then, the sum rate of RS is $R_{\mathrm{sum}}^{RS} = R^c + R^p$. In order to properly design t and $\mathbf{w}_c$, we need to derive the sum rate of RS.

Proposition 4.3. *The average sum rate of RS with the 'OSF + Stat' hybrid precoding is lower bounded as*

$$\begin{aligned}\mathbb{E}(R_{sum}^{RS}) \;\geq\;& \min_k\left\{\log_2\left(1+\frac{e^{-\gamma}\cdot P(1-t)\,\mathbf{w}_c^H\mathbf{R}_{k,eff}\mathbf{w}_c}{1+\frac{Pt}{K}\sum_{j=1}^{K}\mathbf{w}_j^H\mathbf{R}_{k,eff}\mathbf{w}_j}\right)\right\}\\ &+\sum_{k=1}^{K}\log_2\left(1+\frac{e^{-\gamma}\cdot\frac{Pt}{K}\mathbf{w}_k^H\mathbf{R}_{k,eff}\mathbf{w}_k}{1+\frac{Pt}{K}\sum_{j\neq k}\mathbf{w}_j^H\mathbf{R}_{k,eff}\mathbf{w}_j}\right).\end{aligned} \tag{4.30}$$

Proof. See Appendix C.2. □

According to (4.30), the precoder ($\mathbf{w}_c$) and the power splitting ratio (t) can be designed.

4.3.1. Precoder design

With $\mathbf{R}_{k,\mathrm{eff}}$ known at the BS and $\mathbf{w}_j$ designed as (4.11), $\mathbf{w}_c$ can be optimised by solving the following max-min problem subject to a power constraint

$$\mathcal{P}1: \quad \max_{\mathbf{w}_c}\ \min_k\ \frac{1}{\beta_k}\mathbf{w}_c^H\mathbf{R}_{k,\mathrm{eff}}\mathbf{w}_c \tag{4.31}$$

$$\text{s.t.}\quad \mathbf{w}_c^H\,\mathbf{M}\,\mathbf{w}_c = 1, \tag{4.32}$$

1:	**Initialisation**: set $i = 0$ and generate an initial point $\mathbf{z}_0$
2:	**Repeat**
3:	solve (4.36) with $\mathbf{z} = \mathbf{z}_i$ and denote $\mathbf{x}^\star$ as the solution
4:	update $\mathbf{z}_{i+1} = \mathbf{x}^\star$
5:	set $i = i + 1$
6:	**Until convergence**

Table 4.2.: SCA algorithm

where $\mathbf{M} = \mathbf{F}^H\mathbf{F}$ and $\beta_k = 1 + \frac{Pt}{K}\sum_{j=1}^{K}\mathbf{w}_j^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_j$. By replacing $\mathbf{w}_c$ with $\mathbf{x}$ and introducing a slack variable, $\mathcal{P}1$ is equivalently reformulated as

$$\mathcal{P}2: \quad \max_{\mathbf{x},\,t} \; t \tag{4.33}$$

$$\text{s.t.} \quad \mathbf{x}^H\mathbf{R}_{k,\text{eff}}\mathbf{x} \geq \beta_k t \tag{4.34}$$

$$\mathbf{x}^H\,\mathbf{M}\,\mathbf{x} = 1,$$

which is known as NP-hard and can be solved by semi-definite relaxation (SDR) followed by Gaussian randomisation [3]. Rather, we cope with $\mathcal{P}2$ by successive convex approximation (SCA) approach [24] for its lower complexity. In addition, SCA converges to a KKT stationary point for the original problem $\mathcal{P}1$ [93]. For any $\mathbf{z}$ and PSD $\mathbf{R}_{k,\text{eff}}$, we have $(\mathbf{x}-\mathbf{z})^H\mathbf{R}_{k,\text{eff}}\,(\mathbf{x}-\mathbf{z}) \geq 0$ and therefore

$$\mathbf{x}^H\mathbf{R}_{k,\text{eff}}\,\mathbf{x} \geq 2\,\text{Re}\{\mathbf{z}^H\mathbf{R}_{k,\text{eff}}\,\mathbf{x}\} - \mathbf{z}^H\mathbf{R}_{k,\text{eff}}\,\mathbf{z}. \tag{4.35}$$

Substituting the convex constraint of (4.35) into (4.34) leads to the following convex problem

$$\mathcal{P}3: \quad \max_{\mathbf{x},\,t} \; t$$

$$\text{s.t.} \quad 2\,\text{Re}\{\mathbf{z}^H\mathbf{R}_{k,\text{eff}}\,\mathbf{x}\} - \mathbf{z}^H\mathbf{R}_{k,\text{eff}}\,\mathbf{z} \geq \beta_k t \tag{4.36}$$

$$\mathbf{x}^H\,\mathbf{M}\,\mathbf{x} = 1,$$

where the optimal solution $\mathbf{x}^\star$ can be efficiently obtained for given $\mathbf{z}$. Then, $\mathbf{z}$ is iteratively updated by $\mathbf{z} = \mathbf{x}^\star$ and used in the next iteration, yielding a sequence of feasible solutions with non-increasing objective values. The SCA algorithm is summarised in Algorithm 4.2.

4.3.2. Power allocation design

The optimal power splitting ratio t can be determined by maximising the lower bound on average sum rate (4.30) with line search. By contrast, we compute a suboptimal but effective and insightful power allocation. Recall that the common message in RS is dedicated to overcome multiuser interference and its achievable rate is subject to a

1:	BS: sets $\mathbf{F}$ from (4.7) and $\mathbf{R}_{k,\text{eff}} = \mathbf{F}^H \mathbf{R}_k \mathbf{F}$
	computes $\mathbf{w}_k$ from (4.11) and t from (4.37)
2:	If $t < 1$
3:	BS: determines $\mathbf{w}_c$ as the solution of $\mathcal{P}1$ and constructs the TX signal as (4.26)
4:	Users: decode the common message $\rightarrow$ SIC $\rightarrow$ their own private messages
5:	Else
6:	BS: constructs the TX signal as (4.4)
7:	Users: decode directly their own private messages

Table 4.3.: Overall RS transmission with hybrid precoding

minimum constraint (4.29). In the low/non-interference-limited SNR regime, private messages transmission works well and therefore the total power is allocated to the private messages, i.e., $t = 1$. In this case, RS turns into conventional No-RS.

At high SNR, exploiting full power to transmit the private messages only offers marginal sum rate gain by virtue of multiuser interference. Thus, only a fraction of the total power is assigned to the private messages such that the private message decoding can be placed back to the non-interference-limited SNR regime. The basic idea of power allocation for RS is to compute the saturation point of the achievable rate of private message, where the multiuser interference becomes the dominant factor for the performance. Any power beyond is reserved for the common message. Specifically, in (4.30), the interference term of the private message $\Upsilon \triangleq \frac{Pt}{K} \sum_{j \neq k} \mathbf{w}_j^H \mathbf{R}_{k,\text{eff}} \mathbf{w}_j$ and the noise power is 1. The saturation point lies in the regime $\Upsilon > 1$ and can be computed by letting $\Upsilon = K$ for simplicity[4]. Overall, the power splitting ratio between the common and the private messages for arbitrary SNR is therefore designed as

$$t = \min \left\{ \frac{K}{P\Gamma}, \, 1 \right\}, \tag{4.37}$$

where $\Gamma = \min_k \left\{ \frac{1}{K} \sum_{j \neq k} \mathbf{w}_j^H \mathbf{R}_{k,\text{eff}} \mathbf{w}_j \right\}$. The minimum is imposed to guarantee that $\frac{Pt}{K} \sum_{j \neq k} \mathbf{w}_j^H \mathbf{R}_{k,\text{eff}} \mathbf{w}_j \gg 1$ for $\forall k$. The overall RS transmission strategy is presented in Algorithm 4.3.

Remark 4.2. *In general scenarios such as partially overlapped AoDs between users, the sum rate of the baseline (i.e., conventional No-RS strategy with the 'OSF + Stat' hybrid precoding scheme) is interference-limited at high SNR. By contrast, in the 'OSF + Stat' precoded RS strategy, the power allocation* (4.37) *guarantees that the private message transmission of RS achieves almost the same sum rate as the baseline. Meanwhile, the common message transmission of RS with the remaining power leads to rate enhancement which can be quantified as* R_c (4.29)*. Therefore, the sum DoF gain of the 'OSF + Stat' precoded RS over the baseline approaches 1.*

[4]Apparently, the value of Υ satisfying $\Upsilon > 1$ is not unique. Nevertheless, the effectiveness of choosing $\Upsilon = K$ has been demonstrated by [92].

Moreover, the 'TSF + Adp CB' precoded conventional No-RS strategy also enables better performance than the baseline by virtue of extra second-stage feedback. Based on (4.25), *'TSF + Adp CB' requires extra feedback scaling as* $B_{BB} = O(\frac{r-1}{K}\frac{P_{dB}}{3})$ *to achieve a sum DoF of 1. As the rank* r *of the effective channel covariance matrix* $\mathbf{R}_{k,eff} = \mathbf{F}^H\mathbf{R}_k\mathbf{F} \in \mathbb{C}^{K\times K}$ *is constrained to the channel sparsity* L*, we have* $r = \min(L, K)$*. Consequently, the 'OSF + Stat' precoded RS can achieve the same sum DoF as the 'TSF + Adp CB' precoded conventional No-RS that is driven by extra second-stage feedback*[5]*. It implies that RS enables significant saving both in the downlink training and in the CSIT uplink feedback.*

4.4. Performance Evaluation

Numerical results are provided to validate the effectiveness of the proposed hybrid precoding schemes and RS strategy. We consider a system model with ULA equipped at the BS and a channel model with AoDs uniformly distributed in $[0, \pi]$. The codebook $\mathcal{C}_{\text{iid}}$ is designed using Grassmannian line packing.

4.4.1. Multiuser Hybrid Precoding

This section simulates the conventional No-RS transmission strategy in a setup with $M = 64, K = 4, B = 6$. In Fig. 4.3 and Fig. 4.4, we evaluate the performance of the proposed one-stage feedback plus channel statistics-based ('OSF + Stat') and two-stage feedback plus adaptive codebook-based ('TSF + Adp CB') hybrid precoding schemes. The 'OSF + Stat' scheme uses all the feedback bits for the analog beamforming, i.e., $B_{RF} = B, B_{BB} = 0$. By contrast, in 'TSF + Adp CB' scheme, the total feedback overhead B is divided into the analog beamsteering (B_{RF}) and the second-stage channel quantisation (B_{BB}). The optimal feedback division ratio is numerically computed and the corresponding sum rate of 'TSF + Adp CB' is plotted. The effectiveness of the 'OSF + Stat' scheme is verified by comparing with the 'optimised' 'TSF + Adp CB' scheme.

Both SLNR-based and ZF digital precoder designs are considered. The **baseline** ('TSF + RVQ + ZF' [17]) is plotted where RVQ codebook is used to quantise the effective channel and the BS designs the digital precoder as ZF based on the quantised channels. Fig. 4.3 and Fig. 4.4 show that the proposed hybrid precoding schemes largely outperform the baseline due to the exploitation of the channel statistics. It implies that adaptive codebook captures the characteristics of the mmWave channel better than RVQ. Moreover, SLNR-based digital precoder design that maximises a lower bound on the averaged SLNR enables higher rate than ZF, since it takes into account both the desired signal and the interference.

The optimal splitting of the total feedback overhead $B = 6$ in 'TSF + Adp CB' scheme is numerically obtained as $B_{RF} = 5, B_{BB} = 1$. In Fig. 4.3, it can be seen that the

[5] It is worth noting that simply doing TDMA achieves DoF of 1. However, the rate gain of RS over TDMA has been demonstrated in [37, 92, 94].

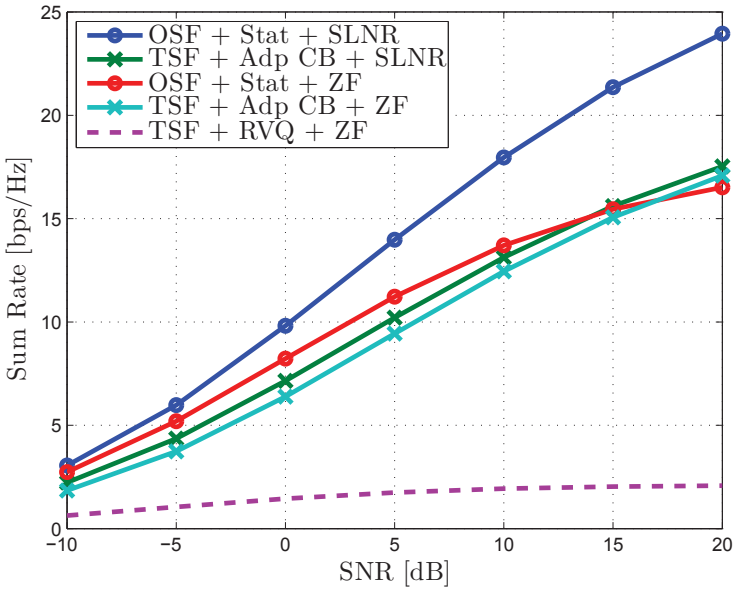

Figure 4.3.: Comparison among various hybrid precoding schemes, in a setup with $B = 6, L = 2$.

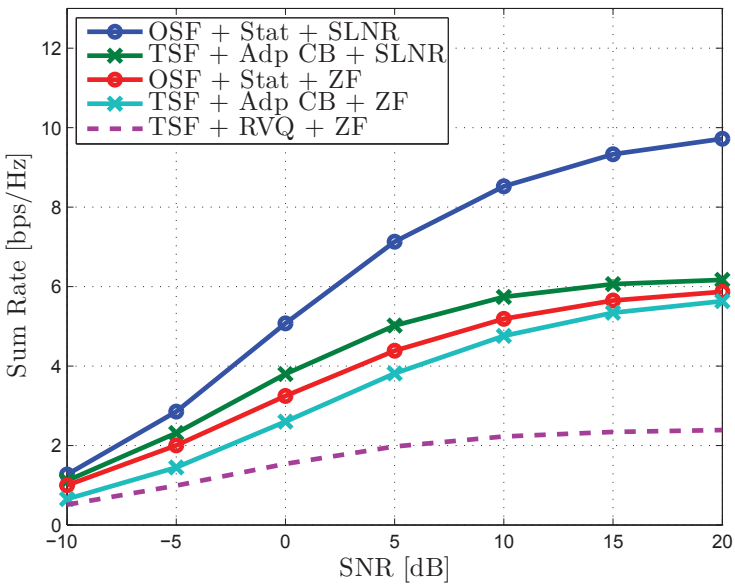

Figure 4.4.: Comparison among various hybrid precoding schemes, in a setup with $B = 6, L = 15$.

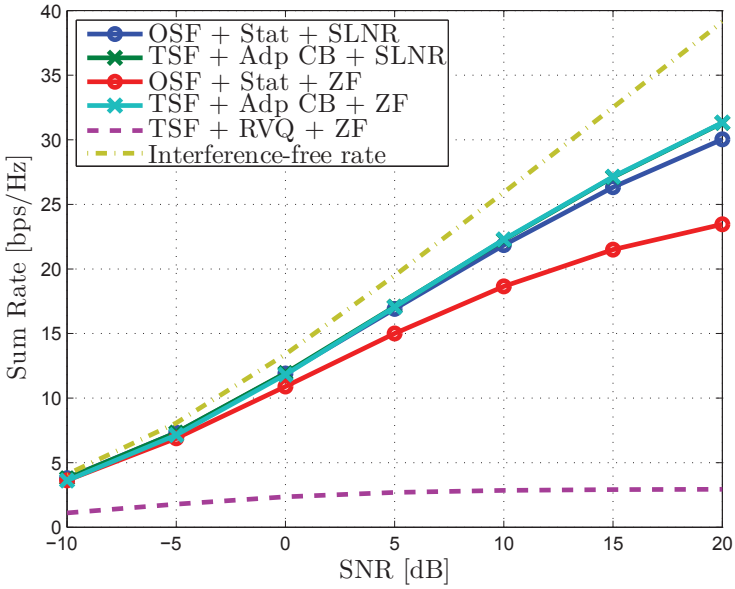

Figure 4.5.: Comparison among various hybrid precoding schemes, in a setup with $B = 10, L = 2$.

proposed 'OSF + Stat + SLNR' scheme achieves much higher sum rate than the 'TSF + Adp CB' scheme. The 'OSF + Stat' scheme utilises all feedback resources to design precise RF beamforming. Meanwhile, the channel covariance matrix is low rank due to channel sparsity ($L = 2$) and leaves nullspace to efficiently mitigate multiuser interference. In Fig. 4.4 where $L = 15$, the rate performance of the 'OSF + Stat' scheme still exceeds the 'TSF + Adp CB' scheme. In 'TSF + Adp CB' scheme, the optimal splitting of total feedback $B = 6$ is numerically obtained as $B_{RF} = 4, B_{BB} = 2$. When a limited amount of feedback bits $B = 6$ is attainable, the splitting operation in 'TSF + Adp CB' scheme leads to coarse analog beamsteering and inaccurate channel quantisation. Therefore, these observations in Fig. 4.3 and Fig. 4.4 verify the comments drawn in Section 4.2.3. Namely, the proposed 'OSF + Stat' scheme exhibits substantial sum rate gain over various baselines for very limited feedback system/very sparse channels.

In Fig. 4.5 and Fig. 4.6, we consider the same setup as Fig. 4.3 but a larger feedback overhead $B = 10$. The sum of interference-free per-user rate $R_{\text{sum}} = \sum_{k=1}^{K} R_k^s$, where R_k^s is given as (4.16), is plotted as an upper bound on the sum rate. In 'TSF + Adp CB' scheme, the optimal splitting of total feedback $B = 10$ is numerically obtained as $B_{RF} = 8, B_{BB} = 2$ in Fig. 4.5 and $B_{RF} = 7, B_{BB} = 3$ in Fig. 4.6. Fig. 4.5 shows that the proposed 'OSF + Stat + SLNR' scheme obtains a rate performance similar to that of the 'optimised' 'TSF + Adp CB' scheme. However, the 'OSF + Stat + SLNR' scheme does not require the downlink training and uplink feedback of the effective channel and therefore is still preferable to the 'TSF + Adp CB' scheme. It can be concluded that multiuser interference is effectively mitigated in 'OSF + Stat' scheme (due to low rank

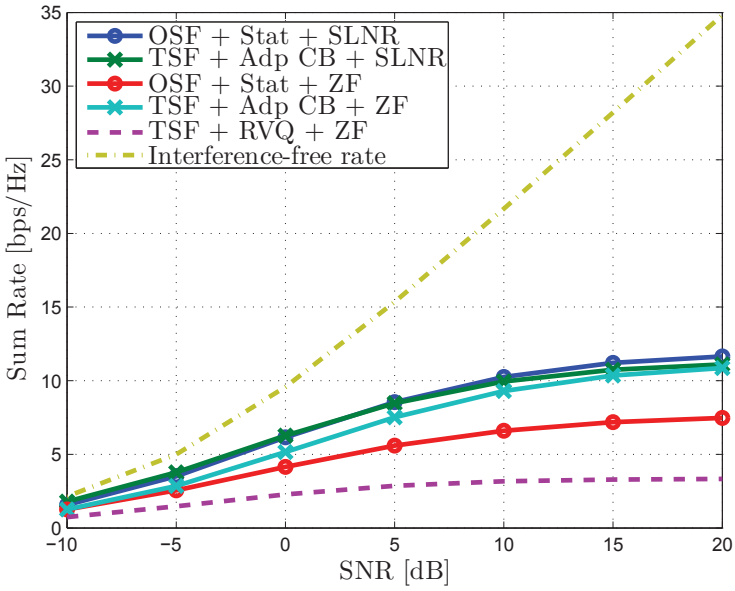

Figure 4.6.: Comparison among various hybrid precoding schemes, in a setup with $B = 10, L = 15$.

channel covariance matrix) and in 'TSF + Adp CB' scheme (owing to a relatively large number of feedback).

In Fig. 4.6, the 'OSF + Stat + SLNR' scheme exceeds various baselines even in not-very-sparse ($L \gg 1$) channels, which confirms the effectiveness of the proposed design. However, compared with Fig. 4.5 where the channels are sparse ($L = 2$), the sum rates of the hybrid precoding schemes with $L = 15$ are highly degraded since one RF chain per user (i.e., $N = K$) captures only one path gain while missing the rest. The sum rates are dominated by multiuser interference at high SNR since the covariance matrices of the effective channels with $L \gg 1$ tend to be full rank.

In Fig. 4.7, we evaluate the effect of the number of transmit antennas on the rate performance of the proposed hybrid precoding schemes at SNR $= -5$ dB. The SLNR-based digital precoding is considered. Each user channel has $L = 4$ paths and the total feedback overhead is $B = 8$. A fixed feedback allocation is used for the 'TSF + Adp CB' scheme, i.e., $B_{RF} = 5, B_{BB} = 3$. Fig. 4.7 shows that the rate gap between the 'OSF + Stat' and 'TSF + Adp CB' schemes enlarges as a larger number of antennas is employed at the BS. It indicates that the proposed 'OSF + Stat' scheme works well in the large-scale array regime. According to Section 4.2.3, the 'OSF + Stat' scheme forms a precise RF beamsteering with the help of all feedback resources. Meanwhile, as M becomes larger but $L = 4$ is fixed, the channel covariance matrices of users have higher dimensional nullspace to efficiently eliminate the multiuser interference. On the contrary, the rate gain of the 'TSF + Adp CB' scheme from increasing M is small due to coarse (low resolution) analog beamforming.

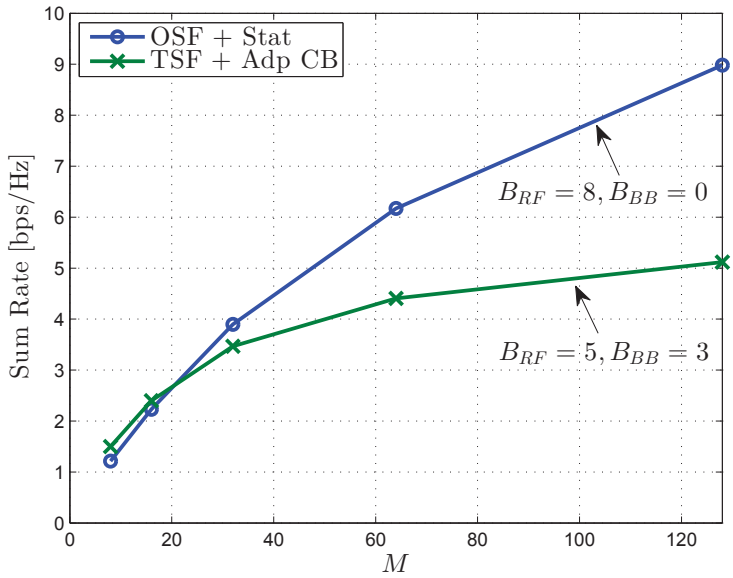

Figure 4.7.: Comparison among the hybrid precoding schemes with various M, in a setup with $L = 4, B = 8$, SNR = -5 dB.

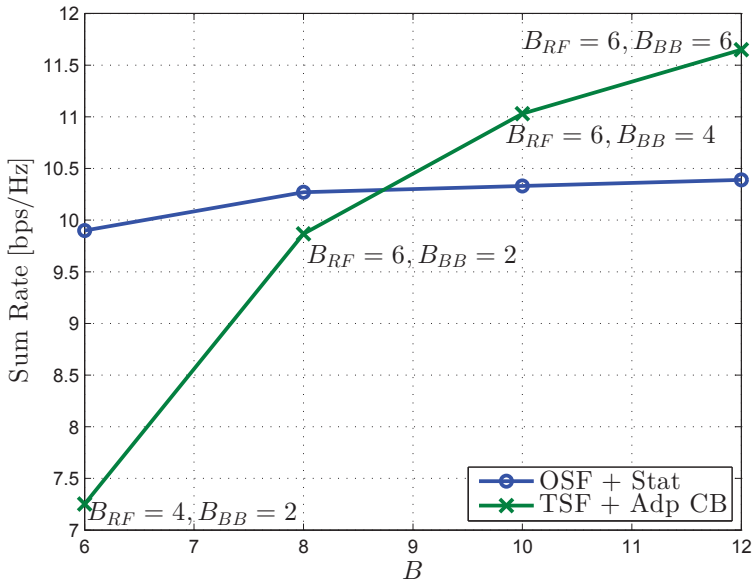

Figure 4.8.: Comparison among the hybrid precoding schemes with various B, in a setup with $L = 4, M = 32$, SNR = 5 dB.

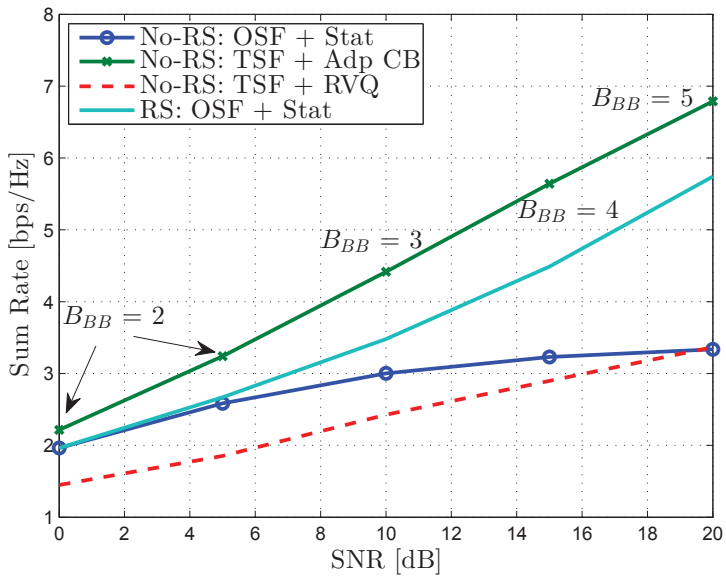

Figure 4.9.: Comparison between No-RS with extra feedback and RS, in a setup with $M = 32, K = 4, L = 8, B_{RF} = 4$ and ZF digital precoder.

In Fig. 4.8, we set $M = 32, L = 4$, SNR = 5 dB and evaluate the performance of the proposed schemes for varying B. Fig. 4.8 shows that when the amount of feedback overhead is small, it is preferable to allocate all resources to the analog beamforming (i.e., 'OSF + Stat'). As B increases, the feedback resources should be divided such that RF codebook with B_{RF} has sufficient resolution to distinguish different channel paths. Moreover, the remaining resources can accurately quantise the effective channel. The digital precoder based on the accurate channel quantisation effectively mitigates the multiuser interference. In the large regime of feedback overhead B, Fig. 4.8 shows that the 'TSF + Adp CB' scheme is preferable over the 'OSF + Stat' scheme, which validates the last observation in Section 4.2.3.

4.4.2. Rate Splitting with Hybrid Precoding

In Fig. 4.9 and Fig. 4.10, we compare the 'OSF + Stat' precoded RS with the 'TSF + Adp CB' precoded No-RS. Both strategies use the same B_{RF} for the analog beamforming while 'TSF + Adp CB' is driven by additional feedback B_{BB} in the second-stage channel quantisation. We also plot the 'OSF + Stat'/'TSF + RVQ' precoded No-RS as references. ZF and SLNR-based digital precoders are respectively considered in Fig. 4.9 and Fig. 4.10 for the proposed strategies. The number of feedback bits in the first-stage beam steering is set as $B_{RF} = 4$ while B_{BB} is annotated in Fig. 4.9 and Fig. 4.10. Following Remark 2, the amount of extra feedback bits should scale as $B_{BB} = O(\frac{K-1}{K}\frac{P_{dB}}{3})$ to achieve sum DoF of 1.

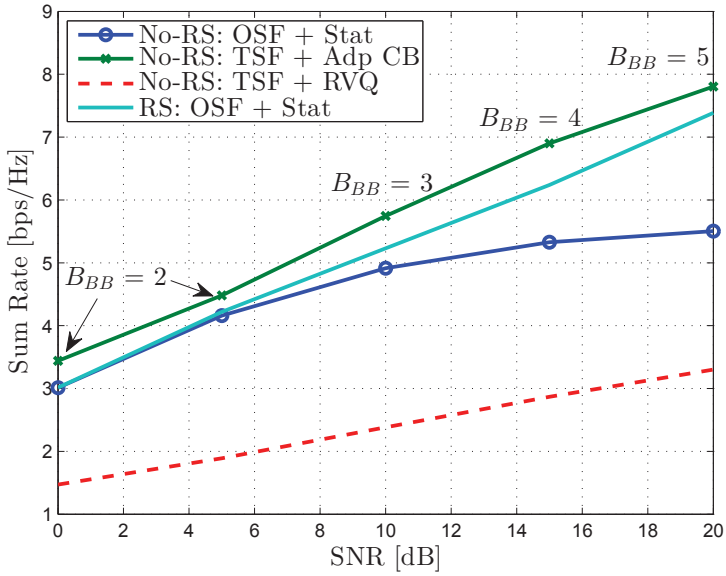

Figure 4.10.: Comparison between No-RS with extra feedback and RS, in a setup with $M = 32, K = 4, L = 8, B_{RF} = 4$ and SLNR-based digital precoder.

Fig. 4.9 and Fig. 4.10 show that the 'OSF + Stat' precoded No-RS is interference-limited beyond SNR = 15 dB, since statistical CSIT-based digital precoding is unable to eliminate multiuser interference. Meanwhile, the 'TSF + Adp CB' or 'TSF + RVQ' precoded No-RS strategies make use of extra SNR-adaptive second-stage feedback so as to achieve a rate performance that linearly increases with the transmit power. Moreover, we note that 'OSF + Stat' precoded RS and 'TSF + Adp CB'/'TSF + RVQ' precoded No-RS achieve almost the same sum DoF (reflected by the slope of the sum rate at high SNR). Even though 'TSF + Adp CB' precoded No-RS has a rate gain over 'OSF + Stat' precoded RS, the latter enables saving both in the second-stage channel training and feedback. Finally, it can be seen that the sum rate of 'OSF + Stat' precoded RS outperforms 'TSF + RVQ' precoded No-RS [17] due to the exploitation of the second-order statistics and the common message transmission.

4.5. Summary and Conclusion

In this chapter, we proposed an 'OSF + Stat' hybrid precoding scheme based on one-stage feedback. Specifically, the 'OSF + Stat' scheme uses all feedback resources to precisely design the analog beamformer while mitigating multiuser interference by statistical CSIT-based digital precoder. As a comparison, the 'TSF + Adp CB' scheme based on two-stage feedback partitions the total overhead into RF feedback and effective channel quantisation. The digital precoder copes with the multiuser interference based on quantised CSIT. Under a fixed total feedback constraint, we showed that the 'OSF + Stat'

scheme outperforms the 'TSF + Adp CB' scheme for very limited feedback and/or very sparse channels. Meanwhile, the 'OSF + Stat' scheme gets rid of the second-stage channel training and feedback and therefore effectively reduces the complexity of signalling and feedback overhead.

Nevertheless, we note that the conventional transmission strategy precoded by either 'OSF + Stat' or 'TSF + Adp CB' is interference-limited at high SNR. Then, we proposed a rate splitting transmission strategy which tackles the residual interference. The idea of RS can be applied with both one-stage/two-stage hybrid precoding schemes to enhance the rate performance. In consideration of the benefits of RS, we particularly showed that given the same amount of feedback for the analog beamforming, the 'OSF + Stat' precoded RS can achieve a rate performance comparable to that of the 'TSF + Adp CB' precoded No-RS with extra second-stage feedback for channel quantisation. By employing a more sophisticated transceiver architecture (i.e., superposition coding at the transmitter and SIC at the receiver), the RS transmission strategy enables significant saving in the second-stage channel training and feedback.

5. A Novel Hybrid Precoding Strategy for Physical Layer Multicasting with Perfect CSIT

In previous section, we have investigated the hybrid precoding strategy for multiuser MISO broadcast channel. In this chapter, we turn to study the hybrid precoding strategy for physical layer multicasting that is also of great interest in 5G cellular networks. Accordingly, we focus on maximising the minimum (max-min) received rate among all users. We propose a low complexity algorithm to compute the RF precoder that achieves near-optimal max-min fairness performance. Moreover, we derive a simple condition under which the hybrid precoding driven by a limited number of RF chains incurs no loss of optimality with respect to the fully digital precoding case.

So far, there has been no investigation on physical layer multicasting with hybrid precoding. To the best of our knowledge, this is the first working dealing with the above problem. For this reason, this work considers perfect CSIT as an initial step, rather than dealing with various imperfect CSIT types in preceding sections. We would leave the imperfect CSIT case to a future work.

5.1. Problem Formulation

Consider the downlink of a multiuser cellular system where a BS equipped with M antennas and N ($N \leq M$) RF chains serves K single-antenna users. Let $\mathbf{h}_k \in \mathbb{C}^M$ denote the frequency-flat quasi-static downlink channel vector of user k and assume that the BS has perfect knowledge of the channel vector for each user. The input-output analytical expression is written as $y_k = \mathbf{h}_k^H \mathbf{x} + n_k$, where $\mathbf{x} = \mathbf{w}s$ represents the transmitted signal with power constraint $\mathbb{E}[||\mathbf{x}||^2] \leq P$. $\mathbf{w} \in \mathbb{C}^M$ is the linear precoder while s is the common message intended to all users with zero-mean and unit variance. $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ denotes the additive white Gaussian noise.

5.1.1. A Sufficient Number of RF Chains

In the case of $N = M$, $\mathbf{w}$ can be fully designed in the digital domain. Let us revisit the traditional max-min SNR fairness problem [3] which can be written as

$$\max_{\mathbf{w}} \min_{k \in \mathcal{K}} \quad |\bar{\mathbf{h}}_k^H \mathbf{w}|^2, \quad \text{s.t.} \ \|\mathbf{w}\|^2 = P, \tag{5.1}$$

where $\bar{\mathbf{h}}_k = \mathbf{h}_k / \sigma_k$ and $\mathcal{K}$ denotes the multicast user set. This NP-hard problem can be approximated in a relaxed or conservative manner. Namely, we can relax problem (5.1) to a convex SDP problem by introducing $\mathbf{W} = \mathbf{w}\mathbf{w}^H$ and ignoring the rank one constraint. If $\mathbf{W}_{\text{opt}}$ has rank of one, its principal eigenvector is the optimal solution to problem (5.1). Otherwise, a randomisation procedure [3] is applied to obtain a feasible solution to the original problem. For example, we perform eigen-decomposition on $\mathbf{W}_{\text{opt}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ and generate a set of candidate vectors as $\mathbf{w} = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{v}$ with $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ and normalise them as $\|\mathbf{w}\|^2 = P$. The best candidate which obtains the largest max-min SNR is chosen as the final solution of problem (5.1).

By contrast, we can exploit the fact that the max-min fairness problem subject to a transmit power constraint is equivalent to the transmit power minimisation problem subject to Quality-of-service (QoS) constraints up to scaling [3]. Then, replacing the non-convex (QoS) constraint by a conservative convex approximation yields a feasible solution to the original problem [60]. Furthermore, [62] designed the precoder as a linear sum of channels of all users which achieves near-optimal max-min fairness. The corresponding details can be found in [60–62].

5.1.2. A Limited Number of RF chains

In the case of $N < M$, $\mathbf{w}$ can be jointly designed in the analog and digital domain, i.e., $\mathbf{w} = \mathbf{F}_{\text{RF}}\mathbf{w}_{\text{BB}}$ where $\mathbf{F}_{\text{RF}} \in \mathbb{C}^{M \times N}$ and $\mathbf{w}_{\text{BB}} \in \mathbb{C}^N$. Since $\mathbf{F}_{\text{RF}}$ is implemented using phase shifting networks, a constant modulus constraint is imposed on its entries. Without loss of generality, we assume that $[\mathbf{F}_{\text{RF}}]_{i,j} = \frac{1}{\sqrt{M}} e^{j\varphi_{i,j}}$. Since P is immaterial with respect to the optimisation problem, we can normalise it and then scale up the solution with $\sqrt{P}$. Proceeding with the design of $\mathbf{F}_{\text{RF}}\mathbf{w}_{\text{BB}}$, problem (5.1) can be stated as

$$\begin{aligned} &\max_{\mathbf{F}_{\text{RF}}, \mathbf{w}_{\text{BB}}} \min_{k \in \mathcal{K}} \quad |\bar{\mathbf{h}}_k^H \mathbf{F}_{\text{RF}} \mathbf{w}_{\text{BB}}|^2 \\ &\text{s.t.} \quad \|\mathbf{F}_{\text{RF}} \mathbf{w}_{\text{BB}}\|^2 = 1, \quad \mathbf{F}_{\text{RF}} \in \mathcal{F}, \end{aligned} \tag{5.2}$$

where $\mathcal{F}$ is the feasible set of $\mathbf{F}_{\text{RF}}$ with constant magnitude entries. The NP-hard problem becomes more challenging in the presence of the non-convex feasibility constraint $\mathbf{F}_{\text{RF}} \in \mathcal{F}$.

5.2. Proposed Solutions

Consider a predefined RF codebook and an optimised $\mathbf{w}_{\mathrm{BB}}$ for a given RF precoder, a low complexity search algorithm is proposed to determine the RF precoder that achieves nearly the same performance as performing an exhaustive search. When it comes to limited scattering channels, the RF precoder can be designed exploiting channel sparsity. We derive a simple condition under which the hybrid precoder driven by a limited number of RF chains achieves the same max-min fairness as with fully digital precoder.

5.2.1. A Low Complexity Algorithm

To resolve the non-convex constraint, the RF precoder can be selected from a predefined codebook $\mathcal{C}$. We denote $\mathcal{C}_{\mathrm{set}}$ as the set of all $M \times N$ matrices whose columns are drawn from N different columns of $\mathcal{C}$. Since the RF precoder belongs to a finite discrete set, the associated combinatorial optimisation of $\mathbf{F}_{\mathrm{RF}}$ can be solved by a high-complexity exhaustive search with $I_{\mathrm{exs}} = \frac{N!}{(N-M)!}$ iterations. For each $\mathbf{F}_{\mathrm{RF}}$ and the associated effective channel $\widehat{\mathbf{h}}_k \triangleq \mathbf{F}_{\mathrm{RF}}^H \bar{\mathbf{h}}_k$, we determine the optimal $\mathbf{w}_{\mathrm{BB}}$ by using methods presented in section 5.1.1. Then, the $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{w}_{\mathrm{BB}}$ that achieve the best max-min SNR performance is adopted as the final solution. To circumvent the issue of complexity, we develop a low complexity search algorithm by leveraging the following observation on the RF precoder design.

Observation: Problem (5.2) can be equivalently formulated as

$$\begin{aligned} \max_{\mathbf{F}_{\mathrm{RF}}, \mathbf{w}_{\mathrm{BB}}} \quad & \min_{k \in \mathcal{K}} \quad \frac{|\bar{\mathbf{h}}_k^H \mathbf{F}_{\mathrm{RF}} \mathbf{w}_{\mathrm{BB}}|^2}{\|\mathbf{F}_{\mathrm{RF}} \mathbf{w}_{\mathrm{BB}}\|^2} \\ \text{s.t.} \quad & \mathbf{F}_{\mathrm{RF}} \in \mathcal{C}_{\mathrm{set}}, \end{aligned} \tag{5.3}$$

where we unify the power constraint into the objective function. For certain $\mathbf{F}_{\mathrm{RF}}$, the optimum value (denoted by t) of the objective function in problem (5.3) is upper bounded, i.e.,

$$t = \min_{k \in \mathcal{K}} \left\{ \frac{\mathbf{w}_{\mathrm{BB}}^H \mathbf{F}_{\mathrm{RF}}^H \bar{\mathbf{h}}_k \bar{\mathbf{h}}_k^H \mathbf{F}_{\mathrm{RF}} \mathbf{w}_{\mathrm{BB}}}{\mathbf{w}_{\mathrm{BB}}^H \mathbf{F}_{\mathrm{RF}}^H \mathbf{F}_{\mathrm{RF}} \mathbf{w}_{\mathrm{BB}}} \right\} \tag{5.4}$$

$$\leq \min_{k \in \mathcal{K}} \left\{ \lambda_{\max} \Big((\mathbf{F}_{\mathrm{RF}}^H \mathbf{F}_{\mathrm{RF}})^{-1} \mathbf{F}_{\mathrm{RF}}^H \bar{\mathbf{h}}_k \bar{\mathbf{h}}_k^H \mathbf{F}_{\mathrm{RF}} \Big) \right\} \tag{5.5}$$

$$= \min_{k \in \mathcal{K}} \left\{ \bar{\mathbf{h}}_k^H \mathbf{F}_{\mathrm{RF}} (\mathbf{F}_{\mathrm{RF}}^H \mathbf{F}_{\mathrm{RF}})^{-1} \mathbf{F}_{\mathrm{RF}}^H \bar{\mathbf{h}}_k \right\}, \tag{5.6}$$

where (5.5) is obtained from the generalised eigenvalue of $\mathbf{F}_{\mathrm{RF}}^H \bar{\mathbf{h}}_k \bar{\mathbf{h}}_k^H \mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{RF}}^H \mathbf{F}_{\mathrm{RF}}$. We assume $\mathcal{C}$ with full column rank and then $\mathbf{F}_{\mathrm{RF}}^H \mathbf{F}_{\mathrm{RF}}$ is invertible.

A selection of $\mathbf{F}_{\mathrm{RF}} \in \mathcal{C}_{\mathrm{set}}$, that leads to a small upper bound, constrains the optimum t to a small value and hence is unlikely to be the optimal RF precoder. In this light, we propose to search $\mathbf{F}_{\mathrm{RF}}$ in descending order of (5.6) and only select a limited number of

1: **Pre-processing**: Select a subset of $\mathbf{F}_{\mathrm{RF}} \in \mathcal{C}_{\mathrm{set}}$ with the largest I values in (5.6)
2: **For** $i \in [1, I]$
3: Compute the effective channel $\widehat{\mathbf{h}}_k^{(i)}, \forall k$ with $\mathbf{F}_{\mathrm{RF}}^{(i)}$
4: Apply algorithm stated in section 5.1.1 to find the optimal $t^{(i)}$ and $\mathbf{w}_{\mathrm{BB}}^{(i)}$
5: **End**
6: The $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{w}_{\mathrm{BB}}$ that obtains the largest t are chosen as the solution of (5.3)

Table 5.1.: Precoding design algorithm

$\mathbf{F}_{\mathrm{RF}}$ that obtains the largest upper bound values. The low complexity search algorithm is outlined in Algorithm 1.

Remark 5.1. *The complexity of Algorithm 1 is approximated by $I \cdot C_d$, where C_d indicates the complexity of algorithm applied in step 4. We note that $I \in [1, I_{exs}]$. The optimal $\mathbf{F}_{RF}$ can be obtained by exhaustive search ($I = I_{exs}$) while the proposed RF precoder design can achieve nearly the same max-min fairness with $I \ll I_{exs}$. The corresponding complexity reduction factor can be approximated by I_{exs}/I.*

Remark 5.2. *The RF codebook design acts as one of the determinant factors for the max-min performance. Though the optimal RF codebook is still unknown, good candidates can be recognised by exploiting second-order channel statistics or array structure. For example, we can normalise each entry of the dominant eigenvectors of channel covariance matrices and collect them as the RF codebook. Otherwise, we can design $\mathcal{C}$ as the steering vectors with uniformly distributed AoDs. Moreover, when we take implementation complexity into account, a DFT-based codebook can be employed where any RF precoder can be implemented by a static (DFT) phase shifting network together with a RF switch.*

5.2.2. Achieving Optimum with a Reduced Number of RF Chains

The proposed algorithm in section 5.2.1 is applicable to arbitrary channels. By imposing a predefined codebook on the RF precoder design, the hybrid precoding can only provide a suboptimal performance compared with the fully digital precoding case. In this section, we consider a physical finite scattering channel model that has been investigated for Massive/Mm-wave MIMO systems [8, 9]. The channel vector from user k is defined as

$$\mathbf{h}_k = \sqrt{\frac{M}{L_k}} \sum_{l=1}^{L_k} g_k^l \, \mathbf{a}(\phi_k^l, \theta_k^l) = \sqrt{\frac{M}{L_k}} \mathbf{A}_k \mathbf{g}_k, \tag{5.7}$$

where $\mathbf{A}_k = [\mathbf{a}(\phi_k^1, \theta_k^1), \cdots, \mathbf{a}(\phi_k^{L_k}, \theta_k^{L_k})] \in \mathbb{C}^{M \times L_k}$ contains L_k steering vectors and $\mathbf{g}_k$ is the path gain vector. The factor $\sqrt{M/L_k}$ is used to normalise the channel, i.e., $\mathbb{E}(\|\mathbf{h}_k\|^2) = M$. Under the plane wave and balanced narrowband array assumptions, the array steering

vector can be written as [5, Ch. 2]

$$\mathbf{a}(\phi_k^l, \theta_k^l) = \frac{1}{\sqrt{M}} \left[e^{-jf_1(\phi_k^l, \theta_k^l)}, \cdots, e^{-jf_M(\phi_k^l, \theta_k^l)} \right]^T, \tag{5.8}$$

where $f_m(\phi, \theta)$ is a function of the AoD azimuth (ϕ) and elevation (θ). Consider first the single-path channels (i.e., $L_k = 1$) and denote a collection of the steering vector of each user by a $M \times K$ matrix $\mathbf{V}$ (e.g., $\mathbf{V} = [\mathbf{a}(\phi_1, \theta_1), \cdots, \mathbf{a}(\phi_K, \theta_K)]$). The following theorem characterises the max-min performance of multicasting driven by a limited number of RF chain.

Theorem 5.1. *When the BS has a prior knowledge of AoDs*[1] *and all channels are single-path, the hybrid precoding with* $\mathbf{F}_{RF} = \mathbf{V}$ *driven by only* K *RF chains achieves the same max-min fairness as with fully digital precoding driven by a sufficient number of RF chains (i.e.,* $N = M$*).*

Proof. We denote the (tall) channel matrix by $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K]$. Focusing on the fully digital precoder design in problem (5.1), we note that $\forall\, \mathbf{w} \in \text{Null}(\mathbf{H})$ makes the value of the objective function zero. Hence, the optimal solution $\mathbf{w}^\star \in \text{span}(\mathbf{H})$ is written as a linear combination of $\mathbf{h}_k$, i.e.,

$$\mathbf{w}^\star = \sum_{k=1}^{K} b_k \mathbf{h}_k = \sum_{k=1}^{K} c_k \mathbf{a}(\phi_k, \theta_k) = \mathbf{F}_{\text{RF}} \mathbf{w}_{\text{BB}}, \tag{5.9}$$

where b_k denote the coefficients and $\mathbf{w}_{\text{BB}} = [c_1, \cdots, c_K]^T$ with $c_k = \sqrt{M} b_k g_k$. Since the steering vector collection matrix $\mathbf{V}$ satisfies inherently the constant modulus constraint, (5.9) implies that the optimal fully digital precoder can be equivalently implemented by a hybrid precoding structure without loss of performance. □

By reusing (5.9), we are able to generalise Theorem 5.1 into multi-path channels (i.e., $L_k \geq 1$) and thereby present the following Corollary without proof.

Corollary 5.1. *When the BS has a prior knowledge of AoDs and at least* N *RF chains with* $N = \sum_k L_k \leq M$*, and let* $\mathbf{V} \in \mathbb{C}^{M \times N}$ *collect the steering vectors of all users, the hybrid precoder with* $\mathbf{F}_{RF} = \mathbf{V}$ *achieves the same max-min fairness as with fully digital precoder.*

5.3. Performance Evaluation

In this section, we numerically compare the proposed low complexity algorithm with several baselines: Baseline 1 assumes a sufficient number of RF chains and exploits fully

[1] Subspace methods (e.g., root-MUSIC and ESPRIT) and compressed sensing algorithms (e.g., OMP and BP) can be applied to identify the distinct path arrivals [95].

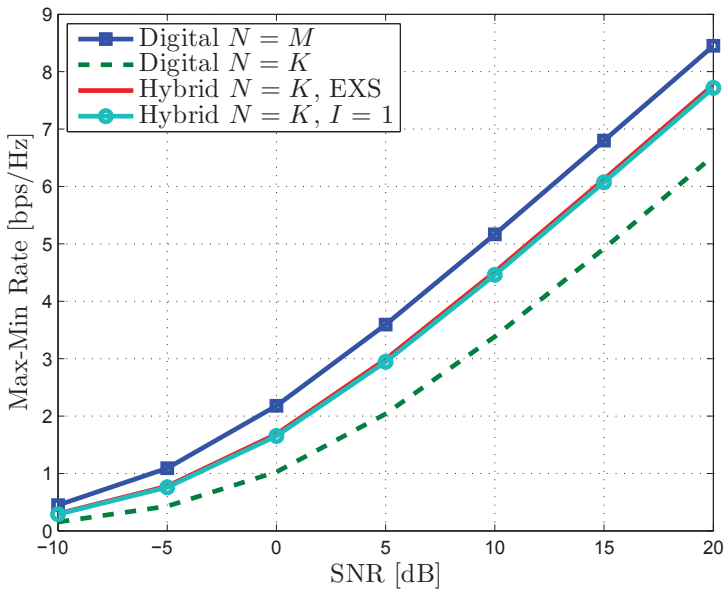

Figure 5.1.: Max-min rate comparison between the proposed hybrid precoding design and various baselines, in a setup with $M = 6, K = 2$.

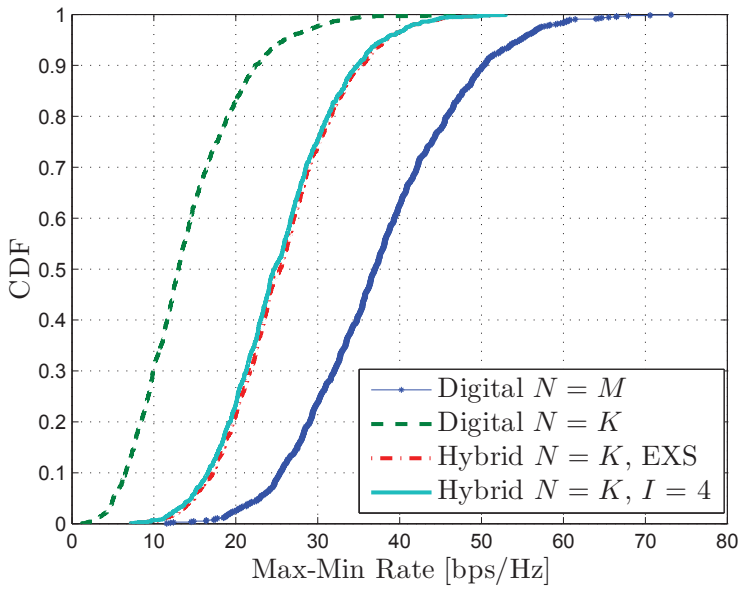

Figure 5.2.: Max-min rate CDF comparison between the proposed hybrid precoding design and various baselines, in a setup with $M = 8, K = 3$.

digital precoding (i.e., $M = N$, denoted by 'Digital $N = M$'). Baseline 2 assumes a random antenna subset selection and performs digital precoding over that subset (i.e., $M' = N = K$, denoted by 'Digital $N = K$'). Intuitively, baselines 1&2 place an upper and lower bound on the performance of hybrid precoder, respectively. Baseline 3 assumes a limited number of RF chains and exploits hybrid precoding with exhaustive search for RF precoder (i.e., $M > N = K$, denoted by 'Hybrid $N = K$, EXS'). DFT-based RF codebook is taken as an example. Since the interest of this letter is the RF precoder design, we compute all the digital precoders by simply using SDR. A Gaussian randomisation procedure is carried out to obtain a feasible solution. Moreover, the max-min rate is investigated by averaging over 1000 random channel realisations.

In Fig. 5.1 and 5.2, we evaluate the max-min rate performance under i.i.d. Rayleigh fading channels. Consider a system with $M = 6, K = 2$, Fig. 5.1 shows that the hybrid precoding structure achieves a max-min rate gain over baseline 2 (e.g., 1 bps/Hz beyond 5 dB SNR). It implies that a cost-efficient and well-designed phase shifter network can reap benefits. Compared with baseline 3 $\left(I_{\text{exs}} = \frac{N!}{(N-M)!} = 30\right)$, the proposed algorithm with $I = 1$ highly reduces the search complexity while keeping the same performance. In Fig. 5.2, we consider $M = 8, K = 3$, SNR = 10 dB and plot the cumulative density function (CDF) of the max-min rate of each approach. Likewise, the proposed low complexity algorithm with $I = 4$ achieves similar performance as with the exhaustive search ($I_{\text{exs}} = 336$). Moreover, we observe that the hybrid precoding with RF precoder selected from a predefined codebook is outperformed by the fully digital precoding. This performance loss is mainly due to the mismatch between the constant magnitude constrained RF precoder and the non-constant magnitude (non-sparse) i.i.d. Rayleigh channels.

Fig. 5.3 examines the max-min rate in single-path channels. Consider a uniform linear array with $M = 10$ isotropic antennas serving $K = 3$ users. The steering vector is given by $\mathbf{a}(\phi_k^l) = \frac{1}{\sqrt{M}}[1, e^{-j2\pi\frac{D}{\lambda}\cos(\phi_k^l)}, \cdots, e^{-j2\pi\frac{(M-1)D}{\lambda}\cos(\phi_k^l)}]^T$, where $D = \lambda/2$ is the half-wavelength antenna spacing and ϕ_k^l is uniformly distributed between 0 and 2π. Fig. 5.3 shows that the AoD-aware RF precoder design (Theorem 5.1) outperforms the low complexity DFT-based approach (Algorithm 1 with $I = 1$). The former dynamic design requires a prior knowledge of AoDs while the latter less-flexible design facilitates ease of implementation. In addition, Theorem 5.1 is validated, namely, the hybrid precoder with a limited number of RF chains can achieve the same max-min rate as with fully digital precoder. It implies that channel sparsity can be exploited to reduce the number of costly RF chains without a compromise on the performance. The intuition behind Theorem 5.1 and Corollary 5.1 is explained as follows. Under a limited scattering model, the $M \times 1$ channel vector is characterised by a few AoDs (ψ^l). The max-min fairness enabled by a high-dimensional (M) digital precoder can be obtained by a $M \times N$ transformation matrix (i.e., the RF precoder) that accurately captures the channel gain and a low-dimensional ($N \geq K$) digital precoder that takes care of the fairness amongst K users.

Moreover, when the channels have multi-path of $\sum_k L_k \geq M$, Theorem 5.1 and Corol-

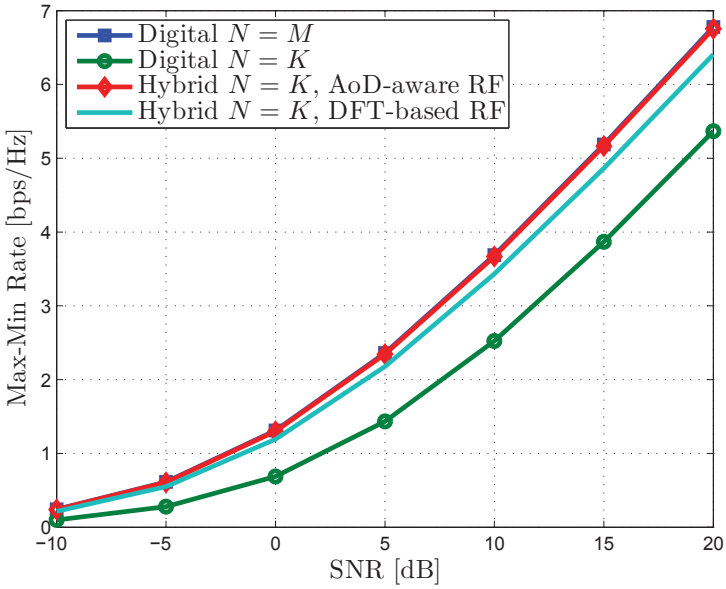

Figure 5.3.: Max-min rate comparison between the proposed hybrid precoding design and various baselines, in a setup with $M = 10, K = 3$.

lary 5.1 are not achievable. In this case, suppose the BS has the knowledge of AoDs, the low complexity algorithm in section 5.2.1 can be applied to select the RF precoder from the codebook that collects the steering vectors of all users. The basic idea is to exclude those selections that constrain the upper bound in (5.6) and to search among the rest. This codebook design would perform better than various codebooks discussed in remark 2, since it mostly captures the directions of the users' channels.

5.4. Summary and Conclusion

In this chapter, we have investigated a hybrid precoding method for multicasting with a limited number of RF chains. We have proposed a low complexity search algorithm to determine the RF precoder and validated its near optimality in terms of max-min rate. The low complexity algorithm can also be applied to finite-resolution phase shifter network. For limited scattering channels, we proved a simple condition under which the hybrid precoding incurs no loss of optimality with respect to the fully digital precoding case. Finally, the hybrid precoding in multi-group/cell multicasting is intriguing and is left for future work.

6. Conclusion and Future work

6.1. Conclusion

This thesis proposed a set of novel transmission strategies and beamforming designs to deal with various imperfect CSIT. We mainly analysed the rate performance of the proposed transmission strategies and addressed the associated precoding design and power allocation optimisation problems. Compared with the conventional baselines, the proposed schemes make better use of the imperfect CSIT and achieve significant rate gain.

In Chapter 2, we considered a two-user MISO broadcast channel with statistical and delayed CSIT. In highly correlated channels and at low SNR, we observed that the conventional transmission strategies employing both statistical and delayed CSIT still underperform some strategies driven by statistical CSIT only. Namely, the statistical CSIT is not fully exploited. To address this issue, we developed a novel and integrated transmission strategy which enables sum rate beyond that achievable with either statistical CSIT strategies or delayed CSIT strategies at any SNR, in arbitrary correlation conditions.

In Chapter 3, we considered a multiuser massive MIMO system with partial and statistical CSIT. The conventional works clustered users into group according to statistical CSIT while serving users in each group based on partial CSIT. However, the rate performance is still subject to inter-group and intra-group interference and therefore highly degraded. Inspired by the idea of RS, we proposed a novel and general framework Hierarchical RS to well tackle the inter-group and intra-group interference.

In Chapter 4, we considered a multiuser mmWave MIMO system with hybrid precoding under statistical and quantised CSIT. With limited feedback, we leveraged statistical CSIT to design digital precoder for interference mitigation while all feedback overhead is reserved for precise analog beamforming. The proposed precoding scheme yields sum rate gain over the conventional schemes in mmWave sparse channels. Moreover, we introduced RS to further overcome multiuser interference. It showed that RS enables remarkable saving in both channel training and feedback overhead compared with conventional multiuser transmission strategies.

In Chapter 5, we considered physical layer multicasting with hybrid precoding driven by a limited number of RF chains. We developed a low complexity algorithm to compute the RF precoder that achieves near-optimal max-min performance. Moreover, we derive a simple condition under which the hybrid precoding driven by a limited number of RF chains incurs no loss of optimality with respect to the fully digital precoding case.

6.2. Future Work

In conclusion of this thesis, some future research directions are pointed out.

- **Rate Splitting for Multi-Cell Downlink Massive MIMO**

  The work in this thesis focused mainly on single-cell downlink system in FDD mode (i.e., channel imperfectness comes from limited feedback). Consider the downlink of a TDD multi-cell massive MIMO system, where the CSIT acquisition model includes the effects of channel estimation error and pilot contamination. A system metric of considerable interest is to maximise the minimum achievable rate (max-min fairness) across all users or to minimise the weighted transmit power under QoS constraint. In these setups, prior works [96, 97] developed efficient precoding and power allocation schemes to optimise the system metrics. Yet, the resulting system performance is severely degraded by multi-cell interference. The idea of RS can be applied to overcome the multi-cell interference, achieving gains over conventional baselines [96, 97]. An interesting problem is to investigate how the power splitting ratio between the common and private messages varies with pilot length, number of cells, SNR, etc. A closed-form analysis is attainable owing to channel hardening effect in the large-scale antenna array regime.

- **Hybrid Precoding vs. Spatial Modulation**

  To alleviate the requirement of multiple transmit RF chains in MIMO systems, Spatial Modulation (SM) is a recently proposed technique using single RF chain [98]. In SM, the information is mapped into a constellation point in signal domain (i.e., a PSK/QAM symbol of size Q) and a constellation point in spatial domain (i.e., the index of the active transmit antenna of size M). Therefore, the information bits conveyed in one channel use is $\log_2 Q + \log_2 M$. The system metric of interest is bit error rate (BER) or symbol error probability (SEP). By contrast, all (M) transmit antennas and single RF chain can be interconnected by phase shifter network and thus used to achieve array gain by analog precoding. SM and precoding techniques may show superiority to each other in different regime of M, channel conditions and SNR. Moreover, this comparison can be extended to multiple (N) RF chains scenario (i.e., $1 < N < M$) [99], where a generalised SM (GSM) is proposed to convey $\log_2 Q + \log_2 \binom{M}{N}$ bits. Alternatively, a hybrid precoding structure can make use of all M transmit antennas and N RF chains. Further extensions can be considered in multiuser [100] and massive MIMO setups [101].

- **Rate Splitting for Multi-Group Multicasting with Hybrid Precoding**

  With limited RF chains, the work in this thesis studied the hybrid precoding design for (single-group) multicasting. It is certainly worth the effort to generalise

to multi-group multicasting. When the number of users served exceeds the number of RF chains, inter-group interference dominates the system performance (i.e., QoS constrained power or power constrained max-min fairness). In this case, RS comes into play. The associated hybrid precoding design and optimal power allocation problems require careful investigation. Prior work relevant to this problem but with fully digital precoding was reported in [4]. Some insightful/closed-form solutions might be obtained in mmWave sparse channels, massive MIMO [66]. An interesting extension lies in spatially correlated channels, where the second-order channel statistics help manage inter-group interference.

Bibliography

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[2] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[3] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, June 2006.

[4] H. Joudeh and B. Clerckx, "A rate-splitting strategy for max-min fair multigroup multicasting," in *Proc. IEEE Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[5] B. Clerckx and C. Oestges, *MIMO Wireless Networks: Channels, Techniques and Standards for Multi-antenna, Multi-user and Multi-cell Systems.* Academic Press (Elsevier), 2013.

[6] V. Raghavan, S. Hanly, and V. Veeravalli, "Statistical beamforming on the Grassmann manifold for the two-user broadcast channel," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6464–6489, Oct. 2013.

[7] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, July 2012.

[8] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, March 2014.

[9] H. Ngo, E. Larsson, and T. Marzetta, "The multicell multiuser MIMO uplink with very large antenna arrays and a finite-dimensional channel," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2350–2361, June 2013.

[10] A. M. Sayeed and V. Raghavan, "Maximizing MIMO capacity in sparse multipath with reconfigurable antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 156–166, June 2007.

[11] J. Mo and R. W. Heath, "High SNR capacity of millimeter wave MIMO systems with one-bit quantization," in *Proc. IEEE Information Theory and Applications Workshop (ITA)*, Feb. 2014, pp. 1–5.

[12] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sept. 2006.

[13] C. Windpassinger, R. F. H. Fischer, T. Vencel, and J. B. Huber, "Precoding in multiantenna and multiuser communications," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1305–1316, July 2004.

[14] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part ii: perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, March 2005.

[15] J. Wang, Z. Lan, C. W. Pyo, T. Baykas, C. S. Sum, M. A. Rahman, R. Funada, F. Kojima, I. Lakkis, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[16] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[17] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, 2015.

[18] J. Wang, S. Jin, X. Gao, K.-K. Wong, and E. Au, "Statistical eigenmode-based SDMA for two-user downlink," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5371–5383, Oct. 2012.

[19] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418–4431, July 2012.

[20] J. Wang, M. Matthaiou, S. Jin, and X. Gao, "Precoder design for multiuser MISO systems exploiting statistical and outdated CSIT," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4551–4564, Nov. 2013.

[21] C. Hao and B. Clerckx, "MISO broadcast channel with imperfect and (un)matched CSIT in the frequency domain: DoF region and transmission strategies," in *Proc. IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2013, pp. 1–6.

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

[23] R. Couillet and M. Debbah, *random matrix methods for wireless communications.* Cambridge University Press, 2011.

[24] O. Mehanna, K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos, "Feasible point pursuit and successive approximation of non-convex QCQPs," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 804–808, July 2015.

[25] W. Hachem, P. Loubaton, and J. Najim, "Deterministic equivalents for certain functionals of large random matrices," *The Annals of Applied Probability*, vol. 17, no. 3, pp. 875–930, June 2007.

[26] J. Hoydis, S. Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[27] J. Oh, S.-J. Kim, R. Narasimhan, and J. M. Cioffi, "Transmit power optimization for Gaussian vector broadcast channels," in *Proc. IEEE International Conference on Communications (ICC)*, May 2005, pp. 2712–2716.

[28] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[29] T. Michel and G. Wunder, "Sum rate iterative water-filling for Gaussian MIMO broadcast channels," in *Proc. IEEE Wireless Personal Multimedia Communications (WPMC)*, San Diego, CA, 2006.

[30] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[31] A. Rajanna and N. Jindal, "Multiuser diversity in downlink channels: When does the feedback cost outweigh the spectral efficiency benefit?" *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 408–418, Jan. 2012.

[32] J. Choi, V. Raghavan, and D. J.Love, "Limited feedback design for the spatially correlated multi-antenna broadcast channel," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec. 2013, pp. 3481–3486.

[33] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 315–328, Jan. 2013.

[34] X. Yi and D. Gesbert, "Precoding methods for the MISO broadcast channel with delayed CSIT," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1–11, May 2013.

[35] D. Vasisht, S. Kumar, H. Rahul, and D. Katabi, "Eliminating channel feedback in next-generation cellular networks," in *ACM Special Interest Group on Data Communication (SIGCOMM)*, Florianopolis , Brazil, 2016.

[36] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[37] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3232–3246, Sept. 2015.

[38] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, June 2010.

[39] H. Joudeh and B. Clerckx, "Sum rate maximization for MU-MISO with partial CSIT using joint multicasting and broadcasting," in *Proc. IEEE International Conference on Communications (ICC)*, June 2015, pp. 4733–4738.

[40] B. Clerckx, G. Kim, and S. Kim, "Correlated fading in broadcast MIMO channels: Curse or blessing?" in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Nov. 2008, pp. 1–5.

[41] ——, "MU-MIMO with channel statistics-based codebooks in spatially correlated channels," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Nov. 2008, pp. 1–5.

[42] Q. Li, G. Li, W. Lee, M.-i. Lee, D. Mazzarese, B. Clerckx, and Z. Li, "MIMO techniques in WiMAX and LTE: A feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, May 2010.

[43] C. Lim, T. Yoo, B. Clerckx, B. Lee, and B. Shim, "Recent trend of multiuser MIMO in LTE-advanced," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 127–135, March 2013.

[44] A. Adhikary, J. Nam, J. Ahn, and G. Caire, "Joint spatial division and multiplexing–the large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[45] J. Park and B. Clerckx, "Multi-user linear precoding for multi-polarized massive MIMO system under imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2532–2547, 2015.

[46] D. Kim, G. Lee, and Y. Sung, "Two-stage beamformer design for massive MIMO downlink by trace quotient formulation," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2200–2211, June 2015.

[47] J. Chen and V. Lau, "Two-tier precoding for FDD multi-cell massive MIMO time-varying interference networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1230–1238, June 2014.

[48] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct. 2014.

[49] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[50] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, June 2014.

[51] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, 2011.

[52] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "Mmwave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, Oct. 2015.

[53] X. Yu, J. C. Shen, J. Zhang, and K. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, April 2016.

[54] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, April 2016.

[55] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, 2014.

[56] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, April 2016.

[57] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014.

[58] T. E. Bogale, L. B. Le, and X. Wang, "Hybrid analog-digital channel estimation and beamforming: Training-throughput tradeoff," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5235–5249, 2015.

[59] B. Gopalakrishnan and N. Sidiropoulos, "High performance adaptive algorithms for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4373–4384, Aug. 2015.

[60] N. Bornhorst and M. Pesavento, "An iterative convex approximation approach for transmit beamforming in multi-group multicasting," in *Proc. IEEE Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2011, pp. 426–430.

[61] L.-N. Tran, M. Hanif, and M. Juntti, "A conic quadratic programming approach to physical layer multicasting for large-scale antenna arrays," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 114–117, Jan. 2014.

[62] J. Joung, H. D. Nguyen, P. H. Tan, and S. Sun, "Multicast linear precoding for MIMO-OFDM systems," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 993–996, June 2015.

[63] E. Karipidis, N. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, March 2008.

[64] D. Christopoulos, S. Chatzinotas, and O. B., "Weighted fair multicast multigroup beamforming under per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5132–5142, Oct. 2014.

[65] ——, "Multicast multigroup beamforming for per-antenna power constrained large-scale arrays," in *Proc. IEEE Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2015.

[66] Z. Xiang, M. Tao, and X. Wang, "Massive MIMO multicasting in noncooperative cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1180–1193, June 2014.

[67] J. Choi, "Iterative methods for physical-layer multicast beamforming," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5185–5196, Sept. 2015.

[68] S. G. Srinivasan and M. K. Varanasi, "Optimal spatial correlations for the noncoherent MIMO rayleigh fading channel," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3760–3769, Oct. 2007.

[69] V. V. Veeravalli, Y. Liang, and A. M. Sayeed, "Correlated MIMO wireless channels: Capacity, optimal signaling, and asymptotics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2058–2072, June 2005.

[70] M. Sadek, A. Tarighat, A. H. Sayed *et al.*, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007.

[71] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imaging Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[72] P. Tseng and S. Yun, "A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization," *J. Optimz. Theory App.*, pp. 513–535, 2009.

[73] I. D. Coope and P. F. Renaud, *Trace inequalities with applications to orthogonal regression and matrix nearness problems.* Department of Mathematics and Statistics, University of Canterbury, 2000.

[74] R. A. Horn and C. R. Johnson, *Matrix analysis.* Cambridge university press, 2012.

[75] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[76] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta numerica*, vol. 4, pp. 1–51, 1995.

[77] R. Fletcher, "The sequential quadratic programming method," in *Nonlinear Optimization.* Springer, 2010.

[78] J. L. Morales, J. Nocedal, and Y. Wu, "A sequential quadratic programming algorithm with an additional equality constrained phase," *IMA J. Numer. Anal.*, vol. 32, no. 2, pp. 553–579, 2012.

[79] F. Boccardi, B. Clerckx, A. Ghosh, E. Hardouin, G. Jongren, K. Kusume, E. Onggosanusi, and Y. Tang, "Multiple antenna techniques in LTE-advanced," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 114–121, March 2012.

[80] B. Clerckx, G. Kim, and S. Kim, "MU-MIMO with channel statistics-based codebooks in spatially correlated channels," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Nov. 2008, pp. 1–5.

[81] C. Wang and R. Murch, "Adaptive downlink multi-user MIMO wireless systems for correlated channels with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2435–2446, Sept. 2006.

[82] E. Bjornson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, July 2014.

[83] P. Billingsley, *Probability and measure*. John Wiley & Sons, Inc., 2008.

[84] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, 2014.

[85] H. Prabhu, J. Rodrigues, O. Edfors, and F. Rusek, "Approximative matrix inverse computations for very-large MIMO and applications to linear pre-coding systems," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 2710–2715.

[86] A. Adhikary, E. A. Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint spatial division and multiplexing for mm-wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, June 2014.

[87] Z. Xiao, T. He, P. Xia, and X. G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[88] M. Dai and B. Clerckx, "Transmit beamforming for MISO broadcast channels with statistical and delayed CSIT," *IEEE Trans. Commun.*, vol. 63, no. 4, pp. 1202–1215, April 2015.

[89] D. J. Love, R. W. Heath, W. Santipach, and M. L. Honig, "What is the value of limited feedback for MIMO channels?" *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 54–59, 2004.

[90] D. J. Love and R. W. Heath, "Limited feedback diversity techniques for correlated channels," *IEEE Trans. Veh. Technol.*, vol. 55, no. 2, pp. 718–722, March 2006.

[91] W. Shen, L. Dai, Y. Zhang, and Z. Wang, "On the performance of channel statistics-based codebook for massive MIMO," in *Proc. IEEE Global Communications Conference (Globecom)*, 2016. [Online]. Available: http://arxiv.org/pdf/1604.04761v2.pdf

[92] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611–4624, July 2016.

[93] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, 2010.

[94] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[95] C. Berger, S. Zhou, J. Preisig, and P. Willett, "Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1708–1721, March 2010.

[96] S. Zarei, J. Aulin, W. Gerstacker, and R. Schober, "Max-min multi-cell aware precoding and power allocation for downlink massive MIMO systems," 2016. [Online]. Available: http://arxiv.org/pdf/1609.06652.pdf

[97] Q. Zhang, S. Jin, M. McKay, D. Morales-Jimenez, and H. Zhu, "Power allocation schemes for multicell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 5941–5955, Nov. 2015.

[98] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2228–2241, July 2008.

[99] T. L. Narasimhan, P. Raviteja, and A. Chockalingam, "Generalized spatial modulation in large-scale multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3764–3779, July 2015.

[100] S. Narayanan, M. J. Chaudhry, A. Stavridis, M. D. Renzo, F. Graziosi, and H. Haas, "Multi-user spatial modulation MIMO," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, April 2014, pp. 671–676.

[101] D. A. Basnayaka and H. Haas, "Spatial modulation for massive MIMO," in *Proc. IEEE International Conference on Communications (ICC)*, June 2015, pp. 1945–1950.

[102] S. H. Rice, "The expected value of the ratio of correlated random variables," 2009. [Online]. Available: http://www.faculty.biol.ttu.edu/Rice/ratio-derive.pdf

[103] H. Seltman, "Approximations for mean and variance of a ratio," 2012. [Online]. Available: http://www.stat.cmu.edu/~hseltman/files/ratio.pdf

[104] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive mimo systems with arbitrary-rank channel means," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.

[105] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

[106] M. Marcus and H. Minc, *A survey of matrix theory and matrix inequalities.* Courier Dover Publications, 1992.

[107] A. Jeffrey and D. Zwillinger, *Table of integrals, series, and products*, 7th ed. Academic Press, Amsterdam, 2007.

[108] M. Matthaiou, C. Zhong, and T. Ratnarajah, "Novel generic bounds on the sum rate of MIMO ZF receivers," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4341–4353, Sept. 2011.

[109] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive mimo systems with arbitrary-rank channel means," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.

A. Proofs for Chapter 2

A.1. Lemma

Consider a non-zero vector $\mathbf{w} \in \mathbb{C}^M$ and $\mathbf{h} = \mathbf{R}^{1/2}\,\mathbf{h}_w$, where $\mathbf{R} \in \Phi_{\mathrm{PD}}$ is a $M \times M$ Hermitian matrix. Then,

$$\mathbb{E}\left[\ln\left(|\mathbf{h}^H\mathbf{w}|^2\right)\right] = \ln\left(\mathbf{w}^H\mathbf{R}\mathbf{w}\right) - \gamma, \tag{A.1}$$

where γ is the Euler constant.

Proof. Define $\mathbf{X} \triangleq \mathbf{R}^{1/2}\mathbf{w}\mathbf{w}^H\mathbf{R}^{1/2}$ and decompose it as $\mathbf{X} = \mathbf{U}_X\mathbf{\Lambda}_X\mathbf{U}_X^H$. Due to rank($\mathbf{X}$) = 1, the diagonal matrix $\mathbf{\Lambda}_X$ has only one (e.g., m-th) non-zero entry, denoted by λ_X. We get

$$\lambda_X = \mathrm{tr}(\mathbf{\Lambda}_X) \overset{(a)}{=} \mathrm{tr}(\mathbf{X}) \overset{(b)}{=} \mathbf{w}^H\mathbf{R}\mathbf{w}. \tag{A.2}$$

where, equalities (a) and (b) can be easily obtained by applying $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$. Then

$$\begin{aligned}
\mathbb{E}\left[\ln\left(|\mathbf{h}^H\mathbf{w}|^2\right)\right] &= \mathbb{E}\left[\ln\left(\mathbf{h}_w^H\mathbf{R}^{1/2}\mathbf{w}\mathbf{w}^H\mathbf{R}^{1/2}\mathbf{h}_w\right)\right] && \text{(A.3)}\\
&\overset{d}{=} \mathbb{E}\left[\ln\left(\mathbf{h}_w^H\mathbf{\Lambda}_X\mathbf{h}_w\right)\right] && \text{(A.4)}\\
&= \mathbb{E}\left[\ln\left(\lambda_X|h_{w,m}|^2\right)\right], && \text{(A.5)}
\end{aligned}$$

where $\overset{d}{=}$ indicates the equivalence in distribution and (A.5) is calculated with the non-zero element in $\mathbf{\Lambda}_X$. Then, (A.1) can be obtained following (A.2) and $|h_{w,m}|^2 \sim \mathrm{Exp}(1)$. □

A.2. Lemma

Suppose x, y are two random variables. $\mathbb{E}(y) \neq 0$, let $f(x,y) = \frac{x}{y}$ and $\mu = (\mathbb{E}(x), \mathbb{E}(y)) = (\mu_x, \mu_y)$. The first order approximation of the expectation of $f(x,y)$ can be written as

$$\mathbb{E}\left(\frac{x}{y}\right) = \frac{\mu_x}{\mu_y} + O\left(\frac{\mathrm{var(y)}\mu_x}{\mu_y^3} - \frac{\mathrm{cov(x,y)}}{\mu_y^2}\right). \tag{A.6}$$

Proof. The closed-form of $\mathbb{E}\left(\frac{x}{y}\right)$ is unknown, however, it can be calculated via bivariate

Taylor expansion at μ

$$\mathbb{E}(f(x,y)) \quad = \quad \frac{\mu_x}{\mu_y} + \sum_{n=1}^{\infty}(-1)^n \frac{\mu_x \cdot \pi_{0,n} + \pi_{1,n}}{\mu_y^{n+1}}, \tag{A.7}$$

where $\pi_{i,j} = \mathbb{E}[(x-\mu_x)^i \cdot (y-\mu_y)^j]$. Take the first order approximation of (A.7) and (A.6) is obtained. Similar results were derived in an alternative manner [102]. However, it is difficult to calculate the high-order terms in (A.7) so that the first and second order approximations were used in [103]. It is assumed here that $\mathbb{E}(\frac{x}{y})$ is bounded and its Taylor expansion converges. Moreover, if x, y are mutually independent nonnegative random variables, the first order approximation is a lower bound, i.e., $\mathbb{E}(\frac{x}{y}) \geq \frac{\mu_x}{\mu_y}$. □

A.3. Proof of Theorem 2.1

The proof relies on deriving a lower bound on the achievable ergodic sum rate. According to (2.4), we can rewrite the ergodic sum rate as

$$\begin{aligned}
R_{\text{sum}} &= \mathbb{E}\Big[\log_2\Big(1+\frac{\rho|\mathbf{h}^H\mathbf{w}|^2}{1+\rho|\mathbf{h}^H\mathbf{q}|^2}\Big)\Big] + \mathbb{E}\Big[\log_2\Big(1+\frac{\rho|\mathbf{g}^H\mathbf{q}|^2}{1+\rho|\mathbf{g}^H\mathbf{w}|^2}\Big)\Big] \\
&= \mathbb{E}\Big[\log_2\Big(1+\exp\Big(\ln\Big(\frac{\rho|\mathbf{h}^H\mathbf{w}|^2}{1+\rho|\mathbf{h}^H\mathbf{q}|^2}\Big)\Big)\Big)\Big] + \mathbb{E}\Big[\log_2\Big(1+\exp\Big(\ln\Big(\frac{\rho|\mathbf{g}^H\mathbf{q}|^2}{1+\rho|\mathbf{g}^H\mathbf{w}|^2}\Big)\Big)\Big)\Big] \\
&\overset{(a)}{\geq} \log_2\Big[1+\exp\Big(\mathbb{E}\left(\ln\left(\rho|\mathbf{h}^H\mathbf{w}|^2\right)\right) - \mathbb{E}\left(\ln\left(1+\rho|\mathbf{h}^H\mathbf{q}|^2\right)\right)\Big)\Big] + \\
&\quad \log_2\Big[1+\exp\Big(\mathbb{E}\left(\ln\left(\rho|\mathbf{g}^H\mathbf{q}|^2\right)\right) - \mathbb{E}\left(\ln\left(1+\rho|\mathbf{g}^H\mathbf{w}|^2\right)\right)\Big)\Big] \\
&\overset{(b)}{\approx} \log_2\left(1+\frac{\rho\mathbf{w}^H\mathbf{R}_A\mathbf{w}}{\rho\mathbf{q}^H\mathbf{R}_A\mathbf{q}}\right) + \left(1+\frac{\rho\mathbf{q}^H\mathbf{R}_B\mathbf{q}}{\rho\mathbf{w}^H\mathbf{R}_B\mathbf{w}}\right) \\
&\overset{(c)}{\geq} \log_2\left(\frac{\mathbf{w}^H\mathbf{R}_A\mathbf{w}}{\mathbf{w}^H\mathbf{R}_B\mathbf{w}}\frac{\mathbf{q}^H\mathbf{R}_B\mathbf{q}}{\mathbf{q}^H\mathbf{R}_A\mathbf{q}}\right).
\end{aligned} \tag{A.8}$$

Since $\log_2(1+re^x)$ is convex in x for $r>0$, we can obtain (a) with Jensen's inequality. At high SNR, (b) can be asymptotically approximated by first dropping '1 +' in the parentheses and applying Lemma A.1. The tightness of (b) has been shown in the asymptotic regime ($M \to \infty$) [32]. Moreover, the lower bound in (c) is tight in high-correlated system with proper beamforming vectors. Interestingly, a recent work [104] independently proved that R_{sum} can be well approximated by (b) in massive MIMO system.

With (c) at hand, we can transform the optimisation problem into

$$\max_{\|\mathbf{w}\|=1,\|\mathbf{q}\|=1} R_{\text{sum,lb}} \triangleq \log_2\left(\frac{\mathbf{w}^H\mathbf{R}_A\mathbf{w}}{\mathbf{w}^H\mathbf{R}_B\mathbf{w}}\frac{\mathbf{q}^H\mathbf{R}_B\mathbf{q}}{\mathbf{q}^H\mathbf{R}_A\mathbf{q}}\right), \tag{A.9}$$

for which the generalised eigenvector structure is the optimal solution [105], as shown in eq. (2.6). $\mathbf{w}$ corresponds to the dominant eigenvector of $\mathbf{R}_B^{-1}\mathbf{R}_A$ while $\mathbf{q}$ corresponds to the

weakest one. The corresponding ergodic sum rate satisfies $R_{\text{sum,lb}} = \log_2\big(\chi\big(\mathbf{R}_B^{-1}\mathbf{R}_A\big)\big)$. Both $\mathbf{R}_B^{-1}\mathbf{R}_A$ and $\mathbf{R}_A^{-1}\mathbf{R}_B$ are positive definite, since $\mathbf{R}_A$,$\mathbf{R}_B \in \Phi_{PD}$ and $\big(\mathbf{R}_B^{-1}\mathbf{R}_A\big)^{-1} = \mathbf{R}_A^{-1}\mathbf{R}_B$. It is easy to find that $\chi\big(\mathbf{R}_A^{-1}\mathbf{R}_B\big) = \chi\big(\mathbf{R}_B^{-1}\mathbf{R}_A\big)$ and thereby we can obtain Theorem 2.1.

A.4. Proof of Proposition 2.1

We can lower bound the mutual information in (2.11) applying Minkowski Determinant Theorem [106]

$$\begin{aligned} I_A &= \log_2 \det\left(\mathbf{I}_{2\times 2} + \rho\,\mathbf{M}\right) && \text{(A.10)}\\ &\geq \log_2 \Big(1 + \rho\,\det(\mathbf{M})^{1/2}\Big)^2 && \text{(A.11)}\\ &= 2\log_2\Big[1 + \rho\exp\big(\frac{1}{2}\ln\det(\mathbf{M})\big)\Big], && \text{(A.12)} \end{aligned}$$

where

$$\begin{aligned} \mathbf{M} &\triangleq \widetilde{\mathbf{H}}^H\mathbf{K}^{-1}\widetilde{\mathbf{H}} && \text{(A.13)}\\ &= \Big[\mathbf{W}^H\mathbf{h}_1, \mathbf{W}^H\mathbf{g}_1\Big]\begin{bmatrix} \frac{|h_{21}|^2}{1+|h_{21}|^2} & 0 \\ 0 & |h_{31}|^2 \end{bmatrix}\begin{bmatrix} \mathbf{h}_1^H\mathbf{W} \\ \mathbf{g}_1^H\mathbf{W} \end{bmatrix} && \text{(A.14)}\\ &= \widetilde{\mathbf{G}}_{2\times 2}\,\mathbf{\Lambda}\,\widetilde{\mathbf{G}}_{2\times 2}^H. && \text{(A.15)} \end{aligned}$$

By applying the convexity of $\log_2(1 + re^x), r > 0$ and Jensen's inequality, the ergodic rate of user A per slot can be lower bounded as

$$\begin{aligned} R_A &\geq \frac{2}{3}\,\mathbb{E}\Big\{\log_2\Big[1 + \rho\exp\big(\frac{1}{2}\ln\det(\mathbf{M})\big)\Big]\Big\} && \text{(A.16)}\\ &\geq \frac{2}{3}\log_2\Big[1 + \rho\exp\big(\frac{1}{2}\mathbb{E}[\ln\det(\mathbf{M})]\big)\Big], && \text{(A.17)} \end{aligned}$$

where $\mathbb{E}[\ln\det(\mathbf{M})] = \mathbb{E}[\ln\det(\mathbf{\Lambda})] + \mathbb{E}[\ln\det(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^H)]$. The first term can be further calculated with equations in [107]

$$\begin{aligned} \mathbb{E}[\ln\det(\mathbf{\Lambda})] &= \mathbb{E}\Big[\ln\Big(\frac{|h_{21}|^2}{1+|h_{21}|^2}\Big)\Big] + \mathbb{E}[\ln(|h_{31}|^2)] && \text{(A.18)}\\ &= e\,\mathrm{Ei}(-1) - 2\gamma, && \text{(A.19)} \end{aligned}$$

where (A.19) is obtained by using the fact that $|h_{jm}|^2 \sim \mathrm{Exp}(1)$. In general, it is nontrivial to evaluate the second term. A special case lies in i.i.d Rayleigh fading channel where $\mathbb{E}[\ln\det(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^H)]$ can be exactly solved by invoking central Wishart distribution [108]. For spatially correlated channel, we use Jensen's inequality to upper bound the second term

as

$$\begin{aligned}
\mathbb{E}\big[\ln\det(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^H)\big] &\leq \ln\mathbb{E}[\det(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^H)\big] \\
&\overset{(a)}{=} \ln\big[\mathbb{E}(\mathbf{h}_1^H\mathbf{W}\mathbf{W}^H\mathbf{h}_1\mathbf{g}_1^H\mathbf{W}\mathbf{W}^H\mathbf{g}_1 - \mathbf{h}_1^H\mathbf{W}\mathbf{W}^H\mathbf{g}_1\mathbf{g}_1^H\mathbf{W}\mathbf{W}^H\mathbf{h}_1)\big] \\
&\overset{(b)}{=} \ln\big[\mathbb{E}(\mathbf{h}_1^H\mathbf{W}\mathbf{W}^H\mathbf{h}_1)\mathbb{E}(\mathbf{g}_1^H\mathbf{W}\mathbf{W}^H\mathbf{g}_1) - \mathbb{E}(\mathbf{h}_1^H\mathbf{W}\mathbf{W}^H\mathbf{g}_1\mathbf{g}_1^H\mathbf{W}\mathbf{W}^H\mathbf{h}_1)\big] \\
&\overset{(c)}{=} \ln(\Theta_{\mathrm{A}}), \qquad \text{(A.20)}
\end{aligned}$$

where Θ_{A} is defined in (2.13). Eq. (a) is obtained with $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ for equal-size square matrices $\mathbf{A}, \mathbf{B}$. Eq. (b) is because $\mathbf{h}_1$ and $\mathbf{g}_1$ are independent Gaussian random vectors. With $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, $\mathbb{E}[\mathrm{tr}(\mathbf{C})] = \mathrm{tr}[\mathbb{E}(\mathbf{C})]$, eq. (c) can be easily calculated. Finally, substituting (A.19) and (A.20) into (A.17) renders an analytical approximation of the ergodic rate of user A and hence completes the proof.

A.5. Proof of Proposition 2.2

For arbitrary M and $\mathbf{w}_2$, it is easy to verify that $\mathbf{w}_2^H\,\mathbf{M}(\mathbf{w}_2)\,\mathbf{w}_1 = 0$, i.e., $\mathbf{M}(\mathbf{w}_2)\,\mathbf{w}_1 \in \mathrm{Null}(\mathbf{w}_2)$. The maximisation of $\Theta_A(\mathbf{w}_1) = \mathbf{w}_1^H\,\mathbf{M}(\mathbf{w}_2)\,\mathbf{w}_1$ leads to the observation that the optimum $\mathbf{w}_1 \in \mathrm{Null}(\mathbf{w}_2)$. Similarly, when we fix $\mathbf{w}_1$ and update $\mathbf{w}_2$, we have the optimum $\mathbf{w}_2 \in \mathrm{Null}(\mathbf{w}_1)$. It implies that the optimal beamforming vectors are always orthogonally chosen ($\mathbf{w}_1 \perp \mathbf{w}_2$). For the special $M = 2$ case, since $\mathbf{w}_1$ is uniquely defined in $\mathrm{Null}(\mathbf{w}_2)$ and vice versa, any two beamforming vectors constituting a unitary matrix are optimal. Moreover, eq. (2.13) becomes constant $\Theta_A = \mathrm{tr}(\mathbf{R}_A)\mathrm{tr}(\mathbf{R}_B) - \mathrm{tr}(\mathbf{R}_A\mathbf{R}_B)$.

A.6. Proof of Proposition 2.3

Define $\mathbf{M} \triangleq \widetilde{\mathbf{H}}_1^H\mathbf{K}^{-1}\widetilde{\mathbf{H}}_1\mathbf{P}_A^2$ and with simple manipulations, we have

$$\begin{aligned}
\mathbf{M} &= [\mathbf{W}^H\mathbf{h}_1, \mathbf{W}^H\mathbf{g}_1]\begin{bmatrix} \frac{1}{k_1} + \frac{P_5|h_{21}|^2}{k_2} & 0 \\ 0 & \frac{P_8|h_{31}|^2}{k_3} \end{bmatrix}\begin{bmatrix} \mathbf{h}_1^H\mathbf{W} \\ \mathbf{g}_1^H\mathbf{W} \end{bmatrix}\mathbf{P}_A^2 \qquad \text{(A.21)} \\
&= \widetilde{\mathbf{G}}_{2\times 2}\,\mathbf{\Gamma}\,\widetilde{\mathbf{G}}_{2\times 2}^H\mathbf{P}_A^2. \qquad \text{(A.22)}
\end{aligned}$$

Rewrite (2.25) as

$$\begin{aligned}
R_{\mathbf{s}_A} &= \mathbb{E}\left[\log_2\det\left(\mathbf{I}_{2\times 2} + \mathbf{M}\right)\right] \qquad \text{(A.23)} \\
&= \mathbb{E}\left[\log_2\left(1 + \mathrm{tr}(\mathbf{M}) + \det(\mathbf{M})\right)\right] \qquad \text{(A.24)} \\
&\leq \log_2[1 + \mathbb{E}\left(\mathrm{tr}\left(\mathbf{M}\right)\right) + \mathbb{E}\left(\det\left(\mathbf{M}\right)\right)]. \qquad \text{(A.25)}
\end{aligned}$$

Eq. (A.23) is obtained with $\det(\mathbf{I} + \mathbf{AB}) = \det(\mathbf{I} + \mathbf{BA})$ while (A.24) makes use of the Cayley-Hamilton theorem. Then, we upper bound (A.24) by (A.25) using Jensen's

inequality. With the help of (A.22) and $\mathbb{E}[\mathrm{tr}(\cdot)] = \mathrm{tr}[\mathbb{E}(\cdot)]$, the first term in (A.25) can be characterised as

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{M}\right)\right] &= \mathrm{tr}\left[\left(\mathbb{E}\left(\Gamma_{11}\right)\mathbf{W}^H\mathbf{R}_A\mathbf{W} + \mathbb{E}\left(\Gamma_{22}\right)\mathbf{W}^H\mathbf{R}_B\mathbf{W}\right)\mathbf{P}_A^2\right] && \text{(A.26)}\\ &\approx \delta_{A1}\left(\tau_{A1}P_1 + \tau_{A2}P_2\right) + \delta_{A2}\left(\lambda_{B1}P_1 + \lambda_{B2}P_2\right), && \text{(A.27)}\end{aligned}$$

where $\lambda_{B1} = \mathbf{w}_1^H\mathbf{R}_B\mathbf{w}_1, \lambda_{B2} = \mathbf{w}_2^H\mathbf{R}_B\mathbf{w}_2, \tau_{A1} = \mathbf{w}_1^H\mathbf{R}_A\mathbf{w}_1,\ \tau_{A2} = \mathbf{w}_2^H\mathbf{R}_A\mathbf{w}_2$ and

$$\mathbb{E}\left(\Gamma_{11}\right) = \mathbb{E}\left(\frac{1}{k_1} + \frac{P_5|h_{21}|^2}{k_2}\right), \quad \mathbb{E}\left(\Gamma_{22}\right) = \mathbb{E}\left(\frac{P_8|h_{31}|^2}{k_3}\right). \tag{A.28}$$

The terms on the right hand side of (A.28) can be further evaluated as follows

$$\mathbb{E}\left(\frac{1}{k_1}\right) = \mathbb{E}\left(\frac{1}{1+|\mathbf{h}_1^H\mathbf{Q}\mathbf{P}_B|^2}\right) \overset{(a)}{\geq} \frac{1}{1+\mathbb{E}\left(|\mathbf{h}_1^H\mathbf{Q}\mathbf{P}_B|^2\right)} = \frac{1}{1+\lambda_{A1}P_3+\lambda_{A2}P_4} \tag{A.29}$$

$$\mathbb{E}\left(\frac{P_5|h_{21}|^2}{k_2}\right) = \mathbb{E}\left(\frac{\mathbf{h}_{w,2}^H\mathbf{A}\mathbf{h}_{w,2}}{1+\mathbf{h}_{w,2}^H\mathbf{B}\mathbf{h}_{w,2}}\right) \overset{(b)}{\approx} \frac{\mathrm{tr}(\mathbf{A})}{1+\mathrm{tr}(\mathbf{B})} = \frac{P_5}{1+P_5+\tau_{A3}P_6+\lambda_{A3}P_7} \tag{A.30}$$

where $\lambda_{A1} = \mathbf{q}_1^H\mathbf{R}_A\mathbf{q}_1,\ \lambda_{A2} = \mathbf{q}_2^H\mathbf{R}_A\mathbf{q}_2,\ \tau_{A3} = \mathbf{w}_3^H\mathbf{R}_A\mathbf{w}_3,\ \lambda_{A3} = \mathbf{q}_3^H\mathbf{R}_A\mathbf{q}_3$. Inequality (a) comes from the fact that $\frac{1}{x}$ is convex in x for $x > 0$. Note that (A.29) can be exactly calculated as an exponential integral function of $\lambda_{A1}P_3,\ \lambda_{A2}P_4$. Nevertheless, such implicit characterisation restrains insightful analysis of the power allocation strategy (for instance, how the power assigned to signal of user B interferes user A). In (A.30), $\mathbf{A} = P_5\mathbf{R}_A^{1/2}\mathbf{x}_1\mathbf{x}_1^H\mathbf{R}_A^{1/2}$ where $\mathbf{x}_1 = [1,\,0]^T$ and $\mathbf{B} = \mathbf{R}_A^{1/2}(P_5\mathbf{x}_1\mathbf{x}_1^H + P_6\mathbf{w}_1\mathbf{w}_1^H + P_7\mathbf{q}_1\mathbf{q}_1^H)\mathbf{R}_A^{1/2}$. Inequality (b) is based on the first order approximation in (A.6). The second (and higher) order approximation would be more accurate, however, rendering the problem too complicated to implement optimisation techniques[1]. Similarly, we can approximate $\mathbb{E}\left(\Gamma_{22}\right)$ as

$$\mathbb{E}\left(\frac{P_8|h_{31}|^2}{k_3}\right) = \mathbb{E}\left(\frac{\mathbf{h}_{w,3}^H\mathbf{C}\mathbf{h}_{w,3}}{1+\mathbf{h}_{w,3}^H\mathbf{D}\mathbf{h}_{w,3}}\right) \approx \frac{\mathrm{tr}(\mathbf{C})}{1+\mathrm{tr}(\mathbf{D})} = \frac{P_8}{1+\tau_{A3}P_9+\lambda_{A3}P_{10}}. \tag{A.31}$$

$\mathbf{D} = \mathbf{R}_A^{1/2}(P_9\mathbf{w}_1\mathbf{w}_1^H + P_{10}\mathbf{q}_1\mathbf{q}_1^H)\mathbf{R}_A^{1/2}$ and $\mathbf{C} = P_8\mathbf{R}_A^{1/2}\mathbf{x}_1\mathbf{x}_1^H\mathbf{R}_A^{1/2}$. The second term in (A.25) can be given by

$$\mathbb{E}\left[\det\left(\mathbf{M}\right)\right] = \mathbb{E}\left[\det\left(\mathbf{\Gamma}\right)\right]\cdot\mathbb{E}\left[\det\left(\widetilde{\mathbf{G}}\,\widetilde{\mathbf{G}}^H\right)\right]\cdot\mathbb{E}\left[\det\left(\mathbf{P}_A^2\right)\right] \approx \delta_{A1}\delta_{A2}\Theta_{\mathrm{A}}P_1P_2, \tag{A.32}$$

where calculation of $\mathbb{E}[\det(\widetilde{\mathbf{G}}\,\widetilde{\mathbf{G}}^H)]$ follows $(a), (b), (c)$ of eq. (A.20). Substituting (A.29) $\sim$ (A.31) into (A.26), we can obtain (A.27). Combining (A.27) and (A.32) with (A.25)

[1]For instance, it is difficult to compute the first/second order derivatives of the objective function which are necessary for various non-linear programming methods.

establishes (2.28). In order to compute $R^p_{\mathbf{s}_A}$, we can reexpress (2.27) as

$$R^p_{\mathbf{s}_A} = \mathbb{E}\Big[\log_2\Big(1+\frac{P_6|\mathbf{h}_2^H\mathbf{w}_3|^2}{1+P_5|h_{21}|^2+P_7|\mathbf{h}_2^H\mathbf{q}_3|^2}\Big)\Big] + \mathbb{E}\Big[\log_2\Big(1+\frac{P_9|\mathbf{h}_3^H\mathbf{w}_3|^2}{1+P_{10}|\mathbf{h}_3^H\mathbf{q}_3|^2}\Big)\Big] \tag{A.33}$$

$$\leq \log_2\Big[1+\mathbb{E}\Big(\frac{P_6|\mathbf{h}_2^H\mathbf{w}_3|^2}{1+P_5|h_{21}|^2+P_7|\mathbf{h}_2^H\mathbf{q}_3|^2}\Big)\Big] + \log_2\Big[1+\mathbb{E}\Big(\frac{P_9|\mathbf{h}_3^H\mathbf{w}_3|^2}{1+P_{10}|\mathbf{h}_3^H\mathbf{q}_3|^2}\Big)\Big] \tag{A.34}$$

$$\approx \log_2\Big(1+\frac{\tau_{A3}P_6}{1+P_5+\lambda_{A3}P_7}\Big) + \log_2\Big(1+\frac{\tau_{A3}P_9}{1+\lambda_{A3}P_{10}}\Big). \tag{A.35}$$

An analytical expression of (A.33) was obtained for the case $M = 2$ in [6], while a lower bound for $M > 2$ case is derived in section 2.2. We here use Jensen's inequality and (A.6) in Lemma A.2 to estimate (A.33), leading to an approximation (A.34) as well as (2.28).

A.7. Proof of Theorem 2.2

At high SNR, problem (2.35) can be rewritten as

$$\begin{aligned}\max_{\{P_i\}} R_{\text{sum}} &\overset{(a)}{\approx} \log_2(\delta_{A1}\delta_{A2}\Theta_A P_1P_2) + \log_2(\delta_{B1}\delta_{B2}\Theta_B P_3P_4) + \log_2\Big(1+\frac{\tau_{A3}P_6}{1+P_5+\lambda_{A3}P_7}\Big)\\ &+ \log_2\Big(1+\frac{\tau_{A3}P_9}{1+\lambda_{A3}P_{10}}\Big) + \log_2\Big(1+\frac{\tau_{B3}P_7}{1+\lambda_{B3}P_6}\Big) + \log_2\Big(1+\frac{\tau_{B3}P_{10}}{1+P_8+\lambda_{B3}P_9}\Big)\\ \text{s.t.}\quad & \bar{P}_{\text{c}} - 3P = 0,\ P_i \geq 0 \quad i = 1,\ldots,10\end{aligned} \tag{A.36}$$

where $\overset{(a)}{\approx}$ comes from the fact that the last terms in (2.28) and (2.30) are dominant at high SNR. Based on KKT necessary conditions, there exist multipliers λ and μ_1, μ_2 such that

$$\begin{cases}\nabla R_{\text{sum}}(P_1) = \lambda(1+\lambda_{B1}P_8) + \mu_1\\ \nabla R_{\text{sum}}(P_2) = \lambda(1+\lambda_{B2}P_8) + \mu_2\\ \mu_1P_1 = 0,\ \mu_2P_2 = 0,\end{cases} \tag{A.37}$$

where $P_1 \neq 0, P_2 \neq 0$, otherwise DoF loss occurs due to $R_{\mathbf{s}_A} = 0$. Therefore, we have $\mu_1 = \mu_2 = 0$ and the first equation in (2.36) can be computed from (A.37). Likewise, the second equation can be obtained.

B. Proofs for Chapter 3

B.1. Proof of Proposition 3.1

By dividing the objective function of (3.10) by M^2 and plugging (3.9) into (3.10), the problem $\mathcal{P}1$ is equivalently transformed to $\mathcal{P}2$

$$\mathcal{P}2: \quad \max_{a_k} \min_k \pi_k (1-\tau_k^2) \cdot a_k^2 \tag{B.1}$$

$$\text{s.t.} \quad \sum_k a_k^2 = \frac{1}{M}. \tag{B.2}$$

From [66, Lemma 2], the optimal solution of problem $\mathcal{P}2$ is obtained when all terms are equal, i.e., $\pi_k (1-\tau_k^2) \cdot a_k^2 = \pi_j (1-\tau_j^2) \cdot a_j^2$, $\forall k \neq j$ and the optimal $\{a_k^\star\}$ are given by (3.11).

B.2. Proof of Proposition 3.2

When $\frac{Pt}{K} (\xi^\circ)^2 \Upsilon_k^\circ \Omega_k > 1$, the equality in (3.22) is nearly established, i.e., the private messages of RS achieve approximately the same sum rate as the conventional No-RS with full power. The power splitting ratio t is then designed as $\frac{Pt}{K} (\xi^\circ)^2 \Upsilon_k^\circ \Omega_k = K$, i.e., $t = K/(P \Upsilon_k^\circ \Omega_k / \Psi^\circ)$. The rationale behind this design is two-fold. On the one hand, the number of users K is generally much larger than 1, which leads to an asymptotically tight approximation. On the other hand, the achievable rate of the common message decreases as K increases due to minimum constraint. This effect can be observed via $\eta = M/K$ in the asymptotic $\text{SINR}^{c,\circ}$. Then, the power allocated to the common message $P(1-t)$ should be reduced as K becomes larger. Otherwise, suppose $P(1-t)$ is constant independent of K. As K increases in (3.23), the achievable rate of the common message $\log_2(1+\text{SINR}^{c,\circ})$ cannot compensate the loss $\sum_{k=1}^K \left(\log_2(1+\text{SINR}_k^{p,\circ}) - \log_2\left(1+\text{SINR}_k^{RZF,\circ}\right)\right)$ incurred from the above approximation. Moreover, Ω_k in (3.14) can be approximated by τ^2 and the approximation is tight when m_k° is large. Thus, $t = K/(P\Gamma_k)$, where $\Gamma_k = (\Upsilon_k^\circ \tau_k^2)/\Psi^\circ$. To establish the equality (3.22) for $\forall k$, the power splitting ratio t is chosen as the largest one. We then obtain (3.24) by truncating t at 1 wherever applicable.

At low SNR, $t = 1$ from (3.24) turns RS into No-RS and leads to $\Delta R^{RS,\circ} = 0$. Namely, transmitting multiple private messages is operated in the non-interference limited SNR regime and thereby a common message is unnecessary. At high SNR, $t < 1$ indicates

that we transmit a common message with remaining power beyond the saturation of the private message transmission. Due to power reduction to the private messages, we first upper bound the rate loss

$$
\begin{aligned}
R_{\text{sum}}^{RZF,\circ} - R_p^{RS,\circ} &= \sum_{k=1}^{K} \Big(\log_2 \big(1 + \frac{S}{P\Gamma_k + 1}\big) - \log_2 \big(1 + \frac{S}{P\Gamma_k + \frac{1}{t}}\big) \Big) && \text{(B.3)} \\
&= \sum_{k=1}^{K} \big(\log_2(S + P\Gamma_k + 1) - \log_2(S + P\Gamma_k + \frac{1}{t}) && \text{(B.4)} \\
&\qquad + \log_2(P\Gamma_k + \frac{1}{t}) - \log_2(P\Gamma_k + 1) \big) && \text{(B.5)} \\
&\overset{(a)}{\leq} \sum_{k=1}^{K} \Big(\log_2(P\Gamma_k + \frac{1}{t}) - \log_2(P\Gamma_k + 1) \Big) && \text{(B.6)} \\
&\overset{(b)}{\leq} \sum_{k=1}^{K} \big(\log_2(1 + \frac{1}{K}) - \log_2(1 + \frac{1}{P\Gamma_k}) \big) && \text{(B.7)} \\
&\overset{(c)}{\leq} K \log_2(1 + 1/K) && \text{(B.8)} \\
&\overset{(d)}{\leq} \log_2(e), && \text{(B.9)}
\end{aligned}
$$

where $S = \frac{P}{K} \left(\xi^\circ\right)^2 \Phi_k$. (a) is obtained since $1/t \geq 1, \forall t \in (0,\, 1]$. By replacing $t = K/(P\Gamma)$ with $\Gamma = \min_k \{\Gamma_k\}$ by $t = K/(P\Gamma_k)$, (a) is lower bounded as (b). Removing $\log_2(1 + \frac{1}{P\Gamma_k})$, we have (c) which is tight at high SNR. (d) is obtained since $K \log_2(1+1/K) \in [1, \log_2(e))$ is an increasing function of K for $K \geq 1$. By plugging (d) into (3.23), we obtain (3.25).

B.3. Proof of Proposition 3.3

The inter-group interference is captured by term $\bar{\mathbf{R}}_{gl} = \mathbf{B}_l^H \mathbf{R}_g \mathbf{B}_l$, $\forall g \neq l$. Firstly, we consider the weak inter-group interference case, i.e., $\bar{\mathbf{R}}_{gl} \approx \mathbf{0}_{b' \times b'}$ and further $\Upsilon_{gl}^\circ \approx 0$, $\forall g \neq l$, i.e., the inter-group interference is sufficiently small and therefore can be negligible. The sum rate of the private messages transmission is limited by intra-group interference. Based on (3.35) to (3.37) and (3.43), the outer common message suffers from more interference while contributing less rate than the inner common messages, since the achievable rate of the outer common message has a pre-log factor of 1 which is smaller than that of the inner common messages ($G > 1$). The optimal β that maximises the sum rate of HRS (3.43) is $\beta = 1$. Then, (3.37) and (3.44) become

$$
\text{SINR}_g^{p,\circ} = \frac{\alpha_g \frac{P}{K} \left(\xi_g^\circ\right)^2 \Phi_g}{\alpha_g \left(\xi_g^\circ\right)^2 \Upsilon_{gg}^\circ \Omega_g + 1}, \quad \text{SINR}_g^{TTP,\circ} = \frac{\frac{P}{K} \left(\xi_g^\circ\right)^2 \Phi_g}{\left(\xi_g^\circ\right)^2 \Upsilon_{gg}^\circ \Omega_g + 1}. \tag{B.10}
$$

Substituting (B.10) into (3.46), the equality is approximately established given that $\alpha_g (\xi_g^\circ)^2 \Upsilon_{gg}^\circ \Omega_g > 1$. Following a similar philosophy of t in Proof of Proposition 3.2, the

intra-group power splitting ratio is designed as $\alpha_g = K_g/\big((\xi_g^\circ)^2\Upsilon_{gg}^\circ\Omega_g\big)$. Otherwise, as K_g increases, the achievable rate of the inner common message $\log_2(1+\mathrm{SINR}_g^{c,\circ})$ cannot compensate the loss $K_g\big(\log_2(1+\mathrm{SINR}_g^{p,\circ})-\log_2(1+\mathrm{SINR}_g^{TTP,\circ})\big)$ incurred from the above approximation. Based on (3.38) to (3.42), α_g is determined by

$$\big(\xi_g^\circ\big)^2\Upsilon_{gg}^\circ\Omega_g = \frac{PK_g}{K}\frac{\mathrm{tr}\big(\bar{\mathbf{R}}_{gg}\mathbf{T}_g\bar{\mathbf{R}}_{gg}\mathbf{T}_g\big)}{\mathrm{tr}\big(\bar{\mathbf{R}}_{gg}\mathbf{T}_g^2\big)}\Omega_g. \tag{B.11}$$

In order to obtain a more insightful understanding of the effects of system parameters, we consider a high SNR approximation of (B.11). At high SNR ($\varepsilon \approx 0$), the RZF matrix in (3.33) converges to the ZF matrix. From [7, Theorem 3], $\mathbf{T}_g$ in (3.42) becomes

$$\mathbf{T}_g = \left(\frac{K_g}{b_g}\frac{\bar{\mathbf{R}}_{gg}}{m_g^\circ} + \mathbf{I}_{b_g}\right)^{-1} \approx \left(\frac{K_g}{b_g}\frac{\bar{\mathbf{R}}_{gg}}{m_g^\circ}\right)^{-1}, \tag{B.12}$$

where $\bar{\mathbf{R}}_{gg}$ is a diagonal matrix from (3.32). Since $\mathbf{B}_g$ lies in the dominant eigenmodes of $\mathbf{R}_g$, the diagonal elements of $\bar{\mathbf{R}}_{gg}$ are much larger than 1 and therefore the approximation in (B.12) is feasible. Moreover, we have $\Omega_g \approx \frac{K_g-1}{K_g}\tau_g^2$ due to the fact $m_g^\circ \geq 1$ in the asymptotic M regime. Plugging (B.12) into (B.11) leads to $\big(\xi_g^\circ\big)^2\Upsilon_{gg}^\circ\Omega_g \approx \frac{P}{K}\frac{b_g(K_g-1)}{\mathrm{tr}\big(\bar{\mathbf{R}}_{gg}^{-1}\big)}\tau_g^2$. Since α is applied to all groups, we choose the largest one to guarantee (3.46)

$$\alpha = \frac{K_g}{P\cdot\Gamma_{IG}}, \quad \Gamma_{IG} = \min_g\left\{\frac{\tau_g^2}{K}\frac{b_g(K_g-1)}{\mathrm{tr}\big(\bar{\mathbf{R}}_{gg}^{-1}\big)}\right\}. \tag{B.13}$$

Secondly, consider the case with $\sum_{l\neq g}(\xi_l^\circ)^2\Upsilon_{gl}^\circ > (\xi_g^\circ)^2\Upsilon_{gg}^\circ$. Since $\Omega_g < 1$ from (3.39), the sum rate of the private messages based on (3.44) is dominated by inter-group interference. Substituting (3.37), (3.44) into (3.46), the equality is approximately established when $\beta\sum_{l\neq g}\big(\xi_l^\circ\big)^2\Upsilon_{gl}^\circ + \beta\alpha\big(\xi_g^\circ\big)^2\Upsilon_{gg}^\circ\Omega_g > 1$ and $\alpha = 1$. Following a similar philosophy of t in Proof of Proposition 3.2, the inter-group power splitting ratio can be designed as $\beta_g = K/\big(\sum_{l\neq g}(\xi_l^\circ)^2\Upsilon_{gl}^\circ + (\xi_g^\circ)^2\Upsilon_{gg}^\circ\Omega_g\big)$. However, we adopt a conservative design of β_g as

$$\beta_g = \frac{K}{\sum_{l\neq g}\big(\xi_l^\circ\big)^2\Upsilon_{gl}^\circ + K_g} \geq \frac{K}{\sum_{l\neq g}\big(\xi_l^\circ\big)^2\Upsilon_{gl}^\circ + \big(\xi_g^\circ\big)^2\Upsilon_{gg}^\circ\Omega_g}, \tag{B.14}$$

which is due to the fact that $(\xi_g^\circ)^2\Upsilon_{gg}^\circ\Omega_g > K_g$ at high SNR (interference regime). The rationale behind this conservative design is two-fold. Larger β is more capable to maintain (3.46). Furthermore, it enables a distributed design of power allocation, i.e., β is determined only by the long-term inter-group interference. By plugging (3.38) to (3.40) and (B.12) into (B.14) and denoting β as the largest β_g, we have

$$\beta = \frac{K}{P\cdot\Gamma_{OG}+K_g}, \quad \Gamma_{OG} = \min_g\left\{\sum_{l\neq g}\frac{K_g}{K}\frac{\mathrm{tr}\big(\bar{\mathbf{R}}_{gl}\bar{\mathbf{R}}_{ll}^{-1}\big)}{\mathrm{tr}\big(\bar{\mathbf{R}}_{ll}^{-1}\big)}\right\} \tag{B.15}$$

Since $0 < \alpha, \beta \leq 1$, we assume implicitly that $\forall \alpha, \beta > 1$ is truncated at 1 wherever applicable.

B.4. Proof of Corollary 3.1

We here provide a sketch proof, since it follows a similar philosophy of $\Delta R^{RS,\circ}$ in Proof of Proposition 3.2. In the weak inter-group interference regime, $\beta = 1$ from (3.47). We first upper bound the rate loss $R_{\text{sum}}^{TTP,\circ} - R_p^{HRS,\circ}$ at high SNR by

$$\begin{aligned} R_{\text{sum}}^{TTP,\circ} - R_p^{HRS,\circ} &= \sum_{g=1}^{G} K_g \Big(\log_2 \big(1 + \frac{S_g}{\Gamma_g + 1}\big) - \log_2 \big(1 + \frac{S_g}{\Gamma_g + \frac{1}{\alpha}}\big)\Big) && \text{(B.16)} \\ &\leq \sum_{g=1}^{G} K_g \log_2(1 + 1/K_g) && \text{(B.17)} \\ &\leq G \log_2(e), && \text{(B.18)} \end{aligned}$$

where $S_g = \frac{P}{K}\big(\xi_g^\circ\big)^2 \Phi_g$ and $\Gamma_g = (\xi_g^\circ)^2 \Upsilon_{gg}^\circ \Omega_g$. The sum rate gain $\Delta R^{HRS,\circ}$ is lower bounded as (3.50). In the strong inter-group interference regime, the rate loss is upper bounded as

$$\begin{aligned} R_{\text{sum}}^{TTP,\circ} - R_p^{HRS,\circ} &= \sum_{g=1}^{G} K_g \Big(\log_2 \big(1 + \frac{S_g}{\Gamma_g + 1}\big) - \log_2 \big(1 + \frac{S_g}{\Gamma_g + \frac{1}{\beta}}\big)\Big) && \text{(B.19)} \\ &\leq \sum_{g=1}^{G} K_g \log_2(1 + 1/K) && \text{(B.20)} \\ &= K \log_2(1 + 1/K) && \text{(B.21)} \\ &\leq \log_2(e), && \text{(B.22)} \end{aligned}$$

where $\Gamma_g = \sum_{l \neq g} \left(\xi_l^\circ\right)^2 \Upsilon_{gl}^\circ + \left(\xi_g^\circ\right)^2 \Upsilon_{gg}^\circ \Omega_g$. Then, the sum rate gain is lower bounded as (3.51).

C. Proofs for Chapter 4

C.1. Proof of Proposition 4.1

Let us first consider the non-overlapping case, i.e., $\|\mathbf{d}_k + \mathbf{d}_j\|_0 = 2L, \forall j \neq k$. Given DFT codebook and $B = \log_2(M)$, the RF beamformer of user k in (4.7) that maximises the signal power is given by $\mathbf{f}_k = \mathbf{e}_{k,1}$. Then, we have $\mathbf{F} = [\mathbf{e}_{1,1}, \cdots, \mathbf{e}_{K,1}]$ and $\mathbf{R}_{k,\text{eff}} = \mathbf{F}^H \mathbf{R}_k \mathbf{F} = \text{diag}\{\mathbf{1}_k\}$. A straightforward calculation of (4.11) gives $\mathbf{w}_k = \mathbf{w}_k^\star = \mathbf{1}_k$. Noting that $\mathbf{h}_k = \sqrt{\frac{M}{L}} \sum_{l=1}^{L} g_{k,l}\, \mathbf{e}_{k,l}$, the achievable rate of user k is thus given by

$$R_k = \log_2\Big(1 + \frac{\rho |\mathbf{h}_k^H \mathbf{f}_k|^2}{1 + \rho \sum_{j \neq k} |\mathbf{h}_k^H \mathbf{f}_j|^2}\Big) \tag{C.1}$$

$$= \log_2\Big(1 + \rho \frac{M}{L} |g_{k,1}|^2\Big), \tag{C.2}$$

which is obtained by the fact that $\mathbf{f}_j = \mathbf{e}_{j,1} \perp \mathbf{e}_{k,l}, \forall k \neq j$ in the non-overlapped scenario.

Secondly, we consider the fully overlapped case where all users share the identical steering vectors set $\{\mathbf{e}_l\}$ and therefore the same channel covariance matrix $\mathbf{R}_k = \mathbf{R} = \frac{M}{L} \sum_{l=1}^{L} \mathbf{e}_l \mathbf{e}_l^H$. Without loss of generality, we have $\mathbf{f}_k \in \mathbf{e}_{l=1,\cdots,L}$ with $\mathbf{f}_k \neq \mathbf{f}_{j|j\neq k}$ from (4.7) and $\mathbf{R}_{k,\text{eff}} = \mathbf{F}^H \mathbf{R}_k \mathbf{F} = \frac{M}{L} \mathbf{I}_K$. It can also be obtained that $\mathbf{w}_k = \mathbf{w}_k^\star = \mathbf{1}_k$. Noting that $\mathbf{h}_k = \sqrt{\frac{M}{L}} \sum_{l=1}^{L} g_{k,l}\, \mathbf{e}_{k,l}$ with $\forall \mathbf{e}_{k,l} \in \mathbf{e}_l$, the achievable rate of user k is given by

$$R_k = \log_2\Big(1 + \frac{\rho |\mathbf{h}_k^H \mathbf{f}_k|^2}{1 + \rho \sum_{j \neq k} |\mathbf{h}_k^H \mathbf{f}_j|^2}\Big) \tag{C.3}$$

$$= \log_2\Big(1 + \frac{\rho \frac{M}{L} |g_{k,1}|^2}{1 + \rho \frac{M}{L} \sum_{j \neq k} |g_{k,l_j(l_j \neq 1)}|^2}\Big), \tag{C.4}$$

where $\mathbf{f}_j$ corresponds to $\mathbf{e}_{k,l_j}$. Then, (C.4) can be further lower bounded by (4.17) and the equality follows $L = K$.

C.2. Proof of Proposition 4.3

The average sum rate of RS can be written as

$$\mathbb{E}(R_{\text{sum}}^{RS}) = \mathbb{E}(R^c) + \sum_{k=1}^{K} \mathbb{E}(R_k^p). \tag{C.5}$$

We first compute the average rate of the private message intended to user k as

$$\begin{aligned}
\mathbb{E}(R_k^p) &= \mathbb{E}\Big[\log_2\Big(1+\frac{P_k\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k|^2}{1+\sum_{j\neq k}P_j\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_j|^2}\Big)\Big] && \text{(C.6)}\\
&= \mathbb{E}\Big[\log_2\Big(1+\exp\Big(\ln\frac{P_k\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k|^2}{1+\sum_{j\neq k}P_j\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_j|^2}\Big)\Big)\Big] && \text{(C.7)}\\
&\geq \log_2\Big[1+\exp\Big(\mathbb{E}\ln(P_k\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k|^2)-\mathbb{E}\ln(1+\sum_{j\neq k}P_j\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_j|^2)\Big)\Big] && \text{(C.8)}\\
&\geq \log_2\Big[1+\exp\Big(\mathbb{E}\ln(P_k\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k|^2)-\ln(1+\sum_{j\neq k}P_j\,\mathbf{w}_j^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_j)\Big)\Big] && \text{(C.9)}\\
&= \log_2\Big[1+\exp\Big(\ln(P_k\,\mathbf{w}_k^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_k)-\gamma-\ln(1+\sum_{j\neq k}P_j\,\mathbf{w}_j^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_j)\Big)\Big] && \text{(C.10)}\\
&= \log_2\Big(1+\frac{e^{-\gamma}\cdot\frac{Pt}{K}\mathbf{w}_k^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_k}{1+\frac{Pt}{K}\sum_{j\neq k}\mathbf{w}_j^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_j}\Big), && \text{(C.11)}
\end{aligned}$$

where (C.8) is due to the convexity of $\log_2(1+e^x)$ in x while (C.9) is obtained by applying Jensen's inequality. Define the rank one matrix $\mathbf{X}_k \triangleq \mathbf{A}_k^H\mathbf{F}\mathbf{w}_k\mathbf{w}_k^H\mathbf{F}^H\mathbf{A}_k$ and decompose it as $\mathbf{X}_k = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{U}_k^H$. Then, we have $|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k|^2 = \mathbf{g}_k^H\mathbf{X}_k\mathbf{g}_k \overset{d}{=} \mathbf{g}_k^H\mathbf{\Lambda}_k\mathbf{g}_k = \lambda_k|g_{k,m}|^2$, where $\overset{d}{=}$ indicates the equivalence in distribution and λ_k is the only non-zero entry of $\mathbf{\Lambda}_k$. Since $|g_{k,m}|^2 \sim \text{Exp}(1)$ and $\lambda_k = \text{tr}(\mathbf{\Lambda}_k) = \text{tr}(\mathbf{X}_k) = \mathbf{w}_k^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_k$, we can obtain (C.10) by $\mathbb{E}[\ln(\mathbf{g}_k^H\mathbf{X}_k\mathbf{g}_k)] = \ln(\mathbf{w}_k^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_k) - \gamma$, where γ is the Euler constant.

A direct calculation of $\mathbb{E}(R^c) = \mathbb{E}[\min_k\,(R_k^c)]$ is technically challenging due to the requirement of the distributions of R_k^c and further $\min_k\,(R_k^c)$. We assume that $\mathbb{E}[\min_k\,(R_k^c)]$ can be well approximated by $\min_k\,\mathbb{E}(R_k^c)$. By following a similar derivation, we can compute the average rate of the common message seen by user k as

$$\begin{aligned}
\mathbb{E}(R_k^c) &= \mathbb{E}\Big[\log_2\Big(1+\frac{P_c\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_c|^2}{1+\sum_{j=1}^{K}P_j\,|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_j|^2}\Big)\Big]\\
&\geq \log_2\Big(1+\frac{e^{-\gamma}\cdot P(1-t)\,\mathbf{w}_c^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_c}{1+\frac{Pt}{K}\sum_{j=1}^{K}\mathbf{w}_j^H\mathbf{R}_{k,\text{eff}}\mathbf{w}_j}\Big). && \text{(C.12)}
\end{aligned}$$

Finally, combining (C.11) and (C.12) completes the proof. Moreover, the effectiveness of Proposition 3 is also supported by [109, Lemma 1]. It states that if $X = \sum_{i=1}^{n_1} x_i$, $Y = \sum_{j=1}^{n_2} y_j$ with random variables $x_i, y_j \in \mathbb{R}_{\geq 0}$, we get $\mathbb{E}[\log_2(1+X/Y)] \approx \log_2[1+\mathbb{E}(X)/\mathbb{E}(Y)]$ and the approximation error decreases as the number of random variables n_1 and n_2 increases. It provides a useful reference calculation of the average rate which can be well approximated by $\mathbb{E}[\log_2(1+S/I)] \approx \log_2[1+\mathbb{E}(S)/\mathbb{E}(I)]$, where S and I represent the signal power and interference plus noise, respectively. Indeed, this average rate approximation is lower bounded by $\log_2[1+e^{-\gamma}\mathbb{E}(S)/\mathbb{E}(I)]$ derived in Proposition 3, where $e^{-\gamma} \approx 0.56$.