Télécom Paris (ENST)
Institut Eurécom

THESIS

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
from Ecole Nationale Supérieure
des Télécommunications

Specialization: Communications and Electronics

Marios Kountouris

Multiuser Multi-antenna Systems with Limited Feedback

President	J.C.Belfiore, ENST (Paris,France)
Reviewers	C.Papadias, AIT (Athens, Greece)
	M.Debbah, Supélec (Gif-sur-Yvette, France)
Examiners	A.I. Pérez-Neira, UPC (Barcelona, Spain)
	T. Sälzer, France Telecom R&D (Paris, France)
Thesis supervisor	D. Gesbert, Eurecom Institute (Sophia-Antipolis, France)

January 10th 2008

Télécom Paris (ENST)

Institut Eurécom

THESE

Présentée pour obtenir le Grade de Docteur
de l'Ecole Nationale Supérieure
des Télécommunications

Spécialité: Communications et Electronique

Marios Kountouris

Systèmes multi-antennes multi-utilisateurs avec voie de retour limitée

Président	J.C.Belfiore, ENST (Paris,France)
Rapporteurs	C.Papadias, AIT (Athènes, Grèce)
	M.Debbah, Supélec (Gif-sur-Yvette, France)
Examinateurs	A.I. Pérez-Neira, UPC (Barcelone, Espagne)
	T. Sälzer, France Telecom R&D (Paris, France)
Directeur de Thèse	D. Gesbert, Eurecom Institute (Sophia-Antipolis, France)

January 10^{th} 2008

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor and friend Prof. David Gesbert for his brilliant supervision and his continual guidance and support throughout the years of my Ph.D. Without his technical insight, creativity and on-going encouragement, this thesis would have never been possible. It has been a real pleasure and privilege to have had David as a mentor.

I would like to acknowledge France Telecom R&D for the financial support of my work. A special and warm thank to my industrial supervisor Dr. Thomas Sälzer, for his support and constructive criticism as well as for providing the proper conditions to pursue my research. I would also like to thank Anne-Gaële Acx for hosting me in her group, as well as all the team members with whom I interacted during my seven-month stay in France Telecom's lab in Paris.

I am very grateful to Prof. Constantinos Papadias and Prof. Mérouane Debbah for taking the time to read the first version of my dissertation and to serve as readers. I would also like to thank Prof. Jean Claude Belfiore and Prof. Ana Pérez-Neira for accepting to be part of my thesis committee. The invaluable feedback of all the Ph.D Jury members is enormously appreciated.

I would like to express my appreciation to my colleagues and friends at Eurecom Institute for the excellent and truly enjoyable ambiance. Special thanks go to Ruben de Francisco, Saad Kiani, Mari Kobayashi, Maxime Guillaud, and Issam Toufik. I am also thankful to my co-authors Prof. Dirk Slock and Ruben de Francisco. Part of this thesis would not have been possible without their stimulating discussions and help. My warmest thanks extend to my dear friends, in France, back in Greece and in many other corners of the globe, for all the unforgettable moments I shared with them over the past years.

Finally, I want to express my gratitude to my family for their unconditional love, support, and encouragement. I am deeply indebted to Teodora for being a boundless source of support, patience, and inspiration. Thank you for bringing so much sincere love and happiness to my life.

Marios Kountouris

Sophia-Antipolis

January 10, 2008

Abstract

The use of multiple antennas has been recognized as a key technology to significantly improve the spectral efficiency of next-generation, multiuser wireless communication networks. In multiuser multiple-input multiple-output (MIMO) networks, the spatial degrees of freedom offered by multiple antennas can be advantageously exploited to enhance the system capacity, by scheduling multiple users simultaneously by means of spatial division multiple access (SDMA). A linear increase in throughput, proportional to the number of transmit antennas, can be achieved even by using linear precoding strategies if combined with efficiently designed scheduling protocols. However, these promising gains come under the often unrealistic assumption of close-to-perfect channel state information at the transmitter (CSIT). Therefore, at the heart of the downlink resource allocation problem lies that of feedback acquisition.

In this thesis, we focus on linear beamforming techniques relying on low-rate partial CSIT. Several methods that allow the base station (BS) to live well even with coarse, limited channel knowledge are identified. One first key idea is based on splitting the design between the scheduling and the final beam design stages, thus taking profit from the fact the number of users to be served at each scheduling slot is much smaller than the total number of active users. This two-stage approach is applied to a scenario in which random beamforming (RBF) is exploited to identify good, spatially separable, users in the first stage. In the second stage, several refinement strategies, including beam power control and beam selection, are proposed, offering various feedback reduction and significant sum rate gains, even in sparse network settings (low to moderate number of users).

In channels that exhibit some form of correlation, either in temporal or in spatial domain, we point out that significant useful information for the SDMA scheduler lies hidden in the channel structure. We show how memory-based RBF can exploit channel redundancy in order to achieve throughput close to that of optimum unitary beamforming with full CSIT for slow time-varying channels. In spatially correlated channels, long-term statistical CSIT, which can be easily obtained with negligible per-slot or no feedback overhead, reveals information about the mean spatial separability of users. A maximum likelihood (ML) channel estimation framework is proposed, which effectively combines slowly varying statistical CSIT with instantaneous low-rate channel quality information (CQI). User selection and beamforming techniques suitable for such settings are also proposed. It is demonstrated that in systems with reasonably limited angle spread at the BS, feeding back a single scalar CQI parameter per user is sufficient to perform SDMA scheduling and beamforming with near optimum performance.

Limited feedback strategies utilizing vector quantization codebooks are also investigated.

In particular, the problem of efficient, sum-rate maximizing CQI design is addressed and several scalar feedback metrics are proposed. These metrics are built upon inter-user interference bounds and can be interpreted as reliable estimates of the received signal-to-interference-plus-noise ratio (SINR) at the receiver side. It is shown that scalar CQI feedback combined with channel directional information (CDI), zero-forcing beamforming, and greedy user selection algorithms can achieve a significant fraction of the capacity of the full CSIT case by exploiting multiuser diversity. An efficient technique that provides the BS the flexibility to switch from multiuser (SDMA) to single-user (TDMA) transmission is provided, exhibiting linear sum-rate growth at any range of signal-to-noise ratio (SNR).

Further feedback compression can be achieved if the CSIT information utilized by the scheduler is represented by ranking-based feedback. We show that an integer value is often sufficient in order to identify users with favorable channel conditions. In parallel, it equalizes the channel access probability in networks where users' channels are not necessarily identically distributed and mobile terminals experience unequal average SNRs due to different distances from the BS and the corresponding different path losses (near-far effects).

Contents

Acknowledgements i
Abstract iii
List of Figures ix
List of Tables xiii
Nomenclature xv

1 Introduction **1**
1.1 Background and Motivation 1
1.2 From Single-user to Multiuser MIMO Communications 2
1.3 Assumptions 3
1.4 Contributions and Outline of the Dissertation 4

2 Multi-antenna Broadcast Channels **9**
2.1 The Wireless Channel 9
2.1.1 Path loss 10
2.1.2 Shadowing 10
2.1.3 Fading 10
2.1.4 Channel Selectivity 11
2.2 Multiple-Input Multiple-Output Channels 13
2.3 Multiuser Multi-Antenna Systems 14
2.3.1 Multi-antenna Channel Modeling 15
2.4 Capacity of MIMO Broadcast Channels 18
2.4.1 Capacity with perfect CSI at the transmitter 18
2.4.2 Capacity with no CSI at the transmitter 20
2.5 Multiuser MIMO Schemes with perfect CSIT 21
2.5.1 Non-linear Precoding 21
2.5.2 Linear Precoding 22
2.6 The cardinal role of Channel State Information 25
2.6.1 Channel Knowledge at the Transmitter 25
2.6.2 Capacity scaling laws in MIMO BC systems 26
2.6.3 Partial Channel State Information 28
2.6.4 Statistical Channel Knowledge at the Transmitter 28
2.7 Scheduling and Multiuser Diversity 29
2.7.1 Asymptotic Sum-rate Analysis with Opportunistic Scheduling 30
2.8 Living with partial CSIT: Limited feedback approaches 32
2.8.1 Quantization-based techniques 32

2.8.2 Dimension reduction and projection techniques . . . 32
2.9 Linear Precoding and Scheduling with Limited Feedback . . . 33
2.9.1 Finite Rate Feedback Model for CDI . . . 33
2.9.2 Codebook design . . . 34
2.9.3 Random Opportunistic Beamforming . . . 36

3 Enhanced Multiuser Random Beamforming 39
3.1 Introduction . . . 39
3.2 Sum-Rate Analysis of Random Beamforming . . . 41
3.3 Capacity scaling laws for high SNR . . . 44
3.4 Two-Stage Scheduling and Linear Precoding . . . 47
3.5 Enhanced Multiuser Random Beamforming . . . 48
3.6 Enhanced Precoding with perfect second-stage CSIT . . . 49
3.7 Beam Power Control with Beam Gain Information . . . 49
3.7.1 Optimum Beam Power Allocation for Two Beams . . . 50
3.7.2 Beam Power Allocation for more than two beams . . . 52
3.7.3 Beam Power Control in Specific Regimes ($\mathcal{B} \geq 2$) . . . 55
3.8 Beam Power Control with SINR feedback . . . 57
3.9 Performance Evaluation . . . 58
3.10 Conclusion . . . 62
3.A Proof of Lemma 3.1 . . . 64
3.B Proof of Lemma 3.2 . . . 64
3.C Proof of Lemma 3.3 . . . 65
3.D Proof of Corollary 3.2 . . . 65
3.E Proof of Theorem 3.1 . . . 65
3.F Proof of Theorem 3.2 . . . 66
3.G Proof of Lemma 3.4 . . . 67
3.H Proof of Lemma 3.5 . . . 67
3.I Proof of Proposition 3.3 . . . 68

4 Exploiting Channel Structure in MIMO Broadcast Channels 69
4.1 Introduction . . . 69
4.2 Exploiting redundancy in time-correlated channels . . . 70
4.2.1 User Selection in time-correlated channels . . . 70
4.2.2 Beamforming and Scheduling exploiting temporal correlation . . . 70
4.2.3 Memory-based Opportunistic Beamforming . . . 71
4.3 Performance evaluation . . . 74
4.4 Exploiting Statistical CSIT in Spatially Correlated Channels . . . 75
4.4.1 System Setting . . . 76
4.4.2 User Selection with ML Channel Estimation . . . 77
4.4.3 ML *coarse* Channel Estimation with CQI Feedback . . . 78
4.4.4 Interference-bounded Multiuser Eigenbeamforming with limited feedback . . . 83
4.4.5 Performance Evaluation . . . 85
4.5 Conclusions . . . 90

4.A Proof of Proposition 4.1 . . . 91

5 Limited Feedback Broadcast Channels based on Codebooks 93
5.1 Introduction . . . 93
5.2 System model . . . 95
5.3 CQI Feedback Design . . . 95
5.3.1 Problem formulation . . . 95
5.3.2 Bounds on average received SINR . . . 96
5.3.3 Lower bound on instantaneous received SINR . . . 98
5.3.4 SDMA/TDMA transition with limited feedback . . . 102
5.4 User Selection Schemes . . . 103
5.4.1 Greedy-SUS algorithm . . . 103
5.4.2 Greedy-US algorithm . . . 104
5.5 Performance Analysis . . . 105
5.5.1 Asymptotic (in K) sum-rate analysis . . . 105
5.5.2 Sum-rate analysis in the interference-limited region . . . 106
5.6 MIMO Broadcast Channels with Finite Sum Rate Feedback Constraint . . . 107
5.6.1 Multiuser Diversity - Multiplexing Tradeoff in MIMO BC with Limited Feedback . . . 107
5.6.2 Finite Sum Rate Feedback Model . . . 108
5.6.3 Problem Formulation . . . 109
5.6.4 Decoupled Feedback Optimization . . . 110
5.7 Performance Evaluation . . . 111
5.8 Conclusion . . . 117
5.A Proof of Theorem 5.1 . . . 119
5.B Proof of Lemma 5.1 . . . 120
5.C Proof of Theorem 5.2 . . . 120
5.D Proof of Lemma 5.2 . . . 121
5.E Proof of Theorem 5.3 . . . 122
5.F Proof of Theorem 5.4 . . . 123

6 Feedback Reduction using Ranking-based Feedback 125
6.1 Introduction . . . 125
6.2 Ranking-based Feedback Framework . . . 127
6.2.1 Two-stage approach . . . 127
6.2.2 Ranking-based CQI Representation . . . 128
6.3 Performance analysis . . . 129
6.3.1 Asymptotic optimality of ranking-based feedback for large window size W . . . 129
6.3.2 Throughput for infite observation window size W . . . 130
6.3.3 Throughput for finite observation window size W . . . 131
6.3.4 Performance reduction bound for finite window size W . . . 132
6.3.5 Window size versus feedback reduction tradeoff . . . 133
6.4 Ranking-based CDI Model . . . 133
6.5 Scheduling with Heterogeneous Users . . . 134

6.6 Performance Evaluation . . . 135
6.7 Conclusion . . . 139
6.A Proof of Proposition 6.1 . . . 140
6.B Proof of Proposition 6.3 . . . 140
6.C Proof of Proposition 6.5 . . . 141

7 System Aspects in Multiuser MIMO Systems 143
7.1 Introduction . . . 143
7.2 Channel State Information Acquisition . . . 144
7.2.1 CSI at the Receiver . . . 144
7.2.2 CSI at the Transmitter . . . 144
7.3 Codebook-based Precoding . . . 145
7.4 CQI feedback metrics and Link Adaptation . . . 147
7.5 Opportunistic Scheduling: System Issues . . . 147
7.6 Fairness . . . 148
7.6.1 Definition of Fairness in Scheduling . . . 148
7.6.2 Proportional Fair Scheduler (PFS) . . . 149
7.6.3 Multiuser Proportional Fair Scheduler (M-PFS) . . . 150

8 Conclusions and Perspectives 153

List of Figures

2.1 Multiple-Input Multiple Output Channel Model. 13
2.2 Downlink of a multiuser MIMO network: A BS/AP communicates simultaneously with several multiple antenna terminals. 15
2.3 Analytical channel model with local scatterers at mobile station 17
2.4 Schematic of Random Opportunistic Beamforming. 38

3.1 Comparison between simulated and analytical achievable sum-rate of RBF with $M = 4$ antennas and SNR = 20 dB. 42
3.2 Achievable sum rate comparison vs. average SNR for RBF with $M = 4$ antennas. Both analytic expressions approximate accurately the simulated performance at high SNR. 43
3.3 Achievable sum rate comparison between simulated and analytical results for RBF with $M = 4$ antennas and SNR = -15 dB. 43
3.4 Sum rate versus the number of users for Optimal Beam Power Control with $M = 2$ transmit antennas and SNR = 20 dB. 59
3.5 Sum rate versus average SNR for Optimal Beam Power Control (strategy 3) with $M = 2$ transmit antennas and $K = 10$ users. 59
3.6 Sum rate comparison of different second-stage precoders (strategy 1) versus the number of users for $M = 2$ and SNR = 10 dB. 60
3.7 Sum rate versus the number of users for Iterative Beam Power Allocation and Optimal Power Control with $M = 2$ transmit antennas and SNR = 10 dB. . . 60
3.8 Sum rate versus the number of users for Iterative Beam Power Allocation with $M = 4$ transmit antennas and SNR = 10 dB. 61
3.9 Sum rate versus the number of users for On/Off Beam Power Control with $M = 2$ transmit antennas and SNR = 20 dB. 61
3.10 Sum rate versus average SNR for On/Off Beam Power Control with $M = 4$ transmit antennas and $K = 25$ users. 62
3.11 Sum rate versus the number of users for On/Off Beam Power Control with $M = 4$ transmit antennas and SNR = 20 dB. 62

4.1 Sum rate vs. the number of transmit antennas M of MOBF with $K = 20$ users and various Doppler spreads. 74
4.2 Sum rate as a function of number of users K of MOBF for different Doppler spreads. 75

4.3 Sum rate performance versus angle spread of proposed ML estimation method for $M = 2$, and $K = 50$ users. Full CSIT is obtained for the selected users at a second step. . . . 86
4.4 Sum rate performance versus the number of users of ML channel estimation method for $M = 2$, and $\sigma_\theta = 0.2\pi$. Full CSIT for the selected users is obtained for precoder design. . . . 86
4.5 Sum rate performance versus angle spread of proposed ML estimation framework for $M = 2$, and $K = 50$ users. Partial CSIT is employed for precoding design. . . . 87
4.6 Sum rate as a function of the number of users for various user selection schemes with $M = 2$, antenna spacing $d = 0.5\lambda$ and $\sigma_\theta = 0.1\pi$. . . . 87
4.7 Sum rate as a function of antenna spacing for various user selection schemes with $M = 2$, $\sigma_\theta = 0.1\pi$ and $K = 50$ users. . . . 88
4.8 Sum rate as a function of angle spread for various user selection schemes with $M = 2$, antenna spacing $d = 0.5\lambda$ and $K = 50$ users. . . . 88
4.9 Sum rate as a function of the number of users for $M = 2$, and $\sigma_\theta = 0.1\pi$. . . . 89
4.10 Sum rate as a function of angle spread for $M = 2$, antenna spacing $d = 0.4\lambda$ and $K = 100$ users. . . . 89

5.1 Finite Sum Rate Feedback Model. . . . 108
5.2 Sum rate versus the average SNR for $B_D = 4$ bits, $M = 2$ transmit antennas and $K = 30$ users. . . . 112
5.3 Sum rate as a function of the number of users for $B_D = 4$ bits, $M = 2$ transmit antennas and SNR = 20 dB. . . . 112
5.4 Sum rate performance as a function of the average SNR for increasing value of the number of users, with $B_D = 4$ bits of feedback per user and $M = 2$ transmit antennas. . . . 113
5.5 Sum rate as a function of the average SNR for increasing codebook size, $M = 2$ transmit antennas, and $K = 50$ users. . . . 114
5.6 Sum rate performance as a function of the number of users for increasing codebook size, $M = 2$ transmit antennas, and SNR = 10 dB. . . . 114
5.7 Sum rate versus the number of users for with SNR = 20 dB, $M = 2$ transmit antennas and 10-bit total feedback bits. $B_D = 5$ bits are used for codebook indexing and ($B_Q = 10 - B_D$ bits) for CQI quantization. For metric IV, 2 bits are used for quantization of the channel norm and 3 bits for the alignment. 115
5.8 Sum rate vs. number of users for $M = 2$ and SNR = 10 dB. . . . 116
5.9 Sum rate vs. number of users for $M = 2$ and SNR = 20 dB. . . . 116
5.10 Sum rate vs. number of users in a system with optimal B_D/B_Q balancing for different SNR values. . . . 117

6.1 Throughput comparison as a function of window size W for single-beam RBF with $M = 2$ antennas, SNR = 10 dB and $K = 10$ active users. . . . 136
6.2 Average rate as a function of the number of users for single-beam RBF with M=2 antennas, SNR = 10 dB and different values of window size W. . . . 137

6.3 Average rate as a function of the number of users for single-beam RBF with $M = 2$ antennas, SNR = 10 dB, W=1000 slots, and ranking-based CQI metric quantized with different resolutions. 137
6.4 Sum rate as a function of the number of users for multi-beam RBF with $M = 2$ antennas, SNR = 10 dB and $W = 1000$ slots. 138
6.5 Sum rate as a function of users for multi-beam RBF in a heterogeneous network in which users' average SNRs range from -10 dB to 30 dB, $M = 4$ antennas and $W = 1000$ slots. 138
6.6 Normalized scheduling probability vs. user index for multi-beam RBF with $M = 4$ antennas and $K = 10$ users. The users are sorted from the lowest to the highest average SNR and the SNR range is from -10 dB to 30 dB. 139

List of Tables

3.1 Iterative Beam Power Control Algorithm for Sum-Rate Maximization 53

4.1 Memory-based Opportunistic Beamforming Algorithm 72
4.2 Greedy User Selection with Statistical CSIT 79
4.3 Resource Allocation Algorithm with Statistical CSIT 84

5.1 Greedy Semi-orthogonal User Selection with Limited Feedback 118
5.2 Greedy User Selection Algorithm with Limited Feedback 118

Nomenclature

In this section, the notational convention of the thesis is summarized. First, we provide a list of abbreviations, followed by an overview of the notation of more general nature. We conclude with the notations that are more specific for this thesis.

Abbreviations and Acronyms

The abbreviations and acronyms used throughout the thesis are summarized here. The meaning of an acronym is usually indicated once, when it first occurs in the text.

3GPP	Third Generation Partnership Project
AMC	Adaptive Modulation and Coding
AoA	Angle of Arrival
AoD	Angle of Departure
AP	Access Point
AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
BD	Block Diagonalization
BER	Bit Error Rate
BF	Beamforming
BGI	Beam Gain Information
bps	bits per second
BS	Base Station
CCI	Channel Covariance Information
CDMA	Code Division Multiple Access
CDF	Cumulative Distribution Function
CDI	Channel Direction Information
CMI	Channel Mean Information
CQI	Channel Quality Information
CSI	Channel State Information
CSIR	Channel State Information at Receiver
CSIT	Channel State Information at Transmitter
DMT	Diversity Multiplexing Tradeoff
DPC	Dirty Paper Coding
EVD	Eigenvalue Decomposition
FDD	Frequency Division Duplex

GEV	Generalized Eigenvalue
HSDPA	High-Speed Downlink Packet Access
i.i.d.	independent and identically distributed
i.ni.d.	independent and non-identically distributed
KKT	Karush-Kuhn-Tucker optimality conditions
l.d.	Limit Distribution
LOS	Line-of-Sight
MAC	Multiple Access Channel
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
ML	Maximum Likelihood
MMSE	Minimum Mean-Square Error
NLOS	Non Line-of-Sight
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PDF	Probability Density Function
PFS	Proportional Fair Scheduling
QoS	Quality of Service
RBF	Random (opportunistic) Beamforming
RHS	Right Hand Side
rms	root mean square
RVQ	Random Vector Quantization
SDMA	Space Division Multiple Access
SINR	Signal-to-Interference-plus-Noise Ratio
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
s.t.	Subject to
STC	Space-Time Code
SVD	Singular Value Decomposition
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
THP	Tomlinson-Harashima Precoding
UCA	Uniform Circular Array
ULA	Uniform Linear Array
UMTS	Universal Mobile Telecommunications System
VQ	Vector Quantization
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
ZF	Zero Forcing
WloG	Without loss of Generality

Notations

The notations used in this dissertation are listed in this section. We use boldface upper (e.g. $\mathbf{X}$) and lower case (e.g. $\mathbf{x}$) letters for matrices and column vectors, respectively. Plain letters are used for scalars and uppercase calligraphic letters (e.g. $\mathcal{S}$) denote sets. No notational distinction is used for a random variable and its realization. Other notational conventions are summarized as follows:

$\mathbb{C}$, $\mathbb{R}$	The sets of complex and real numbers, respectively.
$\|x\|$	The absolute value of a scalar.
$\angle x$	The phase of a complex scalar (in radians).
$\|\mathbf{x}\|$	The Euclidean (ℓ^2) norm of vector $\mathbf{x}$
$\|\mathbf{X}\|_{\mathrm{F}}$	The Frobenius norm of matrix $\mathbf{X}$
$\lceil x \rceil$	The ceiling operator, i.e. the smallest integer not less than x.
$\angle(\mathbf{x}, \mathbf{y})$	The angle between two vectors $\mathbf{x}$ and $\mathbf{y}$.
$\|\mathcal{X}\|$	The cardinality of the set $\mathcal{X}$, i.e. the number of elements in the finite set $\mathcal{X}$.
$\mathbb{E}\{\cdot\}$	The expectation operator.
$\mathcal{CN}(\mathbf{x}, \mathbf{X})$	The circularly symmetric complex Gaussian distribution with mean $\mathbf{x}$ and covariance matrix $\mathbf{X}$.
$(\cdot)^*$	The complex conjugate operator.
$(\cdot)^T$	The transpose operator.
$(\cdot)^H$	The complex conjugate (Hermitian) transpose operator.
$\mathbf{X}^\dagger$	The Moore-Penrose pseudoinverse of matrix $\mathbf{X}$.
$\mathbf{X}^{-1}$	The inverse of matrix $\mathbf{X}$.
$\mathbf{I}$	The identity matrix.
$\mathrm{Tr}(\mathbf{X})$	The trace of matrix $\mathbf{X}$, i.e. the sum of the diagonal elements.
$\mathrm{vec}(\mathbf{X})$	The vector obtained by stacking the columns of $\mathbf{X}$.
$\otimes$	The Kronecker matrix product.
$O(\cdot)$	The big-O notation, i.e. $f(x) = O(g(x))$ as $x \to \infty$ iff $\exists x_0, c > 0$ such that $\|f(x)\| \leq c\,\|g(x)\|$ for $x > x_0$.
$\exp(\cdot)$	The exponential function.
$\log(\cdot)$	The natural logarithm.
$\log_2(\cdot)$	The base 2 logarithm.

Thesis Specific Notations

We summarize here the symbols and notations that are commonly used in this thesis. We have tried to keep consistent notations throughout the document, but some symbols have different definitions depending on when they occur in the text.

M	Number of transmit antennas
N_k	Number of receive antennas at user k.
K	Number of active terminals, i.e. the set of users simultaneously asking for service during one given scheduling window.
$\mathbf{h}_k$	The channel from base station to user k (frequency flat).

$\bar{\mathbf{h}}_k$	The channel of user k normalized by its amplitude, i.e. $\bar{\mathbf{h}}_k = \mathbf{h}_k / \|\mathbf{h}_k\|$.
$\mathbf{W}$	The precoding matrix.
$\mathbf{w}_k$	The beamforming vector of user k.
$\mathbf{Q}$	An isotropically distributed unitary matrix.
$\mathbf{q}$	An orthonormal vector (beam), i.e. column of $\mathbf{Q}$.
$\mathbf{n}_k$	The AWGN noise vector of user k.
$\mathcal{R}_k$	The achievable rate of user k.
P	The maximum transmit power.
$\mathcal{S}$	The set of selected (scheduled) users.
$\mathcal{B}$	The number of active beams.
γ_k	The CQI feedback of user k.
ζ_k	The scheduling (decision) metric for user k.

Chapter 1

Introduction

1.1 Background and Motivation

The last decade the wireless industry has been confronted with a galloping demand for higher data rates and enhanced quality of service (QoS). The applications offered to customers nowadays are no longer limited to voice transmission, but new types of services, such as streaming multimedia, internet browsing, file transfer and video telephony, each with different QoS requirements, are provided. The success story of cellular telephony has opened the way to the development of various types of wireless systems, such as local and metropolitan area networks (LAN, MAN), ad-hoc and sensor networks, short-range wireless protocols, etc. The variety of wireless protocols combined with the increasing demand for data services have amended the wireless service vision to an anywhere-anytime basis.

The introduction of new data services is one of the underlying reasons for the transition from circuit-switched systems to packet-switched networks. Networks accommodating delay-tolerant, best-effort traffic have now evolved, offering flexibility to the resource allocation unit to schedule transmissions in slots where the communication link exhibits favorable channel conditions. This gives rise to the so-called *multiuser diversity gain* [1], which aims at a better utilization of the spectrum inside each cell at the expense of user fairness and delay.

In addition to multiuser diversity, another key technology that efficiently utilizes the scarce bandwidth resource is multi-antenna communications. Multiple-Input Multiple-Output (MIMO) techniques have generated a great deal of interest due to their potential for high spectral efficiency, increased diversity, and interference suppression capabilities. As a result, the use of multiple antennas is envisioned in most of next-generation wireless protocols, including 3GPP Long Term Evolution (LTE) [2], High Speed Downlink Packet Access (HSDPA), IEEE 802.16e (WiMAX) [3], and IEEE 802.11n [4].

1.2 From Single-user to Multiuser MIMO Communications

The high throughput and diversity gains promised by point-to-point (single-user) MIMO communications are essentially achieved via the use of diversity gain-oriented techniques (e.g. space-time coding [5]) combined with rate maximization-oriented techniques (e.g. spatial stream multiplexing). In such a traditional single-user view of MIMO systems, the extra spatial degrees of freedom brought by the use of multiple antennas are exploited to expand the dimensions available for signal processing and detection, thus acting mainly as a physical layer performance booster. In this approach, the link layer protocols for multiple access indirectly reap the performance benefits of MIMO antennas in the form of greater per-user rates, or more reliable channel quality, despite not requiring full awareness of the MIMO capability.

Recently, there has been a vivid interest in the role of multiple antennas in multiuser network settings, and especially in broadcast and multiple access scenarios. The multiple access channel (MAC), also referred to as the uplink, applies to settings where many transmitters send signals to one receiver in the same frequency band. The broadcast channel (BC), also referred to as downlink, models a network in which a base station (BS) communicates (sends data) to many users sharing the same medium. Investigation of the more challenging broadcast channel lies at the core of this thesis. In multiuser MIMO networks, the spatial degrees of freedom offered by multiple antennas can be advantageously exploited to enhance the system capacity, by scheduling multiple users simultaneously by means of Space Division Multiple Access (SDMA). Such a multiple access protocol requires more complex scheduling strategies and transceiver methodologies, but does not involve any bandwidth expansion. In spatial multiple access, the resulting multiuser interference is handled by the multiple antennas, which in addition to providing per-link diversity also give the degrees of freedom necessary to separate users in the spatial domain.

Recent information theoretic advances reveal that the capacity-achieving transmit strategy for the MIMO broadcast channel is the so-called *dirty paper coding* (DPC) [6–8]. However, this optimum transmit strategy, which involves a theoretical pre-interference cancellation technique combined with an implicit user scheduling and power loading algorithm, is highly complex to implement and sensitive to channel estimation errors. The capacity-achieving technique in MIMO broadcast channels revealed the fundamental role played by the spatial dimension on multiple access and scheduling, replacing the simplistic view of MIMO as a pure physical layer technology. This gave rise to the development of the so-called *cross-layer approaches*, which aim at the joint design of the physical layer's modulation/coding and link layer's resource allocation and scheduling protocols.

Multiuser MIMO techniques and their performance have begun to be intensely investigated because of several key advantages over single-user MIMO communications. In particular, multiuser MIMO schemes allow for a linear increase in capacity, proportional to the number of transmit antennas, thanks to their spatial multiplexing capabilities. They also appear more robust with respect to most of propagation limitations plaguing single-user MIMO communications, such as channel rank loss or line-of-sight. Furthermore, the spatial multiplexing gains promised by information theory can be achieved without the need for

multi-antenna terminals, thereby allowing the development of small and cheap terminals while intelligence and cost is kept on the infrastructure side.

As everything good in life, nothing comes for free. All these promising results unfortunately come at the critical assumption of good channel state information at transmitter (CSIT). Multiuser MIMO systems, unlike the point-to-point case, benefit substantially from CSIT, the lack of which may significantly reduce the system throughput. This is because without CSIT, the BS does not know in which direction to send the beams. If a BS with M transmit antennas communicating with K single-antenna receivers has perfect channel state information (CSI), a multiplexing gain of $\min(M, K)$ can be achieved. Although the approximation of close to perfect CSI at the receiver (CSIR) is often reasonable, this assumption is often unrealistic at the transmitter side. If the BS has imperfect channel knowledge, the full multiplexing gain may be reduced, and in settings with complete absence of CSI knowledge, the multiplexing gain collapses to one. CSIT acquisition seems to be the most substantial cost to pay in order to properly serve the spatially multiplexed users and boost the system capacity of multiuser MIMO systems. In systems where channel reciprocity cannot be exploited or is prone to errors, the need for CSIT feedback places a significant burden on uplink capacity, exacerbated in wideband communications (e.g. OFDM) or high mobility systems (such as 3GPP-LTE, WiMAX, etc.).

In this dissertation, we focus on the multi-antenna downlink channel and aim at identifying what kind of partial CSIT, also referred to as limited feedback, can be conveyed to the BS in order to achieve capacity close to that of the full CSIT case. Motivated by recent key findings, which show that linear precoding strategies with partial CSIT can achieve a significant fraction of the full CSIT capacity if combined with efficient scheduling protocols [9–12], we focus on low-complexity, linear beamforming techniques. We try to shed some light on the problem of partial CSIT design by proposing several low-rate feedback strategies that allow the BS to cope well with limited channel knowledge and achieve near-optimal sum rate. As we will see in the following chapters, the role of multiuser diversity and opportunistic scheduling is instrumental in our approaches. Our thesis is that thanks to the multiuser diversity gain, it is generally sufficient to feed back one or two properly designed scalar feedback parameters in order to perform beamforming and user selection that achieves throughput relatively close to the optimum one.

1.3 Assumptions

In an effort to provide a clear and concise framework to this work, we make the following standard assumptions:

- *Single cell network.*
  A single cell is considered and the inter-cell interference is treated as noise.

- *Perfect channel state information at the receiver.*
  Users can estimate perfectly their channels, so that full channel state information at the receiver (CSIR) is always assumed. CSIR is often obtained from pilot symbols and blind channel estimation techniques, especially in downlink channels, where pilot-symbol-based channel estimation is more efficient as the terminals share a common

pilot channel. This assumption may be questioned in high-mobility settings and results in significant overhead in wideband systems.

- *Narrowband channels*
  Flat-fading channels are considered, i.e. the signal bandwidth is much less than the reciprocal of the propagation time of the wavefront across the antenna array. Our proposed methods can be easily applied on a per subcarrier basis in wideband OFDM systems.

- *Ideal link adaptation.*
  Ideal link adaptation protocols are assumed and the continuous-rate, continuous-power Shannon capacity formula is calculated as user throughput measure. This is a reasonable assumption since current powerful coding schemes can perform close to Shannon limit. Furthermore, the SNR-gap if practical coding and modulation schemes are used does not affect the sum-rate scaling of the proposed techniques.

- *Infinite backlogged users.*
  An infinite backlog of packets in each queue is assumed, thus the base station has always data to transmit to the selected (scheduled) users. Since the resource allocation policies are studied from a throughput maximization point of view, queue state information and traffic arrival processes have been neglected.

1.4 Contributions and Outline of the Dissertation

Foreword: *This dissertation stems from an ANRT CIFRE (Convention Industrielle de Formation par la Recherche/Industrial Agreement for Training through Research) agreement between Telecom ParisTech / EURECOM, Sophia-Antipolis, and the Radio Access Networks (RESA) group at France Telecom Research and Development, Paris. The conducted research work was fully funded by France Telecom Research and Development (Orange Labs).*

The main focus of the thesis is user selection and linear precoding in multiuser multi-antenna systems with limited feedback. We provide below an outline of the dissertation and describe the contributions made in each chapter.

Chapter 2 - Multi-antenna Broadcast Channels
In this chapter, we review recent fundamental findings in MIMO broadcast channels. The general multi-antenna system model is introduced and capacity results for the broadcast channel are presented under different assumptions on the quality/amount of CSIT. We emphasize on the cardinal importance of CSIT and the role of multiuser diversity for achieving close to optimum capacity. Capacity scaling laws for opportunistic scheduling under different channel statistical distributions are provided. The capacity growth for networks with path loss and fading is a contribution of this chapter. Finally, we present in detail linear precoding strategies combined with scheduling using limited feedback, which forms the building block of the dissertation. The advantages and drawbacks of this setting are identified, motivating our work and the solutions proposed in the subsequent chapters. Part of this chapter has been published in a tutorial paper:

- D. Gesbert, M. Kountouris, R.W. Heath, Jr., C.-B. Chae, and T. Sälzer, "From Single User to Multiuser Communications: Shifting the MIMO Paradigm," in IEEE Signal Processing Magazine, Special Issue on Signal Processing for Multiterminal Commun. Systems, vol.24, no.5, pp. 36-46, Sept. 2007.

Chapter 3 - Enhanced Multiuser Random Beamforming

The contributions of this chapter are two-fold: In the first part, we provide an unpublished exact sum-rate analysis of conventional random beamforming (RBF) [9]. Capacity scaling laws for the interference-limited region (high SNR) are derived using extreme value theory, showing the cardinal importance of multiuser diversity in this regime. In the second part, a limited feedback-based scheduling and beamforming scenario that builds on RBF is considered. We introduce a two-stage framework that decouples the scheduling and beamforming design problems in two phases. Several refinement strategies, including beam power control and beam selection, are proposed, offering various feedback reduction and performance tradeoffs. The common feature of these schemes is to restore robustness of RBF with respect to sparse network settings (low to moderate number of active users), at the cost of moderate complexity increase.

The work in this chapter has been published in:

- M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in Proc. 6th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2005), pp. 975 - 979, New York, USA, June 5 - 8, 2005 (invited paper).

and will appear in:

- M. Kountouris, D. Gesbert, and T. Sälzer, "Enhanced Multiuser Random Beamforming: Dealing with the not so large number of users case," IEEE Journal on Sel. Areas in Communications (JSAC), Special Issue on Limited Feedback Wireless Comm. Networks, Oct. 2008.

Chapter 4 - Exploiting Channel Structure in MIMO Broadcast Channels

In this chapter, we consider multiuser MIMO channels correlated in either time or spatial domain, and provide several techniques that increase the system throughput by exploiting the channel structure. In time correlated channels, an opportunistic beamforming scheme exploiting channel memory is proposed. This scheme is shown to fill the capacity gap with optimum unitary precoding with full CSIT for slow time-varying channels. In spatially correlated channels, a maximum likelihood (ML) coarse channel estimation framework is established, which effectively combines slowly varying statistical CSIT - assumed available at the transmitter - with instantaneous low-rate feedback. A greedy user selection scheme and a low-complexity SDMA eigenbeamforming technique based on multiuser interference bounds are also proposed and evaluated. It is demonstrated that, in wide-area cellular networks, scalar CSIT feedback is sufficient to achieve near-optimal throughput performance if it is properly combined with long-term statistical knowledge.

The work in this chapter has been published in:

- M. Kountouris and D. Gesert, "Memory-based opportunistic multi-user beamforming," in Proc. of IEEE International Symposium on Information Theory (ISIT 2005), pp. 1426 - 1430, Adelaide, Australia, September 4 - 9, 2005.

- M. Kountouris, D. Gesbert, and L. Pittman, "Transmit Correlation-aided Opportunistic Beamforming and Scheduling," in Proc. of 14th European Signal Processing Conference (EUSIPCO), Florence, Italy, September 4 - 8, 2006 (invited paper).

- D. Gesbert, L. Pittman, and M. Kountouris, "Transmit Correlation-aided Scheduling in Multiuser MIMO Networks," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), Vol.4, pp. 249-252, Toulouse, France, May 14-19, 2006.

- M. Kountouris, R. de Francisco, D. Gesbert, D.T.M. Slock, and T. Sälzer, "Low complexity scheduling and beamforming for multiuser MIMO systems," in Proc. 7th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2006), Cannes, France, July 2 - 5, 2006.

Chapter 5 - Limited Feedback Broadcast Channels based on Codebooks
This chapter deals with limited feedback strategies utilizing vector quantization codebooks. In particular, the problem of efficient, sum-rate maximizing channel quality information (CQI) feedback design is addressed. We proposed several scalar feedback metrics that incorporate information on the channel gain, the channel direction, and the quantization error. These metrics are built upon bounds on the instantaneous inter-user interference, and can be interpreted as reliable estimates of the received SINR. It is shown that scalar CQI feedback combined with channel directional information (CDI) and efficient user selection algorithm can achieve a significant fraction of the capacity of the full CSIT case by exploiting multiuser diversity. An adaptive scheme transiting from SDMA to TDMA transmission mode is proposed and is shown to achieve linear sum-rate growth at any SNR range.

The work in this chapter has been published in:

- M. Kountouris, R. de Francisco, D. Gesbert, D.T.M. Slock, and T. Sälzer, "Efficient metrics for scheduling in MIMO broadcast channels with limited feedback," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Honolulu, USA, April 15 - 20, 2007.

- M. Kountouris, R. de Francisco, D. Gesbert, D.T.M. Slock, and T. Sälzer, "Multiuser diversity - multiplexing tradeoff in MIMO broadcast channels with limited feedback," in Proc. of 40th Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, Oct. 29 - Nov. 1, 2006 (invited paper).

and accepted to:

- M. Kountouris, R. de Francisco, D. Gesbert, D.T.M. Slock, and T. Sälzer, "Exploiting Multiuser Diversity in MIMO Broadcast Channels with Limited Feedback," accepted to IEEE Trans. on Signal Processing, August 2007 (under revision).

Chapter 6 - Feedback Reduction using Ranking-based Feedback
In this chapter, a low-rate representation of CSIT feedback parameters, referred to as ranking-based feedback, is identified as a means to further compress the reported channel feedback. This representation enables the scheduler to identify users that are instantaneously on the highest peak with respect to their own channel distributions, independently of the distribution of the other users. Furthermore, we show that temporal fairness is also

restored in heterogeneous networks with i.ni.d. channel statistics among users. The work in this chapter has been published in:

- M. Kountouris, T. Sälzer, and D. Gesbert, "Scheduling for Multiuser MIMO Downlink Channels with Ranking-based Feedback," EURASIP Journal on Advances in Signal Processing, Special Issue on MIMO Transmission with Limited Feedback, March 2008.

Chapter 7 - System Aspects in Multiuser MIMO Systems
This chapter focuses on several system issues and design challenges that arise in real-world wireless systems. We discuss the main practical and implementation challenges that one may face when deploying techniques as those proposed in Chapters 3-6. Emphasis is put on fairness issues and the proportional fair scheduling (PFS) rule is generalized for multiuser system settings, including OFDM, SDMA, multicell networks, etc. Part of these results has been published in:

- M. Kountouris and D. Gesbert, "Memory-based opportunistic multi-user beamforming," in Proc. of IEEE International Symposium on Information Theory (ISIT 2005), pp. 1426 - 1430, Adelaide, Australia, September 4 - 9, 2005.

Patents
In addition to the above publications, our research work resulted in the following patents:

- PCT WO 2007057568, "Information encoding for a backward channel," (assigned)
- FR 2893474, "Method of information encoding for a backward channel of a SDMA system, user terminal and base station of such a system," (assigned).
- "Feedback communication from a terminal to a transmitter to reduce inter-beam interference," (filed, Jan. 2008).

Chapter 2

Multi-antenna Broadcast Channels

In this chapter, we review multiuser MIMO communications focusing on the more challenging downlink, the so-called broadcast channel (BC). The general multi-antenna system model is introduced and known capacity results for the broadcast channel are presented under different assumptions regarding the amount of CSIT. Information theoretic results shed light on the cardinal importance of CSIT and scheduling, as well as on the role of multiuser diversity for achieving the optimum system capacity. Capacity scaling laws for opportunistic scheduling under different channel models are investigated. Several approaches including non-linear and linear channel-aware precoding are reviewed, discussing design choices and performance tradeoffs. Emphasis is given on low-complexity, linear precoding strategies combined with scheduling using limited feedback, which form the building block of the dissertation. The limited feedback model that we adopt and investigate in subsequent chapters is presented in detail and its limitations are identified.

2.1 The Wireless Channel

The wireless radio channel is a particularly challenging medium for reliable high-rate communications. Apart from being subject to noise, interference and several other impairments, the wireless medium is above all a multipath time-varying channel. A signal transmitted over a radio channel is subject to the physical laws of electromagnetic wave theory, which dictate that multiple paths occur as a result of reflection on large surfaces (e.g. buildings, walls, and ground), diffraction on edges, and scattering on various objects. Therefore, a received signal is a superposition of multiple signals arriving from different directions at different time instances and with different phases and power. These paths may combine constructively or destructively, creating a multi-tap channel impulse response, with each

tap having random phase and time-varying amplitude. We first review the physical phenomena that attenuate the signal power. For a more detailed presentation, the interested reader is referred to [13].

2.1.1 Path loss

Path loss is a range-dependent effect and is due to the distance d between the receiver and the transmitter. In ideal free space, the received signal power is described by the Friis equation and follows an inverse square law power loss. Several deterministic and empirical models have been developed for various cellular environments (microcells, macrocells, picocells, etc.), such as Okumura-Hata, Walfisch-Ikegami, and their COST-231 extensions, plane-earth and clutter factor model [13]. A generic path loss model is given by

$$L = \beta d^{-\epsilon} \tag{2.1}$$

where ϵ is the path loss exponent and β is a scaling factor that accounts for antenna characteristics and average channel attenuation. The path loss exponent varies normally from 2 to 6, depending on the propagation environment. For the case of full specular reflections from ground is 4, while for buildings and indoor environments it can take values from 4 to 6.

2.1.2 Shadowing

Shadowing, also known as macroscopic or long-term fading, results from large obstacles blocking the main signal path between the transmitter and receiver, and is determined by the local mean of a fast fading signal. The random shadowing effects, which are influenced by antenna heights, operating frequency and the features of the propagation environment, may be modeled as log-normal distributed with probability density function (PDF):

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{(\log x - \mu)^2}{2\sigma^2}} \qquad x > 0 \tag{2.2}$$

where μ and σ are the mean and standard deviation of the shadowing's logarithm.

2.1.3 Fading

Fading, often referred to as microscopic or small-scale fading, results from the constructive or destructive superposition of multipaths and describes the rapid signal fluctuations of the amplitudes, phases, or multipath delays. The statistical time varying nature of the received envelope is commonly described by the following three fading distributions:

Rayleigh fading

Rayleigh fading is a reasonable model when there is no dominant propagation path (non line-of-sight, NLOS) between the transmitter and the receiver and is used to describe the amplitude of a signal when there is a large number of independent scattered components. Applying the central limit theorem, the channel impulse response can be considered as a complex-valued Gaussian process irrespective of the distribution of the individual components. In a NLOS configuration, this random process is assumed to have zero mean and

phase evenly distributed between 0 and 2π radians. The envelope of the received signal will therefore be Rayleigh distributed with PDF given by

$$p(x) = \frac{2x}{\Omega} e^{-\frac{x^2}{\Omega}} \qquad x > 0 \tag{2.3}$$

where $\Omega = \mathbb{E}\{x^2\}$ is the average received power.

Ricean fading

If a direct, possibly a line-of-sight (LOS), path exists, the assumption of a zero-mean fading process does no longer hold and the distribution of the signal amplitude is modeled as Ricean. The Ricean distribution is often defined in terms of the Ricean factor K which denotes the ratio of the power in the mean component of the channel (direct path) to the power in the scattered paths. The Ricean PDF is given by

$$p(x) = \frac{2x(K+1)}{\Omega} e^{-K-\frac{(K+1)x^2}{\Omega}} I_0\left(2x\sqrt{\frac{K(K+1)}{\Omega}}\right) \qquad x > 0 \tag{2.4}$$

where $\Omega = \mathbb{E}\{x^2\}$ and $I_0(x)$ is the zero-order modified Bessel function of the first kind defined as

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{-x\cos\theta} d\theta \tag{2.5}$$

Nakagami fading

A general fading distribution that fits well with empirical measured data is the Nakagami distribution given by

$$p(x;m) = \frac{2m^m x^{2m-1}}{\Gamma(m)\Omega} e^{-\frac{mx^2}{\Omega}} \qquad x > 0 \tag{2.6}$$

where Ω is the average received power and $m = \frac{\Omega^2}{\mathbb{E}\{x^2-\Omega^2\}}$. The m factor determines the severity of fading, i.e. for $m = \infty$ there is no fading. For $m = 1$ the distribution in (2.6) reduces to Rayleigh fading, while for $m = (K+1)^2/(2K+1)$ the distribution is approximately Ricean fading with factor K.

2.1.4 Channel Selectivity

Multipath propagation results in the spreading of the signal in different dimensions affecting significantly the received signal. These dimensions are time (Doppler spread), space (angle spread) and frequency (delay spread).

Doppler spread and time selective fading

The motion of the transmitter, the receiver or the scatterers results in time selectivity, i.e. a single tone spreads in frequency over a finite spectral bandwidth. The variations due to Doppler shifts are specific to each path and depend on their angle with respect to the moving direction of the transmitter/receiver. Different Doppler shifts lead to the so-called Doppler spread, which is the maximum frequency spread among all Doppler shifts, and is given by

$$f_m = \frac{v}{\lambda_c} \tag{2.7}$$

where v is the mobile speed and λ_c is the carrier wavelength.

How fast the channel decorrelates with time is specified by the temporal autocorrelation function. The Doppler power spectrum $\rho_d(f_d)$ is defined as the Fourier transform of the temporal autocorrelation function of the channel response to a continuous wave

$$\rho_d(f_d) = \begin{cases} \frac{1}{\pi f_m \sqrt{1-(f_d/f_m)^2}} & \forall f_d \in [-f_m, f_m] \\ 0 & \text{elsewhere} \end{cases} \tag{2.8}$$

The most commonly used model for the autocorrelation function is the Clarke-Jakes' model, which assumes uniformly distributed scatterers on a circle around the antenna

$$\rho_d(\tau) = J_0(2\pi f_m \tau) \tag{2.9}$$

where J_k is the k-th order Bessel function of the first kind and τ is the sampling interval.

A measure of the time selectivity is the channel coherence time T_c, defined as the interval over which the channel remains strongly correlated. The shorter the coherence time, the faster the channel changes over time. The coherence time is a statistical measure and satisfies

$$T_c \sim \frac{1}{f_m} \tag{2.10}$$

As we show in Chapter 4, the scheduler can take advantage of the time selectivity and benefit from the resulting channel redundancy (time diversity), as a means to further compress the channel feedback or successively refine the scheduling decisions.

Delay spread and frequency selective fading

Delay spread is caused when several delayed and scaled versions of the transmitted signal arrive at different time instants at the receiver. The time difference between the maximum multipath delay τ_{max} (typically the arrival time of the LOS component) and the minimum path delay τ_{min} is called delay spread. Delay spread causes frequency selective fading as the channel acts like a tapped-line filter. The range of frequencies over which the channel can be considered 'flat' defines the coherence bandwidth B_c and depends on the form of the power delay spectrum (rms delay spread). A channel is characterized as flat or frequency non-selective if the signal bandwidth B is significantly small compared to the channel coherence time, i.e. $B << B_c = 1/\tau_{max}$. In the subsequent chapters, only flat fading channels are considered.

Angle spread and space-selective fading

Angle spread at the receiver/transmitter refers to the spread in angles of arrival (AoAs) / angles of departure (AoDs) of the multipath component at the receive/transmit antenna array, respectively. The different directions of arrival lead to spatial selectivity that implies that signal amplitude depends on the spatial location of the antenna array. Space selective fading is characterized by the coherence distance d_c, which is the maximum distance between two antenna elements for which the fading remains strongly correlated. An upper bound for the coherence distance is given by

$$d_c \leq \frac{\lambda_c}{2\sin(\Delta\theta_{max}/2)} \tag{2.11}$$

where $\Delta\theta_{max}$ is the maximum angle separation, i.e. the range in which the power azimuth spectrum is non zero.

2.2 Multiple-Input Multiple-Output Channels

Multiple-Input Multiple-Output (MIMO) channels arise in many different scenarios such as multi-antenna wireless systems or wireline systems (e.g. DSL), and can be represented in an elegant, compact, and unified way by a channel matrix. The basic discrete-time, narrowband signal model for a point-to-point MIMO channel with M transmit and N receive antennas is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{2.12}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is the transmitted symbol, $\mathbf{H} \in \mathbb{C}^{N\times M}$ is the channel matrix, $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is the received signal, and $\mathbf{n} \in \mathbb{C}^{N\times 1}$ is the noise vector. We assume zero-mean circularly symmetric complex Gaussian noise with covariance matrix $\mathbf{R}_n$[1]. For convenience, a whitened channel $\tilde{\mathbf{H}} = \mathbf{R}_\mathbf{n}^{-\mathbf{1/2}}\mathbf{H}$ is often used such that the white noise $\mathbf{w} = \mathbf{R}_\mathbf{n}^{-\mathbf{1/2}}\mathbf{n}$ has a unitary covariance matrix, i.e. $\mathbb{E}\{\mathbf{w}\mathbf{w}^H\} = \mathbf{I}$. Due to the noise normalization, the transmit power constraint $P = Tr(\mathbb{E}\{\mathbf{x}\mathbf{x}^H\})$ takes on the interpretation of the average signal-to-noise ratio (SNR) per receive antenna under unity channel gain. Knowledge of the channel gain matrix $\mathbf{H}$ at the transmitter and receiver is referred to as *channel state information at the transmitter* (CSIT) and *channel state information at the receiver* (CSIR), respectively.

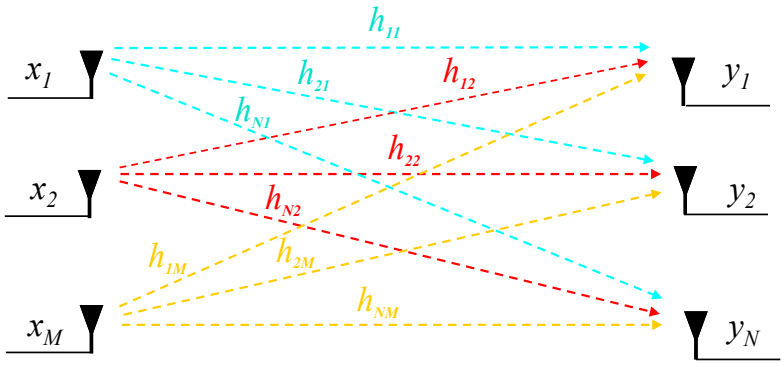

Figure 2.1: Multiple-Input Multiple Output Channel Model.

In the case of a frequency-flat MIMO system, the channel has only one tap and can be represented as a discrete-time channel matrix

$$\mathbf{H}[n] = \begin{pmatrix} h_{11}[n] & h_{12}[n] & \dots & h_{1M}[n] \\ h_{21}[n] & h_{22}[n] & \dots & h_{2M}[n] \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1}[n] & h_{N2}[n] & \dots & h_{NM}[n] \end{pmatrix} \tag{2.13}$$

[1] A complex random vector $\mathbf{x}$ is circularly symmetric if its distribution is the same with the distribution of $e^{j\theta}\mathbf{x}$, $\forall\theta \in [0, 2\pi]$. For $\theta = \pi$ we have $\mathbb{E}\{\mathbf{x}\} = 0$ and for $\theta = \pi/2$, $\mathbf{x}$ is a proper random vector, i.e. $\mathbb{E}\{\mathbf{x}\mathbf{x}^H\} = 0$.

in which $h_{ij}[n]$ is the spatio-temporal signature (channel gain) induced by the j-th transmit antenna across the i-th receive antenna and n is the discrete-time index. Each channel element may have different amplitude and phase due to spatial selectivity.

When the bandwidth-delay spread product of the channel is larger than 0.1, the channel is generally characterized as frequency-selective, and its received signal is given by

$$\mathbf{y}[n] = \sum_{l=0}^{\mathcal{L}} \mathbf{H}[l]\mathbf{x}[n-l] + \mathbf{n}[n] \tag{2.14}$$

where $\mathcal{L}$ is the channel order. To simplify the notation in the subsequent parts of the thesis, we drop the time index n assuming the channel at a given time instant.

When $M = 1$, the MIMO channel reduces to a single-input multiple-output (SIMO) channel, and when $N = 1$, the MIMO channel reduces to a multiple-input single-output (MISO) channel. When both $M = N = 1$, the MIMO channel simplifies to a simple scalar or single-input single-output (SISO) channel.

2.3 Multiuser Multi-Antenna Systems

A multiuser channel is generally any channel that must be shared among multiple users. There are two types of multiuser channels: the uplink and the downlink channel. An uplink channel, also referred to as multiple access channel (MAC) or reverse channel, has many transmitters sending signals to one receiver in the same frequency band. A downlink channel, also referred to as broadcast channel or forward channel, has one transmitter sending signals to many receivers. In this section, we present both multiuser multi-antenna channels (uplink and downlink), however the dissertation focuses solely on the challenges associated with the downlink channel. In a multi-user setting, we consider communication between a BS equipped with M antennas and K active terminals, where each active user k is equipped with N_k antennas. Among all terminals, the set of active users is roughly defined by the set of users simultaneously downloading or uploading packets during one given scheduling window. The length of the scheduling window can be arbitrary but should not exceed the maximum latency expected by the service (likely as small as a few tens of ms to several hundred ms). By all means the active users over one given window will be a small subset of the connected users, themselves forming a small subset of the subscribers.

In the uplink, the received signal at the transmitter can be written as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{H}_k^T \mathbf{x}_k + \mathbf{n} \tag{2.15}$$

where $\mathbf{x}_k \in \mathbb{C}^{N_k \times 1}$ is the k-th user signal vector, possibly encompassing power-controlled, linearly combined, constellation symbols. $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$ represents the channel matrix and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is the complex circularly symmetric additive white Gaussian noise vector (AWGN) at the transmitter. The transpose operator is simply used by convention for consistence with the downlink notation and does not presume a reciprocal link.

In the downlink, illustrated in Fig.2.2, the received signal $\mathbf{y}_k \in \mathbb{C}^{N_k \times 1}$ of the k-th user can be mathematically described as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \mathbf{n}_k \quad \text{for } k = 1, \ldots, K \tag{2.16}$$

where $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$ represents the downlink channel response and $\mathbf{n}_k \in \mathbb{C}^{N_k \times 1}$ is the complex circularly symmetric AWGN at receiver k with $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \sigma_k^2 \mathbf{I})$. The transmitted signal $\mathbf{x}$ is a function of the multiple users' information data, an example of which takes the superposition form

$$\mathbf{x} = \sum_k \mathbf{x}_k \tag{2.17}$$

where $\mathbf{x}_k \in \mathbb{C}^{M \times 1}$ is the transmitted vector signal carrying, possibly non-linearly encoded, message for user k, with covariance $\mathbf{\Sigma}_k = \mathbb{E}\{\mathbf{x}_k \mathbf{x}_k^H\}$. The power allocated to user k is therefore given by $P_k = \mathrm{Tr}(\mathbf{\Sigma}_k)$. Two power constraints are commonly used:

- *individual power constaint*, also referred to as per antenna power constraint, where $P_k^{min} \leq P_k \leq P_k^{max}$, $\forall k$ and $P_k \geq 0$.
- *sum power constraint*, where the power allocation needs to maintain $\sum_k P_k \leq P$.

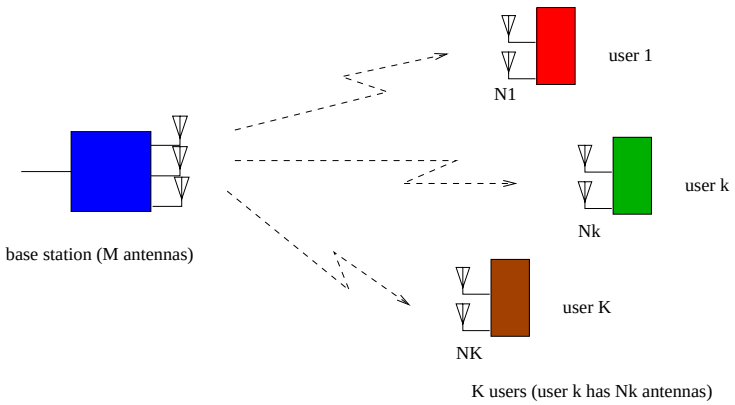

Figure 2.2: Downlink of a multiuser MIMO network: A BS/AP communicates simultaneously with several multiple antenna terminals.

In broadcast channels the available transmit power is divided among the different users, whereas in the uplink each user has an individual power constraint associated with its transmitted signal. In this thesis, unless otherwise stated, we assume a short-term average sum power constraint, which implies that the transmitter has to use the power P at each channel use.

2.3.1 Multi-antenna Channel Modeling

The modeling of MIMO channels is a multi-step procedure of essential importance in system analysis, deployment and network planning since it enables performance prediction and comparison of different system configurations in various propagation environments. The various channel models one can find in the literature can be classified in two categories: propagation-based models and analytical models.

The first category aims at reproducing the physical wave propagation in a deterministic or stochastic way. In deterministic models, the channel matrix is generally generated based on a geometrical description of the propagation environment employing ray-tracing techniques combined with knowledge about the propagation environment. In stochastic models,

the channel behavior is considered as a random variable with a certain statistical distribution depending on the propagation environment. Empirical models, which are based on real channel measurements, also fall into this category.

Analytical channel models focus on modeling only the spatial structure (MIMO channel matrix) of the channel. They are narrowband models since Doppler shifts and delay spreads are neglected. An important category of analytical models is the so-called correlation-based models, presented below.

Correlation-based models

Correlated channels are characterized by the channel correlation matrix which captures the spatial correlation among the elements of MIMO channel matrix $\mathbf{H}$. A full-correlation model is described as

$$\mathbf{H} = \text{unvec}(\mathbf{R}^{1/2}\text{vec}(\mathbf{H}_w)) \tag{2.18}$$

where $\mathbf{H}_w$ is the i.i.d. spatially white (zero-mean circularly symmetric complex Gaussian with unit variance), and $\mathbf{R}$ is the $MN \times MN$ positive semi-definite Hermitian covariance matrix defined as

$$\mathbf{R} = \mathbb{E}\left\{\text{vec}(\mathbf{H})\text{vec}(\mathbf{H})^H\right\} \tag{2.19}$$

The vec$(\cdot)$ operator stacks the columns of a matrix to a vector. An underlying assumption is that vec$(\mathbf{H})$ is Rayleigh distributed.

The full-correlation model is the most accurate - yet very complex - model. For simplicity, the correlation matrix is often assumed to have a less general, separable structure, the so-called Kronecker structure. In this model, the covariance of the scalar channels seen from all the transmit antennas to a receive antenna is assumed to be the same for any receive antenna. The same applies for the receive antenna correlation matrix. The channel model is described as

$$\mathbf{H} = \mathbf{R}_R^{1/2}\mathbf{H}_w\mathbf{R}_T^{1/2} \tag{2.20}$$

where $\mathbf{R}_T = \mathbb{E}\{\mathbf{H}^H\mathbf{H}\}$ and $\mathbf{R}_R = \mathbb{E}\{\mathbf{H}\mathbf{H}^H\}$ is the transmit and receive correlation matrix, respectively. They are related by $\mathbf{R} \approx \frac{1}{\sqrt{Tr(\mathbf{R}_R)}}\mathbf{R}_R \otimes \mathbf{R}_T^T$, where $\otimes$ denotes the Kronecker product. The Kronecker model is satisfied for few antennas or large antenna spacing.

The most simple, yet with no physical relevance, model is the i.i.d. (canonical) model where the channel matrix $\mathbf{H} = \mathbf{H}_w$ is considered i.i.d. spatially white.

LOS component model In the presence of a LOS component, the MIMO channel matrix can be generally modeled as the sum of a fixed or LOS component $\bar{\mathbf{H}}$ and a scattered or NLOS component $\mathbf{H}_w$ given by

$$\mathbf{H} = \sqrt{\frac{K}{K+1}}\bar{\mathbf{H}} + \sqrt{\frac{1}{K+1}}\mathbf{H}_w \tag{2.21}$$

where $\mathbb{E}\{\mathbf{H}\} = \sqrt{K/(K+1)}\bar{\mathbf{H}}$ is the complex channel mean (LOS component) and $K = \frac{\|\bar{\mathbf{H}}\|_F^2}{Tr(\mathbf{R})}$ is the Ricean factor. $K = \infty$ corresponds to non-fading channel and $K = 0$ corresponds to pure fading. The LOS component is assumed to be rank one and generated as

$$\bar{\mathbf{H}} = \mathbf{a}_R(\Omega_R)\mathbf{a}_T^H(\Omega_T) \tag{2.22}$$

where $\mathbf{a}_R(\Omega)$ and $\mathbf{a}_T(\Omega)$ are the receive and transmit array responses, respectively, and Ω_R and Ω_T are the AoAs/AoDs corresponding to the LOS component at the receiver and transmitter sides, respectively.

Propagation-based analytical model

We present here a finite scatterer analytical model that is used for simulating the spatially correlated MIMO channels in Chapter 4. The fundamental assumption of the finite scatterer model is that propagation can be modeled in terms of a finite number P of multipath components. Thus, the channel impulse response is a superposition of P spatially separated paths (rays) given by

$$\mathbf{H} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \phi_p \mathbf{a}_r \left(\theta_p^r\right) \mathbf{a}_t^H \left(\theta_p^t\right) \tag{2.23}$$

where ϕ_p is the gain of the p-th path seen at the receiver, θ_p^t and θ_p^r are the AoDs and AoAs, respectively of the p-th path. The array responses (steering vectors) are given by

$$\mathbf{a}_t \left(\theta_p^t\right) = \left[1, e^{j\Theta_1(\theta_p^t)}, \ldots, e^{j\Theta_{M-1}(\theta_p^t)}\right]^T \tag{2.24}$$

$$\mathbf{a}_r \left(\theta_p^r\right) = \left[1, e^{j\Theta_1(\theta_p^r)}, \ldots, e^{j\Theta_{N-1}(\theta_p^r)}\right]^T \tag{2.25}$$

where Θ_m is the phase shift of the m-th array element with respect to the reference antenna and depends on the array configuration. The two most commonly used uniform array configurations are: the uniform linear array (ULA) and the uniform circular array (UCA). A ULA consists of M elements which are aligned linearly. The spacing between two antenna elements is denoted by d and is identical for all elements. In UCA the elements are uniformly placed on a circle with radius r. ULA facilitates the estimation of the angles of incidence, but it has the drawback that its beamwidth varies with the main direction. Therefore, if a ULA is used for beamforming, it is done so in sectorized systems with a range limited to 120^o. In UCA, the propagation delay between two adjacent elements is not identical. Taking the antenna element 0 as reference point, the transmit steering vector for a ULA is

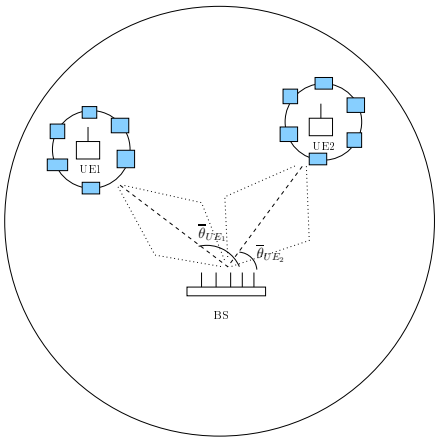

Figure 2.3: Analytical channel model with local scatterers at mobile station

given by

$$\mathbf{a}_t\left(\theta_p^t\right) = \left[1, e^{j2\pi\frac{d}{\lambda}\cos(\theta_p^t)}, \ldots, e^{j2\pi(M-1)\frac{d}{\lambda}\cos(\theta_p^t)}\right]^T \tag{2.26}$$

while for a UCA the transmit array response is given by

$$\mathbf{a}_t\left(\theta_p^t\right) = \left[e^{-j2\pi\frac{r}{\lambda}\cos(\theta_p^t)}, \ldots, e^{j2\pi\frac{r}{\lambda}\cos(\theta_p^t - 2\pi\frac{M-1}{M})}\right]^T \tag{2.27}$$

The AoA and AoD can be well modeled by a truncated Laplacian PDF or a truncated Von Mises distribution. For our simulations in Chapter 4, we assumed that the angles of incidence with respect to the transmitter broadside θ_p follow a Gaussian distribution with 2π-periodic continuation and mean $\overline{\theta}$. The angle spread around its mean is given by the root mean square (rms) deviance $\sigma_\theta = \sqrt{\mathbb{E}\left\{\left|\theta_p - \overline{\theta}\right|^2\right\}}$. The channel gain of each path ϕ_p is assumed to be zero-mean complex Gaussian distributed and all paths have unit variance.

2.4 Capacity of MIMO Broadcast Channels

The complete characterization of the capacity region of multi-antenna broadcast channel was the foremost theoretical challenge in multiuser information theory over the last five years. The analysis of broadcast channels was initiated by Cover [14] and their capacity is generally known only in special cases, where the signals sent to the users can be ordered according to their 'strength'. In contrast to single-user systems where the capacity is a single number, the capacity of a multiuser system with K users is characterized by a *capacity region*, i.e. a K-dimensional rate region, where each point is a vector of rates achievable by all the K users simultaneously. A rate vector is achievable if there exists a coding scheme for which the error probability for all users is arbitrary small as the code block length increases. The maximum of the sum of the communication rates is the so-called *sum-rate point* and lies on the boundary of the capacity region. Clearly, since the K users share the same bandwidth, a tradeoff arises between the reliable communication user rates: if one wants to communicate at a higher rate, the other users may need to lower their rates.

A large class of broadcast channels, known as 'more capable' channels [15], contains two important categories as special cases: 'degraded' and 'less noisy' channels. Roughly speaking, a broadcast channel is degraded when the users can be ordered from the strongest to the weakest in a natural order. For instance, a SISO broadcast channel is degraded, since the users can be ordered according to their $|H_k|^2$, and the capacity region can be achieved by *superposition coding* [14]. However, MIMO broadcast channels are generally non-degraded as there is not a natural way to order channel matrices.

2.4.1 Capacity with perfect CSI at the transmitter

Although the characterization of the general (fading) broadcast capacity region is a long standing problem in multiuser information theory, substantial progress has been made for Gaussian MIMO channels. Despite not being degraded, the Gaussian MIMO BC offers significant structure that can be exploited to characterize its capacity region. The key theoretical tool for characterizing the MIMO BC capacity region with full CSI, the Dirty Paper Coding (DPC), was revealed by the seminal work of Caire and Shamai (Shitz) [7]. Therein, it was shown that the idea of interference pre-subtraction at the transmitter (DPC)

does indeed achieve the capacity of a 2-user MISO broadcast channel. The results of [7] were extended and generalized by [16–18], until the full characterization of MIMO Gaussian BC capacity region (for any compact set of input covariances and not only under a total power constraint) by Weingarten et al. [8], establishing the optimality of DPC as capacity-achieving strategy.

Assuming noise with unit variance and given a set of positive semi-definite matrices $\mathbf{P}_k \geq \mathbf{0}, \forall k$ that satisfy the power constraint $\mathrm{Tr}\left\{\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathbf{P}_{\mathrm{k}} \leq \mathrm{P}\right\}$ and a permutation function π on the user set $\{1, \ldots, K\}$, the following rates are achievable using DPC [8]:

$$\mathcal{C}_k^{DPC}(\pi, \mathbf{P}_{1\ldots K}) = \frac{1}{2} \log \frac{\left|\left(\mathbf{I} + \mathbf{H}_k(\sum_{i=1}^{k} \mathbf{P}_{\pi(i)})\mathbf{H}_k^T\right)\right|}{\left|\left(\mathbf{I} + \mathbf{H}_k(\sum_{i=1}^{k-1} \mathbf{P}_{\pi(i)})\mathbf{H}_k^T\right)\right|} \tag{2.28}$$

The DPC region is given by the convex hull of all the achievable rates as

$$\mathcal{C}_{DPC} = \mathrm{conv}\left\{\bigcup_{\pi} \bigcup_{\mathbf{P}_{1\ldots K}} \mathcal{C}_k^{DPC}(\pi, \mathbf{P}_{1\ldots K})\right\} \tag{2.29}$$

and is shown to be equivalent to the capacity region of MIMO broadcast channel [8].

The capacity expression (2.29) can be simplified as follows:

$$\mathcal{C}_{DPC} = \mathbb{E}_{\mathbf{H}}\left\{\max_{\mathbf{P}_k \geq \mathbf{0}, Tr\left\{\sum_{k=1}^{K} \mathbf{P}_k \leq P\right\}} \log\left|\mathbf{I} + \sum_{k=1}^{K} \mathbf{H}_k^T \mathbf{P}_k \mathbf{H}_k\right|\right\} \tag{2.30}$$

Dirty Paper Coding

The concept of dirty paper coding was introduced by Costa [6], who showed that for a scalar Gaussian channel with AWGN and an interfering Gaussian signal known non-causally at the transmitter (but not at the receiver), the capacity is the same as if there was no additive interference, or equivalently as if the receiver also had knowledge of the interference. In other words, dirty paper coding allows non-causally known interference to be 'pre-subtracted' at the transmitter with no increase in the transmit power. Assume, without loss of generality, that the encoding process is performed in ascending order. The encoder first picks a codeword for i-th receiver, and then chooses a codeword for receiver $(i+1)$-th receiver with full (non-causal) knowledge of the codeword intended for receiver i. Thus, the encoder considers the interference signal caused by users $j < i$ as known non-causally and subsequently, the i-th decoder treats the interference signal caused by users $j > i$ as additional noise.

Uplink-Downlink duality

The main tool that facilitated the extension of the work in [7] and simplified the problem of finding the capacity region of MIMO BC was the *uplink-downlink duality*, introduced in [17–19]. The concept of uplink-downlink duality can be seen, in general, as the equivalence between the performance of a class of receive and transmit strategies when the role of transmitters and receivers are reversed. This equivalence has been observed in seemingly different contexts in the literature. For instance, in point-to-point links, the duality is nothing else but the channel reciprocity. In multiuser information theory, the duality implies

that the capacity region of the MIMO BC, $\mathcal{C}_{DPC}$ with power constraint P is equal to the capacity region of the so-called dual MIMO MAC, $\mathcal{C}_{MAC}$ with sum power constraint P.

$$\mathcal{C}_{DPC}(P, \mathbf{H}_{1\ldots K}) = \bigcup_{Tr\left\{\sum_{k=1}^{K} \mathbf{P}_k \leq P\right\}} \mathcal{C}_{MAC}(\mathbf{P}_{1\ldots K}, \mathbf{H}_{1\ldots K}^{T}) \tag{2.31}$$

where the union is taken over all matrices $\mathbf{P}_k \geq 0\ \forall k$ such that $\mathrm{Tr}\left\{\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathbf{P}_{\mathrm{k}} \leq \mathrm{P}\right\}$.

The major benefit of the uplink-downlink duality is that the capacity region of the downlink can be calculated through the union of regions of the dual uplink, which is convex and whose boundary can be calculated using interior-point methods [20]. An additional benefit is from an optimization theory point of view, since by exploiting the duality the dimensionality of the optimization problem is significantly reduced. In many practical cases, the number of transmit antennas in the broadcast channel is greater than the number of receive antennas of any of the receivers. Therefore, instead of optimizing over K matrices of size $M \times M$, we need to optimize over K matrices of sizes $N \times N$. Note that the uplink-downlink duality only holds under a total power constraint, and extensions of the DPC optimality to general constraint settings (e.g. per-antenna power constraint) are based on the more general concept of min-max duality [8, 21].

On the optimal number of users with non-zero allocated power

Multiuser information theory advocates for transmitting to multiple users simultaneously by properly distributing the spatial dimensions among the best group of users as a means to boost the system throughput. A natural question that arises is *how many users can be simultaneously active, and how the spatial dimensions are distributed among them.* Yu and Rhee [22] obtained a theoretical upper bound on the number of simultaneously active users by counting the number of variables and unknowns in the set of Karush-Kuhn-Tucker (KKT) optimality conditions for the sum-rate maximization problem. This bound indicates that in the downlink channel maximizing the sum rate entails scheduling at most M^2 users simultaneously. In practice, simulations show that typically the number of active users is four times the number of transmit antennas in the high SNR regime using optimum covariance matrices, and that scheduling up to M users, although suboptimal, results to a small capacity loss. In [23], it was independently shown that under certain conditions in a vector downlink with K users and a BS with two transmit antennas, the number of users that can be simultaneously served can be higher than two. The power allocated to the k-th user is no longer a water-filling procedure, but it is found by the KKT conditions. Note that when restricting to linear precoding techniques, as we do in this thesis, the number of served users is directly limited by the number of degrees of freedom at the BS, i.e. M.

2.4.2 Capacity with no CSI at the transmitter

The Gaussian MIMO BC with no CSIT is still degraded no matter whether the receivers have CSIR or not, assuming that the transmitter or the receivers are equipped with multiple antennas [7]. In that case, the capacity region is achieved by superposition coding [24]. When the users have the same number of antennas, it can be shown that superposition coding is the same as time sharing. In this case, the sum capacity is the same as if there is only one user in the system and no gains can be expected from serving multiple users

simultaneously. The capacity region of fading MIMO BC is an open problem of theoretical interest. The capacity region is not explicitly characterized, and only asymptotically tight bounds currently exist. The fading MISO BC is considered in [25] assuming the distribution of the fading coefficients is isotropic. It was shown that the capacity region is equivalent to that of the fading scalar BC, resulting in a multiplexing gain of one. When the transmitter has incomplete CSI on the fading realization, the pre-log factor (multiplexing gain[2]) at high SNR of a two-user real-valued fading MISO BC is upper bounded by 2/3 [26].

2.5 Multiuser MIMO Schemes with perfect CSIT

Although DPC is shown to achieve the entire capacity region of MIMO broadcast channel, this technique, apart from being theoretical and conceptual, it is very difficult to be implemented in practice. One of the major difficulties is that DPC does not indicate how the spatial resources should be shared among users. One class of practical dirty paper codes is the nested lattice codes [27]. Excellent performance on DPC has been also reported in [28,29], and in [30], where a new approach which invokes superposition coding is proposed.

The question of what rate region can be achieved without relying on dirty-paper coding has been widely addressed, mainly in terms of linear and non-linear types of precoding. Many recent publications have shown that for a limited number of users, even techniques that do not invoke DPC are useful, and sometimes provide close to optimum capacity region performance [31]. Precoding works similarly to equalization with the difference that it inverts the fading at the transmitter side instead of the receiver side. The main drawback of precoding is the need for accurate channel estimates of the fading gains of each user at the transmitter side. Although CSIT can be achieved through channel estimation or feedback, it is difficult to be obtained in rapidly-varying channels.

2.5.1 Non-linear Precoding

Several sub-optimal and simplified DPC variants are reported in the literature, such as non-linear scalar versions [32, 33] and high dimensional simplified strategies [34], in which a regularized channel inversion is attempted. An attractive non-linear precoding technique useful for the MIMO BC is proposed in [35] where the processing at the receiver requires a simple one-dimensional modulo operation. Other improved techniques resort on lattice reduction [36] and integer coding [37].

Two popular and representative non-linear precoding methods are based on vector perturbation [38] and on a spatial extension of Tomlinson-Harashima precoding (THP) [27]. Vector perturbation uses a modulo operation at the transmitter to perturb the transmitted signal vector in order to avoid the transmit power enhancement incurred by channel inversion schemes [38]. Finding the optimal perturbation involves solving a minimum distance type of problem and thus can be implemented using sphere encoding or full search based algorithms. THP [39, 40], which is dual to Decision Feedback Equalization (DFE), was originally proposed as a non-linear temporal equalization method that applies a scalar integer offset at the transmitter enabling interference cancellation after application of a modulo

[2]The multiplexing gain m is defined as $m = \lim_{P\to\infty} \frac{\mathcal{C}(P)}{\log_2 P}$

function at the receiver. While in the original THP, a single channel is equalized with respect to time, spatial equalization is required for MIMO channels. THP has generally lower encoder complexity than vector precoding since it computes the components of the translation vector sequentially. For the case of multiuser MISO systems, a THP-based technique, known as trellis precoding, was also proposed by Yu and Cioffi [33].

2.5.2 Linear Precoding

The considerable complexity required by non-linear techniques as well as the fact that linear beamforming combined with efficient user selection exhibits the same asymptotic performance as DPC [7, 11, 12] revitalized the interest for linear precoding schemes. Linear precoding is a generalization of traditional SDMA, where users are assigned different precoding matrices at the transmitter. The precoders are designed jointly based on CSIT from all users and following a number of design and optimization criteria. The transmit precoding optimization problem can be approached under different assumptions, such as power constraints (total or individual), and with different performance criteria (e.g. maximizing SINR, sum rate, error probability, effective bandwidth, assigned SINR targets, minimum power, peak-to-average ratio). The difficulty in designing capacity-optimal downlink precoding, mainly due to the coupling between transmit power, beamforming, and user ordering, has lead to several different approaches ranging from transmit power minimization while maintaining individual SINR constraints to worst-case SINR maximization under a power constraint. Duality and iterative algorithms are often employed in order to provide efficient solutions [41, 42].

Let $\mathbf{s}_k \in \mathbb{C}^{N_k \times 1}$ denote the k-th user (normalized) transmit symbol vector (which is a scalar symbol for beamforming) and $\mathcal{S}$ be the set of selected users (among all K active ones) that will be assigned non-zero rate, with cardinality $|\mathcal{S}| = \mathcal{M} \leq M$. Under linear precoding, the transmitter multiplies the data symbol for each user k by $\mathbf{W}_k$ (or $\mathbf{w}_k$ in the case of beamforming) so that the transmitted signal is a linear function $\mathbf{x} = \sum_{k \in \mathcal{S}} \mathbf{W}_k \mathbf{s}_k$, where $\mathbf{W}_k \in \mathbb{C}^{M \times N_k}$ is the precoding matrix for user k designed to maximize some performance measure. The resulting received signal vector for user k is given by

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{W}_k \mathbf{s}_k + \sum_{j \in \mathcal{S}, j \neq k} \mathbf{H}_k \mathbf{W}_j \mathbf{s}_j + \mathbf{n}_k \tag{2.32}$$

where the second term in (2.32) represents the multiuser or inter-user interference. We assume that each user will decode $S_k \leq N_k$ streams that constitute its data. The goal of linear precoding is to design $\{\mathbf{W}_k\}_{k=1}^{\mathcal{M}}$ based on the channel matrix knowledge, so that a given performance metric is optimized for each stream. If user codes drawn from an i.i.d. Gaussian distribution are used, the achievable rate of user k is

$$\mathcal{R}_k = \mathcal{I}(\mathbf{s}_k; \mathbf{y}_k) = \log_2 \frac{\left| \mathbf{I} + \mathbf{H}_k \left(\sum_{j=1}^{\mathcal{M}} \mathbf{\Sigma}_j \right) \mathbf{H}_k^H \right|}{\left| \mathbf{I} + \mathbf{H}_k \left(\sum_{j \neq k} \mathbf{\Sigma}_j \right) \mathbf{H}_k^H \right|} \tag{2.33}$$

where $\mathbf{\Sigma}_k = \mathbf{W}_k \mathbb{E}\{\mathbf{s}_k \mathbf{s}_k^H\} \mathbf{W}_k^H$ denotes the transmit covariance matrix of user k.

Channel Inversion - ZF Precoding

Precoding design problems incorporating measures such as maximization of SINR or sum rate usually lead to intractable optimization problems. A standard suboptimal approach providing a promising tradeoff between complexity and performance is channel inversion, also known as zero-forcing beamforming (ZFBF). For ease of explanation, we assume $N_k = 1, \forall k$. In ZFBF, the precoder $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_K]$ is designed to achieve zero interference between the users, i.e. $[\mathbf{HW}]_{k,j} = 0$ for $j \neq k$ For a group of selected users $\mathcal{S}$, we denote $\mathbf{H}(\mathcal{S})$ and $\mathbf{W}(\mathcal{S})$ the corresponding submatrices of $\mathbf{H}$ and $\mathbf{W}$ respectively. If $N_k = N \leq M$ and $rank(\mathbf{H}) = N$, the ZFBF matrix is given by the Moore-Penrose pseudoinverse of $\mathbf{H}(\mathcal{S})$

$$\mathbf{W}(\mathcal{S}) = \mathbf{H}(\mathcal{S})^{\dagger} = \mathbf{H}(\mathcal{S})^H (\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H)^{-1} \tag{2.34}$$

The achievable sum rate is given is by

$$\mathcal{R}_{ZF}(\mathcal{S}) = \max_{\sum_{k\in\mathcal{S}} \eta_k P_k \leq P} \sum_{k\in\mathcal{S}} \log_2 (1 + P_k) \tag{2.35}$$

where

$$\eta_k = \frac{1}{\|\mathbf{w}_k\|^2} = \frac{1}{[(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H)^{-1}]_{k,k}} \tag{2.36}$$

can be interpreted as the effective channel gain of the k-th user. The transmit powers can be allocated according to different criteria and depending on the system performance target. If the objective is to maximize the achievable system throughput, the optimum power allocation P_k is given by water-filling

$$P_k = \eta_k \left[\mu - \frac{1}{\eta_k}\right]^+ \quad \forall k \in \mathcal{S} \tag{2.37}$$

where $[x]^+ = \max(0, x)$ and μ is obtained by solving the water-filling equation $\sum_{k\in\mathcal{S}} [\mu - 1/\eta_k]^+ = P$. The sum-rate of ZFBF with optimal power allocation is given by [7]

$$\mathcal{R}_{ZF}(\mathcal{S}) = \sum_{k\in\mathcal{S}} [\log_2(\mu\eta_k)]^+ \tag{2.38}$$

The maximum achievable sum rate of ZFBF is found by exhaustive search, i.e. checking every possible choice of user groups $\mathcal{S}$, however greedy user selection algorithms are shown to achieve near optimal performance [11, 12, 43].

When the channel is ill-conditioned, at least one of the singular values of $(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H)^{-1}$ is very large, resulting in a very low SNR at the receivers. Note also that channel inversion, in contrast to ZF (least-squares) equalization that causes noise enhancement when the channel is nearly rank-deficient, incurs an excess transmission power penalty (signal attenuation at the transmit side). Therefore, the capacity of channel inversion with no user selection does not increase linearly with M, unlike the optimum capacity. User selection offers an important degree of freedom that can be exploited in order to improve the performance of ZFBF by selecting group of users with mutually orthogonal spatial signatures, leading to $rank(\mathbf{H}(\mathcal{S})) = \mathcal{M} \leq M$ (no power penalty). For asymptotically large K, ZFBF with user selection is shown to achieve both the spatial multiplexing and the multiuser diversity gain,

i.e. $\mathcal{R}_{ZF} \sim M \log_2\left(1 + \frac{P}{M} \log K\right)$ [12,44]. Finally, when ZFBF with equal power allocation is performed in a system with K users with $N > 1$ receive antennas each (with $M \geq KN$), it converts the system into KN parallel MISO channels and can be viewed as equivalent (in terms of ergodic sum rate) to performing ZFBF in a channel with KN single-antenna receivers.

Regularized Channel Inversion - MMSE Precoding

For rank-deficient channels, the performance of ZFBF can be improved by a regularization of the pseudo-inverse, which can be expressed as:

$$\mathbf{W}(\mathcal{S}) = \mathbf{H}(\mathcal{S})^H (\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H + \beta \mathbf{I})^{-1} \tag{2.39}$$

where β is the regularization factor. The above scheme is often referred to as Minimum Mean Square-Error (MMSE) precoding due to the analogous with MMSE beamforming weight design criterion if the noise is spatially white. However, at the receiver side the mean-squared error (MSE) between the received vector and the symbol vector is not minimized.

Similarly to MMSE equalization, a non-zero β value results in a measured amount of inter-user interference. The amount of interference is determined by $\beta > 0$ and an optimal tradeoff between the condition of the channel matrix inverse and the amount of crosstalk ought to be found. In practice, the regularization factor is commonly chosen as $\beta = M\sigma^2/P$ motivated by the results in [34] showing that it approximately maximizes the SlNR at each receiver, and leads to linear capacity growth with M. The performance of MMSE is certainly significantly better at low SNR and converges to that of ZF precoding at high SNR. However, MMSE does not provide parallel and orthogonal channels, thus power allocation techniques cannot be performed in a straightforward manner.

Block Diagonalization

If the terminals have each multiple antennas, the additional degree of freedom at the receiver side can be exploited in various ways. For instance, multiple data streams can be transmitted to a user or some level of inter-user interference may remain after precoding, which is canceled using the multiple receive antennas. However, several design challenges arises, such as signal gain and interference cancellation coordination between the transmitter and the receiver, and appropriately allocating resources among all users and all spatial channels of each user.

Block diagonalization (BD) [45] is a generalization of channel inversion techniques when there are multiple antennas at each receiver. When BD is employed, the precoding matrices $\mathbf{W}_j, \forall j$ are chosen such that $\mathbf{H}_k \mathbf{W}_j = \mathbf{0}$, $\forall k \neq j$, thus eliminating the multiuser interference so that $\mathbf{y}_k = \mathbf{H}_k \mathbf{W}_k \mathbf{s}_k + \mathbf{n}_k$. This requires to determine an orthonormal basis for the left null space of the matrix formed by stacking all $\mathbf{H}_j, \forall j \neq k$ matrices together.

Assume that $N_k \geq 1$ with $\sum_{k=1}^{L} N_k = N^{'}$ and up to S_k data streams are transmitted to user k. Define $\tilde{\mathbf{H}}_k$ as a $(N^{'} - S_k) \times M$ matrix

$$\tilde{\mathbf{H}}_k = \left[\begin{array}{ccccccc} \mathbf{H}_1^T & \cdots & \mathbf{H}_{k-1}^T & \mathbf{H}_{k+1}^T & \cdots & \mathbf{H}_L^T \end{array} \right]^T \tag{2.40}$$

then any suitable $\mathbf{W}_k$ lies in the null space of $\tilde{\mathbf{H}}_k$. Let the singular value decomposition (SVD) of $\tilde{\mathbf{H}}_k$ be

$$\tilde{\mathbf{H}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{D}}_k \left[\begin{array}{cc} \tilde{\mathbf{V}}_k^{(1)} & \tilde{\mathbf{V}}_k^{(0)} \end{array} \right]^H \tag{2.41}$$

where $\tilde{\mathbf{U}}_k$ and $\tilde{\mathbf{D}}_k$ are the left singular vector matrix and the matrix of singular values of $\tilde{\mathbf{H}}_k$, respectively, and $\tilde{\mathbf{V}}_k^{(1)}$ and $\tilde{\mathbf{V}}_k^{(0)}$ denote the right singular matrices each corresponding to non-zero singular values and zero singular values ($(M - rank(\tilde{\mathbf{H}}_k))$ singular vectors in the nullspace of $\tilde{\mathbf{H}}_k$)), respectively. Any precoder $\mathbf{W}_k$ that is a linear combination of the columns of $\tilde{\mathbf{V}}_k^{(0)}$ will satisfy the null constraint, since it produces zero interference to the other users. Assuming that $\tilde{\mathbf{H}}_k$ is full rank, the transmitter requires that the number of transmit antennas is at least the sum of all users' receive antennas to satisfy the dimensionality constraint required to cancel interference for each user [45]. The sum rate of block diagonalization can be further enhanced by performing water-filling on each $\tilde{\mathbf{D}}_k$.

2.6 The cardinal role of Channel State Information

Knowledge of the channel state by the transmitter of a communications system has been demonstrated to be beneficial to wireless communications, particularly in multiuser MIMO systems. The major importance of the availability of channel knowledge has been already recognized in [7], by pointing out that lack of perfect CSIT results in total loss of degrees of freedom, in contrast to what happens in single-user or multiple access MIMO schemes. The MIMO BC with no CSIT is degraded no matter whether the receivers have CSIR or not. Hence, when the users have the same number of antennas, it can be shown that superposition coding is the same as time-sharing. Therefore, the sum capacity is the same as if there is only one user in the system and no multiuser diversity gains can be expected. The significant difference on the sum rate behavior between multiuser and single-user MIMO reveals the cardinal role of CSIT in multiuser MIMO downlink systems, presented in detail in the following sections.

2.6.1 Channel Knowledge at the Transmitter

In multiuser MIMO literature it is often assumed that the receiver enjoys close-to-perfect channel knowledge, whereas the transmitter has different levels of CSIT, ranging from no CSIT at all to full CSIT. The assumption that the receiver has accurate channel information is often reasonable especially in the downlink, where pilot symbol-based channel estimation is more efficient since the terminals can share a common pilot channel. Channel acquisition at the transmitter relies on channel measurements at a receiver, since the transmitter is informed by the receiver on the channel state in an implicit or explicit way. The methods available to gather CSI at the transmitter mainly rely on channel reciprocity or feedback. In systems for which channel reciprocity cannot be exploited, the need for CSIT feedback places a significant burden on uplink capacity. The feedback load is further exacerbated in high-mobility systems (such as 3GPP-LTE, WiMAX, etc.) where the channel conditions change rapidly and in wideband systems, where more feedback training is required due to frequency selectivity.

2.6.2 Capacity scaling laws in MIMO BC systems

The dominant role of CSIT in multiuser multi-antenna systems can be identified by studying the asymptotic capacity growth under different assumptions on CSIT. Specifically, the fundamental role played by the multiple antennas in expanding the channel capacity is best apprehended by examining how the sum rate scales with the transmit power and the number of active users.

Full CSI at the Transmitter

High Power Regime: The scaling law of the sum-rate capacity of MIMO BC for fixed $N_k = N$, M, and K and large P is given by [44, 46]

$$\lim_{P\to\infty} \frac{\mathcal{C}_{DPC}}{\log P} = \min(M, \max(N, K)) \tag{2.42}$$

The above result implies that at high SNR, the capacity exhibits linear growth with the number of transmit antennas. Furthermore, the number of receive antennas per user plays very little role in the capacity of MIMO broadcast channels compared to M (provided that $K > M$).

Large K Regime: The scaling law of the sum-rate capacity of MIMO BC for fixed $N_k = N$, M, and P and large K is given by [44]

$$\lim_{K\to\infty} \frac{\mathcal{C}_{DPC}}{\log\log KN} = M \tag{2.43}$$

The result in (2.43) indicates that, with full CSIT, the system can enjoy a multiplexing gain of M, obtained by the BS selecting and sending data to M carefully selected users out of K (multiuser diversity). Since each user exhibits N independent fading coefficients, the total number of degrees of freedom for multiuser diversity is KN, thus giving the extra gain $\log\log KN$.

In contrast, if the BS selects and transmits only to the user with maximum rate, the capacity of time-sharing, $\mathcal{C}_{TS}$, is given by [44]

$$\lim_{K\to\infty} \frac{\mathcal{C}_{TS}}{\min(M, N)\log\log K} = 1 \tag{2.44}$$

From the above results, it is evident why the capacity scaling laws provide the necessary justification for the great appeal of multiuser MIMO systems. The spatial multiplexing gain of M, which is the pre-log factor of the sum rate, implies a linear (in the number of transmit antennas) increase in capacity for no additional power. The corresponding gain is realized by simultaneously transmitting independent data streams in the same frequency band to spatially separable users.

No CSI at the Transmitter

In the absence of CSIT, user multiplexing is generally not possible, as the BS does not know in which 'direction' to form beams.

High Power Regime: The scaling law of the sum-rate capacity of MIMO BC for fixed $N_k = N$, M, and K, satisfies

$$\lim_{P\to\infty} \frac{\mathcal{C}_{DPC}}{\log P} = \min(M, N) \tag{2.45}$$

which implies that at high SNR the capacity is essentially the same as that of a point-to-point MIMO system. In other words, TDMA is optimal in this regime.

Large K Regime: The scaling law of the sum-rate capacity of MIMO BC for fixed $N_k = N$, M, and P and large K is

$$\lim_{K\to\infty} \frac{\mathcal{C}_{DPC}}{\log\log KN} = 0 \tag{2.46}$$

In contrast to (2.43), there is no multiuser gain since the transmitter has no knowledge of the users channels in order to exploit them.

Note that the above results hold under the assumption of perfect CSIR. The impact of lack of CSI at both ends of the MIMO network and in the asymptotically high SNR regime is studied in [25, 47], where it is shown that both the multiuser downlink and the single user capacity scale double logarithmically with the SNR.

Information theoretic design guidelines

The above capacity growth results highlight several fundamental aspects of multiuser MIMO systems, which come in much contrast with the conventional single-user MIMO setting. The design guidelines that can be extracted are summarized as follows:

- Capacity scaling laws advocate for serving multiple users simultaneously in an SDMA fashion, with a suitably chosen precoding scheme at the transmitter. Although the multiplexing gain is limited by the number of transmit antennas, the number of simultaneously served users is in principle arbitrary. How many and which users should *effectively* be served with non-zero power at any given instant of time is the problem addressed by the resource allocation strategy.

- Unlike in the point-to-point MIMO setting, the spatial multiplexing of different data streams can be done while users are equipped with single-antenna receivers, thus enabling the capacity gains of MIMO while maintaining low cost for user terminals. Having multiple antennas at the terminal can thus be viewed as optional equipment allowing extra diversity gain for certain users or giving the flexibility toward interference canceling and multiplexing of several data streams to such users (reducing though the number of other users served simultaneously).

- The multiplexing gain of M in the downlink comes at the condition of close to perfect CSIT. In the absence of CSIT, user multiplexing is generally not possible, as the BS just does not know in which 'direction' to form spatial beams. Thus, the complete lack of CSI knowledge reduces the multiplexing gain to one. This is a key difference with point-to-point MIMO, in which the asymptotic capacity is not sensitive to CSIT, and even in the absence of CSIT, the full multiplexing gain (of one) can be preserved. An

exception lies in scenarios with terminal devices having enough antennas to remove co-stream interference at the receiver ($N_k \geq M$). In the latter case, the BS may decide to either multiplex several streams to a single user or spread the streams over multiple users, achieving an equivalent multiplexing gain in both cases. This is conditioned however on the individual user channels to be full rank.

2.6.3 Partial Channel State Information

The often unrealistic assumption of close to perfect CSIT, as well as the considerable gap between the achievable sum rate of full CSIT compared to the no CSIT case, have motivated research work on schemes employing partial CSIT. *Partial CSIT* or *limited feedback* refers to any possible form of incomplete information on the channel. This term includes, but is not limited to, scalar CQI feedback (e.g. estimate of received SINR), quantized CSIT (quantization of channel vector), channel direction information, statistical CSIT, etc. Multiuser MIMO schemes relying on partial CSIT lie at the heart of this dissertation.

The practical, though suboptimal, approaches described in Section 2.5.1 are shown to be highly sensitive to channel estimation errors, thus difficult to be implemented with partial CSIT. The low-complexity alternative of downlink beamforming and scheduling, despite being less sensitive to CSIT imperfections, requires full CSI as a means to minimize the multiuser interference [12]. Fortunately, work like [9] demonstrates that the optimal capacity scaling of MIMO BC (i.e. $M \log\log K$) assuming K single-antenna users, can be achieved for $K \to \infty$ even though the transmitter relies on scalar CQI. Several schemes based on partial CSIT are shown to achieve close to DPC sum-rate performance in some asymptotic regimes. However, the majority of these approaches become inevitably interference dominated at high SNR since the error introduced (and the increase in inter-user interference) due to partial CSIT scales with SNR. Hence, in the large power regime, such schemes exhibit a sum rate ceiling behavior and fail to achieve full multiplexing gain.

It would have been flawed to conclude that partial CSIT leads necessarily to a collapse of multiplexing gain. This multiplexing gain loss can be mitigated by using a variable - yet finite - rate feedback channel. In [10], Jindal showed that the feedback load per user must increase approximately linearly with the number of transmit antennas as well as with the transmit power (in dB) in order to achieve the full multiplexing gain. In this thesis, we try to shed some light on these issues, by proposing several robust linear beamforming schemes with limited feedback. The interference dominated behavior of such schemes is studied in detail and several of our proposals provide means to circumvent the sum-rate ceiling effect.

2.6.4 Statistical Channel Knowledge at the Transmitter

Another kind of partial channel state knowledge that can be obtained at the BS with little or no feedback overhead is the statistical CSIT. As second-order channel statistics vary much slower in time compared to the channel realization itself, explicit statistical CSIT can be conveyed periodically to the BS resulting in little uplink overhead. Implicit knowledge on the channel statistics can be obtained without any additional feedback by averaging uplink measurements (statistical reciprocity).

In the literature, two common models for statistical CSIT are:

- Channel Mean Information (CMI), which refers to the case where the mean of the channel distribution is available while the covariance matrix is unknown and often assumed as white.
- Channel Covariance Information (CCI), which refers to the case where the mean is assumed zero (as it is assumed to vary rapidly) and the information regarding the relative geometry of the propagation paths is available through a non-white spatial covariance matrix.

Channel knowledge acquisition using covariance feedback can be applied to both time division duplex (TDD) and frequency division duplex (FDD) systems. In contrast to deterministic reciprocity in TDD systems, the channel statistics of the uplink and the downlink remain related in FDD and the difference between the frequency bands can be overcome by using frequency calibration matrix. Long-term statistical channel knowledge is assumed in Chapter 4, where we show how statistical CSIT can be combined with instantaneous low-rate CQI feedback to increase system throughput by selecting spatially compatible users with large channels gains.

2.7 Scheduling and Multiuser Diversity

In Section 2.5, we presented schemes that deal with the optimization of the input covariance matrices or the precoding design. In this section, a different approach is followed and we try to identify the optimal selection of users to be served. Following the seminal work of Knopp and Humblet [1], multiuser diversity received an increase attention in the field of resource allocation for wireless networks, shattering the traditional view of fading as detrimental. In this work, the authors provided novel insights to the question of '*which user should be served in order to maximize the sum rate*' and gave rise to a novel set of techniques, coined as *opportunistic communication*. Simply speaking, opportunism recommends scheduling the best user (i.e. the user with the most favorable channel conditions) in each coherence interval in order to maximize the system throughput.

Consider a MISO K-user broadcast channel, for which the sum rate capacity is upper bounded by

$$\mathcal{C}_{BC} \leq \mathbb{E}\left\{ \max_{P_k \geq 0, \sum_{k=1}^{K} P_k \leq P} \log_2 \left(1 + \sum_{k=1}^{K} P_k \left\| \mathbf{h}_k \right\|^2 \right) \right\} \tag{2.47}$$

Clearly, the sum rate is maximized when only the strongest user is assigned non-zero power $P_k = P$, i.e. $\mathcal{C}_{BC} = \mathbb{E}\left\{ \log_2 \left(1 + P \max_{1 \leq k \leq K} \left\| \mathbf{h}_k \right\|^2 \right) \right\}$.

Traditionally, channel fading was viewed as a source of unreliability that has to be mitigated. An important means to cope with fading is diversity, which can be obtained over time (interleaving of coded bits), frequency (combining of multipaths in spread-spectrum or frequency-hopping systems) and space (multiple antennas). The basic idea is to improve performance by creating several independent signal paths between the transmitter and the receiver. The seminal work of [1] gave the idea that in the context of multiuser diversity, fading can be considered as a source of randomization that can be exploited. This is done by dynamically scheduling transmissions (resources) among the users as a function of the channel state and serve users only when their instantaneous channel qualities are near to

their peaks. In one phrase, under opportunistic transmission, we transmit when and where the channel is good.

2.7.1 Asymptotic Sum-rate Analysis with Opportunistic Scheduling

Multiuser diversity is a form of diversity inherent in wireless networks, provided by the independent time-varying channels across the different users. The multiuser diversity gain comes from the fact that the effective channel gain, denoted as g_k, is improved from g_k to $\max_{1\leq k\leq K} g_k$. The amount of multiuser diversity gain depends crucially on the tail of the distribution of g_k, implying that the heavier the tail, the more likely there is a user with a very strong channel, and the larger the multiuser diversity gain. Therefore, the channel statistics has an impact on system throughput. In the following, we derive the sum-rate growth for different channel gain distribution.

Fading

We consider that the channels of all users are i.i.d. Rayleigh fading, thus g_k is chi-squared distributed with $2M$ degrees of freedom, i.e. $g_k \sim \chi^2_{(2M)}$ if $g_k = \|\mathbf{h}_k\|^2$ or $g_k \sim \chi^2_{(2)}$ if $g_k = |\mathbf{h}_k|^2$. The limiting distribution (l.d.) of a chi-square random variable is of Gumbel type and it can be shown that the maximum value of K i.i.d. $g_k \sim \chi^2_{(2M)}$ random variables satisfies [9]

$$\begin{aligned}\Pr\{ \quad & \log K + (M-2)\log\log K + O(\log\log\log K) \\ & \leq \max_{1\leq k\leq K} g_k \leq \log K + M\log\log K + O(\log\log\log K)\} \\ & \geq 1 - O\left(\frac{1}{\log K}\right)\end{aligned} \tag{2.48}$$

Therefore, for large K, $\max_{1\leq k\leq K} g_k$ behaves as $\log K$ with high probability, thus $\mathcal{R} \approx \log\log K + \log P + o(1)$. The larger the number of users, the stronger tends to be the strongest channel and the larger the multiuser diversity gain.

It can be easily shown that the limiting distribution of Ricean and Nakagami fading is of Gumbel type. However, the multiuser diversity gain is significantly smaller in the Ricean case compared to the Rayleigh case. Exponential and gamma distributions also belong to the maximum domain of attraction of a Gumbel distribution.

Log-normal Shadowing

We consider now that the effective channel gain g_k is dominated by log-normal shadowing. It can be shown that the maximum value of K i.i.d. log-normal distributed r.v. with logarithmic mean μ_s and variance σ_s^2, satisfies [48]

$$\Pr\{b_K - a_K\log\log K \leq \max_{1\leq k\leq K} g_k \leq b_K + a_K\log\log K\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

where $b_K = \exp\{\sqrt{2\log K}\sigma_s + \mu_s\}$ and $a_K = b_K\sigma_s/\sqrt{2\log K}$.
Hence, the throughput scales like $\mathcal{R} \approx \sqrt{2\log K}\sigma_s + \mu_s$

Pathloss

Consider now a more realistic scenario, in which the users are located randomly over a cell given by a disk of radius R around the serving BS. The channel gain consists in the product between a variable representing the path loss and a variable representing the fast fading coefficient, i.e. $g_k = L_k\gamma_k$, where L_k is the path loss between user k and the BS (cf. Sect. 2.1), and γ_k is the corresponding normalized complex fading coefficient.

We consider a uniform distribution of the population in each cell. Thus the distance between user k and the BS, d_k, is a r.v. with non-uniform distribution $f_D(d)$ given by

$$f_D(d) = 2d/R^2, \quad d \in [0, R] \tag{2.49}$$

Further, the random process d_k can be considered i.i.d. across users and cells, if users in each cell are dropped randomly in each disk. The considered coverage region can be assimilated with the inside area of each disk, in a disk-packing region of the 2D plane. Users dropped outside the disks can dropped from the analysis, as these will not affect the scaling law. Assuming $R = 1$ for normalization, the distribution of $L_k = \beta d_k^{-\epsilon}$ is given by

$$f_L(x) = \begin{cases} \frac{2}{\epsilon}(\frac{x}{\beta})^{-\frac{2}{\epsilon}}\frac{1}{x} & \text{with } x \in [\beta, \infty) \\ 0 & \text{with } x \notin [\beta, \infty) \end{cases} \tag{2.50}$$

The distribution of L_k is remarkable in that it differs strongly from fast fading distributions, due to its *heavy tail* behavior. Formally, L_k follows a Pareto-type distribution and is a regularly varying random variable with exponent $-2/\epsilon$, i.e. $\lim_{t\to\infty} \frac{1-F_\alpha(x)}{1-F_\alpha(tx)} \to t^{2/\epsilon}$. An interesting aspect of regularly varying r.v. is that they are stable with respect to multiplication with other independent r.v. with finite moments as pointed out by the following theorem:

Theorem 2.1 [49]: *Let X and Y be two independent r.v. such that X is regularly varying with exponent $-\eta$. Assuming Y has finite moment $\mathbb{E}\{Y^\eta\}$, then the tail behavior of the product $Z = XY$ is governed by:*

$$1 - F_Z(z) \to \mathbb{E}\{Y^\eta\}(1 - F_X(z)) \quad \textit{when } z \to \infty \tag{2.51}$$

The idea behind this theorem is that when multiplying a regularly varying r.v. with another one with finite moment, one obtains a heavy tailed r.v. whose tail behavior is similar to the first one, up to a scaling. Since γ_k has finite moments, the tail behavior of g_k can be characterized by:

$$1 - F_g(x) \to \mathbb{E}\{\gamma_k^{\frac{2}{\epsilon}}\}\left(\frac{\beta}{x}\right)^{\frac{2}{\epsilon}} \quad \text{when } x \to \infty \tag{2.52}$$

Therefore, g_k is also regularly varying with exponent $-\frac{2}{\epsilon}$, which implies that [50]

$$\lim_{K\to\infty} \Pr\{\max_{1\le k\le K} g_k \le \beta\mathbb{E}\{\gamma_k^{2/\epsilon}\}^{\epsilon/2}K^{\frac{\epsilon}{2}}x\} = e^{-x^{-2/\epsilon}} \quad \forall x > 0, \tag{2.53}$$

Using the above result, we can show that the throughput scales for asymptotically large K as

$$\mathcal{R} \approx \frac{\epsilon}{2}\log K \tag{2.54}$$

Observe that a much greater throughput growth than in the case of fading is obtained. This is due to the amplified multiuser diversity gain due to the presence of unequal path

loss across the user locations in the cell. As the distribution of pathloss belongs to the maximum domain of attraction of Fréchet type, a logarithmic capacity growth with K is achieved. However, the scheduling decisions are taken in a quite unfair fashion admittedly, since the scheduler tends to select users closer to the access point as more users are added to the network.

2.8 Living with partial CSIT: Limited feedback approaches

Limited feedback schemes employing SDMA transmission and efficient scheduling are key topics of this dissertation. In this section, we try to categorize the many possible limited feedback strategies and briefly expose the ones that will be extensively discussed in the following chapters.

2.8.1 Quantization-based techniques

Quantization is the first idea that comes to mind when dealing with source compression, in this case the random channel matrices or the corresponding precoders being the possible sources. The amount of CSIT depends on the frequency of feedback reporting (generally a fraction of the coherence time), the number of parameters being quantized, and the resolution of the quantizer. Most research focuses on reducing the number of parameters and the required resolution. The feedback design problem has been studied in single-user MIMO communication systems using a concept known as limited feedback precoding [51]. The key idea of this line of research has been to quantize the MIMO precoder and not simply the channel coefficients. The challenge of extending this work to multiuser channels is that the transmit precoder depends on the channels of the other users in the system. Simplifying the transmit precoding structure, e.g. using ZF or MMSE precoding, is one of the simplest means to reduce feedback requirements.

In approaches assuming single-antenna receivers, the random codebook and Grassmannian quantization ideas are used to quantize the direction of each user's channel [10, 52]. The main observation in [10] is that the feedback load should scale approximately linearly both with the number of transmit antennas and the SNR (in dB), unlike the single-user case. The reason is that quantization error introduces an SINR floor since it prohibits perfect inter-user interference cancellation. Thus this error must diminish for higher SNRs in order to allow for a balancing between the noise and the residual interference due to channel quantization. As we see in Chapter 5, an improvement can be obtained by feeding back the quantized channel vector and a certain SINR-like scalar value that is - among others - a function of the error between the true and quantized channel.

2.8.2 Dimension reduction and projection techniques

Dimension reduction techniques involve projecting the matrix channel onto one or more basis vectors known to the transmitter and receiver. In that way, the CSIT matrix $\mathbf{H}_k$ of size $N_k \times M$ is mapped into an ℓ-dimensional vector with $1 \leq \ell \leq N_k \times M$, thus reducing the dimensionality of the CSIT to ℓ complex scalars (which in turn may be quantized). Once the projection is carried out, the receiver feeds back a metric $\gamma_k = f(\mathbf{H}_k)$ that is typically

related to the square magnitude of the projected signal. For instance, antenna selection methods fall into this category with the projection being carried out by the receiver itself.

Alternatively, the projection can be the result of using a particular precoder $\mathbf{W}$ at the BS. A good example of this approach is given by a class of schemes using unitary precoding matrices. We now review this approach for $N_k = 1$ where the BS designs an arbitrary unitary precoder $\mathbf{W} = \mathbf{Q}$ of size $M \times M$, further scaled in order to satisfy the power constraint. Each terminal identifies the projection of its vector channel onto the precoder by $\mathbf{h}_k\mathbf{Q}$, and reports an index and a scalar metric expressing the SINR measured under an optimal beamforming vector selection:

$$\gamma_k = \max_{1 \leq i \leq M} \frac{|\mathbf{h}_k \mathbf{q}_i|^2}{M\sigma^2/P + \sum_{j \neq i} |\mathbf{h}_k \mathbf{q}_j|^2} \tag{2.55}$$

where $\mathbf{q}_i$ denotes the i-th column of $\mathbf{Q}$. The scheduling algorithm then consists in opportunistically assigning to each beamformer $\mathbf{q}_i$ the user which has selected it and has reported the highest SINR.

When the unitary precoder must be designed without any *a priori* channel knowledge, a scaled identity matrix can be used (per-antenna SDMA scheduling). In this case, the algorithm falls back to assigning a different selected user to each transmit antenna. In the small number of user case, the performance of such scheme is plagued by inter-user interference, however interference tends to decrease as the number of active users becomes large. In low-mobility system settings (slow fading), the use of a fixed set of precoders may result in severe unfairness between the users due to the limited channel dynamics. This problem can be alleviated by the randomization of the precoders. The idea of random opportunistic beamforming (RBF) [9,53], which is presented in detail in Section 2.9.3, can be recast in the context above, assuming that $\mathbf{Q}$ is randomly generated at each scheduling period, according to an isotropic distribution, while preserving the unitary constraint.

2.9 Linear Precoding and Scheduling with Limited Feedback

We review here two of the main building blocks of the dissertation: the finite rate feedback model and random opportunistic beamforming. The first model will be used for the codebook-based SDMA beamforming and scheduling techniques that we propose in Chapter 5, while the latter is the main building block for approaches in Chapters 3 and 4. We also discuss the common characteristics and particularities of both approaches.

2.9.1 Finite Rate Feedback Model for CDI

Probably the most popular partial CSIT model when a bandwidth constraint on the uplink channel is imposed is the so-called *finite rate feedback model* in the multiuser literature. This is often referred to as *limited feedback* model in works focusing on point-to-point MIMO communications. It is initially proposed for single-user MIMO [51, 54–57] and extended to multiuser MIMO settings in [10, 52]. The finite rate feedback model is linked to vector quantization: with a feedback rate constraint of B_D bits, the receiver can report $N_D = 2^{B_D}$ different channel representations. This implies that the channel space at the receiver is

generally partitioned in N_D non-overlapping regions, with each region represented by a distinct codeword. Partial CSIT under finite rate feedback model corresponds to informing the BS in which region the current channel realization lies.

In this approach, a quantization codebook $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_D}\}$ containing $N_D = 2^{B_D}$ unit-norm vectors $\{\mathbf{v}_i\}_{i=1}^{N_D} \in \mathbb{C}^M$ is utilized. The codebook is assumed to be known to both the transmitter and the receivers and we set $N_k = 1\ \forall k$. At each feedback reporting slot t, each receiver k, based on its current channel realization $\mathbf{h}_k$, determines its 'best' vector from the codebook, i.e. the codeword that optimizes a certain cost function. In settings where the BS exploits the quantized CSI to design the downlink beams, it is often assumed that each receiver quantizes its channel to the vector that maximizes the following inner product [10, 54, 56, 58]

$$\hat{\mathbf{h}}_k = \mathbf{v}_n = \arg\max_{\mathbf{v}_i \in \mathcal{V}} |\bar{\mathbf{h}}_k^H \mathbf{v}_i|^2 = \arg\max_{\mathbf{v}_i \in \mathcal{V}} \cos^2(\angle(\bar{\mathbf{h}}_k, \mathbf{v}_i)) \tag{2.56}$$

where the normalized channel vector $\bar{\mathbf{h}}_k = \mathbf{h}_k / \|\mathbf{h}_k\|$ corresponds to the channel direction, and we refer to $\hat{\mathbf{h}}_k$ as the k-th user channel quantization.

Once the channel vector is quantized, each terminal sends the corresponding quantization index n back to the transmitter using $B_D = \lceil \log_2 N_D \rceil$ bits. In the research literature, it is often assumed for simplicity that the feedback reporting stage is accomplished instantaneously and with no errors. The error-free assumption can be well approximated using sufficiently powerful error-correcting codes over the feedback link, whereas the zero-delay assumption may be valid when the processing and feedback delays are small relative to the channel coherence time. However, these assumptions can be challenged in practical scenarios (cf. Chapter 7), e.g. the feedback delay can be significant in fast fading channels with typical user speeds of 30-50 km/h (large Doppler spread).

2.9.2 Codebook design

The performance of a system relying on quantized CSIT depends heavily on the codebook structure and the design criterion considered. The quantization problem exhibits several similarities with classical source coding problems. As the vector $\mathbf{h}_k \in \mathbb{C}^M$ can be represented by a $2M$-dimensional vector of real coefficients, the codebook design is equivalent to a source coding problem, where the encoder describes a random source $\mathbf{s} \in \mathbb{R}^{2M}$ by one of the entries $\hat{\mathbf{s}}_i \in \mathbb{R}^{2M}$ of a finite alphabet codebook. The codebook and the quantizer are designed to minimize the distortion between the source and its unquantized representation. However, there are several key differences when considering the quantization problem in limited feedback MIMO systems.

In point-to-point MIMO systems, the codebook design problem is explicitly related to the Grassmannian line packing problem [59], as a codevector can be viewed as the coordinates of a point on the surface of a hypersphere with unit radius centered around the origin. This point dictates a straight line in a complex space $\mathbb{C}^M$ that passes through the origin. The inner product (2.56) is related to the chordal distance, defined as the distance $d_{chord}(\bar{\mathbf{h}}_k, \mathbf{v}_i) = \sqrt{1 - |\bar{\mathbf{h}}_k^H \mathbf{v}_i|^2} = \sin(\angle(\bar{\mathbf{h}}_k, \mathbf{v}_i))$ between two lines generated by $\bar{\mathbf{h}}_k$ and $\mathbf{v}_i$. In this dissertation, the chordal distance (2.56) is considered as codeword selection criterion (distortion measure), despite the fact that considering an Euclidean distance metric (and

quantizing the non-normalized channel $\mathbf{h}_k$) may result in increased performance. Quantizing the channel direction and using the chordal distance is motivated by the fact that beamforming on the quantized spatial information is generally used. As the transmitter requires information on the channel direction in order to form beams, quantizing directly the channel realization can be viewed as redundant operation.

Another key difference with source problems is that the channel realization and the variable to be quantized may lie in different spaces and may have different dimensions. For instance, one can typically assume that the vector $\bar{\mathbf{h}}_k$ is constrained to be unit-norm and invariant to arbitrary phase rotation $e^{j\theta}$; hence it lies on the unit hypersphere, whereas the channel instantiation $\mathbf{h}_k$ could be anywhere in the $\mathbb{C}^M$ space.

The problem of optimum codebook design is not yet fully solved, and since the optimal channel vector quantizer is generally difficult to obtain and analyze, one typically resorts to approximate or heuristic codebook design. The complexity of the problem lies on the fact that codebook design depends on various system parameters, including the channel properties and statistics, the antenna configuration and correlation, etc. Furthermore, a codebook can be considered as optimum for a specific distortion metric. Apart from the chordal and Euclidean distances, more general non-mean-squared error functions can be considered in limited feedback MIMO systems (e.g. average received SINR or mutual information loss). However, an efficient and general codebook design rule is the following: for random channels with i.i.d. $\mathcal{CN}(0,1)$ entries, $\bar{\mathbf{h}}_k$ is independent of $\|\mathbf{h}_k\|$ and uniformly distributed over the unit-norm sphere $\mathcal{F}_M = \left\{\mathbf{u} \in \mathbb{C}^M : \|\mathbf{u}\| = 1\right\}$, i.e. $\bar{\mathbf{h}}_k \sim \mathcal{U}(\mathcal{F}_M)$. An efficient quantizer has to satisfy the following two conditions:

- Nearest Neighborhood Condition (NNC): For given codevectors $\{\mathbf{v}_i, i = 1, \ldots, N_D\}$, the optimum partition cell (Voronoi region) $\mathcal{H}_i$ of the i-th codevector $\mathbf{v}_i$ satisfies

$$\mathcal{H}_i = \{\bar{\mathbf{h}}_k \in \mathcal{F}_M : |\bar{\mathbf{h}}_k^H \mathbf{v}_i| \geq |\bar{\mathbf{h}}_k^H \mathbf{v}_j|, \forall j \neq i\}, \text{ for } i = 1, \ldots, N_D \tag{2.57}$$

- Centroid Condition (CC): Given the partitions $\{\mathcal{H}_i, i = 1, \ldots, N_D\}$, the optimum codevectors $\mathbf{v}_i$ satisfy

$$\mathbf{v}_i = \arg \max_{\mathbf{v} \in \mathcal{H}_i} \mathbb{E}\{|\bar{\mathbf{h}}_k^H \mathbf{v}|^2 | \bar{\mathbf{h}}_k \in \mathcal{H}_i\} \tag{2.58}$$

In multiuser MIMO systems, simple codebook structures, including random vector quantization (RVQ) [57,60] and approximate cell vector quantization (ACVQ) [56], are often utilized to model the CDI, since the single-user Grassmannian approach has not been extended yet to multi-antenna broadcast channels. In practical systems, several codebook designs have been reported offering good performance under certain channel settings (cf. Chapter 7).

Random Vector Quantization

Random vector quantization has been proposed for CDMA signature optimization with limited feedback in [60] and applied to point-to-point MIMO systems with limited feedback in [57]. In this scheme, each of the N_D codevectors is independently chosen from an isotropic distribution. RVQ provides a lower bound in terms of performance, due to the fact that any structured codebook should perform at least as well as RVQ. The sharpness of the lower bound is decreased, when the codebook size is decreased, due to the fact that a

RVQ codebook does not uniformly cover the M-dimensional space. For the statistics of quantization error, defined as $\sin^2(\angle(\hat{\mathbf{h}}_k, \bar{\mathbf{h}}_k)) = 1 - \left|\hat{\mathbf{h}}_k^H \bar{\mathbf{h}}_k\right|^2$ under RVQ, the interested reader is referred to [10,61]. Nevertheless, performance analysis of multiuser MIMO schemes employing RVQ, despite its simplicity, does not often result in simple calculations and integrals with closed-form solution. In such case, the following codebook design framework might be of interest.

Approximate Cell Vector Quantization

A geometrical framework for vector quantization was presented in [56]. Therein, in order to evaluate the area of no-outage regions, the authors defined spherical caps on the surface of the hypersphere, which yields a good approximation for the area of no-outage regions. Assuming that each codeword is isotropically distributed in $\mathbb{C}^M$, the unit norm sphere $\mathcal{U}$ where a random vector $\bar{\mathbf{h}}_k$ lies is partitioned into N_D 'quantization regions' (decision regions) $\{\bar{\mathcal{H}}_i; i = 1, \ldots, N_d\}$, where $\bar{\mathcal{H}}_i = \{\bar{\mathbf{h}}_k \in \mathcal{U} : |\bar{\mathbf{h}}_k^H \mathbf{v}_i|^2 \geq |\bar{\mathbf{h}}_k^H \mathbf{v}_j|^2, \forall j \neq i, 1 \leq j \leq N_D\}$. If the channel $\bar{\mathbf{h}}_k \in \bar{\mathcal{H}}_i$, the receiver k feeds back the index i. Approximate cell vector quantization results assuming that each quantization cell is a Voronoi region of spherical cap with the surface area $1/N_D$ of the total surface area of the unit sphere [62]. Since $\bar{\mathbf{h}}_k$ is uniformly distributed over $\mathcal{U}$, we have that $\Pr\{\bar{\mathbf{h}}_k \in \bar{\mathcal{H}}_i\} \approx 1/N_D, \forall i$, and the (approximate) quantization cell is given by [55, 56, 63, 64]

$$\tilde{\mathcal{H}}_i = \{\bar{\mathbf{h}}_k \in \mathcal{U} : 1 - |\bar{\mathbf{h}}_k^H \mathbf{v}_i|^2 \leq \delta\}, \quad \forall i, k$$

for $\delta = (1/N_D)^{\frac{1}{M-1}} = 2^{-B_D/(M-1)}$. Although generally there are overlaps in the approximate quantization cells, this approximation is shown through numerical results to be quite accurate even for small N_D [63]. Furthermore, it can be shown that ACVQ yields an accurate lower bound to the quantization error for any vector quantization codebook [55, 56].

2.9.3 Random Opportunistic Beamforming

If we consider that each user is allowed to use only $B_D = \log_2 M$ bits for CDI quantization, the optimal choice for a randomly generated codebook is one that contains orthonormal vectors. Therefore, the above vector quantization-based techniques can be viewed as extension to a popular, alternative low-rate feedback scheme, coined as random opportunistic beamforming (RBF).

In RBF, $1 \leq \mathcal{B} \leq M$ mutually orthogonal random beams are generated at the transmitter. The single-beam RBF ($\mathcal{B} = 1$) was proposed in [53], while multi-beam RBF ($\mathcal{B} = M$) is proposed in [53] and analyzed in [9]. A unitary precoding matrix $\mathbf{Q}$ is generated randomly according to an isotropic distribution. Its M columns (vectors) $\mathbf{q}_m \in \mathbb{C}^{M \times 1}$ can interpreted as random orthonormal beams. An isotropically distributed (i.d.) unitary matrix can be generated by first generating a $M \times M$ random matrix $\mathbf{X}$ whose elements are independent circularly symmetric complex normal $\mathcal{CN}(0, 1)$, and then perform the QR decomposition $\mathbf{X} = \mathbf{QR}$, where $\mathbf{R}$ is upper triangular and $\mathbf{Q}$ is an i.d. unitary matrix. At time slot t the transmitted signal is given by

$$\mathbf{x}(t) = \sum_{m=1}^{\mathcal{B}} \mathbf{q}_m(t) s_m(t) \tag{2.59}$$

where $s_m(t)$ is a scalar signal intended for the user served on beam m. The SINR of user k in beam m is equal to

$$\mathrm{SINR}_{k,m} = \frac{|\mathbf{h}_k \mathbf{q}_m|^2}{\sum_{j \neq m} |\mathbf{h}_k \mathbf{q}_j|^2 + \mathcal{B}\sigma^2/P} \qquad m = 1, \ldots, \mathcal{B} \tag{2.60}$$

Each user, say the k-th, calculates the SINRs over all beams, i.e. $\mathrm{SINR}_{k,m}$ for $m = 1, \ldots, \mathcal{B}$, finds the beam b_k that provides the maximum SINR, i.e. $b_k = \arg \max_{1 \leq m \leq \mathcal{B}} \mathrm{SINR}_{k,m}$, and feeds back the value of SINR_{k,b_k} in addition to the corresponding beam index b_k. An underlying assumption here is that the users know their own channel coefficients. In turn, the transmitter assigns each beam m to the user k_m with the highest corresponding SINR, i.e. $k_m = \arg \max_{1 \leq k \leq K} \mathrm{SINR}_{k,m}$. Since the users have i.i.d. channels, the CDF of the SINR of a selected user (after scheduling) $F_s(x)$ is given by [9]:

$$F_s(x) = (F_{SINR}(x))^K = \left(1 - \frac{e^{-x\mathcal{B}\sigma^2/P}}{(1+x)^{\mathcal{B}-1}}\right)^K \tag{2.61}$$

The achievable sum rate (assuming Gaussian signaling) is given by

$$\mathcal{R}_{RBF} \approx \mathbb{E}\left\{\sum_{m=1}^{\mathcal{B}} \log_2(1 + \max_{1 \leq k \leq K} \mathrm{SINR}_{k,m})\right\} \tag{2.62}$$

where the approximation is used since there is a probability that user may be the strongest user for more than one beam.

Asymptotic sum-rate analysis showed that, for fixed M, P and $K \to \infty$, the average sum rate of RBF scales as $M \log\log K$, which is the same as the scaling law of the capacity when perfect CSI is available. This is due to the fact that the $\max_{1 \leq k \leq K} \mathrm{SINR}_{k,m}$ behaves like $\log K$, which is the behavior of the numerator (maximum of K i.i.d. $\chi^2_{(2)}$ r.v.'s), as the interference terms become arbitrary small. In other words, in the large K regime, RBF with partial CSIT does not suffer any capacity loss due to inter-user interference despite relying on imperfect (scalar) CSIT. The intuition behind that scheme is that for large K, there exists almost surely a user well-aligned to each beam, as well as with very little interference from other beams. Thus, we have M data streams being transmitted simultaneously in orthogonal spatial directions and as a result, full spatial multiplexing gain is exploited. Furthermore, the authors in [9] show that if $M = O(\log K)$, then a linear capacity scaling with M is guaranteed, and fairness is achieved as a byproduct. Here, the term fairness implies that the probability of choosing users with unequal SNRs is equalized.

A limitation of [9] is that it is optimal for a very large, typically unrealistic number of users. The performance is quickly degrading with decreasing number of users. Furthermore, this degradation is amplified when the number of transmit antennas increases. The reason is intuitive: as the number of active users decreases and M increases, it becomes more and more unlikely that M randomly generated, equipowered beams will match well the vector channels of any set of M users in the cell. In Chapters 3 and 4 we propose several enhancements in order to restore robustness and increase the sum-rate performance of RBF in sparse networks. Moreover, RBF is highly sub-optimal at high SNR, i.e. $\lim_{P \to \infty} \frac{\mathcal{R}_{RBF}}{\log P} = 0$, as it becomes interference dominated. As interference scales with P and cannot be eliminated due

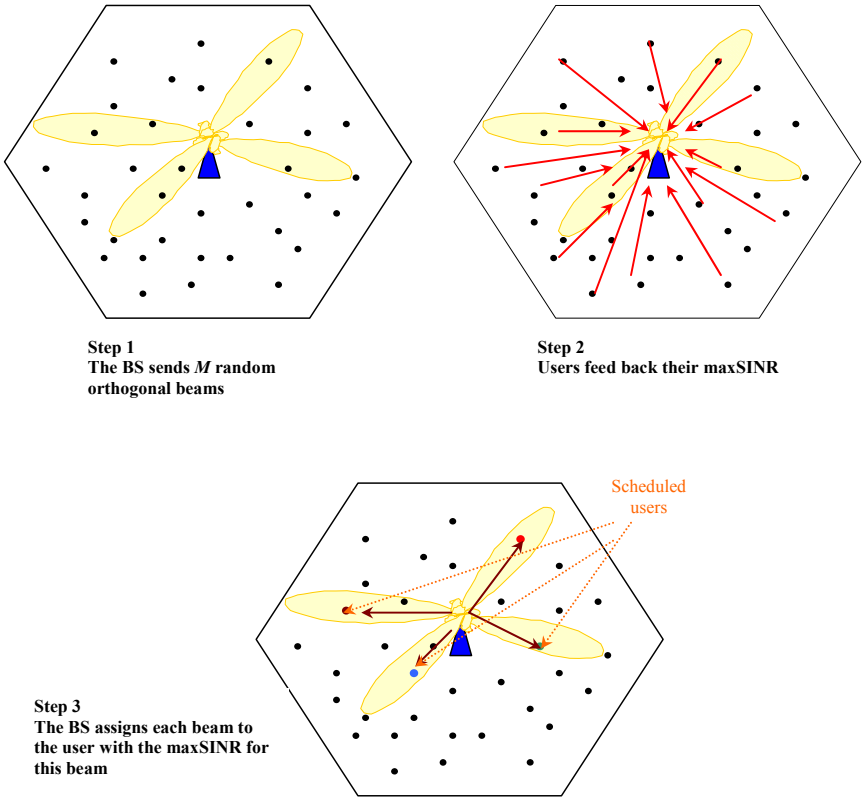

Figure 2.4: Schematic of Random Opportunistic Beamforming.

to partial channel knowledge of fixed rate, the multiplexing gain of M cannot be achieved at high SNR.

Chapter 3

Enhanced Multiuser Random Beamforming

3.1 Introduction

In this chapter, we consider the downlink of a wireless system with a M-antenna base station and K single-antenna users. A limited feedback-based scheduling and beamforming scenario is studied that builds upon the multi-beam RBF framework [9] presented in detail in Section 2.9.3. The popularity of RBF has been spurred by the fact that it yields the same capacity scaling, in terms of multiplexing and multiuser diversity gain, as the optimal full CSIT-based precoding scheme. The optimal capacity scaling of $M \log\log K$ is achieved when the number of users K is arbitrary large, with only little feedback from the users, i.e. in the form of individual SINR. RBF-based approaches have in fact evolved in a topic of research in its own right and many possible strategies can be pointed out [65–68].

The intuition behind the RBF concept is that although the beams are generated randomly and without any a priori CSIT, for large K, the selected group of users exhibit large channel gains as well as good spatial separability, and the probability that the random beam direction is nearly matched to certain users is increased. However, a major drawback of this technique is that its performance is quickly degrading with decreasing K. Furthermore, this degradation is amplified when the number of transmit antennas increases. As the number of active users decreases and M increases, it becomes more and more unlikely that M randomly generated, *equipowered* beams will closely match the vector channels of any set of M users. This situation could easily be faced as traffic is normally bursty with frequent silent periods in data-access networks, thus the scheduler may not count on a large number of simultaneously active users at all times. Another limitation of RBF is that it becomes interference dominated at high SNR, and its multiplexing gain vanishes since interference - which scales with SNR - cannot be eliminated with fixed-rate partial CSIT.

In the first part of this chapter, we provide analytic sum-rate expressions for conventional random beamforming [9] and derive capacity scaling laws at high SNR. Main implication of our results is that in certain asymptotic regimes, it is beneficial to reduce the number of active beams, i.e. the beams allocated non-zero power.

In the second part of this chapter, we investigate solutions to circumvent the limitations of RBF for K decreasing. We introduce a new class of random unitary beamforming-inspired schemes that exhibits robustness in cells with - practically relevant - low to moderate number of users (sparse networks), while preserving the limited feedback and low-complexity advantages of RBF. One first key idea is based on splitting the design between the scheduling and the final beam computation (or "user serving") stages, thus taking profit from the fact the number of users to be served at each scheduling slot is much less than the number of active users (i.e., $\mathcal{B} \leq M << K$). In the scheduling phase, a finite feedback rate scheduling scheme is presented exploiting the concept of RBF. We use the SINR reported by all users, which is measured upon the initial precoding matrix as a basis on which to further improve the design of the final beams that will be used to serve the selected users. In general, the initial precoder can be designed based on any a priori channel knowledge; however here we assume that the first-stage beams are generated at random as in [9] since no a priori CSIT is assumed. Once the group of $\mathcal{B}(1 \leq \mathcal{B} \leq M)$ users is pre-selected using the SINR feedback on the random beams, additional CSIT may be requested to only the selected user group in order to design the final precoder. More specifically, we make the following proposals and contributions:

- The second-stage precoding matrix may require variable levels of additional CSIT feedback to be computed, depending on design targets, and the final beams will improve over the random beamforming used in [9]. In particular, while we expect little gain over [9] for large K, significant throughput gain appears for sparse networks in which the initial random beamformer may not provide satisfactory SINR for all M users.

- If we restrict ourselves to the case that the initial beam directions do not change, we propose then to adapt the power and the number of active beams according to the number of users, the average SNR and the number of transmit antennas as a means to maximize the system throughput.

- In one variant of the proposed designs, we study a power allocation scheme across the $\mathcal{B}$ (initially equipowered) random beams showing substantial capacity improvement over [9] for a wide range of values of K. The scheme requires $\mathcal{B} \leq M$ real-valued scalar values to be fed back from each of the $\mathcal{B}$ pre-selected users. For a 2-beam system, the global optimal beam power solution is provided in closed-form, whereas for the general $\mathcal{B}$-beam case, solutions based on iterative algorithms are proposed and numerically simulated.

- In another proposed robust variant of RBF, no additional CSIT feedback is required during the second stage. Instead, we exploit the SINR information obtained under the random beams in the first stage in order to not only perform scheduling but also to refine the beamforming matrix itself. An on/off beam power control is proposed as a low-complexity solution, yielding a dual-mode scheme switching from TDMA transmission (only one beam is allocated non-zero power) to SDMA where all beams

are active with equal power. The throughput gains over [9] are shown to be substantial for high SNR and low K values.

3.2 Sum-Rate Analysis of Random Beamforming

We first consider the conventional multiuser random beamforming [9], for which Sharif and Hassibi provide capacity scaling laws for asymptotically large K using extreme value theory. In this section, we complement their throughput analysis by calculating analytically the average sum rate for any values of K and M. In addition to an exact throughput characterization, a simple, closed-form expression is provided that approximates very accurately the throughput for relatively high and low SNR levels. Furthermore, using extreme value theory, we derive the capacity growth in P up to the second order revealing the beneficial role of multiuser diversity in the interference-limited region ($P \to \infty$).

Exact throughput of multiuser RBF

We consider the system model described in Section 2.9.3 and for notation convenience we define $\rho = \frac{P}{\sigma^2 M}$.

Lemma 3.1: *For any values of P, M, and K, the average sum rate of multiuser RBF satisfies*

$$\mathcal{A} - \alpha \leq \mathcal{R}_{\text{RBF}} \leq \mathcal{A} \tag{3.1a}$$

with

$$\mathcal{A} = \frac{M}{\log 2} \sum_{k=1}^{K} \binom{K}{k} (-1)^{k+1} e^{\frac{k}{2\rho}} \left(\frac{k}{\rho}\right)^{\frac{(M-1)k-1}{2}} \mathcal{W}_{\frac{k(1-M)-1}{2}, \frac{k(1-M)}{2}}\left(\frac{k}{\rho}\right) \tag{3.1b}$$

$$\alpha = \frac{M}{\log 2} \sum_{k=1}^{K} \binom{K}{k} (-1)^{k+1} e^{\frac{k}{\rho}} \left(\frac{k}{\rho}\right)^{(M-1)k} \left(\Gamma(k(1-M), \frac{k}{\rho}) - \Gamma(k(1-M), \frac{k}{2\rho})\right) \tag{3.1c}$$

where $\mathcal{W}_{k,m}(z)$ is the Whittaker function and $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$ is the upper incomplete gamma function.

Proof. The proof is given in Appendix 3.A. □

Approximate throughput of multiuser RBF

Although the closed-form expression (3.1a) is accurate, it is unfortunately involved and offers no insight. For that, we derive the following simple, approximate expression for the average sum rate, which proves to be accurate.

Lemma 3.2: *For any values of P, M, and K, the average sum rate of multiuser RBF is approximately given by*

$$\mathcal{R}_{\text{RBF}} \approx \frac{M}{\log 2} \frac{\rho H_K}{(M-1)\rho + 1} \tag{3.2}$$

where $H_K = \sum_{k=1}^{K} \frac{1}{k}$ is the K-th harmonic number.

Proof. The proof is given in Appendix 3.B. □

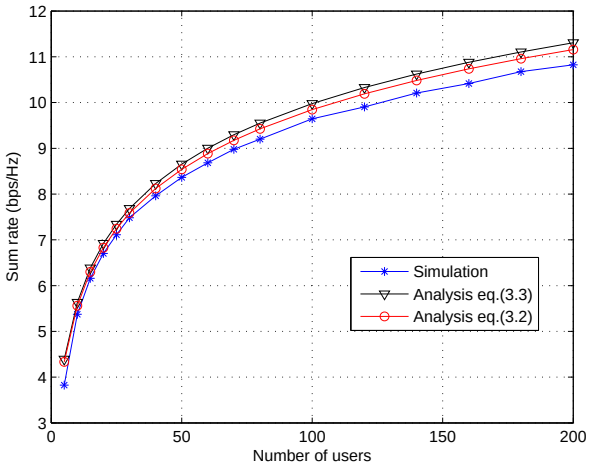

Figure 3.1: Comparison between simulated and analytical achievable sum-rate of RBF with $M = 4$ antennas and SNR = 20 dB.

Average sum rate at high SNR

In the high power regime ($P \to \infty$), the throughput is given by the following corollary, which is a direct result of Lemma 3.2 for $\rho \to \infty$ (i.e, $\mathcal{R}_{high} = \lim_{P\to\infty} \mathcal{R}_{\text{RBF}}$).

Corollary 3.1: *The average sum rate of multiuser RBF for any K, M at high SNR is upper bounded by*

$$\mathcal{R}_{high} \approx \frac{M}{M-1} H_K \log_2 e \tag{3.3}$$

The upper bound is sharp for asymptotically high SNR values. Similar result can be found in [68]. The above corollary can be alternatively derived by approximating the received SINR as $\text{SINR}_{k,m} \approx \frac{|\mathbf{h}_k \mathbf{q}_m|^2}{\sum_{j\neq m} |\mathbf{h}_k \mathbf{q}_j|^2}$ with CDF given by $F(x) = 1 - \frac{1}{(1+x)^{M-1}}$. The approximate average sum rate is given by

$$\mathcal{R}_{high} \approx \int_0^\infty \log_2(1+x) dF^K = \frac{M}{M-1} \int_0^1 \log_2 \frac{1}{1-z^{1/K}} dz = \frac{M}{M-1} H_K \log_2 e \tag{3.4}$$

The tightness of the approximate closed-form expressions (3.2) and (3.3) is compared with simulated results in Figures 3.1 and 3.2.

Average sum rate at low SNR

In the low power regime ($P \to 0$), the throughput is characterized by the following lemma:

Lemma 3.3: *The average sum rate of multiuser RBF for any K, M at low SNR is given by*

$$\mathcal{R}_{low} \approx \frac{MK}{\log 2} \sum_{k=0}^{K-1} \binom{K-1}{k} (-1)^{k+1} \frac{e^{\frac{k+1}{\rho}}}{k+1} \text{Ei}(-\frac{\text{k}+1}{\rho}) \tag{3.5}$$

with $\text{Ei(x)} = -\int_{-\text{x}}^{\infty} \frac{\text{e}^{-\text{t}}}{\text{t}} \text{dt}$ *is the exponential integral.*

Proof. The proof is given in Appendix 3.C. □

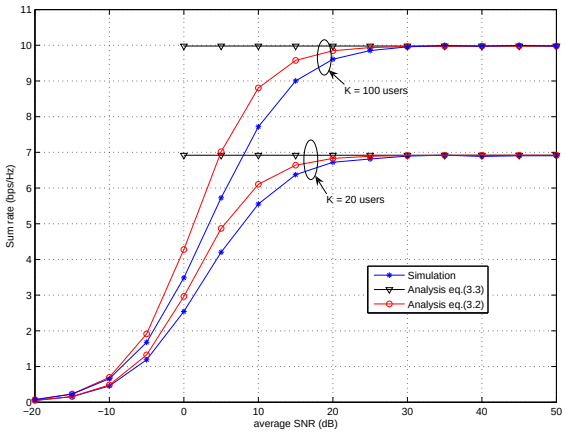

Figure 3.2: Achievable sum rate comparison vs. average SNR for RBF with $M = 4$ antennas. Both analytic expressions approximate accurately the simulated performance at high SNR.

Corollary 3.2: *In the low power regime, the average sum rate of RBF can be approximated as*

$$\mathcal{R}_{low} \approx \frac{\rho}{\log 2} H_K \left(1 - \frac{\rho}{2}(1 + H_K)\right) \leq \frac{\rho}{\log 2} H_K \tag{3.6}$$

Proof. The proof is given in Appendix 3.D. □

The tightness of the above sum-rate approximation is examined in Figure 3.3.

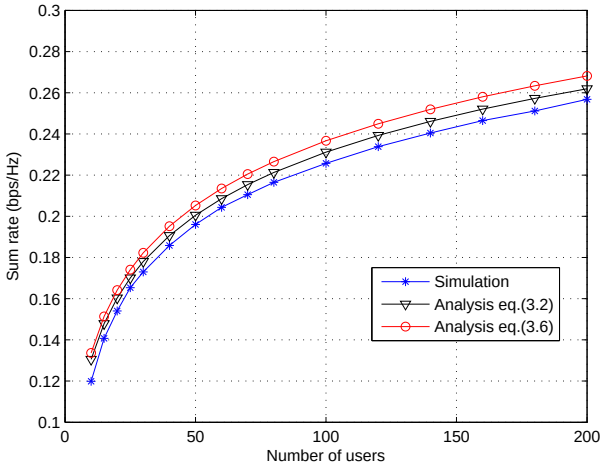

Figure 3.3: Achievable sum rate comparison between simulated and analytical results for RBF with $M = 4$ antennas and SNR = -15 dB.

On the Optimal Number of Active Beams

From the above closed-form sum-rate expressions, we can conclude that the achievable throughput is not always an increasing function with the number of beams for all average SNR ranges. In this section, we provide the optimal number of active beams, i.e. the beams that are assigned non-zero power, for different operating average SNRs. The obtained results provide an additional motivation for the techniques presented in the subsequent parts of this chapter, in which we adjust the number of active beams and/or the power allocated to them as a means to maximize the achievable throughput (beam selection).

We denote the number of active beams as $\mathcal{B}$ and we try to identify the optimal value of $\mathcal{B}^*$ that maximizes the sum rate for fixed K, i.e.

$$\mathcal{B}^* = \max_{1 \leq \mathcal{B} \leq M} \mathcal{R}_{\mathrm{RBF}}(\mathcal{B}) \tag{3.7}$$

When RBF operates at low SNR, then

Proposition 3.1: *At low SNR ($P \to 0$), it is optimal to allocate power to all beams, i.e. $\mathcal{B}^* = M$.*

Proof. Differentiating (3.5) with respect to $\mathcal{B}$, we see that $\frac{\partial \mathcal{R}_{low}}{\partial \mathcal{B}} > 0$ which implies that $\mathcal{R}_{low}$ is increasing with $\mathcal{B}$. The result can be shown alternatively considering the CDF of SINR at low SNR, i.e. $F_{low}(x) = 1 - e^{-\sigma^2 x \mathcal{B}/P}$ and showing

$$\begin{aligned} \frac{\partial \mathcal{R}_{low}}{\partial \mathcal{B}} &= \frac{\partial}{\partial \mathcal{M}} \left\{ \mathcal{B} \int_0^\infty \log_2(1+x) dF_{low}^K \right\} \\ &> -\frac{P^2}{\mathcal{B}^2} \int_0^1 \log_2^2(1 - x^{1/K}) dx > 0 \end{aligned} \tag{3.8}$$

□

On the other side, when the system operates at asymptotically high SNR, we have that

Proposition 3.2: *At high SNR ($P \to \infty$), it is optimal to allocate non-zero power to only one beam, i.e. $\mathcal{B}^* = 1$.*

Proof. As $\frac{\partial \mathcal{R}_{high}}{\partial \mathcal{B}} < 0$, we have that $\mathcal{R}_{high}$ is a monotonically decreasing function with $\mathcal{B}$. □

To summarize, in the low power regime, it is beneficial from a sum-rate maximizing point of view to allocate non-zero power to a higher number of beams, whereas in the interference-limited region ($P \to \infty$) with fixed K, scheduling only one user (TDMA) is the transmission strategy that maximizes the system throughput.

3.3 Capacity scaling laws for high SNR

The asymptotic throughput analysis in [9] was focused on the large K regime with fixed P. However, when P is increasing, random beamforming is highly sub-optimal since it becomes interference dominated. The achievable sum rate saturates, as it does not scale logarithmically with the power. Therefore, the multiplexing gain collapses to zero, i.e.

$\lim_{P\to\infty} \frac{\mathcal{R}_{\text{RBF}}}{\log P} = 0$. Here we investigate the asymptotic behavior of $\max_{1\le k\le K} \text{SINR}_{k,m}$ in high power regime. In this regime, the SINR becomes

$$\lim_{P\to\infty} \text{SINR}_{k,m} = \text{SIR}_{k,m} = \frac{|\mathbf{h}_k\mathbf{q}_m|^2}{\sum_{j\neq m}|\mathbf{h}_k\mathbf{q}_j|^2} \sim \frac{\chi^2_{(2)}}{\chi^2_{(2M-2)}} \tag{3.9}$$

which is an F-distributed r.v. as it is the ratio of two independent chi-squared random variables. Let $X_{k,m} = \text{SIR}_{k,m}$ for $k = 1,\ldots,K$ be a sequence of K i.i.d. r.v. with common parent distribution $F(x) = 1 - \frac{1}{(1+x)^{M-1}}$ and PDF $f(x) = \frac{M-1}{(1+x)^M}$. Let $F_{j:K}(x)$ denote the CDF of the j-th largest r.v. among $\{X_1,\ldots,X_K\}$, denoted as $X_{j:K}$, where the beam index is omitted for notation convenience. The asymptotic sum rate performance depends on the limiting distribution of the variate $X_{K:K} = \max_{1\le j\le K} X_{j:K}$, whose CDF is given by $F_{K:K}(x) = [F(x)]^K$. The distribution $F(x)$ is of Pareto-type and it belongs to the class of regularly varying functions.

Definition 3.1: *A non-negative r.v. X and its distribution are said to be regularly varying with index $\alpha \geq 0$ if the right distribution tail $\overline{F}(x) = 1 - F(x)$ is regularly varying with index $-\alpha$, i.e.,*

$$\lim_{x\to\infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = t^{-\alpha} \quad \forall t > 0$$

Since $F_X(x)$ is a regularly varying function at ∞ with exponent $-(M-1)$, the necessary and sufficient condition for maximal attraction to the limit law of the Fréchet type $\mathcal{D}(G_1)$, i.e. $F_X(x) \in \mathcal{D}(G_1)$, is satisfied [50]. Hence, the distribution $F(x)$ belongs to the maximal domain of attraction of Fréchet type, with limit distribution (l.d.)

$$G_1\ (x; M-1) = \begin{cases} e^{-x^{-(M-1)}} & x > 0, M > 1 \\ 0 & x \leq 0 \end{cases}$$

meaning that there is a sequence $a_K > 0$ such that

$$\lim_{K\to\infty} \Pr\{X_{K:K} \leq a_K x\} = \lim_{K\to\infty} F^K(a_K x) \to G_1\ (x; M-1) \tag{3.10}$$

The fact that the SIR distribution lies in the domain of attraction for maxima of Fréchet type can be alternatively proved using Smirnov's theorem [69,70]. For normalizing sequences $a_K = K^{1/(M-1)}$ and $b_K = -1$, we have that

$$\lim_{K\to\infty} F_{K:K}(a_K x + b_K) = \Upsilon_1^K(x) \tag{3.11}$$

with $\Upsilon_1^K(x) = G_1\ (x; M-1) \sum_{i=0}^{K-1} \frac{(-\log G_1\ (x; M-1))^i}{i!}$

In order to derive the second-order terms of the capacity growth, we need to measure the rate of convergence of the distribution of the sample maximum. For that, the uniform distance metric, defined as $d_K = \sup_x \left|F_X^K(a_K x) - G_1\ (x; M-1)\right|$, is considered, resulting in the following theorem.

Theorem 3.1: *Let $X_{j:K}$ denote the j^{th} largest random variable in a random sample of K, then*

$$\Pr\left\{\left(\frac{\log\sqrt{K}}{K}\right)^{-\frac{1}{M-1}} - 1 \leq X_{j:K} \leq \left(K\log\sqrt{K}\right)^{\frac{1}{M-1}} - 1\right\} \geq 1 - O\left(\frac{1}{\log K}\right) \quad (3.12)$$

Proof. The proof is given in Appendix 3.E □

Therefore, it readily follows that at large K, the sum rate of multiuser random beamforming in the interference-limited region scales (and a fortiori its average) as

$$\mathcal{R}_{high} \sim \frac{M}{M-1}\log_2 K + \frac{M}{M-1}\log_2\log\sqrt{K} + O(1) \quad (3.13)$$

Theorem 3.1 establishes rigorously the sub-optimality of RBF in the high power regime. As the interference scales with P, the scheme becomes interference dominated at high SNR, and the multiplexing gain vanishes. Interestingly, multiuser diversity gain becomes more important in this regime, since the sum rate exhibits logarithmic growth with K (in contrast to the double logarithmic $\log\log K$). Although only a fraction of the spatial multiplexing gain is achieved ($r = \frac{M}{M-1}$), multiuser diversity increases the sum rate by a factor of $\log K$, compensating thus for the loss in degrees of freedom. Simply speaking, having more active users to choose from, it 'pushes' the interference-limited region to higher SNR values. Another implication of the above theorem is the *optimality of TDMA at high SNR*: as $\frac{\partial\mathcal{R}_{high}}{\partial M} < 0$ (either (3.4) or (3.13)), the throughput is a monotonically decreasing function with M, implying that at high SNR the sum rate is maximized by using only one beam.

The first-order term in capacity growth of RBF with respect to P can be alternatively derived using the following, more intuitive way. From the convergence of $F^K(x)$ to a Fréchet distribution we have

$$\Pr\{X_{K:K} \leq u_K x\} \to G_1\ (x; M-1) = e^{-x^{-(M-1)}} \quad (3.14)$$

with normalizing sequence $u_K = F^{-1}(1-1/K) = K^{1/(M-1)} - 1$. The average sum-rate at high SNR is given as

$$\mathcal{R}_{high} = M\int_0^\infty \log_2(1+x)dF^K(x) = \frac{M}{\log 2}\int_0^\infty \frac{1-e^{-(u_K/x)^{M-1}}}{1+x}dx \quad (3.15)$$

where the RHS of the equation is obtained through integration by parts. By using the change of variable $y = 1/x$ for $x \in (0,\infty)$ and the approximation $e^{-c} \approx 0$ for some positive value c, we have

$$\mathcal{R}_{high} \approx \frac{M}{\log 2}\left(\int_0^{\frac{c^{\frac{1}{M-1}}}{u_K}} \frac{1-e^{-(u_K x)^{M-1}}}{y(1+y)}dy + \int_{\frac{c^{\frac{1}{M-1}}}{u_K}}^\infty \frac{1}{y(1+y)}dy\right) \quad (3.16)$$

Therefore, for K asymptotically large, $\lim_{K\to\infty}\frac{c^{\frac{1}{M-1}}}{u_K} = 0$ as $u_K \overset{K\to\infty}{\to} \infty$ and

$$\lim_{K\to\infty}\frac{\mathcal{R}_{high}}{M\log_2(u_K+1)} = \lim_{K\to\infty}\frac{\frac{M}{\log 2}\log\left(\frac{u_K}{c^{1/(M-1)}}+1\right)}{M\log_2(u_K+1)} = 1 \quad (3.17)$$

which implies that the average sum-rate scales (in the ratio convergence sense) as $\mathcal{R}_{high} \sim M \log_2(u_K + 1) = \frac{M}{M-1} \log_2 K$.

3.4 Two-Stage Scheduling and Linear Precoding

In this section, we propose a MIMO downlink scheduling and beamforming framework in which the design is split into two stages. In the first stage, a coarse beamforming matrix is used (possibly selected even at random) and user group (of size $|\mathcal{S}|$) selection is performed among all K active users. In the second stage, possibly additional channel quality information is collected for the selected user group, and an improved beamforming matrix is designed to serve them. The fact that $|\mathcal{S}| << K$ is instrumental in reducing the total feedback requirement in this scenario. The two-stage framework can be described as follows:

Stage 1: User Selection

The transmitter generates a linear precoding matrix $\mathbf{W}$ based on any a priori channel information it may has. Here, since we consider that the channel conditions of the users are not known a priori, a $\mathcal{B} \times \mathcal{B}$ ($\mathcal{B} \leq M$) unitary precoding matrix $\mathbf{Q}$ is drawn randomly and equal power allocation is used ($P_m = \frac{P}{\mathcal{B}}, \forall m$), as a means to reduce the feedback burden and complexity requirements, i.e. $\mathbf{W} = \mathbf{Q} = [\mathbf{q}_1 \dots \mathbf{q}_\mathcal{B}]$. The $\mathcal{B}$ columns $\mathbf{q}_m \in \mathbb{C}^{M \times 1}$ of the precoder can be interpreted as random orthonormal beams, generated according to an isotropic distribution, as proposed in [9]. Each of the K users, say the k-th, calculates the SINRs over all equipowered beams, i.e.

$$\text{SINR}_{k,m} = \frac{|\mathbf{h}_k \mathbf{q}_m|^2}{\sum_{j \neq m} |\mathbf{h}_k \mathbf{q}_j|^2 + \mathcal{B}\sigma^2/P} \qquad m = 1, \dots, \mathcal{B} \tag{3.18}$$

finds the beam b_k that provides the maximum SINR, and feeds back $\gamma_k = \text{SINR}_{k,b_k}$ in addition to the corresponding beam index. A simple and low-complexity user selection scheme is employed at the BS by selecting the users that have the highest SINR on each beam $\mathbf{q}_m$. The group of selected users is denoted as $\mathcal{S}$. In [9] $\mathcal{B} = M$ beams are activated. In the general case however, we could decide to activate the $\mathcal{B} \leq M$ best beams only.

Stage 2: Final Precoding design

In our proposed framework, we follow up with a second stage where the $\mathcal{B}$ users in $\mathcal{S}$ may be allowed to report back to the BS additional limited feedback, denoted as $\gamma_k^{'}$, $k \in \mathcal{S}$. Based on the feedback information, the transmitter designs the final precoding matrix $\mathbf{W}^{'}(\mathcal{S}) = f(\gamma_k^{'})$, where $f(\cdot)$ is some feedback-based beamforming design function. Note that in [9] there is no second stage, in other words $\mathbf{W}^{'}(\mathcal{S}) = \mathbf{Q}$. The second-stage feedback can take on among others the following forms, depending on the system feedback rate constraint:

- *Strategy 1*: $\gamma_k^{'} = \mathbf{h}_k$ (full CSIT)
- *Strategy 2*: $\gamma_k^{'} = \hat{\mathbf{h}}_k$ (quantized channel vector)
- *Strategy 3*: $\gamma_k^{'} = |\mathbf{h}_k \mathbf{q}_m|^2$ (BGI: beam gain information)
- *Strategy 4*: $\gamma_k^{'} = \gamma_k$ (no additional feedback)

Note that anyone of these two-stage schemes represents an efficient feedback reduction strategy considering the number of selected users $\mathcal{B}$ is typically very small in comparison with K. For instance $\mathcal{B} = 2$ or 3 in practical standardized systems while K could be a few tens even for moderately sparse networks. The optimal way of splitting the feedback load across the stage 1 (scheduling) and the stage 2 (beam design) is an interesting open problem, beyond the scope of the thesis, although some design rules for ZFBF systems where $\gamma_k^{'}$ is given by a quantized version of the quantization error of the channel and ZFBF have been already suggested [71].

Note that the design of a two-stage feedback scheme will inevitably introduce a longer hand-shaking delay before the actual payload data can be sent to the mobile. For an efficient operation of feedback-based approach (whether single stage or two-stage), the total duration spent on feedback together with payload transmission must be significantly less that the coherence time of the channel T_{coh}. Therefore, for the 2-stage approach to be applicable, we envision a framing structure that encompasses the two stages of feedback, back to back, as an overall feedback preamble, prior to payload transmission. This preamble (minislot) of short duration τ_m, during which users report their feedback messages is thus followed by a larger slot of duration $\tau_s >> \tau_m$, which is dedicated to pilot and data transmission. The total framing interval duration should be kept less than the coherence time of the channel, i.e. $\tau_s + \tau_m \leq T_{coh}$. Note that the second stage of feedback collects fresh CSIT, so that the precoder design does not suffer from extra outdating degradation (compared with a single stage feedback).

3.5 Enhanced Multiuser Random Beamforming

When the number of active users K is large (dense networks), RBF can benefit from multiuser diversity by scheduling users with favorable channel conditions (highest SINR), improving thus the system capacity. The selected group of users exhibit large channel gains and good spatial separability among them and the probability that the random beam direction is closely matched to certain users is increased with increasing K. For low to moderate number of users (sparse networks), the probability that all $\mathcal{B}$ users enjoy a reasonable SINR is lower since the selected users may not be fully separable under a randomly generated unitary beamforming matrix $\mathbf{Q}$. Nevertheless, we point out that this user set, the user group selected by the scheduler under the initial random orthogonal beams, is likely to exhibit good separability conditions relative to the rest of the users, since it is at least the best user group for *one* orthogonal precoder $\mathbf{Q}$. Therefore, we argue that a design based on random $\mathbf{Q}$ could be kept for the purpose of *scheduling*. In strategies 1-3, we propose to augment the random beamforming step (stage 1) with an additional yet low-rate CSIT feedback (stage 2), as a means to restore robustness and improve sum-rate performance. Note that the second stage only involves the $\mathcal{B}$ pre-selected users. In this chapter, we present results for strategy 1, but we focus on *strategies 3 and 4* in particular due to their low-rate feedback merits. Results for *strategy 1* are also presented in the following section.

3.6 Enhanced Precoding with perfect second-stage CSIT

We first consider the case where, once the set of scheduled users is determined, perfect CSIT feedback is requested for the $\mathcal{B}$ selected users (*strategy 1*). Note that this results in an overall feedback requirement much inferior to that of [12]. Based on the second-stage CSIT, for any set of transmission powers $\mathbf{P} = [P_1, ..., P_{\mathcal{B}}]$, the beamforming matrix $\mathbf{W}'(\mathcal{S})$ that maximizes the SINR of each user is given by

$$\mathbf{W}'(\mathcal{S}) = \mathbf{H}(\mathcal{S})^H(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H + \beta\mathbf{I})^{-1}$$

Note that the optimal precoding matrix in the downlink is derived from the uplink MMSE beamformer, based on the uplink-downlink duality. Therefore, using the RBF as a user pre-selection scheme, a set of quasi-orthogonal users is revealed to the transmitter. The BS in turn applies MMSE precoding in order to serve the selected users. The suboptimality of this strategy depends on the sparsity of the system. The more users are in the cell, the more likely is to select an orthogonal user group at the first step. Note that the performance of MMSE downlink precoder can be enhanced using power allocation. However, the solution to this optimization problem is not trivial, even if the duality is exploited. Another key message of this technique is the effective channel (SINR) is a powerful user selection metric, since it reveals the set of users with high channel gains and quasi-orthogonal channel directions.

3.7 Beam Power Control with Beam Gain Information

We consider now that *strategy 3* is adopted during the second stage, thus the scheduler gains knowledge of $\eta_{k_m m} = |\mathbf{h}_{k_m}\mathbf{q}_m|^2$ for each $k_m \in \mathcal{S}$. Without loss of generality (WloG), we order the users such that $\eta_{k_i i} \geq \eta_{k_j j}, \forall i < j$ is assumed, and unless otherwise stated $\mathcal{B} = M$. Note that the extra feedback load is minimal because it concerns only $\mathcal{B}$ users. If a moderate number of users exist, some of the random beams may not reach a target. This is measured at the BS in terms of the BGI $\eta_{k_m m}$. In turn, the beam power control is used to reduce the resource allocated to the low-quality beams, to the benefit of the good-quality beams. As a result, we choose not to change the direction of the initial random beams. Based on this *beam gain information* (BGI) $\eta_{k_m m}$ we propose to design the beamforming matrix by applying a power allocation strategy across the beams of $\{\mathbf{q}_m\}_{m=1}^M$, i.e. $\mathbf{w}_m = \sqrt{P_m}\mathbf{q}_m$.

Define the vector of transmit powers $\mathbf{P} = [P_1 \dots P_M]$ where P_m is the transmit power on beam m. The SINR of the selected user $k_m \in \mathcal{S}$ over its preferred beam m can be expressed as:

$$\text{SINR}_{k_m,m}(\mathbf{P}) = \frac{P_m \eta_{k_m m}}{\sigma^2 + \sum_{j \neq m} P_j \eta_{k_m j}} \tag{3.19}$$

The beam power allocation problem for RBF in order to maximize the sum rate subject to a total power constraint can be formulated as:

$$\begin{aligned} \max_{\mathbf{P}} \mathcal{R}(\mathcal{S}, \mathbf{P}) &= \max_{\mathbf{P}} \sum_{m=1}^{M} \log_2 \left(1 + \text{SINR}_{k_m,m}(\mathbf{P})\right) \\ s.t. \sum_{m=1}^{M} P_m &\leq P,\ P_m \geq 0,\ m = 1, \dots, M \end{aligned} \tag{3.20}$$

We first remark that the power constraint is always satisfied with equality. This is easily verified by noting that any power vector $\mathbf{P}^{'}$ with $\sum_m P_M^{'} < P$ cannot be the optimum power vector. For any $\epsilon > 1$, a power vector $\mathbf{P}$ with $P_m = \epsilon P_m^{'}$, $m = 1, \dots M$ such that $\sum_m \epsilon P_M = P$ increases the sum rate $\mathcal{R}(\mathcal{S}, \mathbf{P})$, since it increases all user rates.
In what follows we search for the optimal beam power allocation (power vector $\mathbf{P}^*$) by finding

$$\mathbf{P}^* = \arg \max_{\mathbf{P} \in \mathcal{P}^M} \mathcal{R}(\mathcal{S}, \mathbf{P}) \tag{3.21}$$

where $\mathcal{P}^M = \{\mathbf{P} | \sum_m P_m \leq P, P_m \geq 0, m = 1, \dots, M\}$ is the constraint set, which is a closed and bounded set. Although the sum rate function is concave in SINR, it is not strictly concave in power. Thus, the optimization problem is hard to solve due to non-convexity of the objective function, plus no transformation into convex by relaxation seems doable. This problem is however typical of sum-rate maximizing power control [72]. In the following sections, we investigate a closed-form optimal solution for a 2-beam system and iterative solutions for the general case. Moreover, the above beam power control setup can be seen as an instance of the interference channel, the analysis of which is a famously difficult problem in information theory. Our power allocation solutions can be therefore used to any communication network that can be modeled as an interference channel.

3.7.1 Optimum Beam Power Allocation for Two Beams

For RBF scheme with $\mathcal{B} = 2$ beams, the optimum beam power allocation policy under strategy 3 can be derived analytically. The sum rate for user set $\mathcal{S} = \{k_1, k_2\}$ is given in terms of $P_1 \in [0, P]$ by:

$$\begin{aligned} \mathcal{R}(\mathcal{S}, P_1) &= \sum_{m=1}^{2} \log_2 \left(1 + \mathrm{SINR}_{k_m, m}\right) \\ &= \log_2 \left[\left(1 + \frac{P_1 \eta_{k_1 1}}{\sigma^2 + (P - P_1)\eta_{k_1 2}}\right) \left(1 + \frac{(P - P_1)\eta_{k_2 2}}{\sigma^2 + P_1 \eta_{k_2 1}}\right) \right] \end{aligned} \tag{3.22}$$

Since the logarithm is a monotonically increasing function, we can consider the following objective function:

$$\begin{aligned} \mathcal{J}(P_1) &= \left(1 + \mathrm{SINR}_{k_1 1}\right) \left(1 + \mathrm{SINR}_{k_2 2}\right) \\ &= \left(1 + \frac{P_1 \eta_{k_1 1}}{\sigma^2 + (P - P_1)\eta_{k_1 2}}\right) \left(1 + \frac{(P - P_1)\eta_{k_2 2}}{\sigma^2 + P_1 \eta_{k_2 1}}\right) \end{aligned} \tag{3.23}$$

By Fermat's theorem, the necessary conditions for maxima of the continuous objective function can occur either at its critical points or at points on its boundary. Therefore, the global maximizer of the above generally non-convex optimization problem is given by the following alternatives:

- boundary points of $\mathcal{P}^2$: $P_1 = 0$ or $P_1 = P$.
- extreme points on the boundary of $\mathcal{P}^2$: i.e., the values $P_1 \in [0, P]$ resulting from $\frac{\partial \mathcal{J}(P_1)}{\partial P_1} = 0$.

Specifically, we have the following result:

Theorem 3.2: *For the two-beam RBF, the optimum sum-rate maximizing beam power allocation* $\mathbf{P}^* = (P_1^*, P_2^*)$ *is given by:*

$$\begin{cases} P_1^* = \arg \max\limits_{\mathrm{P_1}=\{0,\mathrm{P},\mathrm{P}'\}} \mathcal{J}(\mathrm{P_1}) \\ P_2^* = P - P_1^* \end{cases} \tag{3.24}$$

where $P_1 \in [0, P]$ *and*

$$P' = \begin{cases} (-B \pm \sqrt{B^2 - 4A\Gamma})/2A & \text{if } A \neq 0 \\ -\Gamma/B & \text{if } A = 0 \end{cases} \tag{3.25a}$$

$$A = \eta_{k_1 1}\eta_{k_2 1}(\eta_{k_2 1} - \eta_{k_2 2})(P\eta_{k_1 2} + \sigma^2) + \eta_{k_2 2}\eta_{k_1 2}(\eta_{k_1 1} - \eta_{k_1 2})(P\eta_{k_2 1} + \sigma^2)$$

$$\begin{aligned} B &= (P\eta_{k_1 2} + \sigma^2)\eta_{k_1 1}(P\eta_{k_2 1}\eta_{k_2 2} + 2\eta_{k_2 1}\sigma^2 - \eta_{k_2 2}\sigma^2) \\ &+ \eta_{k_2 2}(2\eta_{k_1 2} - \eta_{k_1 1})(P\eta_{k_2 1} + \sigma^2)(P\eta_{k_1 2} + \sigma^2) \end{aligned}$$

$$\Gamma = \eta_{k_1 1}\sigma^2(P\eta_{k_1 2} + \sigma^2)(P\eta_{k_2 2} + \sigma^2) - \eta_{k_2 2}(P\eta_{k_2 1} + \sigma^2)(P\eta_{k_1 2} + \sigma^2)^2 \tag{3.25b}$$

Proof. The proof is given in Appendix 3.F. □

Hence, the optimal power control is either TDMA-mode (only one user/beam is allocated non-zero power) or SDMA-mode in which the transmit power values to multiple users are positive and allocated according to (3.25a).

Beam power control in extreme interference cases

To gain more intuition on the optimal power allocation scheme, we investigate two extreme cases in terms of interference. Define the interference factors $\alpha_{k_m} = \frac{\sum_{j \neq m} \eta_{k_m j}}{\eta_{k_m m}}$. In the 2-beam case, we have $\alpha_{k_1} = \frac{\eta_{k_1 2}}{\eta_{k_1 1}}$ and $\alpha_{k_2} = \frac{\eta_{k_2 1}}{\eta_{k_2 2}}$. For non-interfering beams (i.e., $\alpha_{k_1} = \alpha_{k_2} = 0$), the optimal beam power allocation is given by the water-filling power allocation

$$P_1^* = \min\left(P, \left[\frac{P}{2} + \frac{(\eta_{k_1 1} - \eta_{k_2 2})\sigma^2}{2\eta_{k_1 1}\eta_{k_2 2}}\right]^+\right) \quad \text{and} \quad P_2^* = P - P_1^* \tag{3.26}$$

where $[x]^+ = \max(0, x)$. Note that SDMA with equal power allocation is optimal when both users experience the same channel conditions ($\eta_{k_1 1} = \eta_{k_2 2}$).
In the case of fully-interfering beams (i.e., $\alpha_{k_1} = \alpha_{k_2} = 1$), TDMA mode is of course optimal as the solution to (3.20) under the assumption wlog $\eta_{k_1 1} > \eta_{k_2 2}$ is

$$P_1^* = P \quad \text{and} \quad P_2^* = 0 \tag{3.27}$$

Optimality conditions for TDMA transmission mode

The beam power solution stated in Theorem 3.1 implies that the optimum transmission mode is either TDMA ($P_1 = 0$ or P) or SDMA with $P_1 = P'$. It is therefore interesting to identify the region of TDMA optimality and provided the relevant conditions. We first derive conditions requiring knowledge of the interference factors $\alpha_{k_i} \in (0, 1]$ only. These conditions can be used as practical design rules, especially in distributed resource allocation

scenarios. Formally, we have that

Lemma 3.4: *If $\alpha_{k_i} \geq 0.5$, the optimum power allocation is $P_1^* = P$ and $P_2^* = 0$ (TDMA transmission mode).*

Proof. The proof is given in Appendix 3.G. □

Corollary 3.3: *A sufficient condition for TDMA optimality is*

$$\alpha_{k_1} + \alpha_{k_2} \geq 1 \text{ or equivalently } \left(\frac{1}{\cos^2\theta_1}\right)^2 + \left(\frac{1}{\cos^2\theta_2}\right)^2 \geq 3 \tag{3.28}$$

where $\theta_i = \angle\left(\bar{\mathbf{h}}_{k_i}, \mathbf{q}_i\right)$ is the angle (misalignment) between the direction of the normalized channel $\bar{\mathbf{h}}_{k_i} = \mathbf{h}_{k_i} / \|\mathbf{h}_{k_i}\|$ and beam $\mathbf{q}_i$.

Proof. The first condition is a trivial result of Lemma 3.4 by summing up the interference factors and the equivalent second relation is derived by using $\alpha_{k_i} = \tan^2\theta_i$. □

Additionally, if BGI knowledge is allowed (strategy 3), a (sharper) sufficient TDMA optimality condition is the following:

Lemma 3.5: *The optimum power allocation is TDMA mode ($P_1^* = P$) if*

$$\frac{P\eta_{k_1 1}}{\sigma^2} \geq \frac{1 - \alpha_{k_1} - \alpha_{k_2}}{\alpha_{k_1}\alpha_{k_2}} \tag{3.29}$$

Proof. The proof is given in Appendix 3.H. □

3.7.2 Beam Power Allocation for more than two beams

For the general case of $\mathcal{B} > 2$ beams, an analytical treatment of (3.20) does not unfortunately seem tractable, because of the lack of convexity. Therefore, we propose here a suboptimal - yet efficient - iterative algorithm that aims to increase system throughput by allocating power over the beams. The algorithm tries to identify the extreme points of the sum rate and find the power vector $\mathbf{P}$ that maximizes (3.20). The extremum of the sum rate function can be found analytically using Lagrangian duality theory and considering the Karush-Kuhn-Tucker (KKT) conditions. Let WloG $\mathcal{B} = M$ and define the objective function

$$\mathcal{G}(\mathbf{P}) = \sum_{m=1}^{M} \log_2\left(1 + \frac{P_m\eta_{k_m m}}{\sigma^2 + \sum_{j\neq m} P_j\eta_{k_j j}}\right) \tag{3.30}$$

In order to solve the optimization problem

$$\max_{\mathbf{P}\in\mathcal{P}^M} \mathcal{G}(\mathbf{P}), \quad \text{subject to } \mathbf{P} \geq \mathbf{0}, \sum_{m=1}^{M} P_m = P \tag{3.31}$$

we may formulate the Lagrangian function as

$$\mathcal{L}(\mathbf{p}, \mu, \nu) = \mathcal{G}(\mathbf{p}) + \sum_{m=1}^{M} \nu_m P_m - \mu\left(\sum_{m=1}^{M} P_m - P\right) \tag{3.32}$$

where $\nu \geq \mathbf{0}$ and $\mu \geq 0$ are dual variables. The cost function is neither convex not concave with respect to $\{P_m\}_{m=1}^M$, therefore a global optimal solution for any channel model is hard to obtain. However, KKT conditions are necessary for extremum, whether local or global, of $\mathcal{G}(\mathbf{P})$. By differentiating with respect to P_m, we find

$$\begin{aligned} \frac{\partial \mathcal{G}(\mathbf{P})}{\partial P_m} + \nu_m - \mu = 0,\ 1 \leq m \leq M \\ P_m \geq 0,\ 1 \leq m \leq M \\ P - \sum_m P_m \geq 0 \end{aligned} \tag{3.33}$$

The KKT conditions are necessary and sufficient if and only if the Hessian of (3.32) is a negative definite matrix. For such class of channels, a global maximum is identified through the KKT conditions above. For general channels, the KKT points can be a global or local maximum, a saddle-point, or even a global or local minimum.

Iterative Beam Power Control Algorithm

Performing transformation of the primal problem (3.20) into its dual and solving the latter by KKT conditions does not guarantee global optimal primal solution. As the primal is not a convex optimization problem, there could be a duality gap. Nevertheless, we propose an iterative algorithm, inspired by the iterative water-filling (IWF) algorithm [73] and the KKT solution of (3.31), as a means to identify the extreme points on the boundary of $\mathcal{P}^M$. In this Iterative Beam Power Control Algorithm, each user iteratively maximizes its own rate by performing single-user water-filling and treating the multiuser interference from all the other users (beams) as noise. Clearly, our algorithm does not seek to find a global optimum, however it can provide significant sum-rate improvement.

Algorithm I Let $\mathbf{P}^{(0)}$ be the initial point and $\mathcal{I}(\mathbf{P}^{(i)}) = \sigma^2 + \sum_{j \neq m} P_j^{(i)} \eta_{k_j j}$ be the interference function at i-th iteration. The steps of the algorithm are summarized in Table 3.1.

Iterative Beam Power Control Algorithm
Step 1 (*Initialization*) Set $\mathbf{P}^{(0)} = \mathbf{0}$
Step 2 For iteration $i = 1, 2, \ldots$, compute $\forall k_m \in \mathcal{S}$: $\lambda_{k_m}^{(i)} = \frac{\eta_{k_m m}}{\mathcal{I}(\mathbf{P}^{(i-1)})} = \frac{\eta_{k_m m}}{\sigma^2 + \sum_{j \neq m} P_j^{(i-1)} \eta_{k_j j}}$
Step 3 (*Water-filling*): let $\pi^{(i)}$ be the solution of: $\pi^{(i)} = \arg \max_{\pi \geq 0, \sum_m \pi_m \leq P} \sum_{k_m \in \mathcal{S}} \log_2 \left(1 + \pi_m \lambda_{k_m}^{(i)}\right)$
Step 4 (*Update*): let $\mathbf{P}^{(i)} = \pi^{(i)}$

Table 3.1: Iterative Beam Power Control Algorithm for Sum-Rate Maximization

Some observations are in order:
At each iteration i, once $\lambda_{k_m}^{(i)} = \frac{\eta_{k_m m}}{(\sigma^2 + \sum_{j \neq m} P_j^{(i-1)} \eta_{k_m j})}$ is calculated for each user k_m using $P_j^{(i-1)}, j \neq m$, it is kept fixed and treated as noise. Given the total power constraint P, the

'water-filling step' is a convex optimization problem similar to multiuser water-filling with common water-filling level. Thus, all transmit powers in $\mathbf{P}$ assigned to beams are calculated simultaneously in order to maintain a constant water-filling level. The algorithm computes iteratively the beam power allocation that leads to sum rate increase and converges to a limit value greater or equal to the sum rate of equal power allocation. Formally, the power assigned to beam m at iteration i yields $P_m^{(i)} = [\mu - 1/\lambda_{k_m}^{(i)}]^+$, with $\sum_{k_m \in \mathcal{S}} [\mu - 1/\lambda_{k_m}^{(i)}]^+ = P$, where μ is the common water-filling level. The beam power control for strategy 3 assigns transmit powers over the beams according to the iterative solution when the achieved sum rate is higher than that of the boundary points.

Convergence Issues As stated before, no global maximum is guaranteed due to the lack of convexity of sum-rate maximization problem. Therefore, we do not expect that the convergence point of the iterative algorithm be generally a global optimal power solution. Interestingly, it can be shown that the convergence leads to a Nash equilibrium, when considering that each user participates in a non-cooperative game. The convergence to an equilibrium point can be guaranteed since $\mathcal{I}(\mathbf{P})$ is a standard interference function [73, 74]. The proof of existence of Nash equilibrium follows from an easy adaptation of the proof in [75]. However, the uniqueness of these equilibrium points cannot be easily derived for the case of arbitrary channels.

Let us now derive analytically the convergence point of the 2-beam case using the iterative algorithm and compare it with the optimal beam power solution given by Theorem 3.2. At the steady state, say iteration s, we have that

$$\begin{cases} P_1^{(s)} = \mu - 1/\lambda_{k_1}^{(s)} \\ P_2^{(s)} = \mu - 1/\lambda_{k_2}^{(s)} \end{cases} \quad \text{with} \quad \begin{cases} \lambda_{k_1}^{(s)} = \frac{\eta_{k_1 1}}{P_2^{(s-1)}\eta_{k_1 2}+\sigma^2} \\ \lambda_{k_2}^{(s)} = \frac{\eta_{k_2 2}}{P_1^{(s-1)}\eta_{k_2 1}+\sigma^2} \end{cases} \tag{3.34}$$

and $\mu = \frac{P}{2} + \frac{1}{2\lambda_{k_1}} + \frac{1}{2\lambda_{k_2}}$ from the sum power constraint. Upon convergence of the algorithm, we have that $P_i^{(s)} = P_i^{(s-1)}, i = 1, 2$, which results into a system of equations $\mathbf{A}\mathbf{P}^T = \mathbf{b}$ with

$$\mathbf{A} = \begin{bmatrix} 2 - \eta_{k_2 1}/\eta_{k_2 2} & \eta_{k_1 2}/\eta_{k_1 1} \\ \eta_{k_2 1}/\eta_{k_2 2} & 2 - \eta_{k_1 2}/\eta_{k_1 1} \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} P + \sigma^2 \left(\frac{1}{\eta_{k_2 2}} - \frac{1}{\eta_{k_1 1}} \right) \\ P + \sigma^2 \left(\frac{1}{\eta_{k_1 1}} - \frac{1}{\eta_{k_2 2}} \right) \end{bmatrix}$$

For $\det(\mathbf{A}) \neq \mathbf{0} \rightarrow \alpha_{k_1} \neq 1$ and $\alpha_{k_2} \neq 1$, we have that $\mathbf{P}^T = \mathbf{A}^{-1}\mathbf{b}$, giving the following 'water-filling' solution

$$P_1 = \frac{P\eta_{k_2 2}\left(\eta_{k_1 1} - \eta_{k_1 2}\right) + \sigma^2\left(\eta_{k_1 1} - \eta_{k_2 2}\right)}{2\eta_{k_1 1}\eta_{k_2 2} - \eta_{k_2 2}\eta_{k_1 2} - \eta_{k_2 1}\eta_{k_1 1}} \quad \text{and} \quad P_2 = P - P_1 \tag{3.35}$$

It can be observed that (3.35) is different from (3.25a). Fortunately, it still provides a heuristic power allocation algorithm and as shown through simulations in Section 3.9, there is not a significant reduction in sum rate by allocating the power over beams using this algorithm.

Reinterpretation in terms of Successive Convex Approximation

In this section, we resort to Geometric Programming (GP) [72] which represents the state of the art in continuous power control for non-convex problems. The GP approach has

become a very popular and powerful technique as it provides efficient solutions in power control problems with non-linear objective functions and specific SINR constraint, by revealing the hidden convexity structure. Furthermore, the proposed solutions are very fast and numerically efficient, often exhibiting polynomial time complexity. In particular, we capitalize on the so-called *successive convex approximation* (SCA) technique [72, 76], which is shown to be convergent and turns out that it often computes the globally optimal power allocation. Interestingly, the heuristic iterative algorithm proposed in Table 3.1 finds an equivalent interpretation, since applying SCA to our beam power control problem results in the same iterative algorithm. We first lower bound $\log(1+\text{SINR})$ in the objective function for some a and b [76]:

$$\log(1+\text{SINR}) \geq a\log(\text{SINR}) + b \tag{3.36}$$

Applying (3.36) into the optimization problem (3.20) results in the relaxation

$$\max_{\mathbf{P}} \frac{1}{\log 2}\sum_{m=1}^{M}\left(a_m \log\left(\text{SINR}_{k_m m}(\mathbf{P})\right) + b_m\right) \quad \text{subject to} \quad \sum_{m=1}^{M} P_m \leq P \tag{3.37}$$

which still remains a non-convex problem since the objective function is not concave in $\mathbf{P}$. However, using the transformation $\tilde{P_m} = \log(P_m)$ we have the following concave maximization problem:

$$\max_{\tilde{\mathbf{P}}} \sum_{m=1}^{M}\left(a_m \log\left(\text{SINR}_{k_m m}(e^{\tilde{\mathbf{P}}})\right) + b_m\right) \quad \text{subject to} \quad \sum_{m=1}^{M} e^{\tilde{P}_m} \leq P$$

Defining the Lagrangian function as

$$\mathcal{D}(\tilde{\mathbf{P}}, \lambda) = \sum_{m=1}^{M}\left(a_m \log\left(\text{SINR}_{k_m m}(e^{\tilde{\mathbf{P}}})\right) + b_m\right) - \lambda\sum_{m=1}^{M}\left(e^{\tilde{P}_m} - P\right) \tag{3.38}$$

we consider the dual problem (3.38) that is $\min_{\lambda}\max_{\tilde{\mathbf{P}}}\mathcal{D}(\tilde{\mathbf{P}}, \lambda)$. The dual solution of the inner maximization problem is given by the stationary point of the Lagrangian function (3.38) with λ fixed. Differentiating wrt $\tilde{P}_m$ and applying the inverse transformation $P_m = e^{\tilde{P}_m}$ we form the following fixed-point equation

$$\frac{\partial\mathcal{D}}{\partial\tilde{P}_m} = 0 \Rightarrow P_m = \frac{a_m}{\lambda + a_m\frac{\eta_{k_m m}}{\sum_{j\neq m}P_j\eta_{k_m j}+\sigma^2}} \tag{3.39}$$

Remarkably, this fixed point-equation provides the same power allocation algorithm as in Table 3.1 for $a_m = 1, \forall m$ (wlog) where the powers can be updated iteratively using (3.39). However, we note that a zero duality gap cannot be guaranteed formally due to lack of convexity, implying that no theoretical argument can show convergence to the global optimum for general class of channels.

3.7.3 Beam Power Control in Specific Regimes ($\mathcal{B} \geq 2$)

The apparent non-convexity of the $\mathcal{B}$-beam case can be alleviated in certain SINR/interference regimes, as a hidden convexity of the beam power allocation problem appears. We shall consider the beam power allocation for $\mathcal{B} = M$ beams in four cases: 1) the high SINR regime, 2) the low SINR regime, 3) approximation by the arithmetic-geometric means inequality, and 4) the symmetric interference regime.

High SINR regime

In the high SINR regime, which corresponds to SINR values higher than 0 dB, the approximation $\log(1+x) \approx \log(x)$ can be applied. In that case, the objective function $\mathcal{G}(\mathbf{P})$ becomes

$$\begin{aligned}\mathcal{G}(\mathbf{P}) &\approx \frac{1}{\log 2}\sum_{m=1}^{M}\log\left(\mathrm{SINR}_{k_m m}\right) = \log_2\left(\prod_{m=1}^{M}\mathrm{SINR}_{k_m m}\right) \\ &= \log_2\left(\prod_{m=1}^{M}\frac{P_m\eta_{k_m m}}{\sum_{j\neq m}P_j\eta_{k_m j}+\sigma^2}\right)\end{aligned} \tag{3.40}$$

A similar result has previously observed in [72] in the case of code division multiple access (CDMA) power control. The optimum power allocation solution can be found using Geometric Programming, as the approximate high-SINR sum rate is a concave function of $\log P_m$.

Low SINR regime

In the low SINR regime, the sum rate is approximated by applying Taylor first-order series expansion, i.e. $\log(1+x) \approx x$. In that case, the objective function becomes

$$\begin{aligned}\mathcal{G}(\mathbf{P}) &= \sum_{m=1}^{M}\log_2\left(1+\mathrm{SINR}_{k_m m}\right) \approx \log_2 e\sum_{m=1}^{M}\mathrm{SINR}_{k_m m} \\ &= \log_2 e\sum_{m=1}^{M}\frac{P_m\eta_{k_m m}}{\sum_{j\neq m}P_j\eta_{k_m j}+\sigma^2}\end{aligned} \tag{3.41}$$

The objective function (3.41) is convex in each variable P_m since

$$\frac{\partial^2}{\partial P_m^2}\left(\sum_{m=1}^{M}\frac{P_m\eta_{k_m m}}{\sum_{j\neq m}P_j\eta_{k_m j}+\sigma^2}\right) = \sum_{m\neq i}\frac{2P_m\eta_{k_m m}\eta_{k_m i}^2}{\left(\sum_{j\neq m}P_j\eta_{k_m j}+\sigma^2\right)^3} \geq 0 \tag{3.42}$$

Therefore, the optimal beam power control strategy is found by the KKT conditions and can be solved numerically using efficient interior-point methods [20].

Arithmetic-geometric means approximation

From the arithmetic-geometric means inequality [77], the sum rate can be upper bounded as

$$\begin{aligned}\mathcal{R}(\mathbf{P}) &= \log_2\left(\prod_{m=1}^{M}\left(1+\mathrm{SINR}_{k_m,m}(\mathbf{P})\right)\right) \\ &\leq M\log_2\left(1+\frac{1}{M}\sum_{m=1}^{M}\frac{P_m\eta_{k_m m}}{\sigma^2+\sum_{j\neq m}P_j\eta_{k_m j}}\right) = \mathcal{G}_{AGM}(\mathbf{P})\end{aligned} \tag{3.43}$$

where the inequality is sharp for $\mathrm{SINR}_{k_i i} = \mathrm{SINR}_{k_j j}, \forall i,j \in \mathcal{S}$. Since the logarithm is a monotonically increasing function and the argument of the log-function of $\mathcal{G}_{AGM}(\mathbf{P})$ is

convex wrt each P_m (similarly to the low SINR regime), a closed-form global optimal solution can be derived. The sharpness of the above sum-rate approximation is quantified by the difference $\delta = \mathcal{G}_{AGM}(\mathbf{P}) - \mathcal{R}(\mathbf{P})$. For $h = \frac{\max_i(1+\text{SINR}_{k_i i})}{\min_j(1+\text{SINR}_{k_j j})} > 1$ the following inequality stands

$$0 \leq \delta \leq \frac{M}{\log_2} K^{'}(h,1) \tag{3.44}$$

where $K^{'}(h,1) = \log\left(\frac{h^{\frac{1}{h-1}}}{e\log h^{\frac{1}{h-1}}}\right)$ is the first derivative of the Kantorovich constant [78]. The upper bound is tight for equal SINR values, and the approximation is better when the spread of (1+SINR) values is small ($h \to 1$).

Symmetric interference regime

We restrict here ourselves to the case of symmetric interference networks, in which all users have the same interfering beam gains. This scenario corresponds to the case where the selected users are situated at about the same distance from the interfering beams. Hence, for $\eta_{k_m j} = \eta_{k_m i}, \forall i, j \neq m$, the achievable sum rate is given by

$$\mathcal{R}(\mathcal{S},\mathbf{P}) = \sum_{m=1}^{M} \log_2\left(1 + \frac{P_m}{\sigma^2/\eta_{k_m m} + \alpha_{k_m}\sum_{j\neq m} P_j}\right) \tag{3.45}$$

with $\alpha_{k_m} = \frac{\eta_{k_m j}}{\eta_{k_m m}}, j \neq m$. Since the objective function (3.45) is concave in P_m and the feasible region is convex, the KKT conditions imply that there exist a unique Nash equilibrium that can be achieved using iterative water-filling.

3.8 Beam Power Control with SINR feedback

Suppose now that we have a harder rate constraint for the second-stage feedback. Specifically, we adopt *strategy 4* in the second stage, assuming thus that the scheduler has access only to the same amount of feedback information as in [9], namely $\gamma_k^{'} = \gamma_k = \text{SINR}_{k_m m}$ (1 scalar). Nevertheless, we further exploit this scalar information in view of rendering the precoding matrix more robust with respect to cases where not all M users can be served satisfactorily simultaneously with the same amount of power. This can be viewed as a low-complexity, low-feedback variant of the two-stage linear beamforming framework. The major challenge here is that when only SINR feedback is available, the transmitter does not have access to BGI and thus it cannot estimate the precise received SINR and inter-user interference if the transmit beam powers had been allocated differently. Therefore, it cannot explicitly maximize the instantaneous sum rate by allocating the power unequally over the beams. We then resort to a power control strategy based on the maximization of the expected sum rate.

On/Off Beam Power Control

We propose a simple power allocation scheme, coined as On/Off Beam Power Control, in which the transmitter takes a binary decision between:

- TDMA mode toward one selected user (the one with maximum γ_k from stage 1).

- SDMA where all random equipowered beams are active, as in [9].

The scheduler, based only on SINR feedback, compares the instantaneous achievable SDMA sum rate with the expected TDMA rate, and selects the transmission mode that maximizes the system throughput.

Let $\mathcal{R}_{\text{SDMA}} = \sum_{m=1}^{M} \log(1 + \text{SINR}_{k_m m})$ denote the achievable SDMA sum rate that can be explicitly calculated at the BS, and $\mathcal{R}_{\text{TDMA}}$ denote the expected TDMA transmission rate. The expected TDMA rate can be efficiently calculated by considering the statistics of the BGI of the user, say k_1, with maximum γ_{k_1} conditioned to the feedback information γ_{k_1}. Formally, the distribution function of $s = \frac{P}{\sigma^2}\eta_{k_1 1}$ (BGI of the highest SINR user) is given by

$$F_s(x) = \Pr\{s \leq |\gamma_{k_1}\} = \frac{F_Y(\frac{\sigma^2}{P}(x/\gamma_{k_1} - M))}{F_{\text{SINR}}(\gamma_{k_1})} \tag{3.46}$$

where $F_Y(x)$ is the CDF of the interference $Y = \sum_{j \neq 1} \eta_{k_1 j}$ and $F_{\text{SINR}}(x) = 1 - \frac{e^{-x\sigma^2/\rho}}{(1+x)^{M-1}}$.
The On/Off Beam power control scheme results in the following binary mode decision denoted as $\mathcal{F}$:

$$\mathcal{F} = \begin{cases} \text{TDMA} & \text{if } \Delta\mathcal{R} > 0 \\ \text{SDMA} & \text{if } \Delta\mathcal{R} \leq 0 \end{cases} \tag{3.47}$$

where $\Delta\mathcal{R} = \mathcal{R}_{\text{TDMA}} - \mathcal{R}_{\text{SDMA}}$.
For the expected TDMA rate $\mathcal{R}_{\text{TDMA}} = \mathbb{E}\left\{\log_2\left(1 + \frac{P}{\sigma^2}\eta_{k_1 1}\right)\right\}$, where $F_{\eta_{k_1 1}}(x) = (1 - e^{-x})^K$, the following closed-form expression can be derived:

Proposition 3.3: *For any values of P, M, and K, the average rate of TDMA-based random beamforming is given by*

$$\mathcal{R}_{\text{TDMA}} = \frac{1}{\log 2}\sum_{k=1}^{K}\binom{K}{k}(-1)^k e^{k\sigma^2/P}\text{Ei}(-\text{k}\sigma^2/\text{P}) \tag{3.48}$$

where $\text{Ei(x)} = -\int_{-\text{x}}^{\infty}\frac{\text{e}^{-\text{t}}}{\text{t}}\text{dt}$ *is the exponential integral.*

Proof. The proof is given in Appendix 3.I. □

3.9 Performance Evaluation

We evaluate the sum-rate performance of the proposed beam power control algorithms through Monte Carlo simulations assuming i.i.d. flat fading Rayleigh channels across users and transmit antennas. We also consider that $\mathcal{B} = M$ beams are generated. The achieved sum rate is compared with conventional SDMA-based random beamforming [9] where equal power is allocated over the beams.

We first assess the performance of enhanced RBF with perfect second-stage CSIT feedback (strategy 1). In Figure 3.6 we compare the sum rate performance of the two-approach in which the second-stage precoding is calculated based on full CSIT. As expected, the MMSE precoder applied to a set of quasi-orthogonal users outperforms significantly the single-stage random beamforming. The performance gain of 1.7 bps/Hz of MMSE beamformer can be further increased if optimal power allocation is used.

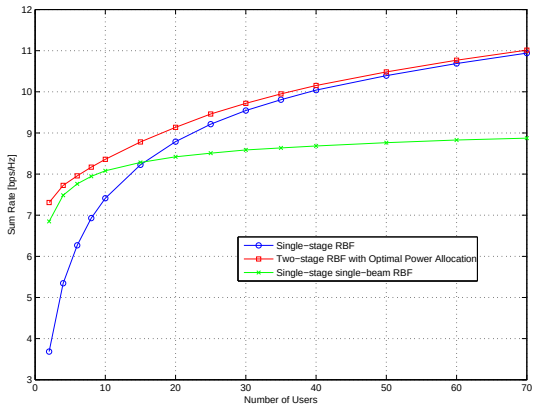

Figure 3.4: Sum rate versus the number of users for Optimal Beam Power Control with $M = 2$ transmit antennas and SNR = 20 dB.

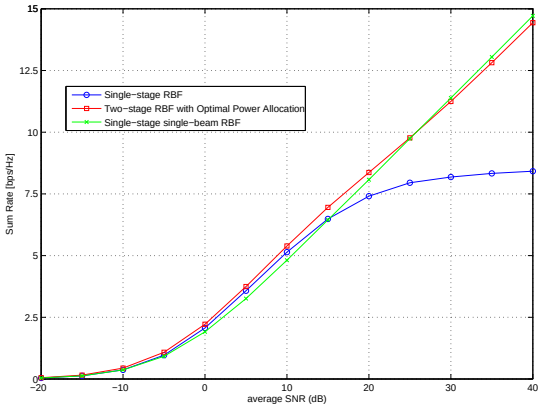

Figure 3.5: Sum rate versus average SNR for Optimal Beam Power Control (strategy 3) with $M = 2$ transmit antennas and $K = 10$ users.

We then assess the performance of beam power control with BGI second-stage feedback (strategy 3). In Figure 3.4 we present the sum rate achieved using optimal power allocation versus the number of active users K for the 2-beam case and SNR = 20 dB. Single-beam random beamforming refers to the scheme proposed in [53] where only one random beam is generated (TDMA) at each slot. The gains of optimally allocating power across beams are more pronounced for systems with low to moderate number of users (up to 30), whereas for K increasing, the benefits of beam power control vanishes as the optimal solution advocates expectedly the use of equipowered beams. Figure 3.5 shows a sum-rate comparison as a function of the average SNR for $K = 10$ users, illustrating that beam power allocation prevents the system from becoming interference-limited. Power control allows us to switch off beams, thus keeping a linear capacity growth in the interference-limited regime at high SNR by converging to TDMA.

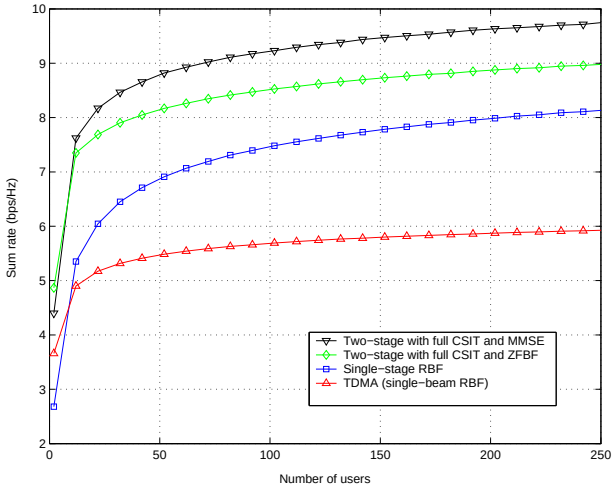

Figure 3.6: Sum rate comparison of different second-stage precoders (strategy 1) versus the number of users for $M = 2$ and SNR $= 10$ dB.

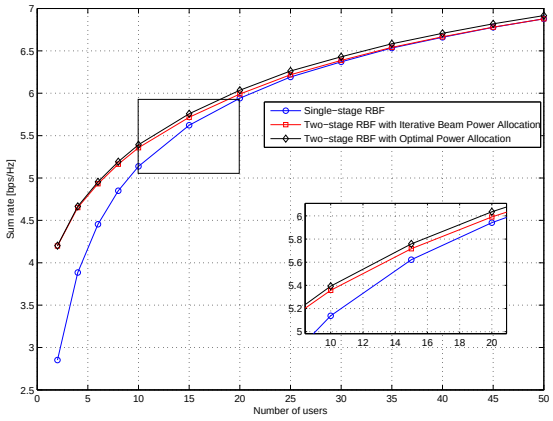

Figure 3.7: Sum rate versus the number of users for Iterative Beam Power Allocation and Optimal Power Control with $M = 2$ transmit antennas and SNR $= 10$ dB.

In Figure 3.7 we compare the achieved sum rate difference between the optimal power allocation and the power solution given by our iterative algorithm at SNR $= 10$ dB. Use of the iterative algorithm, despite suboptimal, results in negligible throughput loss at all ranges of K. The performance of the iterative power control is further evaluated in Figure 3.8 for a 4-beam downlink showing substantial sum-rate enhancements for practical number of users.

We then evaluate the results of the on/off beam power control (strategy 4), which uses the same amount of feedback as the conventional RBF [9]. In Figure 3.9 we plot the sum rate versus the number of users for $M = 2$ transmit antennas and SNR $= 10$ dB. The scheme is switching from TDMA mode at low K (all transmit power is given to the highest SINR_{k_m} user) to SDMA-based RBF with equal power allocation. We also observe that the sum-rate gap between the optimal power control (with second-stage feedback)

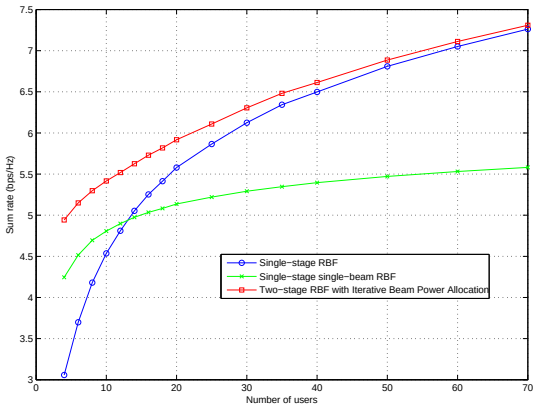

Figure 3.8: Sum rate versus the number of users for Iterative Beam Power Allocation with $M = 4$ transmit antennas and SNR = 10 dB.

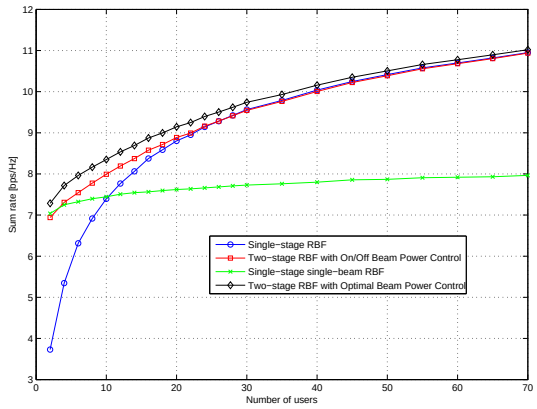

Figure 3.9: Sum rate versus the number of users for On/Off Beam Power Control with $M = 2$ transmit antennas and SNR = 20 dB.

and on/off power control (no additional feedback) for $K < 20$ users is approximately 0.4 bps/Hz. In Figures 3.10 and 3.11 we consider a 4-beam RBF scheme and show the sum rate performance of on/off beam power control as a function of average SNR and the number of users, respectively. Although the throughput curve of conventional RBF converges to a finite ceiling at high SNR, the TDMA-SDMA binary decision capability of the beam on/off scheme provides a simple means to circumvent the interference-limited behavior of RBF with no extra feedback. We note also that TDMA mode is generally preferable from a sum-rate point of view in sparse networks, and the range of K in which TDMA is beneficial increases for SNR increasing.

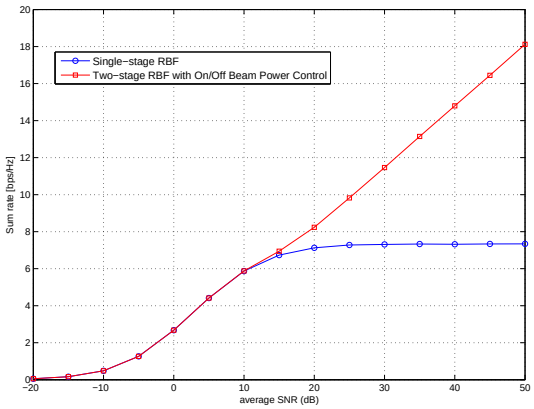

Figure 3.10: Sum rate versus average SNR for On/Off Beam Power Control with $M = 4$ transmit antennas and $K = 25$ users.

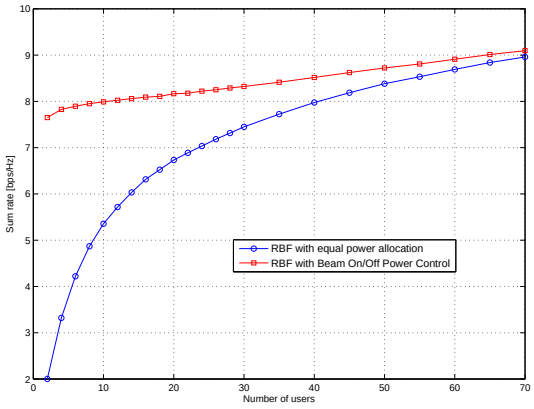

Figure 3.11: Sum rate versus the number of users for On/Off Beam Power Control with $M = 4$ transmit antennas and SNR $= 20$ dB.

3.10 Conclusion

This chapter focused on SDMA-based random beamforming techniques. We first studied conventional random beamforming and provided an exact characterization of the expected sum-rate, as well as of the capacity growth in the interference-limited region. Main outcome of this analysis is that the number of beams that should be allocated non-zero power has to be adapted depending on the system average SNR and the number of active users K in the cell.

Then, we introduced a two-stage scheduling and linear precoding framework, which divides the scheduling and the precoding design stages into two steps. Based on this decoupled approach, we proposed a scheme coined as enhanced random beamforming. In the scheduling phase, RBF is exploited to identify good, spatially separable performing low-rate feedback user selection. In the second stage, additional finite rate CSIT may be requested to only the pre-selected users in order to refine the final precoder. Several beam power con-

trol strategies, with various levels of complexity and feedback load, are proposed in order to restore robustness of RBF in sparse networks. Their sum-rate performance is assessed, revealing substantial gains compared to RBF for systems with low to moderate number of users, at a moderate or zero cost of extra feedback.

Throughout this chapter, the users' channels were considered temporally and spatially uncorrelated. In the following chapter, we investigate how information and redundancy hidden on the channel structure can be exploited by the scheduler in temporally and spatially-correlated channel.

APPENDIX

3.A Proof of Lemma 3.1

From Lemma 3 in [9], we have that for any values of P, M, and K, the average sum rate of multi-beam RBF satisfies

$$I_1 \leq \mathcal{R}_{\text{RBF}} \leq I_2 \tag{3.49a}$$

with

$$I_1 = M \int_1^\infty \log_2(1+x) dF^K(x) \qquad I_2 = M \int_0^\infty \log_2(1+x) dF^K(x) \tag{3.49b}$$

We first evaluate the upper bound as follows

$$\begin{aligned} I_2 &= M \int_0^\infty \log_2(1+x) d(F^K(x)-1) \overset{(a)}{=} \frac{M}{\log 2} \int_0^\infty \frac{1-F^K(x)}{1+x} dx \\ &= \frac{M}{\log 2} \int_0^\infty \frac{1-\left(1-\frac{e^{-x/\rho}}{(1+x)^{M-1}}\right)^K}{1+x} dx \\ &\overset{(b)}{=} \frac{M}{\log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^{k+1} \int_0^\infty \frac{e^{-xk/\rho}}{(1+x)^{(M-1)k}} dx \end{aligned} \tag{3.50}$$

where (a) is obtained by using the integration by parts and (b) follows from binomial expansion. The closed-form expressions of the integral in (3.50), which then gives (3.1b), can be obtained by the following formula (Schlömilch function) [79]:

$$S(\nu, z) = \int_0^\infty (1+t)^{-\nu} e^{-zt} dt = z^{\nu-1} e^z \Gamma(1-\nu, z) = z^{-\nu/2-1} e^{z/2} \mathcal{W}_{-\nu/2,(1-\nu)/2}(z) \tag{3.51}$$

where $\mathcal{W}_{k,m}(z)$ is the Whittaker function and $\Gamma(a,x)$ the upper incomplete gamma function. To obtain a lower bound, we use the fact that $I_1 = I_2 - M\int_0^1 \log_2(1+x) dF^K(x) = \mathcal{A} - \alpha$, which results in (3.1c) using similar steps as for I_2.

3.B Proof of Lemma 3.2

Starting from (3.50), we have

$$\begin{aligned} \mathcal{R}_{\text{RBF}} &\approx \frac{M}{\log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^{k+1} \int_0^\infty \frac{e^{-xk/\rho}}{(1+x)^{(M-1)k}} dx \\ &\overset{(a)}{\leq} \frac{M}{\log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^{k+1} \int_0^\infty \frac{(1+x)^{-k/\rho}}{(1+x)^{(M-1)k}} dx \\ &= \frac{M}{\log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^{k+1} \frac{\rho}{k((M-1)\rho+1)} = \frac{M}{\log 2} \frac{\rho H_K}{(M-1)\rho+1} \end{aligned}$$

where (a) follows from $(1+x)^r \leq e^{rx}$ for any real $x, r > 0$.

3.C Proof of Lemma 3.3

When $P \to 0$, the approximation $\mathrm{SINR}_{k,m} \approx \rho \left|\mathbf{h}_k \mathbf{q}_m\right|^2$ with CDF $F_l(x) = 1 - e^{-x/\rho}$. The average sum rate is given as follows

$$\begin{aligned} \mathcal{R}_{low} &= M \int_0^\infty \log_2(1+x) dF_l^K(x) = \frac{MK}{\rho \log 2} \int_0^\infty \log_2(1+x) e^{-\frac{x}{\rho}} (1 - e^{-\frac{x}{\rho}})^{K-1} dx \\ &\overset{(a)}{=} \frac{MK}{\rho \log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^k \int_0^\infty \log_2(1+x) e^{-\frac{x(k+1)}{\rho}} dx \\ &= \frac{MK}{\log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^{k+1} \frac{e^{-\frac{k+1}{\rho}} \mathrm{Ei}(-\frac{\mathrm{k+1}}{\rho})}{k+1} \end{aligned} \tag{3.52}$$

where (a) follows from binomial expansion.

3.D Proof of Corollary 3.2

Expanding the logarithm in (3.52) to second-order Taylor series, i.e., $\log(1+x) \approx x - x^2/2$ we have

$$\begin{aligned} \mathcal{R}_{low} &= \frac{MK}{\rho \log 2} \sum_{k=0}^{K} \binom{K}{k} (-1)^k \int_0^\infty \left(x - \frac{x^2}{2} \right) e^{-x(k+1)/\rho} dx \\ &= \frac{\rho}{\log 2} H_K \left(1 - \frac{\rho}{2}(1 + H_K) \right) \overset{(a)}{\leq} \frac{\rho}{\log 2} H_K \end{aligned}$$

where (a) is obtained by neglecting the second-order term.

3.E Proof of Theorem 3.1

We use Dziubdziela's Theorem [80] with $a_K = K^{1/(M-1)}$ and $b_K = -1$. We first evaluate the following functions:

$$\tilde{\delta}_K(x) = 1 - F(a_K x + b_K) = \frac{1}{K x^{M-1}} \tag{3.53}$$

and

$$g(j, K\tilde{\delta}_K(x)) = \begin{cases} e^{-K\tilde{\delta}_K(x)} & j = 1 \\ e^{-K\tilde{\delta}_K(x)} \left(\frac{\left[K\tilde{\delta}_K(x)\right]^{j-1}}{(j-1)!} - \frac{\left[K\tilde{\delta}_K(x)\right]^{j-2}}{(j-2)!} \right) & j \geq 2 \end{cases} \tag{3.54}$$

and

$$\Theta(x) = \left| \frac{1}{(j-1)!} \int_{K\tilde{\delta}_K(x)}^{-\log G_1(x)} \omega^{j-1} e^{-\omega} d\omega \right| = \left| \frac{1}{(j-1)!} \int_{x^{-M+1}}^{x^{-M+1}} \omega^{j-1} e^{-\omega} d\omega \right| = 0 \tag{3.55}$$

In the following, we apply the theorem to find out how $F_{j:K}(x)$ is close to its l.d. at $x = (\log \sqrt{K})^{1/(M-1)}$ and $x = (\log \sqrt{K})^{-1/(M-1)}$.
Substituting $x = (\log \sqrt{K})^{1/(M-1)}$ and $x = (\log \sqrt{K})^{-1/(M-1)}$ in $\Upsilon_1^j(x)$, we obtain

$$\Upsilon_1^j \left((\log \sqrt{K})^{\frac{1}{M-1}} \right) = e^{-\frac{1}{\log \sqrt{K}}} \sum_{i=1}^{j-1} \frac{1}{i! (\log \sqrt{K})^i} = 1 - O(1/\log K) \tag{3.56a}$$

and

$$\Upsilon_1^j\left((\log\sqrt{K})^{-\frac{1}{M-1}}\right) = e^{-\log\sqrt{K}}\sum_{i=1}^{j-1}\frac{(\log\sqrt{K})^i}{i!} = O\left(\frac{(\log\sqrt{K})^i)}{\sqrt{K}}\right) \tag{3.56b}$$

Thus,

$$\left|\Upsilon_1^j\left((\log\sqrt{K})^{\frac{1}{M-1}}\right) - \Upsilon_1^j\left((\log\sqrt{K})^{-\frac{1}{M-1}}\right)\right| \geq 1 - O(1/\log K) \tag{3.56c}$$

Then, for $x = (\log\sqrt{K})^{1/(M-1)}$, we have $\tilde{\delta}_K((\log\sqrt{K})^{1/(M-1)}) = \frac{1}{K\log\sqrt{K}}$, hence

$$K\tilde{\delta}_K^2((\log\sqrt{K})^{1/(M-1)})g(j, K\tilde{\delta}_K((\log\sqrt{K})^{1/(M-1)})) = o(1/K)$$

Therefore, we have

$$\left|F_{j:K}\left((K\log\sqrt{K})^{\frac{1}{M-1}} - 1\right) - \Upsilon_1^j((\log\sqrt{K})^{\frac{1}{M-1}}) + o\left(\frac{1}{K}\right)\right| = O\left(\frac{1}{K^2}\right) \tag{3.57}$$

In the same way, for $x = (\log\sqrt{K})^{-1/(M-1)}$, we have $\tilde{\delta}_K((\log\sqrt{K})^{-1/(M-1)}) = \frac{\log\sqrt{K}}{K}$, hence

$$K\tilde{\delta}_K^2((\log\sqrt{K})^{-1/(M-1)})g(j, K\tilde{\delta}_K((\log\sqrt{K})^{-1/(M-1)})) = o(1/K)$$

Therefore, we have

$$\left|F_{j:K}\left(\left(\frac{\log\sqrt{K}}{K}\right)^{-\frac{1}{M-1}} - 1\right) - \Upsilon_1^j((\log\sqrt{K})^{-\frac{1}{M-1}}) + o\left(\frac{1}{K}\right)\right| = O\left(\frac{(\log\sqrt{K})^3}{K}\right) \tag{3.58}$$

Using (3.56c),(3.57), and (3.58), we obtain

$$\left|F_{j:K}\left((K\log\sqrt{K})^{\frac{1}{M-1}} - 1\right) - F_{j:K}\left(\left(\frac{\log\sqrt{K}}{K}\right)^{-\frac{1}{M-1}} - 1\right)\right| \geq 1 - O\left(1/\log K\right) \tag{3.59}$$

or equivalently

$$\Pr\left\{\left(\frac{\log\sqrt{K}}{K}\right)^{-\frac{1}{M-1}} - 1 \leq X_{j:K} \leq \left(K\log\sqrt{K}\right)^{\frac{1}{M-1}} - 1\right\} \geq 1 - O\left(\frac{1}{\log K}\right) \tag{3.60}$$

3.F Proof of Theorem 3.2

Since $\mathcal{J}(P_1)$ is not always concave in P_1, the P_1^* that maximizes the objective function is either the boundary points ($P_1 = 0$ and $P_1 = P$) or the solutions corresponding to $\partial\mathcal{J}/\partial P_1 = 0$. By differentiating the objective function with respect to P_1, we have

$$\frac{\partial\mathcal{J}}{\partial P_1} = AP_1^2 + BP_1 + \Gamma \tag{3.61}$$

where

$$A = \eta_{k_1 2}\eta_{k_2 2}(\eta_{k_1 1} - \eta_{k_1 2})(P\eta_{k_2 1} + \sigma^2) + \eta_{k_1 1}\eta_{k_2 1}(\eta_{k_2 1} - \eta_{k_2 2})(P\eta_{k_1 2} + \sigma^2)$$

$$\begin{aligned} B &= \eta_{k_1 1}(P\eta_{k_1 2} + \sigma^2)(P\eta_{k_2 1}\eta_{k_2 2} + 2\eta_{k_2 1}\sigma^2 - \eta_{k_2 2}\sigma^2) \\ &+ \eta_{k_2 2}(P\eta_{k_1 2} + \sigma^2)(P\eta_{k_2 1} + \sigma^2)(2\eta_{k_1 2} - \eta_{k_1 1}) \end{aligned}$$

$$\Gamma = \eta_{k_1 1}\sigma^2(P\eta_{k_1 2}+\sigma^2)(P\eta_{k_2 2}+\sigma^2) - \eta_{k_2 2}(P\eta_{k_2 1}+\sigma^2)(P\eta_{k_1 2}+\sigma^2)^2$$

Setting $\frac{\partial \mathcal{J}}{\partial P_1} = 0$, the *possible* values of P_1 that maximize the throughput are the real-valued roots of the second-order polynomial $AP_1^2 + BP_1 + \Gamma = 0$ (for $A \neq 0$) that satisfy the constraint $P_1 \in [0, P]$ or $P_1 = -\Gamma/B$ for $A = 0$. Hence, the optimum P_1^* is the value among the boundary points ($P_1 = 0$ and $P_1 = P$) and the extreme points (roots of the polynomial) that maximizes $\mathcal{J}(P_1)$, which concludes the proof.

3.G Proof of Lemma 3.4

Let $\mathcal{J}_i(P_i)(i = 1, 2)$ represent the individual rate of user k_i given as

$$\mathcal{J}_i(P_i) = \log_2\left(1 + \frac{P_i\eta_{k_i i}}{\sigma^2 + (P - P_i)\eta_{k_i j}}\right) = \log_2\left(1 + \frac{P_i}{\sigma^2/\eta_{k_i i} + \alpha_{k_i}(P - P_i)}\right),\ j \neq i \tag{3.62}$$

The sum-rate maximizing beam power allocation problem can be rewritten as

$$\max_{\mathbf{P}\in\mathcal{P}^2} \mathcal{J}_1(P_1) + \mathcal{J}_2(P_2) \text{ subject to } P_1 + P_2 = P$$

We investigate now the behavior of the individual user rate objective function. By calculating the first and second derivative of $\mathcal{J}_i(P_i)$ we have

$$\frac{\partial \mathcal{J}_i(P_i)}{\partial P_i} = \frac{\Delta + \alpha_{k_i}P_i}{\Delta(\Delta + P_i)} > 0 \qquad \frac{\partial^2 \mathcal{J}_i(P_i)}{\partial P_i^2} = \frac{d_1(\Delta + \alpha_{k_i}P_i)}{d_2} \tag{3.63}$$

with $\Delta = \alpha_{k_i}(P - P_i) + \sigma^2/\eta_{k_i i}$, $d_1 = (2\alpha_{k_i} - 1)\Delta + \alpha_{k_i}P_i$, and $d_2 = \Delta^2(\Delta + P_i)^2$. The sign of d_1 determines the convexity or concavity of $\mathcal{J}_i(P_i)$. If $d_1 > 0 \rightarrow P_i > \left(\frac{1}{\alpha_{k_i}} - 2\right)\Delta$, $\mathcal{J}_i(P_i)$ is a convex function of P_i, and concave otherwise. Since $\Delta > 0$, for $\alpha_{k_i} \geq 0.5$ the objective function $\mathcal{J}_i(P_i)$ is convex $\forall i$, i.e. $\frac{\partial \mathcal{J}_i(P_i)}{\partial P_i} > 0$, hence the sum of two convex functions $\mathcal{J}_1(P_1) + \mathcal{J}_2(P_2)$ is maximized for $P_1^* = P$ and $P_2^* = 0$.

3.H Proof of Lemma 3.5

Let $\mathcal{R}_{\text{TDMA}} = \log_2\left(1 + \frac{R\eta_{k_1 1}}{\sigma^2}\right)$ denote the system throughput for TDMA mode. TDMA is optimal when $\mathcal{R}_{\text{TDMA}} \geq \mathcal{R}(\mathbf{P}) \Rightarrow \log_2\left(\frac{\mathcal{B}(P_1)}{\mathcal{C}(P_1)}\right) \geq 0$, where

$$\mathcal{B}(P_1) = (1 + P\eta_{k_1 1}/\sigma^2)(P\alpha_{k_1}\eta_{k_2 2} + \sigma^2)((P - P_1)\alpha_{k_1}\eta_{k_1 1} + \sigma^2)$$
$$\mathcal{C}(P_1) = (P\alpha_{k_1}\eta_{k_1 1} + P_1\eta_{k_1 1}(1 - \alpha_{k_1}) + \sigma^2)(P\eta_{k_2 2} + P_1\eta_{k_2 2}(\alpha_{k_2} - 1) + \sigma^2)$$

The region of TDMA optimality depends on the convexity of $\Psi(P_1) = \mathcal{B}(P_1) - \mathcal{C}(P_1)$. By differentiating twice we have that

$$\frac{\partial^2 \Psi(P_1)}{\partial P_1^2} = -2\eta_{k_1 1}\eta_{k_2 2}\left(\frac{P\eta_{k_1 1}}{\sigma^2}\alpha_{k_1 1}\alpha_{k_2 2} + \alpha_{k_1 1} + \alpha_{k_2 2} - 1\right) \tag{3.64}$$

For $\frac{\partial^2 \Psi(P_1)}{\partial P_1^2} \leq 0$, $\Psi(P_1)$ is concave with respect to P_1 ($\Psi(P_1) \geq 0$), since $\Psi(0) \geq 0$ and $\Psi(P) = 0$, which results in (3.29).

3.I Proof of Proposition 3.3

The average sum rate of TDMA-based random opportunistic beamforming is given by

$$\mathcal{R} = \mathbb{E}\left\{\log_2(1+\max_{1\le k\le K}\frac{P\left|\mathbf{h}_k\mathbf{q}_1\right|^2}{\sigma^2}\right\} = \int_0^\infty \log_2(1+x)dF_s^K(x)dx =$$

$$\overset{(a)}{=} \frac{1}{\log 2}\int_0^\infty \frac{1}{1+x}(1-F_s^K(x))dx = \log_2 e\int_0^\infty \frac{1}{1+x}(1-(1-e^{-x\sigma^2/P})^K)dx$$

$$\overset{(b)}{=} -\frac{1}{\log 2}\sum_{k=0}^{K}\binom{K}{k}(-1)^k\int_0^\infty \frac{e^{-\frac{xk\sigma^2}{P}}}{1+x}dx = \frac{1}{\log 2}\sum_{k=1}^{K}\binom{K}{k}(-1)^k e^{\frac{k\sigma^2}{P}}\mathrm{Ei}(-\frac{\mathrm{k}\sigma^2}{\mathrm{P}})$$

where integration by parts is applied to obtain (a) and (b) follows from binomial expansion.

Chapter 4

Exploiting Channel Structure in MIMO Broadcast Channels

4.1 Introduction

Exploiting multiuser diversity by selecting at each scheduling window the user(s) with the most favorable channel realizations is known to maximize the sum rate of multiuser systems. However, several practical implications may limit the applicability of such opportunistic scheduling schemes. Several opportunistic schemes required complete channel knowledge in order to fully benefit from multiuser diversity gains. This may lead to prohibitive feedback requirements in FDD systems and/or lack of robustness to CSIT errors in TDD setups. The significant feedback overhead in the uplink channel can be alleviated by feeding back coarse, quantized CSIT. The feedback load can be also reduced by allowing only users likely to be selected, i.e. users with large CQIs, to access the feedback channel.

For simplicity, many contributions in the limited feedback literature adopt a spatially white, block fading channel model, in which each channel realization remains constant over one block and changes independently in the next block. Nevertheless, the block fading channel model is rather pessimistic in practice, since temporal and spatial correlation often exists. In this chapter, we focus on such correlated channel scenarios and show that this channel structure, either in time or in space domain, can be seen as an additional degree of freedom to be exploited during the scheduling phase. We show this additional channel information can be used for significant throughput increase and/or feedback reduction and compression.

In time-varying channel configurations, the inherent temporal redundancy can be exploited for:

- Feedback aggregation: information derived from low-rate feedback channel can be cumulated over time to approach the performance of full CSIT scenario.

- Feedback compression: the channel can be seen as a Markov source and redundancy is exploited to reduce feedback close to rate of innovation.

In spatially-correlated channels, long-term statistical channel knowledge can reveal information about the mean spatial separability of users, thus it contains relevant information for the SDMA scheduler. For instance, two users in very different areas of the cell are more likely to be separable than closely located users because their channels lie in two distinct cones of energy as seen by the BS, if reasonably limited angle spread at the BS is assumed. Note that the angle information is implicit in the transmit correlation matrix of the user's channel and needs not be estimated. Moreover, statistical CSIT can be easily obtained by the mobile and fed back to the transmitter while causing almost negligible per-slot feedback overhead.

The remainder of the chapter is organized as follows: in the first part, we focus on time-correlated channels and address the question how temporal correlation can significantly improve user scheduling decisions and achieve near optimal sum rate. Specifically, we propose a scheme that builds on random multi-beam beamforming [9], in which channel memory is exploited as a means to successively refine the random precoder selections. In the second part, we address the problem of SDMA scheduling and beamforming with limited feedback in spatially-correlated channels. Several user selection strategies exploiting statistical CSIT are investigated. We show how second-order statistical information is combined with instantaneous CQI and derive a coarse channel estimation framework. Finally, we propose a low-complexity, interference-bounded SDMA eigenbeamforming scheme, which relies on multi-user interference estimates (bounds).

4.2 Exploiting redundancy in time-correlated channels

4.2.1 User Selection in time-correlated channels

Consider that the channel exhibits correlation from one scheduling time slot to the other. Evidently, in such configurations, the scheduling decisions exhibit in turn some form of correlation over successive intervals. In other words, channel correlation in the time domain creates temporal redundancy, which can be exploited as means to either reduce feedback rate or increase the system throughput. If the channel varies slowly, then clearly the best user in terms of channel quality at current time slot τ is highly likely to be the best user at the subsequent time slots $\tau + T_c$. Therefore, the fact that previously selected users are highly likely to remain good can be further exploited during the user selection process. Temporal correlation has been exploited in packet switch design, either by using a maximum weight matching algorithm [81] or by a randomized algorithm exploiting temporal correlation of queue states [82]. In [83], the authors proposed a randomized scheduler that exploits temporal correlations in slow fading channels.

4.2.2 Beamforming and Scheduling exploiting temporal correlation

Since scheduling and linear precoding is the leitmotiv of this dissertation, we address the problem how to exploit temporal correlation and enhance the sum-rate performance of low-complexity multiuser transmission techniques in MISO broadcast channels. We present a

novel SDMA scheduling/precoding scheme, coined as *Memory-based Opportunistic Beamforming* (MOBF). The scheme builds on multi-beam random beamforming [9] presented in Section 2.9.3, and exploits memory in the channel as a means to fill the gap to sum-rate optimality.

In a nutshell, MOBF replaces the random selection of precoding matrices with a combination of random and past feedback-aided beamforming matrices that are kept in memory. The scheme can be seen as successive refinement of the precoding matrix inside the coherence time of the channel. When the coherence time of the channel is high (e.g. large Doppler spread), MOBF approaches the sum capacity of optimal unitary precoder with perfect CSIT. For uncorrelated i.i.d. channels, the performance of the proposed scheme remains superior to that of [9] at the expense of moderate additional feedback (two SINR values per user instead of one).

Interestingly, the scheme can be seen to also relate to recent useful results [84], presented to improve the delay performance of the single-beam opportunistic beamforming [53]. In [84] scheduling is limited to one user and temporal channel correlation is exploited through the use of a fixed set of beams determined in advanced. This scheme does not automatically reach the performance of a full CSIT scenario, since the temporal correlation was used to restore fairness and users with long waiting times are prioritized. Another scheme that exploits temporal correlation in orthogonal frequency division multiple access (OFDMA) systems have been proposed in [85].

4.2.3 Memory-based Opportunistic Beamforming

As stated before, MOBF builds on random beamforming (cf. Section 2.9.3), in which the transmitter generates at each time slot t a $\mathcal{B} \times \mathcal{B}$ ($\mathcal{B} \leq M$) unitary precoding matrix $\mathbf{Q}(t)$ randomly, as a means to reduce the feedback burden and complexity requirements, i.e. $\mathbf{W}(t) = \mathbf{Q}(t) = [\mathbf{q}_1(t) \dots \mathbf{q}_{\mathcal{B}}(t)]$. In conventional RBF [9], a new random unitary precoding is generated and used for serving the selected users at each time slot. Hence, any kind of structure in the physical channel is not exploited. Memory-based Opportunistic Beamforming attempts to exploit memory in the channel by making at each time slot an improved selection of the unitary precoding matrix based on past CQI information. Temporal correlation is exploited by memorizing the previous best scheduling decision(s), i.e. the group of selected users $\mathcal{S}$ for a random precoder $\mathbf{Q}(t)$, and comparing it with the next random matchings $\mathbf{Q}(t+i)$ for $i = 1, \dots, T_c$.

Specifically, we consider that the BS has a codebook (set) of 'preferred' unitary matrices of size U:

$$\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_U\} \tag{4.1}$$

with $\mathcal{Q} \subseteq \mathcal{U}(M, M)$, where $\mathcal{U}(M, M)$ denotes the unitary group of degree M, i.e. the group of $M \times M$ unitary matrices defining the complex Stiefel manifold. The notion of 'preferred' is used in the sense of (relative) maximization of the sum rate among past used random beamforming matrices.

At each time instant t, the unitary matrix of the preferred set, denoted $\tilde{\mathbf{Q}}$ and defined as the precoder that has provided the highest sum rate in previous time slots, is applied and its sum rate is measured (updated) under current channel conditions. The achievable sum rate of $\tilde{\mathbf{Q}}$ at time slot $t+1$ is compared with that of a new, randomly generated unitary matrix $\mathbf{Q}_r$,

and the beamforming matrix that offers the highest sum rate is selected for transmission. In the phase of updating the codebook $\mathcal{Q}$, the sum rate value of $\tilde{\mathbf{Q}}$ in the codebook is updated, and the newly generated random precoder $\mathbf{Q}_r$ is added into the codebook if and only if its sum rate is higher than the sum rate of the codebook matrix with the minimum sum rate. Let $\mathcal{S}_t$ denote the set of selected users at each scheduling window t and $\mathbf{H}(\mathcal{S}_t)$ be the corresponding submatrix of $\mathbf{H} = [\mathbf{h}_1^T \dots \mathbf{h}_K^T]^T$. With $\mathcal{R}(\mathbf{Q}, \mathcal{S}_t)$ we denote the sum rate when unitary beamforming matrix $\mathbf{Q}$ is used for serving the users belonging to $\mathcal{S}_t$. The steps of the proposed algorithm are outlined in Table 4.1.

Table 4.1: Memory-based Opportunistic Beamforming Algorithm

Memory-based Opportunistic Beamforming (MOBF) Algorithm

***First phase*:** ('best' unitary matrix selection)

Step 0 Initialize codebook $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_U\}$,

each with sum rate $\mathcal{R}(\mathbf{Q}_i), i = 1, \dots, U$

At each time slot t,

Step 1 Generate a new random precoder $\mathbf{Q}_r$

Step 2 Select $\tilde{\mathbf{Q}} \in \mathcal{Q}$: $\tilde{\mathbf{Q}} = \arg\max_{\mathbf{Q}_i \in \mathcal{Q}} \mathcal{R}(\mathbf{Q}_i)$

Step 3 Apply $\tilde{\mathbf{Q}}$, collect updated feedback from the users

and calculate $\mathcal{R}(\tilde{\mathbf{Q}}, \mathcal{S}_{t+1})$

Step 4 If $\mathcal{R}(\tilde{\mathbf{Q}}, \mathcal{S}_{t+1}) > \mathcal{R}(\mathbf{Q}_r, \mathcal{S}_t)$, $\mathbf{Q}^* \rightarrow \tilde{\mathbf{Q}}$, else $\mathbf{Q}^* \rightarrow \mathbf{Q}_r$

***Second phase*:** (Update of codebook $\mathcal{Q}$)

Step 5 Update the value $\mathcal{R}(\tilde{\mathbf{Q}})$ in the set $\mathcal{Q}$

Step 6 If $[\mathcal{R}(\mathbf{Q}_r) > \mathcal{R}(\mathbf{Q}_{min})]$, $\mathbf{Q}_{min} \rightarrow \mathbf{Q}_r$, where $\mathbf{Q}_{min} = \arg\min_{\mathbf{Q}_i \in \mathcal{Q}} \mathcal{R}(\mathbf{Q}_i))$

Some comments are in order: The algorithm outlined in Table 4.1 presents a general framework for memory-based, randomized scheduling in slow time-varying channels. First, in practice, at each time slot t the set $\mathcal{Q}$ contains only one precoder matrix ($U = 1$), i.e. the one that has provided the highest system throughput up to the current time instant. Secondly, although MOBF is based on RBF for precoding and user selection, our proposed scheme is not only restricted to such systems. The idea of memory-based precoding can be also applied to systems where the users utilize a codebook to quantize their channels and feed back quantized CDI. If the channel is strongly correlated, the above concept can be used to reduce the feedback load by decreasing the feedback reporting rate. At each slot, additional CDI is then fed back only if it is sufficiently different than the one previously reported. Alternatively, if we enforce CDI reporting at each time instant, users may have the possibility to refine their CDI information at each time slot, using hierarchical codebooks.

Performance Analysis

The underlying idea behind the sum-rate analysis of MOBF is the following: the process of memorizing at each scheduling slot the sum-rate maximizing precoding matrix can be seen as a random search of beamforming configurations in the space of orthogonal unitary precoders. Evidently, the performance of such scheme depends on the distribution of the sum rate conditioned to a certain channel realization $\mathbf{H}(\mathcal{S})$ for the selected group of users $\mathcal{S}$. To simplify our analysis, we fix the channel of the selected users to a certain realization $\mathbf{H}$ and we analyze the properties of $X_i = \mathcal{R}(\mathbf{Q}_i, \mathbf{H})$, which represents the sum rate provided by random unitary matrices $\mathbf{Q}_i$ for a given channel $\mathbf{H}$. Therefore, $\{X_i\}_{i=1}^{\infty}$ is a random process whose distribution depends on the underlying random variable $\mathbf{Q}_i$. For fixed channel realization, X_i is i.i.d. for i with associated PDF $f_X(\cdot)$ and CDF $F_X(\cdot)$. If the channel is quasi-static, memory-based beamforming aims at finding the unitary beamforming matrix $\mathbf{Q}^*$ from the feasible set of unitary matrices $\mathcal{U}$ that maximizes the sum rate. This can be mathematically written as:

$$\mathbf{Q}^* = \arg \max_{\mathbf{Q}_i \in \mathcal{U}} \mathcal{R}(\mathbf{Q}_i) = \arg \max_{1 \leq i \leq |\mathcal{U}|} X_i \tag{4.2}$$

Note that this optimization returns one out of possibly many global maximizers $\mathbf{Q}^*$ since the global maximizer is not unique, i.e. $\mathcal{R}(\mathbf{Q}^*) = \mathcal{R}(\mathbf{Q}^*\mathbf{Q}'^H)$, for any $\mathbf{Q}' \in \mathcal{U}$. However, the maximum value of the sum rate, $X^* = \mathcal{R}(\mathbf{Q}^*, \mathbf{H})$, is unique over the set $\mathcal{U}$.

Assuming that the set of unitary matrices $\mathcal{U}$ is finite with cardinality $|\mathcal{U}|$, then for $|\mathcal{U}|$ i.i.d. random unitary matrices $\{\mathbf{Q}_i\}_{i=1}^{|\mathcal{U}|}$, the achievable sum rate X^* is given by

$$X^* = \max_{1 \leq i \leq |\mathcal{U}|} X_i = \int_0^{\infty} x dF_X^{|\mathcal{U}|}(x) \tag{4.3}$$

For asymptotically large $|\mathcal{U}|$, the distribution of $\max_{1 \leq i \leq |\mathcal{U}|} X_i$ converges - after proper shifting and scaling - to a limiting distribution (l.d.) of Gumbel, Fréchet or Weibull type. However, as the exact form of the CDF $F_X(x)$ is difficult to obtain, the exact l.d. is difficult to be inferred. Hence, we resort to the following result in order to derive the asymptotic (in $|\mathcal{U}|$) convergence of our algorithm.

Proposition 4.1: *Consider a channel with memory $\mathcal{L} = \frac{T_c}{T_s}$, where T_c is the channel coherence time, and T_s is the slot duration. For $\mathcal{L} \to \infty$, the sum rate of memory-based beamforming $\mathcal{R}_{MOBF}$ converges to the capacity of optimum unitary beamforming $\mathcal{R}^*$ for a given channel $\mathbf{H}$:*

$$\mathcal{R}_{MOBF} \to \mathcal{R}^* = \max_{\mathbf{Q} \in \mathcal{U}} \mathcal{R}(\mathbf{Q}_i, \mathbf{H}) \tag{4.4}$$

Proof. The proof is given in Appendix 4.A. □

The above result implies that the maximum of the sum rate offered by using various precoders $\mathbf{Q}_i$ converges asymptotically to the optimum capacity of unitary beamforming $\mathcal{R}^*$. As a result, the corresponding unitary precoding matrix, denoted $\mathbf{Q}^*$, which corresponds to the matrix that maximizes the sum rate converges to one of the possibly many optimum unitary precoders. Therefore, if the channel is quasi-static (very large $\mathcal{L}$), the codebook of MOBF will contain an optimal beamforming matrix, i.e. a unitary matrix that maximizes the sum rate for a certain channel realization.

4.3 Performance evaluation

For the evaluation of MOBF, we consider a time-varying Rayleigh fading channel where the fading channels $\mathbf{h}_k(t)$ are i.i.d. among users and transmit antennas. We consider that the channel evolves according to the Clark-Jake's Doppler model, with autocorrelation function $\mathbb{E}\{\mathbf{h}(t)\mathbf{h}(t+\ell T_s)\} = J_0(2\pi f_d \ell T_s)$ where f_d denotes the one-sided Doppler bandwidth (in Hz). We set T_s=1 ms and carrier frequency equal to 2GHz. The average SNR is set to 0 dB for all users.

In Figure 4.1, we plot the sum rate of MOBF versus the number of transmit antennas M for different Doppler spreads and $K = 20$ active users. Expectedly, the capacity of MOBF increases as the channel order (memory) increases. Furthermore, MOBF exhibits the same capacity scaling as that of RBF. The worst performance is achieved for a rapidly time-varying channel with memory $\mathcal{L} = 1$, where the probability that the 'preferred' matrix will be valid if reapplied falls to $1/2$. In this case, MOBF benefits from selection diversity gain as compared to conventional RBF. This means that MOBF is equivalent to a RBF scheme where two randomly generated precoders are generated and the one with the highest sum rate is applied. The sum rate of MOBF is also plotted for a static channel ($\mathcal{L} \to \infty$). In that case, the tracking capability of our algorithm is increased and the transmitter is able to successively 'learn' the channel directions of users, approaching thus the case of complete CSIT. Note also that MOBF achieves high sum rate even for fixed, but not necessarily large, number of users.

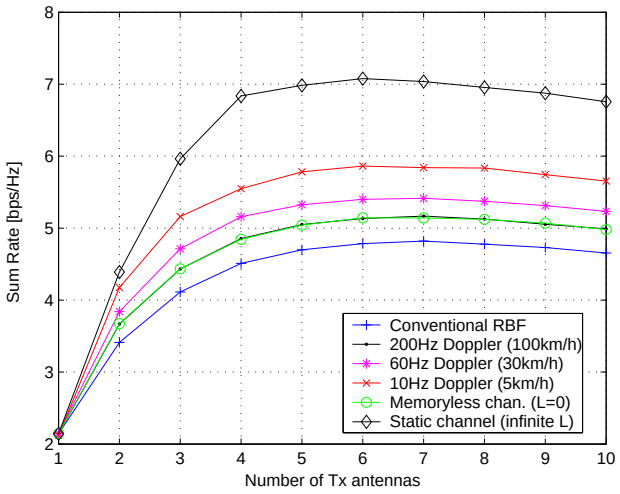

Figure 4.1: Sum rate vs. the number of transmit antennas M of MOBF with $K = 20$ users and various Doppler spreads.

In Figure 4.2 we evaluate the system throughput as a function of number of users for M = 8 antennas. As expected, the gap between MOBF and RBF is bigger for small number of users. For K increasing, the sum rate of RBF improves as it is more likely that the random beams will find users with high channel gains and closely aligned with the random beam directions.

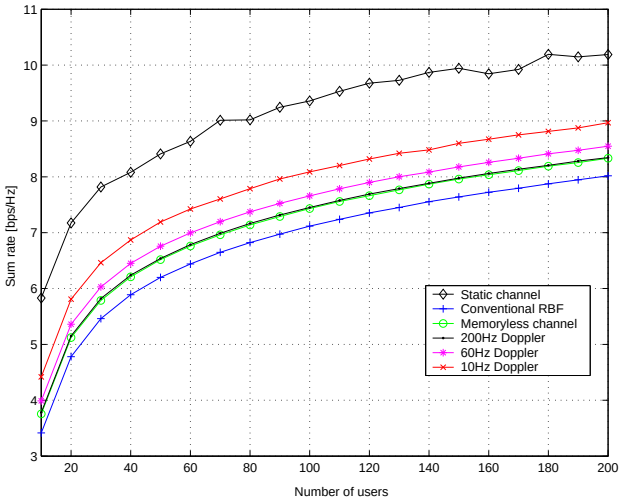

Figure 4.2: Sum rate as a function of number of users K of MOBF for different Doppler spreads.

4.4 Exploiting Statistical CSIT in Spatially Correlated Channels

Apart from exhibiting temporal correlation, in real wireless systems, users' channels are often correlated in the space domain. In the following sections, we consider an outdoor cellular (wide-area) network, for which the i.i.d. spatially white channel model used so far in this dissertation does not hold. In practice, each user tends to exhibit different spatial channel statistics, which is captured by its channels correlation matrix. For ease of exposition, no temporal channel correlation is considered below.

We assume that the transmitter has statistical CSIT, i.e. information of the statistics of the wireless stochastic propagation channel. This is a reasonable and practical assumption since statistical CSIT has the advantage of longer coherence time as compared to that of the fading channel, thus it can be easily obtained by the mobile and fed back to the BS at low rate. Furthermore, several forms of statistical CSIT are even reciprocal, e.g. the second-order correlation matrix, the power of Ricean component, etc., and do not necessitate any feedback to be revealed to the transmitter. A key observation here is that useful information relevant to the scheduler lies untapped in the long-term statistical information of the user's channels. Second-order statistical channel knowledge reveals a great deal of information on the macroscopic nature of the underlying channel, including the multipath's mean angle of arrival/departure and its angular spread.

On the other hand, in order to exploit multiuser diversity during the scheduling procedure, the transmitter must have some form of instantaneous CQI for each user as a means to distinguish favorable from unfavorable channel realizations. The question we try to answer here is which type of low-rate CQI is relevant and sufficient in order to minimize the feedback load, while allowing the scheduler to extract multiuser diversity gain. A generic maximum likelihood (ML) coarse channel estimation framework is established, which let the BS to

efficiently select users combining statistical CSIT and instantaneous CQI. Low-complexity user selection metrics and algorithms are also proposed. Finally, in order to better estimate the inter-user interference, we augment the per-slot CQI feedback with instantaneous scalar CDI on beamforming alignment. We demonstrate the merit of channel/beamforming alignment information and propose SDMA eigenbeamforming based on inter-user interference estimates.

Combining the second-order channel statistics with instantaneous CQI for resource allocation was also considered in [86] for point-to-point systems. In [87, 88], Hammarwall et al., proposed a minimum mean squared error (MMSE) estimation framework for combining CQI and long-term CSIT. The signal/interference power estimates, which are computed by the conditional moments of the channel, are used for SINR estimation, scheduling and transmission.

4.4.1 System Setting

We consider the downlink of a cellular FDD system with single-antenna mobiles and correlation between the channels gains of different antennas. This scenario models an environment where transmit antennas are placed for instance at an elevated high-point base station, i.e. the near-field scattering at the transmitter is limited [89]. We assume that the receivers are located in a rich-scattering surrounding, thus correlation appears only at the transmitter side.

Channel Model

The channel vector of k-th user is modeled as satisfies $\mathbf{h}_k \sim \mathcal{CN}(\bar{\mathbf{h}}_k, \mathbf{R}_k)$. This means that the complex random vector $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ is circularly-symmetric Gaussian distributed, with mean $\bar{\mathbf{h}}_k = \mathbb{E}\{\mathbf{h}_k\}$ and covariance matrix $\mathbf{R}_k = \mathbb{E}\{(\mathbf{h}_k - \bar{\mathbf{h}}_k)(\mathbf{h}_k - \bar{\mathbf{h}}_k)^H\}$. Its multivariate PDF is given by

$$f_{\mathbf{h}}(\mathbf{h}_k) = \frac{1}{\pi^M |\mathbf{R}_k|} \exp\left\{-(\mathbf{h}_k - \bar{\mathbf{h}}_k)^H \mathbf{R}_k^{-1} (\mathbf{h}_k - \bar{\mathbf{h}}_k)\right\} \tag{4.5}$$

The correlation matrix $\mathbf{R}_k \in \mathbb{C}^{M\times M}$, which is perfectly known at both ends of the link, is assumed to be dominated by one or a few eigenvalues. This is a valid assumption since the statistical channel information changes slower than the small-scale fading of the channel, and can be obtained with low or no additional feedback.

Instantaneous CQI Feedback

At each scheduling slot, the users feed back instantaneous information on their channel quality (CQI), denoted as γ_k. A general representation of CQI utilized in this chapter is

$$\gamma_k = \left\|\mathbf{h}_k^H \mathbf{Z}_k\right\|^2 \tag{4.6}$$

where $\mathbf{Z}_k \in \mathbb{C}^{M\times T}$ can be seen as a training matrix containing T vectors $\{\mathbf{z}_{ki}\}_{i=1}^{T}$, resulting in a weighted norm of the channel vector. The CQI feedback can take on among others the following forms, depending on the system feedback rate and pilot signaling overhead constraints:

- *Strategy 1*: $\gamma_k = \left|\mathbf{h}_k^H \mathbf{z}_1\right|^2$ (beam gain information - T=1)
- *Strategy 2*: $\gamma_k = \|\mathbf{h}_k\|^2$ (channel norm feedback - $\mathbf{Z}_k = \mathbf{I}$)

In what follows, we focus on the above two CQI feedback strategies.

4.4.2 User Selection with ML Channel Estimation

Optimal User Selection

If we restrict ourselves to the case of joint linear beamforming and scheduling, the optimal user selection policy is to exhaustively search over all the user sets for all combination of feasible beamformers and select the one that maximizes the system throughput. Formally, the optimal group of selected users is determined as

$$\mathcal{S}^* = \arg \max_{\forall \mathcal{S}, \mathbf{W}} \mathcal{R}(\mathcal{S}, \mathbf{W}) \tag{4.7}$$

where $\mathcal{R}(\mathcal{S}, \mathbf{W})$ is the achievable sum rate when the user set $\mathcal{S}$ is served using precoder $\mathbf{W}$. The problem may be extremely complex for dense networks, since the search complexity increases exponentially with the number of users. The complexity can be reduced by taking into account a smaller group of pre-selected users. This group may be defined based on coarse channel knowledge, which is obtained by very low-rate feedback.

MSE User Selection Metric

Here we reduce the complexity of optimal user selection by restricting the choice of the precoding matrix to be the one that minimizes the mean-square error (MSE) between the received and symbol vectors. Thus, the objective here is to find the optimal group of users under MMSE precoding $\mathbf{W}_{\text{MMSE}}(\mathcal{S})$. We first derive the solution assuming full CSIT in order to gain insight. Let $\mathbf{H} \in \mathbb{C}^{K \times M}$ denote the concatenation of all channels, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]^H$, where the k-th row is the channel of the k-th receiver ($\mathbf{h}_k^H$). Mathematically the problem can be expressed as

$$\mathbf{W}_{\text{MMSE}}(\mathcal{S}) = \arg \min_{\substack{\mathbf{W} \\ \|\mathbf{W}\|_F^2 \leq P}} \mathbb{E}\left\{||\mathbf{s}(\mathcal{S}) - \mathbf{y}(\mathcal{S})||^2\right\} \tag{4.8}$$

which results in the following optimal precoding matrix

$$\mathbf{W}_{\text{MMSE}}(\mathcal{S}) = \left(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H + \beta \mathbf{I}\right)^{-1} \tag{4.9}$$

where β is the non-negative Lagrange multiplier tuned to fulfill the transmit power constraint. When the criteria is to maximize the sum rate, the regularization constant takes the value $\beta = M\sigma^2/P$ [34]. Inserting this solution into to the MSE minimization problem (with cost function $\mathcal{J}_{\text{MMSE}}$, it can be shown that the MMSE level is given by

$$\begin{aligned} \mathcal{J}_{\text{MMSE}}(\mathcal{S}) &= M - 2\text{Re}\{\text{Tr}\left(\mathbf{W}_{MMSE}(\mathcal{S})\mathbf{H}(\mathcal{S})\right) \\ &+ \text{Tr}(\mathbf{W}_{MMSE}\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H(\mathcal{S})\mathbf{W}_{MMSE}(\mathcal{S})^H) + M\sigma^2\} \end{aligned}$$

where $\text{Re}\{\cdot\}$ denotes the real part. We should remark that the downlink MMSE precoders do not in fact minimize the MSE at the receiver side, since the precoder affects all received

signals before noise is introduced. The MSE user selection metric that minimizes the MSE for the selected group of users $\mathcal{S}^*$ is given by

$$\begin{aligned}\mathcal{S}^* &= \arg\max_{\mathcal{S}} 2\mathrm{Re}\left\{\mathrm{Tr}\left\{\mathbf{\Psi}(\mathcal{S})\left(\mathbf{\Psi}(\mathcal{S})+\beta\mathbf{I}\right)^{-1}\right\}\right\} \\ &- \mathrm{Tr}\left\{\left(\mathbf{\Psi}(\mathcal{S})\left(\mathbf{\Psi}(\mathcal{S})+\beta\mathbf{I}\right)^{-1}\right)^2\right\} \\ &= \arg\min_{\mathcal{S}} \mathrm{Tr}\left\{\left(\left(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H+\beta\mathbf{I}\right)^{-1}\right)^2\right\} \end{aligned} \quad (4.10)$$

where $\mathbf{\Psi}(\mathcal{S}) = \mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H$.
If we consider that the scheduler has only statistical knowledge of the channels, i.e. relies only on the correlation matrices, eq. (4.10) can be approximated by replacing $\mathbf{\Psi}(\mathcal{S})$ by its statistical estimate $\hat{\mathbf{\Psi}}(\mathcal{S})$. In this chapter, we consider that $\hat{\mathbf{\Psi}}(\mathcal{S})$ may take the following two forms:

- $\hat{\mathbf{\Psi}}(\mathcal{S}) = \mathbf{R}(\mathcal{S})$ if no additional instantaneous CQI is available at the scheduler. The concatenated correlation matrix is defined $\mathbf{R}(\mathcal{S}) = \sum_{k\in\mathcal{S}} \mathbf{R}_k = \mathbb{E}\{\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H\}$.
- $\hat{\mathbf{\Psi}}(\mathcal{S}) = \hat{\mathbf{H}}(\mathcal{S})\hat{\mathbf{H}}(\mathcal{S})^H$, where $\hat{\mathbf{H}}$ is the concatenation of channel estimates $\hat{\mathbf{h}}_k$ combining long-term statistical knowledge and instantaneous CQI feedback.

In that case, the MSE minimizing group of users $\mathcal{S}^*_{CE}$ based on channel estimates (CE) is given by

$$\mathcal{S}^*_{CE} = \arg\min_{\mathcal{S}} \mathrm{Tr}\left\{\left(\left(\hat{\mathbf{\Psi}}(\mathcal{S})+\beta\mathbf{I}\right)^{-1}\right)^2\right\} \quad (4.11)$$

Greedy User Selection

In the previous section, determining the optimal set of spatially separable users, $\mathcal{S}^*$, requires exhaustive search over the entire user set. However, when is K large, the complexity of optimal user selection becomes prohibitively high, since the size of its search space $\sum_{i=1}^{M}\binom{K}{i}$ is large. A suboptimal, yet efficient, greedy user selection scheme can be used instead, similar to the approach in [11]. Here we extend this scheme for MMSE linear beamforming with long-term spatial information and instantaneous scalar CQI feedback γ_k. The proposed greedy user selection algorithm is given in Table 4.2. In this algorithm, users are added one by one to the set. The complexity can be reduced by considering only users exceeding a threshold γ_{th}. The user with the highest CQI is examined at each time, and it is added to the set of scheduled users $\mathcal{S}$ only if it results in sum-rate increase. We should note that the overall performance of greedy user selection depends heavily on whether the precoding matrix can be reprocessed each time a user is added, which in turn depends on the form of the channel feedback available at the transmitter.

4.4.3 ML *coarse* Channel Estimation with CQI Feedback

As stated before, the correlation matrix provides useful information about the spatial channel characteristics, especially if it is ill-conditioned, however it does not reveal any information about the quality of the current channel realization. In order to exploit multiuser diversity, the scheduler requires properly designed instantaneous low-rate feedback γ_k, which

Table 4.2: Greedy User Selection with Statistical CSIT

Greedy User Selection with statistical CSIT

At each time slot t

1. Initialize $\mathcal{S} = \emptyset$ and $\mathcal{G} = \emptyset$.
2. Select the users that exceed the threshold γ_{th}
$$\mathcal{G} = \{\forall k \in \{1, \ldots, K\} | \gamma_k \geq \gamma_{th}\}$$
3. Select the user with the highest CQI value
$$k_{max} = \arg \max_{1 \leq k \leq K} \gamma_k$$
$$\mathcal{S} \leftarrow \mathcal{S} \cup \{k_{max}\}, \quad \mathcal{G} \leftarrow \mathcal{G} \setminus \mathcal{S}$$
4. Repeat
$$k^* = \arg \min_{k \in \mathcal{G}} \mathrm{Tr} \left\{ \left(\left(\hat{\mathbf{\Psi}}(\mathcal{S}) + \beta \mathbf{I} \right)^{-1} \right)^2 \right\}$$
$$\mathcal{S} \leftarrow \mathcal{S} \cup \{k^*\}, \quad \mathcal{G} \leftarrow \mathcal{G} \setminus \mathcal{S}$$
until $|\mathcal{S}| = M$
5. Return user set $\mathcal{S}$

can be a measure of the quality of the current channel. In this section, we restrict ourselves to Rayleigh fading correlated channels, i.e. $\bar{\mathbf{h}}_k = 0$, and we propose a simple framework in which long-term statistical channel knowledge is combined with short-term partial CSIT as a means to provide a coarse channel estimate at each slot.

ML Estimation with Beam Gain Information

We adopt here the feedback *strategy 1* and consider that each user k feeds back the squared magnitude of the channel with a beamforming vector $\mathbf{z}_k \in \mathbb{C}^{M \times 1}$, i.e. $\gamma_k = \left| \mathbf{h}_k^H \mathbf{z}_k \right|^2$. The beamforming vectors can be interpreted as pilot signals during the training phase or as the preferred beamformer in a two-stage precoding and scheduling approach (see Section 3.4). This beamformer can be chosen randomly or it can be optimized based on long-term statistical information.

Optimized training vectors As the correlation matrix of each user is known at the transmitter side, the training vectors $\mathbf{z}_k$ can be optimized. Briefly speaking, an efficient training codebook can contain $N = N_p + N_l + N_r$ vectors, where the indices p,l,r indicate principal, local, and random, respectively as explained below. The codebook construction follows a three-step procedure:

1) Based on each user's statistical CSIT, the codebook will contain N_p principal eigenvectors of the covariance matrix $\mathbf{R}_k$ ($N_p = 1$ for MISO channels).

2) In this step, we select N_l vectors in the local area of each principal statistical direction $\mathbf{v}_k$ as a means to account for those channel realizations that steer the principal singular vector in a locality of the principal statistical direction. The local area of the principal statistical direction is defined by a cone around $\mathbf{v}_k$ and is characterized by the angle between the training and the principal statistical vectors.

3) During the third step, we generate N_r vectors that are outside the cone defined in step 2 and account for the channel realizations in which the direction of the principal right singular vector (or vector channel) is far from the statistical (mean) channel direction. These vectors can be chosen randomly or as the ones that covers optimally the remaining space, outside the cone, which is related to the Grassmannian line packing problem. In the ideal case, the size of N_r should be adapted based on the strength of correlation, as it gives a measure on the frequency that these deviations occur.

Random training vectors For simplicity, we rather adopt a low-complexity approach and consider a random opportunistic beamforming setting [9]. In this setting, we assume that the vectors $\mathbf{z}_k$ are isotropically distributed and chosen randomly, i.e. $\mathbf{z}_k = \mathbf{q}_m$ where $\{\mathbf{q}_m\}_{m=1}^{M}$ are the columns of the unitary matrix $\mathbf{Q}$. We combine the information extracted from the correlation matrix with a scalar instantaneous feedback in the form of $\gamma_k = |\mathbf{h}_k^H \tilde{\mathbf{q}}_k|^2$, where the vector $\mathbf{z}_k = \tilde{\mathbf{q}}_k$ is chosen by user k as

$$\tilde{\mathbf{q}}_k = \arg \max_{m=1,\ldots,M} |\mathbf{h}_k^H \mathbf{q}_m|^2 \tag{4.12}$$

Clearly, this type of scalar CQI provides a joint *instantaneous* measure of the quality of the current channel realization and its direction of the channel instantaneously. Although the amount of spatial information encapsulated into this metric cannot be decomposed from the channel gain information, it is particularly useful for users with strong channels, i.e. users that are very likely to be scheduled. It can be also shown that the choice of $\tilde{\mathbf{q}}_k$ is equivalent to selecting the beam over which user k experience the highest received SINR_k in [9]. Assume that user k has its maximum SINR on beam i out of the $j \in \{1, \ldots, M\}$, defined as:

$$i = \arg\max_j \frac{x_j}{c - x_j} \tag{4.13}$$

where $x_j = \left|\mathbf{h}_k^H \mathbf{q}_j\right|^2$ with $0 < x_j < c$, and $c = \sum_{m=1}^{M} \left|\mathbf{h}_k^H \mathbf{q}_m\right|^2 + M\sigma^2/P$ is a positive constant. Defining the function $f(x) = \frac{x}{c-x}$, we have that $\lim_{x\to 0} f(x) \to 0$ and $\lim_{x\to c} f(x) \to \infty$. Since $f(x)$ is always monotonous positive for $x \in (0, c)$, we have that

$$i = \arg\max_j f(x_j) = \max_j x_j \tag{4.14}$$

or equivalently $\arg\max_j \mathrm{SINR}_{k,j} = \arg\max_j \left|\mathbf{h}_k^H \mathbf{q}_j\right|^2$.

Constrained Maximum Likelihood Optimization

We propose a ML estimation framework that combines long-term statistical knowledge and instantaneous CSIT provided by the feedback metric γ_k. This feedback allows us to pick

users whose channels span spatially separated cones of multipath and have good channel gains. This so-called Constrained Maximum Likelihood (CML) channel estimate is the one that maximizes the log-likelihood function of the PDF (4.5) conditioned to the scalar feedback constraint $\gamma_k = |\mathbf{h}_k^H \tilde{\mathbf{q}}_k|^2$:

$$\hat{\mathbf{h}}_k = \arg\max f(\mathbf{h}|\gamma_k) \tag{4.15}$$

This results to the following optimization problem:

$$\begin{array}{ll} \max\limits_{\mathbf{h}_k} & \mathbf{h}_k^H \mathbf{R}_k \mathbf{h}_k \\ s.t. & |\mathbf{h}_k^H \tilde{\mathbf{q}}_k|^2 = \gamma_k \end{array} \tag{4.16}$$

It can be easily shown that (4.16) is equivalent to solving the following generalized eigenvalue problem (GEV): $\mathbf{R}_k \mathbf{h}_k = \lambda \mathbf{\Phi}_k \mathbf{h}_k$, where $\mathbf{\Phi}_k = \tilde{\mathbf{q}}_k \tilde{\mathbf{q}}_k^H$. The maximum generalized eigenvalue of the Hermitian matrix pair $(\mathbf{R}_k, \mathbf{\Phi}_k)$, with $\mathbf{\Phi}_k > \mathbf{0}$ is defined as

$$\lambda_{max}(\mathbf{R}_k, \mathbf{\Phi}_k) = \sup\{\lambda | \det(\lambda \mathbf{\Phi}_k - \mathbf{R}_k) = 0\} = \sup_{\mathbf{q} \neq \mathbf{0}} \frac{\mathbf{h}_k^H \mathbf{R}_k \mathbf{h}_k}{\mathbf{h}_k^H \mathbf{\Phi}_k \mathbf{h}_k} \tag{4.17}$$

The solution of (4.16), in the view of the generalized Rayleigh-Ritz quotient, is given by

$$\hat{\mathbf{h}}_k = \arg\max_{\mathbf{h}_k} \frac{\mathbf{h}_k^H \mathbf{R}_k \mathbf{h}_k}{\mathbf{h}_k^H \mathbf{\Phi}_k \mathbf{h}_k} \tag{4.18}$$

which corresponds to the dominant generalized eigenvector, denoted as $\mathbf{u}_k$, associated with the largest positive generalized eigenvalue of the Hermitian matrix pair $(\mathbf{R}_k, \mathbf{\Phi}_k)$. Therefore, the ML channel estimate is given by

$$\hat{\mathbf{h}}_k = \frac{\sqrt{\gamma_k}}{|\tilde{\mathbf{q}}_k^H \mathbf{u}_k|} \mathbf{u}_k \tag{4.19}$$

Orthogonal Basis expansion

The solution of the CML estimate as a generalized eigenvalue problem requires the computation of the principal generalized eigenvector at each time slot, thus it may exhibit remarkable computational complexity in practice. In order to facilitate the calculation of the coarse estimate, we derive an equivalent channel estimation framework in which the channel of the k-th user is expressed as a linear combination of orthogonal vectors. Although any orthogonal basis can be used, in the case of random training vectors it is more natural to choose the beamforming vectors $\{\mathbf{q}_m\}_{i=1}^M$ as our orthonormal basis. In that case, the channel vector can be expressed as

$$\mathbf{h}_k^H = \sum_{m=1}^{M} \alpha_m \mathbf{q}_m^H \tag{4.20}$$

where α_m are the (complex) weights of the orthogonal expansion.

Consider, without loss of generality, that $\mathbf{q}_1$ corresponds to the best beam chosen by user k. Substituting (4.20) into (4.16), and solving the optimization problem (4.16) using Lagrange multipliers, we obtain that the optimal weights $\mathbf{b}_{opt} = [\alpha_2, \cdots, \alpha_M]^T$ equal to

$$\mathbf{b}_{opt} = -\alpha_1 \mathbf{A}^{-1} \mathbf{c} \tag{4.21}$$

where

$$\mathbf{c} = \left[\mathbf{q}_2^T\mathbf{R}_k^{-1}\mathbf{q}_1^*,\ \cdots\ ,\mathbf{q}_M^T\mathbf{R}_k^{-1}\mathbf{q}_1^*\right]^T$$

$$\mathbf{A} = \begin{bmatrix}\mathbf{q}_2^T\mathbf{R}_k^{-1}\mathbf{q}_2^* \\ \mathbf{q}_3^T\mathbf{R}_k^{-1}\mathbf{q}_2^*\end{bmatrix}$$

and $\alpha_1 = \sqrt{\gamma_k}$ so that the instantaneous CQI feedback constraint is satisfied.
Observing the similarity in the structure of matrix $\mathbf{A}$ with that of $\mathbf{Q}^T\mathbf{R}_k^{-1}\mathbf{Q}^*$, the computational complexity of the matrix inversion of $\mathbf{A}$ can be further reduced through use of block matrix decomposition. Denote $\mathbf{F} = \mathbf{Q}^T\mathbf{R}_k^{-1}\mathbf{Q}^*$, then

$$\mathbf{F} = \begin{bmatrix}\mathbf{q}_1^T\mathbf{R}_k^{-1}\mathbf{q}_1^* & \mathbf{q}_1^T\mathbf{R}_k^{-1}\mathbf{q}_2^* & \cdots & \mathbf{q}_1^T\mathbf{R}_k^{-1}\mathbf{q}_M^* \\ \mathbf{q}_2^T\mathbf{R}_k^{-1}\mathbf{q}_1^* & & & \\ \vdots & & \mathbf{A} & \\ \mathbf{q}_M^T\mathbf{R}_k^{-1}\mathbf{q}_1^* & & & \end{bmatrix}$$

The inverse $\mathbf{A}^{-1}$ can be easily obtained using the equation:

$$S_A^{-1}\begin{bmatrix}1 & -\mathbf{c}^H\mathbf{A}^{-1} \\ -\mathbf{A}^{-1}\mathbf{c} & S_A^{-1}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{c}\mathbf{c}^H\mathbf{A}^{-1}\end{bmatrix} = \mathbf{F}^{-1}$$

where $S_A = \mathbf{q}_1^T\mathbf{R}_k^{-1}\mathbf{q}_1^* - \mathbf{c}^H\mathbf{A}^{-1}\mathbf{c}$ is the Schur complement of $\mathbf{A}$ and $\mathbf{F}^{-1} = \mathbf{Q}^T\mathbf{R}_k\mathbf{Q}^*$ as $\mathbf{Q}$ is unitary.

ML Channel Estimation with Channel Norm Feedback

Consider now that the instantaneous CQI metric takes on the form of the channel norm, i.e. $\gamma_k = \|\mathbf{h}_k\|^2$. Clearly, the above direction independent CQI feedback provide less instantaneous spatial information than $\gamma_k = \left|\mathbf{h}_k^H\mathbf{z}_k\right|^2$. However, for users with large channel gain, thus for users that are more likely to be selected, channel norm feedback provide some form of additional spatial information (especially in Ricean channels). Moreover, the larger the channel gain, the more accurate this channel directional information. There is however a difference between Rayleigh and Ricean channels. In the Rayleigh case, the sign ambiguity on the direction cannot be eliminated, whereas in Ricean channels (non-zero mean) there is additional CDI on the sign of large channel realizations.

Similarly to Section 4.4.3, we formulate a *coarse* ML channel estimate assuming that the channel norm of user k is known, which results in the following constrained optimization problem:

$$\begin{aligned}\max_{\mathbf{h}_k}\quad & \mathbf{h}_k^H\mathbf{R}_k\mathbf{h}_k \\ s.t.\quad & \|\mathbf{h}_k\|^2 = \gamma_k\end{aligned} \tag{4.22}$$

The solution of (4.22) is given by

$$\hat{\mathbf{h}}_k = \sqrt{\gamma_k}\mathbf{u}_k \tag{4.23}$$

where $\mathbf{u}_k$ is the eigenvector associated with the largest eigenvalue of $\mathbf{R}_k$ and γ_k is chosen such that the constraint on the instantaneous channel norm is satisfied.

4.4.4 Interference-bounded Multiuser Eigenbeamforming with limited feedback

In the previous sections, we dealt with the problem of defining an efficient type of instantaneous CQI to be combined with long-term statistical channel knowledge. The proposed coarse ML channel estimate framework is mainly useful for the purpose of user selection. Although precoding design based on the channel estimates is feasible, providing good performance for small angle spreads, it is in general sensitive and prone to sign ambiguities.

In this paragraph, we exploit the long-term statistical information in a different way for the problem of joint scheduling and beamforming with limited feedback and focus on a practical, low-complexity scheme. In brief, each user k has a fixed, predefined beamforming vector, matched to the principal eigenvector of its channel correlation matrix $\mathbf{R}_k$. At each scheduling slot, the users are allowed to feed back two scalar values: the alignment between the channel and their predefined beamforming vectors and their channel norms. In turn, the scheduler selects the group of users that maximizes the system throughput using greedy user selection and by estimating of the received SINR based on inter-user interference bounds. Once the users to serve are identified, the precoding matrix contains the preferred beamforming vectors (principal eigenvectors) of the selected users. The proposed scheduling and precoding algorithm is outlined in Table 4.3.

Feedback Strategy We propose that each user feeds back the following two scalar values:

- its channel norm $\gamma_k^{(1)} = \|\mathbf{h}_k\|$.
- the alignment (angle) between its instantaneous channel vector and a preset normalized beamforming vector $\mathbf{w}_k$, i.e. $\gamma_k^{(2)} = \frac{\left|\mathbf{h}_k^H \mathbf{w}_k\right|}{\|\mathbf{h}_k\|}$.

The intuition behind this feedback policy is two-fold: in MIMO BC with partial CSIT an efficient scheduling set should contain users with large instantaneous channel gains and mutually quasi-orthogonal channel spatial signatures, as means to achieve both spatial multiplexing and multiuser diversity gains. The first scalar CQI $\gamma_k^{(1)}$ allows to select users with high channel gains as a means to benefit from multiuser diversity. In contrast to a feedback metric of type $\gamma_k = \left|\mathbf{h}_k^H \mathbf{w}_k\right|^2$, large $\gamma_k^{(1)}$ clearly identifies the users with the most favorable conditions, whereas the latter metric can be large even for users with moderate gains but whose vector channels are perfectly aligned with their beamforming vectors. The second scalar metric $\gamma_k^{(2)}$ provides a measure of the misalignment between the channel and the beamformer, and can be interpreted as a measure of the channel quantization error due to limited CSIT knowledge. In single-user settings, the quantization error affects only the received signal and is translated to a power offset. However, in multiuser SDMA settings, it can be shown that $\gamma_k^{(2)}$ plays a vital role in the estimation of the inter-user interference. Therefore, both $\gamma_k^{(1)}$ and $\gamma_k^{(2)}$ can be used as a means to estimate the inter-user interference due to limited feedback.

User Selection If a perfectly orthogonal set of beamforming vectors can be found, the above limited feedback is sufficient to achieve the same asymptotic sum rate as that of DPC. However, in practice, this is highly unlikely to be fulfilled and the remaining interference

Table 4.3: Resource Allocation Algorithm with Statistical CSIT

At each time slot

At receiver side

Compute & Feedback $\gamma_k^{(1)} = \|\mathbf{h}_k\| \rightarrow \text{BS} \qquad \forall k = 1, \cdots, K$

$\gamma_k^{(2)} = \frac{|\mathbf{h}_k^H \mathbf{w}_k|}{\|\mathbf{h}_k\|} \rightarrow \text{BS}$

At transmitter side

User selection

Step 1 Preselect users with $\gamma_k^{(1)} \cdot \gamma_k^{(2)} > \mu_{th}$, $\mathcal{Q} \rightarrow \mathcal{Q}'$

Set $\mathcal{R}_{LB}^* = 0$ and $\mathcal{S}^* = \emptyset$

For all $\mathcal{S} \in \mathcal{Q}'$ repeat

Step 2 Compute

$$\overline{I}_{UB_k}(\mathcal{S}) = (\gamma_k^{(2)})^2 \alpha_k(S) + \left(1 - (\gamma_k^{(2)})^2\right) \beta_k(S) + 2\rho_k \sqrt{1 - (\gamma_k^{(2)})^2}\, \delta_k(S)$$

Step 3 Compute $SINR_k^{LB}(\mathcal{S}) = \frac{\frac{P}{M}\left(\gamma_k^{(1)} \gamma_k^{(2)}\right)^2}{\frac{P}{M}\left(\gamma_k^{(1)}\right)^2 \overline{I}_{UB_k}(\mathcal{S}) + 1}$

Step 4 Compute $\mathcal{R}_{LB} = \sum_{k \in \mathcal{S}} \log_2 \left[1 + SINR_k^{LB}(\mathcal{S})\right]$

Step 5 If $\mathcal{R}_{LB} > \mathcal{R}_{LB}^*$, $\mathcal{R}_{LB} \rightarrow \mathcal{R}_{LB}^*$ and $\mathcal{S} \rightarrow \mathcal{S}^*$

Beamforming

Construct beamforming matrix $\mathbf{W}(\mathcal{S})$

cannot be calculated explicitly. For that, approximate expressions and bounds on the inter-user interference based on limited channel knowledge are of interest. For user $k \in \mathcal{S}$, the interference can be expressed as $I_k(\mathcal{S}) = \sum_{i \in \mathcal{S}, i \neq k} P_i |\mathbf{h}_k^H \mathbf{w}_i|^2 = \|\mathbf{h}_k\|^2 \overline{I}_k(\mathcal{S})$, where $\overline{I}_k(\mathcal{S})$ denotes the interference over the normalized channel $\overline{\mathbf{h}}_k$. Let $\overline{I}_k^{UB}(\mathcal{S})$ denote an upper bound on $\overline{I}_k(\mathcal{S})$, a lower bound on the SINR assuming is given by

$$\text{SINR}_k^{LB}(\mathcal{S}) = \frac{P_k \|\mathbf{h}_k\|^2 \cos^2(\angle \mathbf{h}_k, \mathbf{w}_k)}{\|\mathbf{h}_k\|^2 \overline{I}_k^{UB}(\mathcal{S}) + \sigma^2} = \frac{P_k \left(\gamma_k^{(1)} \gamma_k^{(2)}\right)^2}{\|\mathbf{h}_k\|^2 \overline{I}_k^{UB}(\mathcal{S}) + \sigma^2} \tag{4.24}$$

where $\overline{I}_k^{UB}(\mathcal{S})$ is also a function of $\gamma_k^{(1)}$ and $\gamma_k^{(2)}$. The scheduler aims to select the group of users that maximizes a lower bound on the sum rate as follows

$$\mathcal{S}^* = \arg\max_{\mathcal{S}} \sum_{k \in \mathcal{S}} \log\left(1 + SINR_k^{LB}(\mathcal{S})\right) \tag{4.25}$$

Analytic low and upper bounds on the inter-user interference under linear precoding are presented in detail in the following chapter. At this point, we propose to use the following upper bound [90]

$$\overline{I}_{UB_k}(\mathcal{S}) = (\gamma_k^{(2)})^2 \alpha_k(S) + \left(1 - (\gamma_k^{(2)})^2\right) \beta_k(S) + 2\rho_k \sqrt{1 - (\gamma_k^{(2)})^2}\, \delta_k(S) \tag{4.26}$$

where $\alpha_k(\mathcal{S}) = \mathbf{w}_k^H \left(\sum_{i \in S, i \neq k} \mathbf{w}_i \mathbf{w}_i^H\right) \mathbf{w}_k$, $\beta_k(\mathcal{S})$ denotes the largest eigenvalue of the matrix $\mathbf{U}_k^H \left(\sum_{i \in S, i \neq k} \mathbf{w}_i \mathbf{w}_i^H\right) \mathbf{U}_k$ and $\delta_k(\mathcal{S}) = \left\| \mathbf{U}_k^H \left(\sum_{i \in S, i \neq k} \mathbf{w}_i \mathbf{w}_i^H\right) \mathbf{w}_k \right\|$, where $\mathbf{U}_k \in \mathbb{C}^{M \times (M-1)}$ is an orthonormal basis spanning the null space of $\mathbf{w}_k$.

Linear Precoding Let the eigenvalue decomposition of the transmit correlation matrix be $\mathbf{R}_k = \mathbb{E}\{\mathbf{h}_k\mathbf{h}_k^H\} = \mathbf{V}_k\mathbf{\Sigma}_k\mathbf{V}_k^H$, where $\mathbf{\Sigma}_k$ is a diagonal matrix with the eigenvalues of $\mathbf{R}_k$ in descending order and $\mathbf{V}_k$ is a unitary matrix with the eigenvectors of $\mathbf{R}_k$. As a low complexity approach, we propose a system where each user has a preferred beamforming vector known both by the BS and the mobile terminal. As shown in [91], for single-user MIMO communications, given a certain user k with correlation matrix $\mathbf{R}_k$ the average rate is maximized by matching the beamforming vector to the principal eigenvector of its correlation matrix, $\mathbf{w}_k = \mathbf{v}_k^1$ (eigenbeamforming). Hence, we design each user's beamforming vector inspired by this single-user strategy. This multiuser eigenbeamforming transmission scheme can be seen as an equivalent codebook-based system where each user has a trivial codebook of size one. The codebook contains a single codevector, i.e. the principal eigenvector, and is updated at very low rate equal to the coherence time of the second-order statistics. We should also remark that under the prism of channel estimation framework, the interference-bounded eigenbeamforming can be seen as a method where the transmitter designs the precoder based on a coarse channel estimate given by

$$\hat{\mathbf{h}}_k = \|\mathbf{h}_k\| \cos(\angle\mathbf{h}_k, \mathbf{v}_k^1)\mathbf{v}_k^1 = \gamma_k^{(1)}\gamma_k^{(2)}\mathbf{v}_k^1 \tag{4.27}$$

4.4.5 Performance Evaluation

For the system evaluation, we assume that the channel evolves according to a specular model where the channel impulse response is a superposition of a finite number of paths, as described in Section 2.3.1. We consider ULA at the transmitter with antenna spacing $d = 0.4\lambda$, where $\lambda = 0.15m$ is the wavelength (here for 2GHz). We consider a narrowband, flat-fading Rayleigh (spatially correlated) channel where each user k has a different covariance matrix $\mathbf{R}_k$. The assumption that the receivers do not have the same correlation matrix is well motivated by the fact that in broadcast channels, the angle-of-arrival is different for each user because they are not physically co-located. The covariance matrix is computed using the assumption of Gaussian distributed scattering with angular spread σ_θ (standard deviation of the distribution) and is averaged over 60 time slots. Note that the angular spread corresponds to an angular spread sector of $2\sigma_\theta$ degrees. Unless otherwise stated, the BS is equipped with $M = 2$ antennas and the transmit SNR is set to 10dB.

ML Estimation with Beam Gain Information

We compare the sum rate achieved using the coarse ML channel estimate with BGI with that of optimal MMSE beamforming with full CSIT and with a random beamforming-based scheduling approach [9].

Figures 4.3 and 4.4 show the performance comparison as a function of the angle spread and the number of users, respectively. Once the group of selected users $\mathcal{S}$ is identified based on each user's coarse channel estimate, the transmitter obtains full CSIT *only* for the selected M users and designs the MMSE precoding matrix of user set $\mathcal{S}$. In RBF approach, the users are selected based on the maximum SINR [9]. We observe that the scheduler, despite using only coarse only channel estimate, is able to identify a better group of users than RBF for all angle spreads. When the angle spread is close to zero, our method closes the throughput gap with respect to the MMSE precoding with full CSIT. Note also that both

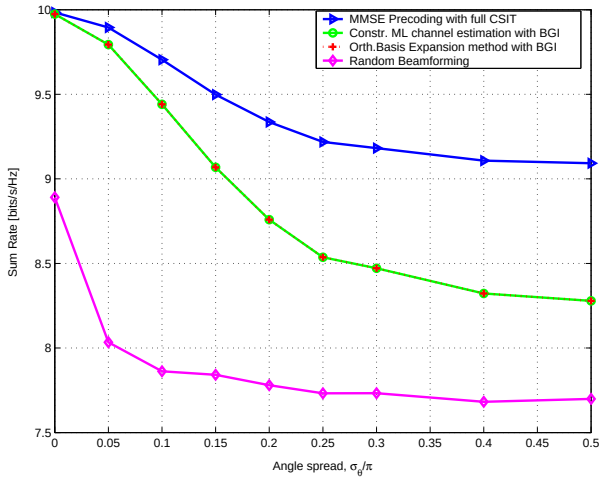

Figure 4.3: Sum rate performance versus angle spread of proposed ML estimation method for $M = 2$, and $K = 50$ users. Full CSIT is obtained for the selected users at a second step.

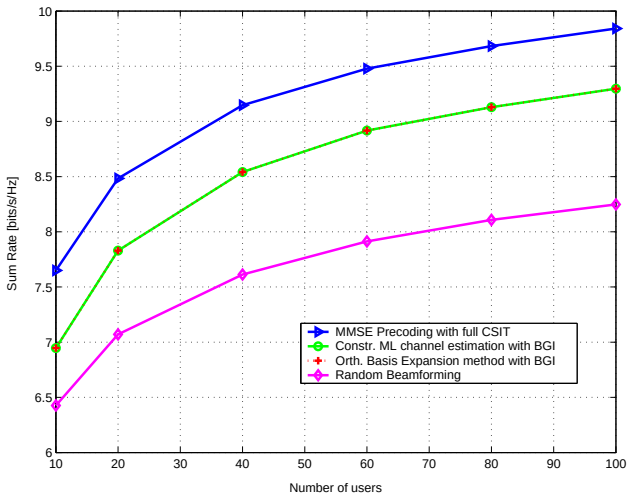

Figure 4.4: Sum rate performance versus the number of users of ML channel estimation method for $M = 2$, and $\sigma_\theta = 0.2\pi$. Full CSIT for the selected users is obtained for precoder design.

estimation methods exhibit exactly the same performance as they are equivalent solutions for the same optimization problem, differing only in terms of computational complexity.

In Figure 4.5, we evaluate the performance of the channel estimation methods when user selection and beamforming design are performed in one step based on coarse channel estimates. Evidently, MMSE precoding design based on the estimated channel is robust only in highly correlated channels, for which the channel estimate is closer to the real channel. Nevertheless, both estimation methods show - with no additional feedback - a significant throughput gain over RBF for angle spread less than 35 degrees.

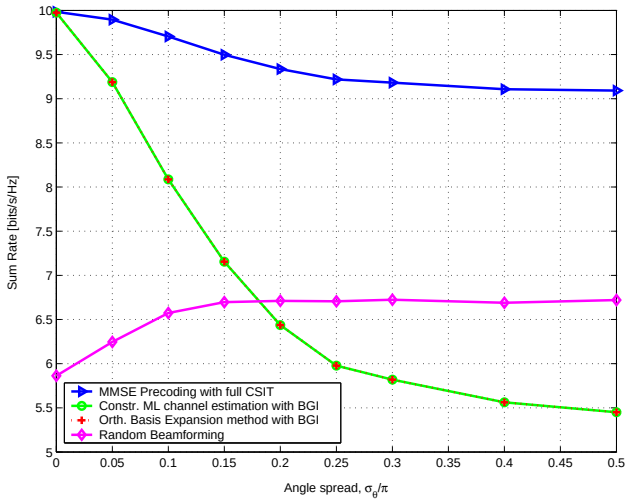

Figure 4.5: Sum rate performance versus angle spread of proposed ML estimation framework for $M = 2$, and $K = 50$ users. Partial CSIT is employed for precoding design.

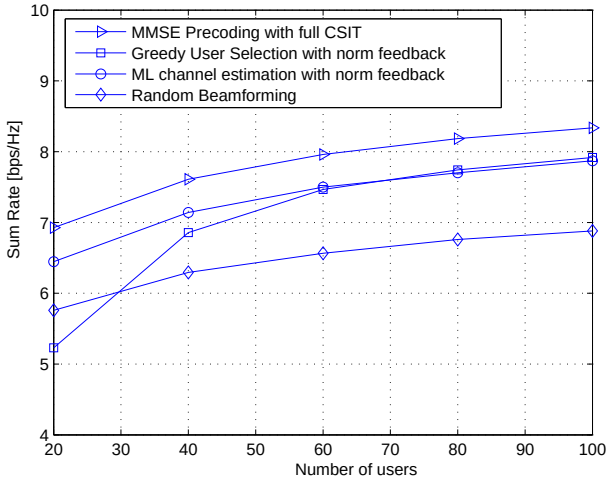

Figure 4.6: Sum rate as a function of the number of users for various user selection schemes with $M = 2$, antenna spacing $d = 0.5\lambda$ and $\sigma_\theta = 0.1\pi$.

ML Estimation with Norm Feedback

In Figures 4.6 - 4.8 we evaluate the ML channel estimation framework with norm feedback and greedy user selection algorithm (Table 4.2) as a function of K, antenna spacing d, and the angle spread σ_θ, respectively. As a benchmark, we also plot the sum rate of MMSE beamforming with full CSIT and RBF. In all methods, once the group of users to be scheduled is identified, the BS obtains full CSIT for the selected users in order to design the MMSE precoding matrix. Our methods show a clear gain over RBF for angle spread less than 35 degrees making it practical approach for cellular outdoor systems, as typical measurements in outdoor networks report angle spreads in the region less than 5-20 degrees at

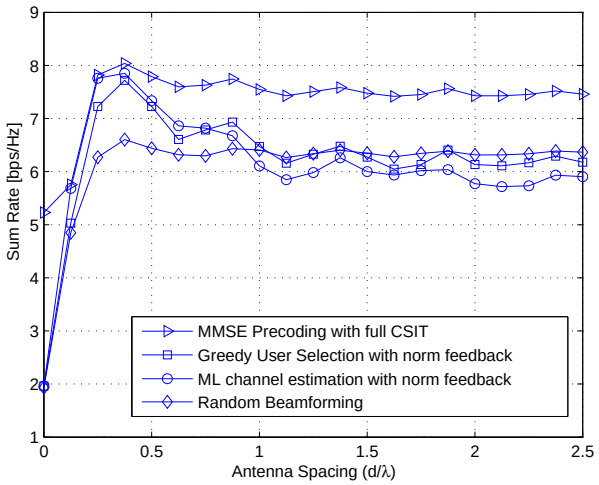

Figure 4.7: Sum rate as a function of antenna spacing for various user selection schemes with $M = 2$, $\sigma_\theta = 0.1\pi$ and $K = 50$ users.

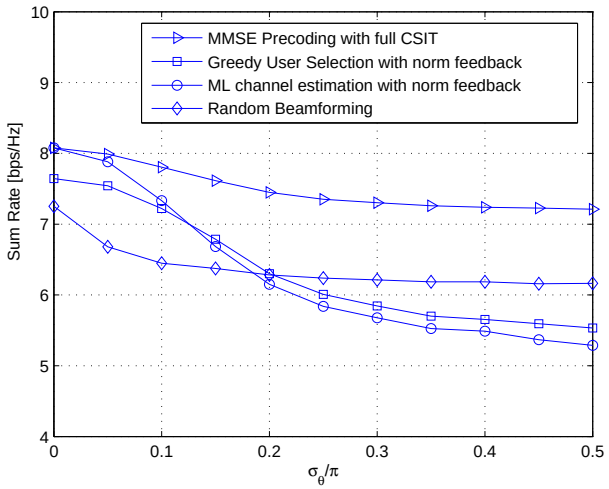

Figure 4.8: Sum rate as a function of angle spread for various user selection schemes with $M = 2$, antenna spacing $d = 0.5\lambda$ and $K = 50$ users.

the BS [89]. Interestingly, the antenna spacing can be optimized and it is found that about 0.4λ gives optimal results, as it gives the best tradeoff between resolution and suppression of spatial aliasing. Note that a small antenna spacing reduces transmit antenna diversity, however multiuser diversity can compensate for that during the phase of scheduling.

Interference-bounded Eigenbeamforming

We evaluate now the performance of interference-bounded multiuser eigenbeamforming (Interf.-bounded MU EigenBF). Figures 4.9 and 4.10 show the achieved sum rate of our proposed scheme as a function of the number of users and the angle spread, respectively.

For comparison, we also plot the performance of optimal MU eigenbeamforming with perfect CSIT and that of interference-bounded multiuser eigenbeamforming using full CSIT for user selection. As we can see, the performance of the proposed low-complexity scheme exceeds that of RBF but depends on the level of antenna correlation, i.e. angle spread σ_θ. Expectedly, gains are more pronounced for angle spread less than 45 degrees (outdoor cellular networks).

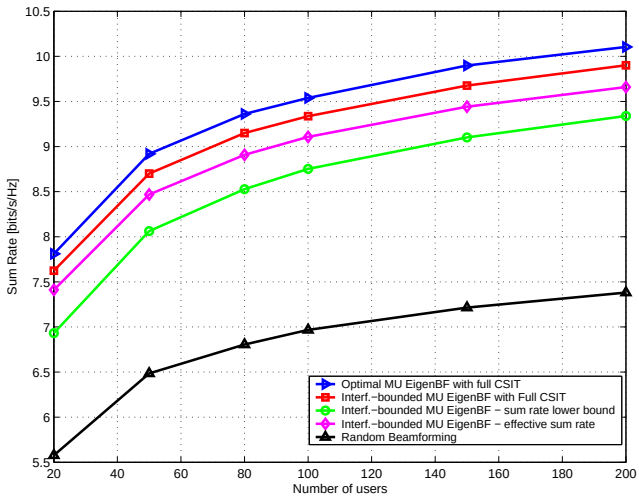

Figure 4.9: Sum rate as a function of the number of users for $M = 2$, and $\sigma_\theta = 0.1\pi$.

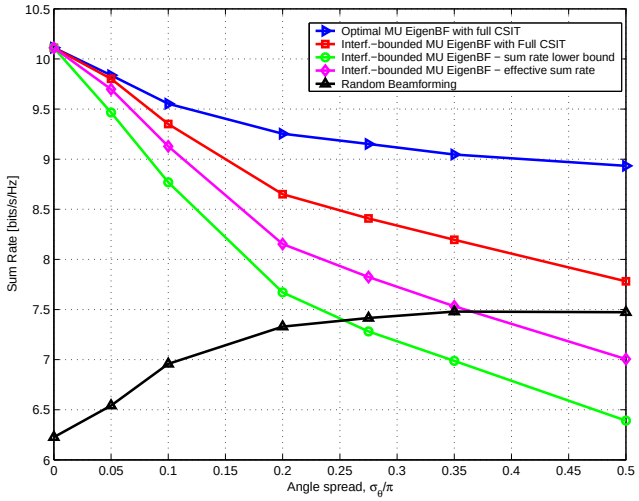

Figure 4.10: Sum rate as a function of angle spread for $M = 2$, antenna spacing $d = 0.4\lambda$ and $K = 100$ users.

4.5 Conclusions

In this chapter, we showed that the redundancy that arises in temporally and spatially correlated channels can be exploited in order to increase the system throughput by optimizing the SDMA scheduling decisions. In the first part, motivated by the fact that the performance of random beamforming degrades severely with low number of users, we show how exploiting channel time correlation we can alleviate this problem at minimal cost. The proposed memory-based opportunistic beamforming provides a way to close the gap to optimality for arbitrary number of users when the channel coherence time is large, e.g. in low mobility (indoor) settings. In the second part, we investigated spatially-correlated MISO channels and showed how statistical channel knowledge can be efficiently combined with instantaneous scalar channel feedback for the purpose of scheduling and linear precoding. Specifically, it was demonstrated that, in SDMA systems with channel-aware scheduling, it is sufficient to feed back a single scalar CSIT parameter - either the channel norm or beam gain information - in order to achieve near optimal sum-rate performance. We derived new scheduling metrics that have the advantage of accommodating statistical channel information and limited instantaneous channel feedback. A ML channel estimation framework has been established that is suitable for resource allocation in wide-area multi-antenna cellular systems. Finally, a low complexity precoding/scheduling algorithm, based on interference-bounded SDMA eigenbeamforming for spatially correlated MISO channels. All the above schemes exhibit performance close to that of complete CSI when the multipath angular spread per user at the BS is small enough, making these approaches suitable to wireless systems with elevated BS such as outdoor cellular networks, in which the elevation of the BS above the clutter decreases the angle spread of the multipath.

APPENDIX

4.A Proof of Proposition 4.1

To prove this statement, we can equivalently show that for the set of i.i.d. random unitary matrices, $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_{|\mathcal{U}|}\} \subset \mathcal{U}$, $\max\limits_{1 \leq i \leq |\mathcal{U}|} X_i$ converges to $\mathcal{R}^*$ for $|\mathcal{U}|$ sufficiently large and fixed number of users K. Thus, we want to show that $\forall \epsilon, \delta > 0$, $\exists\, |\mathcal{U}|$ such that $\Pr\{\max_{1 \leq i \leq |\mathcal{U}|} \leq \mathcal{R}^* - \epsilon\} \leq \delta$
As the sequence of unitary matrices $\{\mathbf{Q}_i\}_1^{|\mathcal{U}|}$ are i.i.d. r.v.'s, and $\{X_i\}_1^{|\mathcal{U}|}$ are also i.i.d. for i, using order statistics we have that

$$\Pr\left\{\max_{1 \leq i \leq |\mathcal{U}|} X_i \leq \mathcal{R}^* - \epsilon\right\} = [F_X(\mathcal{R}^* - \epsilon)]^{|\mathcal{U}|} \tag{4.28}$$

For a channel with memory $\mathcal{L}$, it is evidently meaningful to have $|\mathcal{U}| \geq \mathcal{L}$. As $0 \leq F_X(x) \leq 1$, asymptotically for $\mathcal{L} \to \infty$, we have that

$$\Pr\left\{\max_{1 \leq i \leq |\mathcal{U}|} X_i \leq \mathcal{R}^* - \epsilon\right\} \to 0 \tag{4.29}$$

Chapter 5

Limited Feedback Broadcast Channels based on Codebooks

5.1 Introduction

In the previous chapters, we investigated limited feedback approaches that can be mainly categorized as dimension reduction or projection techniques (cf. Section 2.8.2). The majority of our proposed solutions were built on - although not limited to - a random beamforming context. Conventional RBF [9] was mainly employed as a pre-scheduling technique, while the random beamformer were optimized during the precoding design phase. A limitation of RBF is that the resolution of CDI is fixed to $B_D = log_2 M$, thus the scheme cannot be extended for the case where additional CDI bits can be utilized. On the other hand, recent findings suggest that CDI is of particular importance in limited feedback multiuser multi-antenna systems, especially in the high power regime. As it was shown in [10], if channel inversion (zero-forcing) is employed as transmission strategy, the feedback load per user must increase approximately linearly with M and the transmit power (in dB) in order to achieve the full multiplexing gain. Moreover, up to now, we considered schemes where a random unitary precoder is first generated with *no a priori* CSIT and the BS collects low-rate (scalar) CQI from each user as a means to select a group of good users and potentially re-design the precoding matrix for the selected group.

In this chapter, we take on a quantization-based approach (cf. 2.8.1). The precoding matrix is not pre-designed (before the feedback phase), but is generated based on partial CSIT obtained by all active users. In other words, each user first reports some form of quantized CSIT, which in turn is used at the BS for user selection and precoding design. Several limited feedback approaches, imposing a bandwidth constraint on the feedback channel have been studied in point-to-point MIMO systems [54–56, 58]. In this context, each user feeds back finite precision (quantized) CSIT on its channel direction by quantizing its

normalized channel vector to the closest vector contained in a predetermined codebook. In this chapter, we consider a multi-antenna broadcast channel with $K \geq M$ users, in which each user is allowed to report feedback back to the BS via a finite rate feedback channel. This CSIT consists of B_D-bit quantized information on its channel vector direction, referred to as CDI, complemented with additional instantaneous CQI. CDI information is mainly employed for the purposes of precoding design, while CQI serves as a means to intelligently select M spatially separable users with large channel gains. This approach can be seen as an extension of RBF to a codebook containing $N_D > M$ beamforming vectors (not necessarily orthonormal). It has the ability to tune the feedback load per user, providing more flexibility in realistic finite rate feedback scenarios, in which the few feedback bits need to be split between channel directional and channel quality information. Our model is on the lines of work in [62] which extended the finite feedback rate model [10, 52] for the case of $K \geq M$. As transmission strategy, several beamforming methods have been investigated in the literature, including orthogonal unitary beamforming [92], transmit matched-filtering [93], and zero-forcing beamforming [64,93–95]. Note that in the above contributions, the channel gain feedback is considered unquantized for analytical simplicity.

A major part of this chapter focuses on the following question: "What type of scalar CQI information needs to be conveyed in order to achieve close-to-optimum performance?" Recent results show that if the scalar CQI contains information only on the channel norm, the sum rate growth is independent of the average SNR and the number of active users K [62, 64]. Therefore, the system becomes interference-limited for high SNR, and fails to achieve the optimum sum rate growth, even when the number of users goes to infinity (no multiuser diversity gain). This is due to the fact that an estimate on the inter-user interference is needed, and thus additional knowledge in the form of channel quantization error is necessary in order to achieve both multiplexing and multiuser diversity gains and approach the capacity with perfect CSIT.

The problem of efficient CQI design for sum-rate maximization with scheduling and linear precoding in the above finite rate feedback setting is addressed here. Our main contributions can be summarized as follows:

- We propose several scalar feedback metrics based on inter-user interference bounds, which encapsulate information on the channel gain, the channel direction, as well as on the quantization error. These metrics can be interpreted as estimates of the received SINR, which is generally unknown to the individual users that have knowledge only on their own channels.

- We employ these metrics in a system employing linear ZF beamforming on the quantized channel directions and greedy user selection. For that, we extend the greedy scheduling algorithm of [11] for the limited feedback case. This algorithm has the advantages that it does not depend on any a priori defined system parameter (such as quantized channels' orthogonality [62]) and is able to switch from multiuser to single-user transmission.

- Using the above precoding setting, we derive upper bounds on the instantaneous multiuser interference that allows us to analytically predict the worst case interference and a SINR lower bound in a system employing zero-forcing on the quantized channel directions.

- The system throughput is analyzed and its asymptotic optimality in terms of capacity growth (i.e. $M \log\log K$) is shown for $K \to \infty$. Sum rate upper bounds for the high SNR regime are also derived.
- Scheduling metrics suitable for switching the transmission mode from multiuser (SDMA) to single-user (TDMA) are proposed, based on a refined feedback strategy. We show that expectedly single-user mode is preferred as the average SNR increases, whereas multiuser mode is favored when the number of users increases.

5.2 System model

We consider a multi-antenna broadcast channel consisting of M antennas at the transmitter and $K \geq M$ single-antenna receivers. The finite rate feedback model presented in Section 2.9.1 is adopted and users quantize their channel directions using (2.56). The channel quantization of user k is denoted as $\hat{\mathbf{h}}_k$. For analytical simplicity, we adopt the ACVQ codebook design [63, 64] (cf. Section 2.9.2).

As linear precoding scheme, we use ZF beamforming on the quantized channel directions available at the BS. The beamforming matrix is then given by

$$\mathbf{W}(\mathcal{S}) = \hat{\mathbf{H}}(\mathcal{S})^{\dagger} = \hat{\mathbf{H}}(\mathcal{S})^{H} \left(\hat{\mathbf{H}}(\mathcal{S}) \hat{\mathbf{H}}(\mathcal{S})^{H} \right)^{-1} \tag{5.1}$$

where $\hat{\mathbf{H}}(\mathcal{S})$ is a matrix whose columns are the quantized channels $\hat{\mathbf{h}}_k$ (codevectors) of the users belonging to the group of selected users, denoted by $\mathcal{S}$. The normalized beamforming vector intended for the k-th user is denoted by $\mathbf{w}_k$ and equal power allocation across users is assumed. Clearly, non-linear precoding schemes or regularized inversion (MMSE precoding) can achieve a better sum rate than ZFBF. However, we use ZFBF for two main reasons. First, ZFBF is a linear precoding technique that can be implemented with reduced complexity and is asymptotically optimal at high SNR or for large K [11, 12]. Secondly, a significantly simpler and more tractable theoretical analysis can be accomplished using ZFBF, resulting in closed-form expressions for performance.

Some terms that will be used extensively in the following sections are:

- channel direction (normalized channel): $\bar{\mathbf{h}}_k = \mathbf{h}_k / \|\mathbf{h}_k\|$
- quantized channel: $\hat{\mathbf{h}}_k$
- quantization error: $\sin^2 \phi_k = \sin^2(\angle(\hat{\mathbf{h}}_k, \bar{\mathbf{h}}_k))$
- channel alignment: $\cos\theta_k = \left| \bar{\mathbf{h}}_k \mathbf{w}_k \right|$

5.3 CQI Feedback Design

5.3.1 Problem formulation

In multiuser SDMA downlink systems with more active users than transmit antennas ($K > M$), user selection has to be performed based on some properly chosen channel side information. The scheduling decisions depend in turn on the optimization criteria considered,

e.g. maximization of system throughput, maximization of user rates, fairness, delay minimization, etc. If the sum-rate maximization is considered as optimization criterion, the scheduled users need to exhibit:

- mutually orthogonal channel directions
- high channel gains

for close to optimum throughput performance. The spatial separability among users allows the BS to form non-interfering beams with no significant power penalty, whereas the importance of CQI is two-fold: it is used for identifying users with favorable channel conditions and it indicates the rate (coding and modulation order) at which the BS can transmit data to a particular user (link adaptation).

One challenge when designing feedback metrics is that information on received SINR is in principle not available to the individual users that only have knowledge of their own channels. The SINR measurement depends, among others, on the channel as well as on the number of other mobiles being simultaneously scheduled along with the user making the measurement. As user cooperation is not considered, the number of simultaneous users and the available power for each of them will generally be unknown at the mobile. However, in the large number of user case, simplifications arise, which give the user the possibility of estimating its SINR. This SINR estimate feedback enables the scheduler to identify users with large channel norms, as well as small quantization errors. In the following paragraphs, we study the problem of efficient design of channel quality feedback. Our objective is to derive scalar feedback metrics, denoted as γ_k, that allow us to exploit the multiuser diversity and achieve close to optimum sum-rate performance.

5.3.2 Bounds on average received SINR

The SINR of user $k \in \mathcal{S}$ under equal power allocation and ZFBF on the quantized channels is given by

$$\mathrm{SINR}_k = \frac{P|\mathbf{h}_k\mathbf{w}_k|^2}{\sum_{j\in\mathcal{S}\setminus\{k\}} P|\mathbf{h}_k\mathbf{w}_j|^2 + M} = \frac{P\left\|\mathbf{h}_k\right\|^2|\bar{\mathbf{h}}_k\mathbf{w}_k|^2}{\sum_{j\in\mathcal{S}\setminus\{k\}}\left(P\left\|\mathbf{h}_k\right\|^2|\bar{\mathbf{h}}_k\mathbf{w}_j|^2\right) + M} \tag{5.2}$$

The channel direction $\bar{\mathbf{h}}_k$ can be expressed in reference to its quantized version via the cross correlation indicator $\pi_k = \sin^2\phi_k = 1 - \left|\hat{\mathbf{h}}_k\bar{\mathbf{h}}_k^H\right|^2$ as $\bar{\mathbf{h}}_k = \sqrt{1-\pi_k}\hat{\mathbf{h}}_k + \sqrt{\pi_k}\hat{\mathbf{h}}_k^{\perp}$, where $\hat{\mathbf{h}}_k^{\perp}$ is the normalized projection of $\bar{\mathbf{h}}_k$ onto the orthogonal complement of $\hat{\mathbf{h}}_k$. Note that the actual phase information in $\bar{\mathbf{h}}_k$ is omitted since it is not relevant for SINR computation. Then, for the terms that appear in the interference we have that

$$|\bar{\mathbf{h}}_k\mathbf{w}_j|^2 = (1-\pi_k)|\hat{\mathbf{h}}_k\mathbf{w}_j|^2 + \pi_k|\hat{\mathbf{h}}_k^{\perp}\mathbf{w}_j|^2 = \pi_k|\hat{\mathbf{h}}_k^{\perp}\mathbf{w}_j|^2, \quad \forall k \neq j \tag{5.3}$$

since the ZF beamforming vector $\mathbf{w}_j$ is chosen orthogonal to the quantized channel vectors of all other users, i.e. $\hat{\mathbf{h}}_k\mathbf{w}_j = 0$ for all $k \neq j$, $k \in \mathcal{S}$. Then, using (5.3), eq. (5.2) can be written as

$$\mathrm{SINR}_k = \frac{P\left\|\mathbf{h}_k\right\|^2|\bar{\mathbf{h}}_k\mathbf{w}_k|^2}{P\left\|\mathbf{h}_k\right\|^2\pi_k\sum_{j\in\mathcal{S}\setminus\{k\}}|\hat{\mathbf{h}}_k\mathbf{w}_j|^2 + M} \tag{5.4}$$

Lower bound on the average received SINR

The received SINR can be normally measured at the received side. However, in a multiuser system, mobile terminals cannot calculate their received SINR in advance. This is due to the fact that each receiver k cannot estimate the inter-user interference since it does not have access to the beamforming vectors $\mathbf{w}_j$, $j \in \mathcal{S}$ and the group of selected users. Although the received SINR cannot be calculated explicitly at the receiver side, as all beamforming vectors $\mathbf{w}_j$, $j \neq k$ would lie in the null space of $\hat{\mathbf{h}}_k$, each receiver can calculated a bound on the expected interference caused by the other users. Therefore, a lower bound on the expected SINR with respect to the expected inter-user interference can be derived. Conditioned on $\mathbf{h}_k$ and $\hat{\mathbf{h}}_k$ and taking the expectation with respect to the interference terms $\mathbf{w}_j$, $j \in \mathcal{S} \setminus \{k\}$ we have [64]:

$$\begin{aligned}
\mathbb{E}\{\mathrm{SINR}_k\} &= \mathbb{E}\left\{ \frac{P|\mathbf{h}_k\mathbf{w}_k|^2}{\sum_{j\in\mathcal{S}\setminus\{k\}} P|\mathbf{h}_k\mathbf{w}_j|^2 + M} \right\} \\
&= \mathbb{E}\left\{ \frac{P\left\|\mathbf{h}_k\right\|^2 |\bar{\mathbf{h}}_k\mathbf{w}_k|^2}{P\left\|\mathbf{h}_k\right\|^2 \pi_k \sum_{j\in\mathcal{S}\setminus\{k\}} |\hat{\mathbf{h}}_k\mathbf{w}_j|^2 + M} \right\} \\
&\overset{(a)}{\geq} \frac{P\left\|\mathbf{h}_k\right\|^2 \mathbb{E}\left\{|\bar{\mathbf{h}}_k\mathbf{w}_k|^2\right\}}{P\left\|\mathbf{h}_k\right\|^2 \pi_k \mathbb{E}\left\{\sum_{j\in\mathcal{S}\setminus\{k\}} |\hat{\mathbf{h}}_k\mathbf{w}_j|^2\right\} + M} \\
&\overset{(b)}{\geq} \frac{P\left\|\mathbf{h}_k\right\|^2 \mathbb{E}\left\{|\bar{\mathbf{h}}_k\mathbf{w}_k|^2\right\}}{P\left\|\mathbf{h}_k\right\|^2 \sin^2\phi_k + M}
\end{aligned} \tag{5.5}$$

where (a) results from applying Jensen's inequality. The unit vectors $\hat{\mathbf{h}}_k^{\perp}$ and $\mathbf{w}_j$ are both isotropically distributed on the $(M-1)$ dimensional hyperplane orthogonal to $\hat{\mathbf{h}}_k$. As the distribution of $\mathbf{w}_j$ on this hyperplane depends only on $\hat{\mathbf{h}}_i$ for $i \in \mathcal{S} \setminus \{j,k\}$, then $\mathbf{w}_j$ is independent of $\hat{\mathbf{h}}_k^{\perp}$, for $j \neq k$. Thus, the inner product $|\hat{\mathbf{h}}_k\mathbf{w}_j|$ follows a beta distribution $B(1, M-2)$. Hence, the expected interference is given by

$$\mathbb{E}\left\{ \sum_{j\in\mathcal{S}\setminus\{k\}} |\hat{\mathbf{h}}_k\mathbf{w}_j|^2 \right\} = (|\mathcal{S}|-1)\cdot\frac{1}{M-1} \leq 1, \quad \text{for } |\mathcal{S}| \leq M \tag{5.6}$$

When M users are scheduled simultaneously, i.e., $|\mathcal{S}| = M$, inequality (b) (cf. eq.5.5) becomes tight.

Upper bound on the average received SINR

The inter-user interference is minimized by performing orthogonal transmission and selecting users with near-orthogonal quantized channel directions. In that case, we have that $\angle(\hat{\mathbf{h}}_k, \mathbf{w}_k) \approx 0$, and the average received SINR can be upper bounded by

$$\mathbb{E}\left\{\mathrm{SINR}_k\right\} \leq \frac{P\left\|\mathbf{h}_k\right\|^2 \cos^2\phi_k}{P\left\|\mathbf{h}_k\right\|^2 \sin^2\phi_k + M} \tag{5.7}$$

The above upper bound becomes tight when a set of perfectly orthogonal users can be found, in which case the received SINR is given by

$$\mathrm{SINR}_k = \frac{P\left\|\mathbf{h}_k\right\|^2 \cos^2\phi_k}{P\left\|\mathbf{h}_k\right\|^2 \sin^2\phi_k + M} \tag{5.8}$$

This is the actual received SINR under the assumption that $\mathcal{M} = M$ perfectly orthogonal users are scheduled.

CQI feedback metric I

In the previous paragraph, we saw that although the receivers do not have knowledge of the scheduling decisions and thus of $\mathbf{w}_j$, simple (upper) bounds on the expected received SINR can be obtained. Motivated by that, we consider that that each user can calculate and feed back information on its effective channel (SINR) by feeding back the following scalar metric

$$\gamma_k^I = \frac{P \|\mathbf{h}_k\|^2 \cos^2 \phi_k}{P \|\mathbf{h}_k\|^2 \sin^2 \phi_k + M} \tag{5.9}$$

proposed in parallel in [62, 94, 96, 97]. This type of CQI encapsulates information on the channel gain as well as the CDI quantization error, $\sin^2 \phi_k$. The above metric results from an upper bound on the average received SINR, which in turn is calculated based on the expected value of the inter-user interference due to quantized CSIT and using an upper bound on the expected received signal power. This CQI metric can be interpreted as an upper bound on each user's received SINR under the assumption that exactly M users will be served by M equipowered beams, designed based on quantized CDI. We should remark that this CQI value cannot be used directly for link adaptation. Clearly, it is not achievable and the only case where the received SINR equals the one predicted by (5.9) is when the M beamforming vectors at the transmitter are perfectly orthogonal (i.e. the columns of $\hat{\mathbf{H}}(\mathcal{S})$ are orthogonal), i.e. $\hat{\mathbf{H}}(\mathcal{S})$ is unitary and $\mathbf{W}(\mathcal{S}) = \hat{\mathbf{H}}(\mathcal{S})^H$. Despite this design limitation, it does however provide an efficient estimate of the multiuser interference at the receiver side and of the average SINR, allowing the scheduler to identify users with large channel gains and near-orthogonal channel directions. Moreover, this bound becomes more accurate when the number of active users K is increasing.

5.3.3 Lower bound on instantaneous received SINR

In the previous section, we studied bounds on the average received SINR and identified an efficient CQI metric. However, from a practical point of view, metric I has the limitation that is not achievable (upper bound), since in general the beamforming vectors are not perfectly orthogonal, especially in networks with low to moderate number of users. As a result, metric I may be useful for user selection purposes; however it cannot be employed for rate adaptation. If the system matches the coding rate and modulation order based on the γ_k^I value (cf. eq. (5.9)), the link will suffer from significant outage events since CQI metric I overestimates the received SINR. To circumvent that, the BS is required to ask for additional feedback from the selected users to perform rate allocation. This second step of feedback may be detrimental in terms of signaling overhead and protocol delays, and it is rather impractical in fast time-varying channels.

In order to avoid the need for this second step and to guarantee outage-free transmissions, we aim at finding a feedback metric that can be efficiently utilized for both scheduling and rate allocation simultaneously. For that, we propose to feedback a *lower bound* on the SINR rather than an upper bound. Additionally, we derive bounds on the *instantaneous* SINR and not on the average one. Our lower bound is based on:

- a lower bound on the received signal power.
- an upper bound on the *actual* multiuser interference.

Note also that the SINR estimated by (5.9) does not take into account the fact that a specific precoder is used for transmission scheme. Therefore, it neglects the effect of precoding and that of the corresponding misalignment between the quantized channel direction and the beamforming vector. In this paragraph, assuming that ZFBF is employed, we derive interference bounds that incorporate the power loss introduced by the misalignment between the instantaneous channel and the ZF beamformers.

Notation: The following orthogonality constraints, which that are used extensively below, can be imposed:

- Two quantized channel vectors $\hat{\mathbf{h}}_i$ and $\hat{\mathbf{h}}_j$ are ϵ-orthogonal if $\left|\hat{\mathbf{h}}_i \hat{\mathbf{h}}_j^H\right| \leq \epsilon$.
- The orthogonality between the quantized channel and the zero-forcing beamformer is defined as: $\xi \leq \left|\hat{\mathbf{h}}_k \mathbf{w}_k\right|$.
- The worst-case orthogonality between two zero-forcing beamformers is defined as $\epsilon_{ZF} = \max_{i,j \in \mathcal{S}} \left|\mathbf{w}_i^H \mathbf{w}_j\right|$.

Lower bound on received signal power

The quantity $|\bar{\mathbf{h}}_k \mathbf{w}_k|^2 = \cos^2(\angle(\bar{\mathbf{h}}_k, \mathbf{w}_k))$ that appears in the numerator of (5.4) can be bounded as follows: using the inequality $\angle(\bar{\mathbf{h}}_k, \mathbf{w}_k) \leq \angle(\bar{\mathbf{h}}_k, \hat{\mathbf{h}}_k) + \angle(\hat{\mathbf{h}}_k, \mathbf{w}_k)$, and the fact that the function $\cos x$ is monotonically decreasing in x for the interval of interest, we have $\cos^2(\angle(\bar{\mathbf{h}}_k, \mathbf{w}_k)) \geq \cos^2(\angle(\bar{\mathbf{h}}_k, \hat{\mathbf{h}}_k) + \angle(\hat{\mathbf{h}}_k, \mathbf{w}_k))$. Therefore, the received signal power of user k, denoted as S_k, can be lower bounded as

$$S_k = P|\mathbf{h}_k \mathbf{w}_k|^2 \geq P \left\|\mathbf{h}_k\right\|^2 \cos^2(\phi_k + \angle(\hat{\mathbf{h}}_k, \mathbf{w}_k)) = \tag{5.10}$$

Note that if the BS is able to find perfectly orthogonal user channels, the quantized channel direction $\hat{\mathbf{h}}_k$ and zero-forcing beamforming vector $\mathbf{w}_k$ coincide, and hence $\angle(\hat{\mathbf{h}}_k, \mathbf{w}_k) = 0$, yielding the following simple expression for the lower bound in (5.10): $S_k = P \left\|\mathbf{h}_k\right\|^2 \cos^2 \phi_k$. As the above lower bound cannot be calculated explicitly at the receiver side, we are obliged to use the orthogonality constraint ξ, which results in the following lower bound:

$$S_k^{LB1} \geq P \left\|\mathbf{h}_k\right\|^2 \cos^2(\phi_k + \arccos(\xi)) \tag{5.11}$$

A different lower bound on the received signal can be derived as follows:

$$\cos^2 \theta_k = |\bar{\mathbf{h}}_k \mathbf{w}_k|^2 = \left|\sqrt{1-\pi_k}\hat{\mathbf{h}}_k \mathbf{w}_k + \sqrt{\pi_k}\hat{\mathbf{h}}_k^{\perp} \mathbf{w}_k\right|^2$$

$$\overset{(a)}{\geq} \left||\sqrt{1-\pi_k}\hat{\mathbf{h}}_k \mathbf{w}_k| - |\sqrt{\pi_k}\hat{\mathbf{h}}_k^{\perp} \mathbf{w}_k|\right|^2$$

$$= (1-\pi_k)|\hat{\mathbf{h}}_k \mathbf{w}_k|^2 - 2\sqrt{(1-\pi_k)\pi_k}|\hat{\mathbf{h}}_k \mathbf{w}_k||\hat{\mathbf{h}}_k^{\perp} \mathbf{w}_k| + \pi_k|\hat{\mathbf{h}}_k^{\perp} \mathbf{w}_k|^2$$

where for (a) the inverse triangle inequality $||x| - |y|| \leq |x - y|$ is used.

Since the receivers cannot calculate the above lower bound (as they do not have access in the quantity $|\hat{\mathbf{h}}_k \mathbf{w}_k|$), the received signal power needs to be further bounded as:

$$S_k^{LB2} \geq P \left\|\mathbf{h}_k\right\|^2 \left(\xi^2 \cos^2 \phi_k - \xi\sqrt{1-\xi^2}|\sin(2\phi_k)|\right) \tag{5.12}$$

Bounds on Instantaneous Multiuser Interference

The inter-user interference of the k-th user can be expressed as:

$$I_k(\mathcal{S}) = P\left\|\mathbf{h}_k\right\|^2 \sum_{j\in\mathcal{S},j\neq k} \left|\overline{\mathbf{h}}_k\mathbf{w}_j\right|^2 = P\left\|\mathbf{h}_k\right\|^2 \overline{\mathrm{I}}_k(\mathcal{S}) \tag{5.13}$$

where $\overline{\mathrm{I}}_k(\mathcal{S})$ denotes the multiuser interference experienced by the k-th user over the normalized channel. Since the zero-forcing beamformers satisfy the orthogonality constraint $\hat{\mathbf{h}}_k\mathbf{w}_j = 0, \forall j \neq k$, we have that

$$\overline{\mathrm{I}}_k(\mathcal{S}) = \pi_k \sum_{j\in\mathcal{S},j\neq k} \left|\hat{\mathbf{h}}_k^{\perp}\mathbf{w}_j\right|^2 = \pi_k \overline{\mathrm{I}}_k^{\perp}(\mathcal{S}) \tag{5.14}$$

In order to bound the inter-user interference, we need to bound either the term $\overline{\mathrm{I}}_k^{\perp}(\mathcal{S})$ or $\overline{\mathrm{I}}_k(\mathcal{S})$ that cannot be calculated at the receiver side (since $\mathbf{w}_j$ are not known in advance to mobile terminals). Let us define the matrix $\mathbf{\Psi}_k(\mathcal{S}) = \sum_{j\in S,j\neq k}\mathbf{w}_j\mathbf{w}_j^H$, the operator $\lambda_{max}\{\cdot\}$, which returns the largest eigenvalue, and $\mathbf{U}_k \in \mathbb{C}^{M\times(M-1)}$ an orthonormal basis spanning the null space of $\mathbf{w}_k$. We recall that the ZF beamformers are considered as unit-norm vectors.

Theorem 5.1: *Given a set of normalized beamforming vectors* $\{\mathbf{w}_k\}, k \in \mathcal{S}$*, the normalized interference term* $\overline{I}_k(\mathcal{S})$ *is upper bounded by*

$$\overline{I}_k(\mathcal{S}) \leq \cos^2\theta_k\alpha_k(\mathcal{S}) + \sin^2\theta_k\beta_k(\mathcal{S}) + 2\sin\theta_k\cos\theta_k\omega_k(\mathcal{S}) \tag{5.15}$$

where

$$\begin{cases} \alpha_k(\mathcal{S}) = \mathbf{w}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{w}_k \\ \beta_k(\mathcal{S}) = \lambda_{max}\left\{\mathbf{U}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{U}_k\right\} \\ \omega_k(\mathcal{S}) = \left\|\mathbf{U}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{w}_k\right\| \end{cases} \tag{5.16}$$

Proof. The proof is given in Appendix 5.A. □

For notation simplification, we drop below the dependence on $\mathcal{S}$.

Lemma 5.1: *The worst-case orthogonality of a set of* M *zero-forcing beamforming vectors and the alignment with the normalized channel* ($\cos\theta_k$) *are bounded as a function of* $\cos\phi_k$ *for* $\epsilon < \frac{1}{M-1}$ *as follows:*

$$\epsilon_{ZF} \leq \vartheta \tag{5.17}$$

$$\cos\theta_k \geq \frac{\left|\cos\phi_k - \sqrt{\vartheta}\right|}{1+\vartheta} \tag{5.18}$$

$$\text{with } \vartheta = \frac{\epsilon}{1-(M-1)\epsilon}$$

Proof. The proof is given in Appendix 5.B. □

It is worth noting that the above lemma provides another lower bound on the received signal power, given by:

$$S_k^{LB3} \geq \frac{P}{(1+\vartheta)^2}\left\|\mathbf{h}_k\right\|^2\left(\cos\phi_k - \sqrt{\vartheta}\right)^2 \tag{5.19}$$

Based on this result, we have:

Theorem 5.2: *Given a group of ϵ-orthogonal users with cardinality $|\mathcal{S}| = M$, the received SINR in a system employing zero-forcing beamforming is lower bounded by*

$$SINR_k \geq \frac{P \|\mathbf{h}_k\|^2 \cos^2 \theta_k}{P \|\mathbf{h}_k\|^2 \overline{I}_k^{UB1} + M} \tag{5.20}$$

where

$$\overline{I}_k^{UB1} = (M-1)\left(\vartheta \cos\theta_k + \sin\theta_k\right)^2 - (M-2)(1-\vartheta)\sin^2\theta_k \tag{5.21}$$

with $\cos\theta_k = \frac{|\cos\phi_k - \sqrt{\vartheta}|}{1+\vartheta}$ *and* $\vartheta = \frac{\epsilon}{1-(M-1)\epsilon}$.

Proof. The proof is given in Appendix 5.C. □

An additional inter-user interference upper bound $\overline{I}_k^{UB2}$ can be derived by trying to upper bound the term $\overline{\mathrm{I}}_k^{\perp}(\mathcal{S})$.

CQI feedback metric II

Motivated by the above lower bound on the instantaneous SINR (cf. Theorem 2), we propose that each user feeds back to the BS the following scalar metric

$$\gamma_k^{II} = \frac{S_k^{LBx}}{\overline{I}_k^{UBx} + M} \tag{5.22}$$

where the wildcard 'x' can be replaced by 1, 2 or 3 for the received signal (LB) and 1 or 2 for the interference (UB). In the numerical results section, we only simulate the following metric:

$$\gamma_k^{II} = \frac{S_k^{LB3}}{\overline{I}_k^{UB1} + M} = \frac{\frac{P}{(1+\vartheta)^2} \|\mathbf{h}_k\|^2 (\cos\phi_k - \sqrt{\vartheta})^2}{P \|\mathbf{h}_k\|^2 \overline{I}_k^{UB1} + M} \tag{5.23}$$

In order to calculate (5.23), the receiver has to know the orthogonality system parameters ϵ and ξ and to assume that $\mathcal{M} = M$ exactly users will be scheduled. The basic difference between (5.9) and (5.23) is on the estimation of the inter-user interference and the received signal power. In (5.9) the interference is replaced by an upper bound on its average value, i.e. $\mathbb{E}\left\{\sum_{j\in\mathcal{S}\setminus\{k\}} \frac{P}{M} \|\mathbf{h}\|^2 |\bar{\mathbf{h}}_k \mathbf{w}_j|^2\right\} \leq \frac{P}{M} \|\mathbf{h}\|^2 \sin^2\phi_k$, where for CQI metric II an upper bound on the instantaneous multiuser interference (cf. eq. 5.21) is used instead.

CQI metric I can be viewed as an estimation of received SINR assuming that the quantized channel $\hat{\mathbf{h}}_k$ and the zero-forcing beamformer $\mathbf{w}_k$ coincide, i.e. $\angle(\hat{\mathbf{h}}_k, \mathbf{w}_k) = 0$. This assumption becomes valid for large number of users K. Therefore, in metric I, the approximation $\cos^2(\angle(\bar{\mathbf{h}}_k, \mathbf{w}_k)) \approx \cos^2(\angle(\bar{\mathbf{h}}_k, \hat{\mathbf{h}}_k))$ is used, whereas in CQI metric II the power loss introduced by the angle shift due to the misalignment of $\hat{\mathbf{h}}_k$ and $\mathbf{w}_k$ is taken into account (usong Lemma 5.1).

Evidently, the two proposed metrics coincide for $\epsilon = 0$ since we have $\cos^2\theta_k = \cos^2\phi_k$, thus $\bar{I}_{UB_k} = \sin^2\theta_k = \sin^2\phi_k$ and $\frac{(\cos\phi_k - \sqrt{\vartheta})^2}{(1+\vartheta)^2} = \cos^2\phi_k$. Hence, metric II (5.23) takes exactly the form of metric I (5.9).

5.3.4 SDMA/TDMA transition with limited feedback

In the previous paragraphs, we tried to derive efficient scalar CQI metrics. An upper bound on the expected SINR as well as a lower bound on the actual received SINR have been proposed as useful metrics that allow the BS to benefit from multiuser diversity and achieve near-optimal sum rate. A common underlying assumption of both $\gamma_k^{(I)}$ and $\gamma_k^{(II)}$ is that $\mathcal{M} = M$ users are necessarily scheduled. However this can be a major drawback as in MIMO broadcast channels with partial CSIT, it is not guaranteed that multiuser transmission (full SDMA) always outperforms single-user transmission (TDMA). There are several contexts in which it is beneficial from a capacity point of view to softly transit to TDMA by switching off beams and communicating with $\mathcal{M} < M$ users, especially in the high SNR regime and/or for low number of users. The inaccuracy in the multiuser interference calculation introduced by limited channel knowledge is detrimental in the high power regime, in which the system becomes interference-limited and its sum rate saturates. Motivated by the above claim, we are interested here to find a feedback strategy that offers the desirable flexibility between SDMA of various orders and TDMA, as a means to achieve linear capacity growth at any SNR range.

CQI feedback strategy for adaptive SDMA/TDMA

In order to obtain flexibility on estimating the resulting inter-user interference and hence the users' SINRs for various values of $\mathcal{M}$, a different form of CQI feedback needs to be considered. In [90] we already presented the idea of decomposing the CQI feedback in two scalar values, which was further exploited in [94]. In addition to the codevector index (CDI), we propose that each user feeds back:

- the channel norm $\gamma_k^{(1)} = \|\mathbf{h}_k\|$
- the square of the alignment $\gamma_k^{(2)} = \cos^2 \phi_k$

The decomposition of the CQI into two scalars enables the BS to calculate more accurate SINR estimates for any set of scheduled users with cardinality $\mathcal{M} \leq M$. This is due to the ability of calculating more accurately the inter-user interference by having the CQI in the form of channel gain and quantization error. Note that under a certain finite and fixed rate feedback constraint, each scalar value is quantized with reduced accuracy compared to the case of only one scalar CQI metric (e.g. metric I and II). The effect of CQI quantization is studied through simulations in Section 5.7, where it can be seen that the reduced precision of the two scalar CQIs does not reduce the sum-rate performance compared to the one scalar CQI case.

Scheduling metrics

At the transmitter side, the scheduler based on the decomposed CQI and CDI information estimates the received SINR. User selection can be performed based on the following scheduling metric, referred to as *metric III*:

$$\zeta_k^{III} = \frac{P \|\mathbf{h}_k\|^2 \rho_k^2}{P \|\mathbf{h}_k\|^2 \overline{I}_{UBd_k} + \mathcal{M}} \tag{5.24}$$

where

$$\rho_k^2 = \cos^2(\phi_k + \angle(\hat{\mathbf{h}}_k, \mathbf{w}_k)) \tag{5.25}$$

and

$$\overline{I}_{UBd_k} = \rho_k^2 \alpha_k(\mathcal{S}) + \left(1-\rho_k^2\right) \beta_k(\mathcal{S}) + 2\rho_k \sqrt{1-\rho_k^2}\omega_k(\mathcal{S}) \tag{5.26}$$

which can be explicitly calculated at the transmitter using (5.16).

The scheduling decision metrics are denoted with ζ_k in order to distinguish them from the CQI feedback metrics denoted with γ_k. The values ζ_k^{III} and ζ_k^{IV} are calculated on the BS and are not fed back to the BS from the users, whereas γ_k^I and γ_k^{II} are reported back by the mobile and also serves as user selection decision metrics from the scheduler.

In the ideal case of $\epsilon \to 0$, we have that $\bar{\mathrm{I}}_{UBd_k} \to \sin^2\phi_k$, and when $\epsilon = 0$ the following *scheduling metric IV*, interpreted as an upper bound on the received SINR, can be used at the BS

$$\zeta_k^{IV} = \frac{P\left\|\mathbf{h}_k\right\|^2 \rho_k^2}{P\left\|\mathbf{h}_k\right\|^2 \sin^2\phi_k + \mathcal{M}} \tag{5.27}$$

Actually, setting ϵ to be inversely proportional to K, it can be seen from Lemma 5.1 that as $K \to \infty$, $\epsilon_{ZF} \to 0$, and $\cos\theta_k \to \cos\phi_k$. Thus, for $K \to \infty$, $\bar{\mathrm{I}}_{UB_k} = \sin^2\phi_k$ and hence (5.27) converges to (5.9) for $\mathcal{M} = M$.

Note that scheduling metric (5.24) provides a more accurate SINR estimate compared to (5.23) as $\rho_k^2 \geq \frac{(\cos\phi_k - \sqrt{\vartheta})^2}{(1+\vartheta)^2}$ and $\bar{I}_{UBd_k} \leq \bar{I}_{UB_k}$. Furthermore, as $\rho_k^2 \leq \cos^2\phi_k$, we have that $\gamma_k^I \geq \gamma_k^{IV} \geq \zeta_k^{III} \geq \zeta_k^{II}$. An important difference with practical implications is that γ_k^{II} and ζ_k^{III} calculate SINR values that can be supported by the user channel and can be used for outage-free rate allocation, whereas γ_k^I and ζ_k^{IV} are upper bounds that are not achievable in general. A major advantage using the decomposed CQI feedback strategy is that the BS can adapt the number of scheduled user $\mathcal{M}$ depending on the average SNR, the number of users K and the amount of multiuser interference. This results in a adaptive multi-mode scheme where the transmitter switches between single-user transmission mode (TDMA with $\mathcal{M} = 1$) and multiuser mode (SDMA with $2 \leq \mathcal{M} \leq M$).

5.4 User Selection Schemes

At the transmitter side, the CQI metrics proposed in Section 5.3 are employed in order to select users with favorable channel conditions and orthogonality properties. We present here two user selection algorithms for scheduling in systems employing linear beamforming. Our optimization objective is to maximize the system capacity, therefore the optimum scheduling policy is to select through exhaustive search, the $\mathcal{M} \leq M$ among K users that maximize the system throughput. Nevertheless, since the complexity of such a combinatorial optimization problem is prohibitively high for large K, we resort to low-complexity scheduling strategies based on greedy user selection (see e.g. [11,12,62]).

5.4.1 Greedy-SUS algorithm

We first review a heuristic scheduling algorithm based on semi-orthogonal user selection (SUS) proposed in [12,62]. Using CQI_k defined in equations (5.9), (5.23), (5.24), and (5.27),

and $\text{CDI}_k = \hat{\mathbf{h}}_k$, $k = 1, \ldots, K$, the BS selects up to M out of K users at each time slot. The algorithm is outlined in Table 5.1. The first user is selected from the set $\mathcal{Q}^0 = \{1, \ldots, K\}$ of cardinality $\left|\mathcal{Q}^0\right| = K$ as the one having the highest channel quality, i.e. $k_1 = \arg\max_{k \in \mathcal{Q}^0} \text{CQI}_k$. The $(i+1)$-th user, for $i = 1, \ldots, M-1$, is selected as $k_{i+1} = \arg\max_{k \in \mathcal{Q}^i} \gamma_k$ among the user set $\mathcal{Q}^i$ with cardinality $\left|\mathcal{Q}^i\right| \leq K$, defined as $\mathcal{Q}^i = \left\{k \in \mathcal{Q}^{i-1} \mid |\hat{\mathbf{h}}_k \hat{\mathbf{h}}_j^H| \leq \epsilon \ \forall j \in \mathcal{S}\right\}$. The orthogonality ϵ between the quantized channels is system parameter that has to be set in advance. Evidently, if ϵ is very large, the selected user group may experience significant multiuser interference, reducing the system sum rate. Conversely, if ϵ is too small, the scheduler cannot find enough semi-orthogonal users to transmit to, and less than M users are multiplexed.

We should remark that greedy user selection results in multiuser diversity reduction. The metric CQI_{k_i} of the selected user at the i-th step of the algorithm, k_i is not always selected among K users. At each step, CQI_{k_i} is equal to the maximum of $\mathcal{K}_i = \left|\mathcal{Q}^{i-1}\right|$ i.i.d. random variables with common CDF $F_\gamma(x)$. Obviously, the multiuser diversity gain of $\log|\mathcal{Q}^0| = \log K$ is experienced only from the first selected user and decreases with the user index.

5.4.2 Greedy-US algorithm

A limitation of the previous scheduling algorithm is that it does not generally adapt the number of selected users and forces to select M users. As a result, full SDMA transmit mode is always supported independently of the system operating points, namely K and SNR. Therefore, it is more appropriate to be used with metrics of the type of γ_k^I and γ_k^{II}. In contrast with MIMO broadcast channels with complete CSIT, in limited feedback systems it is not guaranteed that multiuser transmission (SDMA) always outperforms single-user transmission (TDMA). There are several contexts in which it is beneficial from a capacity point of view to softly transit to TDMA by switching off beams and communicating with $\mathcal{M} < M$ users. Soft SDMA/TDMA switching can be realized by feeding back two scalar values (strategy 3 and 4). In order to exploit the flexibility of this decoupled feedback approach and adapt the number of scheduled users, we need to modify the greedy selection procedure. For that, we generalize a standard greedy user selection (GUS) algorithm with perfect CSIT [11] for the case of quantized CSIT, summarized in Table 5.2. We denote $\mathcal{S}_i$ the set of selected users up to the i-th step, and $\mathcal{R}(\mathcal{S}_i) = \sum_{k \in \mathcal{S}_i} \log_2(1 + \text{CQI}_k)$, with CQI_k being: γ_k^I, γ_k^{II}, ζ_k^{III} or ζ_k^{IV}. The user with the highest rate (equivalently SINR metric) among K users is first selected, and at each iteration, a user is added only if the sum rate (based on the estimated SINR) is increased. At each step, it is important to re-process the set of previously selected users (thus, re-calculating the zero-forcing beamformers) once a user is added to the set $\mathcal{S}_i$. We should note that if γ_k^I, γ_k^{II} are used, the algorithm becomes trivial and coincides with greedy-SUS algorithm, since the one scalar CQI information does not allow us to re-process the precoding strategy each time a user is added.

As stated before, the value of the orthogonality constraint ϵ affects the performance of the greedy-SUS algorithm. If ϵ is set too small, the multiuser diversity gain decreases, and the user set $\mathcal{Q}^i$ can be empty before M quasi-orthogonal users are found. The optimal value decreases with K, as the probability of finding M semi-orthogonal users among K is

larger, however it is difficult to be optimized analytically. A main advantage of Greedy-US algorithm compared to Greedy-SUS is that it does not require to predetermine any system parameter ϵ, as it can be calculated and optimized at each step based on the feedback values $\gamma_k^{(1)}$, $\gamma_k^{(2)}$ and the candidate users.

5.5 Performance Analysis

We analyze the sum-rate performance of the above CQI feedback metrics combined with user selection algorithms under the ollowing system configurations:

- *Strategy 1*: CQI feedback and scheduling metric γ_k^I combined Greedy-SUS algorithm.
- *Strategy 2*: CQI feedback and scheduling metric γ_k^{II} combined Greedy-SUS algorithm.
- *Strategy 3*: CQI feedback metrics $\gamma_k^{(1)}$ and $\gamma_k^{(2)}$ combined Greedy-US algorithm and scheduling metric ζ_k^{III}.
- *Strategy 4*: CQI feedback metrics $\gamma_k^{(1)}$ and $\gamma_k^{(2)}$ combined Greedy-US algorithm and scheduling metric ζ_k^{IV}.

Closed-form throughput expressions can be derived using similar tools as in Section 3.2; however little or no insight is gained from these involved expressions. For that, we focus on two practically relevant regimes: the large number of users regime ($K \to \infty$) and the high power regime ($P \to \infty$). XX We decide to investigate the performance using the lower bound on the received SINR γ_k^{II} , since it provides a lower bound on the achievable sum rate of *strategy 1* as well.

5.5.1 Asymptotic (in K) sum-rate analysis

We consider the asymptotic case of $K \to \infty$ and M fixed. As (5.23) is a lower bound on the user's SINR, the exact received SINR that can be supported by the channel is unknown at the BS (but higher than γ_k^{II}). Thus, the expected sum rate $\mathcal{R}$ of *strategy 2* is lower bounded as

$$\mathcal{R} \geq \mathbb{E}\left\{\sum_{i=1}^{M} \log_2\left(1+\gamma_{k_i}^{II}\right)\right\} = \mathbb{E}\left\{\sum_{i=1}^{M} \log_2\left(1+\max_{k\in\mathcal{K}_i}\gamma_k^{II}\right)\right\} \tag{5.28}$$

where $\mathcal{K}_i = \left|\mathcal{Q}^{i-1}\right|$ captures the multiuser diversity gain reduction due to Greedy-SUS algorithm. A bound on the cardinality of $\left|\mathcal{Q}^{i}\right|$ can be calculated through the probability that a user i in $\mathcal{Q}^i$ is ϵ-orthogonal to users in $\mathcal{Q}^{i-1}$, which is equal to $I_{\epsilon^2(i,M-i)}$, where $I_x(a,b)$ is the regularized incomplete beta function. The k_i-th user is the one that has the maximum CQI metric among $\mathcal{Q}^{i-1}$, whose cardinality converges to the following value (by using the law of large numbers) [98, 99]:

$$\left|\mathcal{Q}^{i-1}\right| \approx KPr\{\mathbf{h}\in\mathcal{Q}^{i-1}\} \geq KI_{\epsilon^2}(i-1, M-i+1)$$

with $\left|\mathcal{Q}^{0}\right| = K$.

Note that for large number of users K and choosing $\epsilon = 1/\log K$, so that $\lim_{K\to\infty} KI_{\epsilon^2}(i-1, M-i+1) = \infty$ and $\lim_{K\to\infty} \epsilon = 0$, we have that $\gamma_k^{II} \to \gamma_k^I$. Therefore, before establishing the asymptotic sum-rate optimality of strategy 2, we need to derive the statistics of γ_k^I.

Distribution of γ_k^I For the statistics of the upper bound on the expected received SINR we have:

Lemma 5.2: *The distribution function of $F_\gamma(x)$ of the CQI feedback metric γ_k^I is given by*

$$F_\gamma(x) = \begin{cases} 1 - N_D \frac{e^{-Mx/P}}{(1+x)^{M-1}} & x \geq \frac{1-\delta}{\delta} \\ 1 - N_D \frac{e^{-Mx/P}}{(1+x)^{M-1}} + \mathcal{T} & 0 \leq x < \frac{1-\delta}{\delta} \end{cases} \tag{5.29}$$

where $\mathcal{T} = \frac{1}{\Gamma(M-1)} \left[N_D \frac{e^{-Mx/P}}{(1+x)^{M-1}} \left(\Gamma\left(M-1, \delta(x+1)v\right) - \Gamma\left(M-1, v\right) \right) \right]$, $v = \frac{Mx}{P(1-\delta-\delta x)}$, and $\Gamma(a,x)$ is the (upper) incomplete gamma function.

Proof. The proof is given in Appendix 5.D. □

Note that the first branch of the CDF was first derived in [62]. In the Appendix, we provide a different proof for $x \geq \frac{1-\delta}{\delta}$ as well as the expression of $F_\gamma(x)$ for $x < \frac{1-\delta}{\delta}$.

Asymptotic Sum-rate Optimality If we denote $\beta = \frac{1}{N_D} \cdot (P/M)^{M-1}$, the following results the asymptotic optimality of the proposed limited feedback scheme (strategy 2):

Theorem 5.3: *The sum rate of the proposed scheme $\mathcal{R}$ converges to the optimum capacity of MIMO broadcast channel $\mathcal{R}_{opt}$, for $K \to \infty$, i.e.*

$$\lim_{K\to\infty} (\mathcal{R}_{opt} - \mathcal{R}) = \lim_{K\to\infty} \left[M \log_2 \frac{1 + \frac{P}{M} \log K}{1 + \frac{P}{M} \log\left(\frac{K}{\beta}\right)} \right] = 0 \tag{5.30}$$

with probability one.

Proof. The proof is given in Appendix 5.E. □

The above theorem implies that the optimal $M \log\log K$ capacity growth can be achieved for $K \to \infty$ by using the proposed metric (5.23) with greedy user selection algorithm and ZF beamforming on the channel quantizations. Note also that this notion of sum rate convergence is stronger than that capacity ratio convergence, i.e. $\lim_{K\to\infty} \frac{\mathcal{R}}{\mathcal{R}_{opt}} = 1$, as the latter cannot guarantee that there is unbounded SINR gap between the proposed scheme and the optimal one (full CSIT case).

5.5.2 Sum-rate analysis in the interference-limited region

In this section, we study the sum rate achieved by strategy 2 in the high-power regime (interference-limited region). For $P \to \infty$, it can be shown that

Theorem 5.4: *The sum rate of strategy 2 at high SNR with finite B_D and K is upper bounded by*

$$\mathcal{R} \leq \frac{M}{M-1} \left(B_D + \frac{1}{\log 2} H_K \right) \tag{5.31}$$

where $H_K = \sum_{k=1}^{K} \frac{1}{k}$ is the harmonic number (K-th partial sum of the harmonic series).

Proof. The proof is given in Appendix 5.F. □

The above theorem implies that the system becomes interference-limited and its sum rate converges to a constant value at high SNR, even for arbitrary large but finite B_D and K. This behavior is inherited to all finite fixed-rate feedback-based MISO systems due to the quantization error, which results to loss of the multiplexing gain at high SNR. Furthermore, as $\partial\mathcal{R}/\partial M < 0$, the sum rate is a monotonically decreasing function with M, implying that at high SNR the sum rate is maximized by using $M = 1$ beam.

The asymptotic behavior of H_K is given by the standard Euler expansion as $H_K \sim \log K + \gamma_{em} - \sum_{n=1}^{\infty} \frac{\mathcal{B}_n}{K^n} \sim \gamma + \log K + \frac{1}{2K} - \frac{1}{12K^2} + O(\frac{1}{120K^4})$, where $\gamma_{em} \approx 0.57721566...$ is the Euler-Mascheroni constant and $\mathcal{B}_n$ denotes the n-th Bernoulli number. A sharp lower and upper bound of the harmonic sequence for any natural $K \geq 1$ is derived in [100] as follows:

$$\frac{1}{2K + \frac{1}{1-\gamma_{em}} - 2} \leq H_K - \log K - \gamma_{em} < \frac{1}{2K + \frac{1}{3}} \tag{5.32}$$

Therefore, for large number of users ($K \to \infty$), $\lim_{K\to\infty} H_K = \log K + \gamma_{em}$. Thus, the sum rate at high SNR and $K \to \infty$ exhibits logarithmic growth with K due to the multiuser diversity gain. In other words, for fixed B_D, although only a fraction of the full multiplexing gain is achieved ($r = \frac{M}{M-1}$), the sum rate scales as $\log K$, compensating for the loss in degrees of freedom and 'shifting' the interference-limited region to higher SNR values.

5.6 MIMO Broadcast Channels with Finite Sum Rate Feedback Constraint

In the previous paragraphs, the term quantization refers to the CDI feedback since we implicitly consider that the reported CQI values are not quantized. In other words, each user uses B_D bits for CDI feedback and infinite number of bits for reporting the scalar CQI value. In this section, we impose a finite sum rate feedback constraint, which implies that each user can only utilize B_{tot} bits to report both CDI and CQI channel knowledge.

5.6.1 Multiuser Diversity - Multiplexing Tradeoff in MIMO BC with Limited Feedback

We present here a tradeoff between multiuser diversity and spatial multiplexing gain that arises in SDMA downlink with finite sum rate feedback constraint, where each user sends CDI (based on a codebook) and CQI feedback. This is mainly due to the following fact: on one hand, CDI is sufficient to achieve the full multiplexing gain, but cannot simultaneously exploit multiuser diversity gain of order $\log\log K$. Furthermore, CDI feedback load needs to scale appropriately depending on system parameters (e.g., operating SNR, number of active users, etc.) in order to guarantee throughput that scales linearly with the number of transmit antennas [10]. On the other hand, in order to achieve the optimal double logarithmic capacity scaling with K, CQI has to be conveyed at the transmitter as a means to perform efficient user selection and control the effect of CDI quantization error. Therefore, in a system where only a finite number of feedback bits per user can be conveyed, the amount of bits used for CSIT quantization has to be shared between CDI (multiplexing gain) and CQI quantization (multiuser diversity). While CDI quantization incurs in loss of multiplexing

gain, CQI quantization leads to a degradation of the multiuser diversity benefit. Therefore, assuming that each user is allowed to feed back a finite number of bits results in a tradeoff between the spatial multiplexing gain

$$r = \lim_{P\to\infty} \frac{\mathcal{R}(P)}{\log P} \tag{5.33}$$

and the multiuser diversity gain

$$m = \lim_{K\to\infty} \frac{\mathcal{R}(P,K)}{r \log\log K} \tag{5.34}$$

Although the term is inspired by the popular diversity-multiplexing tradeoff (DMT) in MIMO point-to-point systems [101], there are several fundamental differences. The multiuser diversity differs from single-user diversity in the sense that the latter refers to the ability for the multiple antennas to receive the same information across different paths, while in multiuser systems, different information is transmitted and received by different users. The multiuser diversity gain increases with the number of active users in the cell, while the available multiplexing gain remains equal to $min(M,K)$, regardless of the value of K. Hence, with full CSIT both multiuser diversity and multiplexing gain can be attained since they scale with different magnitudes, K and SNR respectively. In contrast, in the DMT for single-user MIMO systems, both diversity and multiplexing gain scale with the SNR, thus the above two gains cannot be fully achieved simultaneously.

5.6.2 Finite Sum Rate Feedback Model

We present here a general framework which is referred to as finite sum rate feedback model. Each receiver k is constrained to have a limited total number of feedback bits B_{tot}, available for quantizing its channel vector and feeding back its quantized CSIT back to the BS. From this total amount of bits, B_D bits are used to represent the CDI $\bar{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|$ based on a predetermined codebook, and B_Q bits are used for scalar quantization of the real-valued CQI. This model is depicted in Fig. 5.1. In [10] it was shown that channel directional

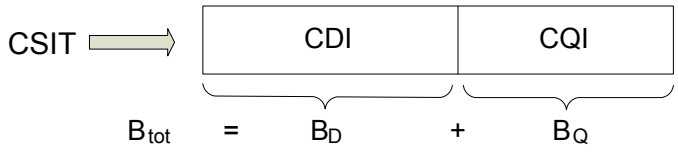

Figure 5.1: Finite Sum Rate Feedback Model.

information can be used to achieve the full multiplexing gain when the feedback load B_D scales appropriately. In a multiuser context with $K > M$, the CDI does not provide any information on users' channel gains, thus it is not sufficient to be used for efficient user selection and to exploit multiuser diversity gain. Hence, additional instantaneous, low-rate CQI is required. We try here to reveal the interplay between K, SNR, and feedback load B_D and B_Q, in order to exploit in the best possible way the degrees of freedom available in a multiuser MIMO downlink, i.e. the multiuser diversity and spatial multiplexing gain. We aim at characterizing the tradeoff that results from the sum feedback rate constraint per

user (B_{tot}), by identifying the optimal feedback rate allocation (split) in order to achieve both gains. Simply speaking, we try to quantify how many feedback bits are worth CDI and CQI.

CQI Quantization

As channel quality indicator, we consider instantaneous scalar feedback, denoted as γ_k, which can take on various forms and is evidently a certain function of the current channel realization $\mathbf{h}_k$ (i.e., $\gamma_k = f(\mathbf{h}_k)$). We assume that γ_k are i.i.d. random variables with probability density function (PDF) $f_\gamma(\gamma)$.

Let $\mathcal{X} = \{q_0 < q_1 < \ldots < q_{N_Q}\}$ and $\mathcal{Y} = \left\{\gamma_{q_0} < \ldots < \gamma_{q_{N_Q-1}}\right\}$ be the input decision levels and the output representative levels (reconstruction values), respectively, of an N_2-level quantizer $\mathcal{Q}(\cdot)$ defined as:

$$\mathcal{Q}(\gamma) = \gamma_{q_i} \quad \text{if} \;\; q_i \leq \gamma < q_{i+1} \;\; 0 \leq i \leq N_Q - 1$$

with $q_0 = 0$ and $q_{N_Q} = \infty$. A partition region (quantization level) is defined as $Q_i = [q_i, q_{i+1}), 0 \leq i \leq N_Q - 1$. Each user sends the corresponding quantization level index i back to the transmitter using $B_Q = \lceil \log_2 N_Q \rceil$ bits. In order to minimize the outage probability, we assume the following conservative but reliable quantization rule $\gamma_{q_i} = q_i$.

The distortion $\mathcal{D}$ introduced by the quantizer is given by

$$\mathcal{D}_{N_Q} = \mathbb{E}\left[e(\gamma, \mathcal{Q}(\gamma))\right] = \sum_{i=0}^{N_Q-1} \int_{Q_i} e(\gamma, \gamma_{q_i}) f_\gamma(\gamma) d\gamma \tag{5.35}$$

where $e(\cdot,\cdot)$ is an error weighting function. Necessary conditions for optimal quantizer $\mathcal{Q}$:

$$\frac{\partial \mathcal{D}_{N_Q}}{q_i} = 0 \quad i = 0, \ldots, N_Q$$

$$\frac{\partial \mathcal{D}_{N_Q}}{y_{q_i}} = 0 \quad i = 0, \ldots, N_Q - 1$$

5.6.3 Problem Formulation

Our objective is to dynamically allocate bits to CDI and CQI feedback (as shown in Fig. 5.1) given a total amount of feedback bits B_{tot}, so that the capacity of the multiuser MIMO downlink $\mathcal{R}(B_D, B_Q)$ is maximized. In the described finite sum rate feedback model, the optimal feedback rate allocation that maximizes the capacity can be formulated in the following constrained optimization problem:

$$\left.\begin{array}{r} \max\limits_{B_D, B_Q} \mathcal{R}(B_D, B_Q) \\ s.t. \;\; B_D + B_Q = B_{tot} \end{array}\right\} \tag{5.36}$$

Let $\mathcal{W}_{k,m}$ be the event that a user k is selected for transmission among K users over beam m. Capitalizing on the analysis of [102], we calculate the probability of this event conditioned on the fact that γ_k falls into the quantization level Q_j

$$\Pr\left(\mathcal{W}_{k,m} | \gamma_k \in Q_j\right) = \sum_{n=0}^{K} \frac{1}{n+1} \cdot \left(\begin{array}{c} K-1 \\ n \end{array} \right) \cdot \mathcal{P}_1 \cdot \mathcal{P}_2$$

where

$$\mathcal{P}_1 = \Pr\{n \text{ users other than user } k \in Q_j\} = (\Pr(\gamma \in Q_j))^n$$

and

$$\mathcal{P}_2 = \Pr\{(K-n-1) \text{ users other than user } k \in Q_w, w < j\}$$
$$= \left(\Pr\left(\gamma \in \bigcup_{w<j} Q_w\right)\right)^{K-n-1}$$

We assume here that if more than one user lie in Q_j, a random user is scheduled for transmission. Note also that for i.i.d. channels, $\Pr(\mathcal{W}_{k,m}|\gamma_k \in Q_j)$ is not dependent on k and m. Using that $(\Pr(\gamma \in Q_j)) = F_\gamma(q_{j+1}) - F_\gamma(q_j)$, and after some manipulations, one can show that

$$\Pr(\mathcal{W}_{k,m}|\gamma_k \in Q_j) = \frac{[F_\gamma(q_{j+1})]^K - [F_\gamma(q_j)]^K}{K(F_\gamma(q_{j+1}) - F_\gamma(q_j))} \tag{5.37}$$

Consider now that the quality indicator γ is a function of each user's SINR. In that case, the effect of CDI quantization will be reflected on the distribution of γ. Hence, the CQI contains information both on channel gain and CDI quantization error. For instance, the value γ can be a lower or an upper bound on the achievable SINR or even the achievable SINR value itself. Suppose now that the metric γ represents a lower bound on the SINR. Then, the rate of the selected user k, $\mathcal{R}_k$ is given by

$$\mathcal{R}_k \geq \sum_{j=0}^{N_Q-1} \int_{\gamma \in Q_j} \Pr(\mathcal{W}_{k,m}|\gamma_k \in Q_j) \log_2(1+\gamma) f_\gamma(\gamma) d\gamma$$

$$= \sum_{j=0}^{N_Q-1} \int_{Q_j} \log_2(1+\gamma) \cdot \frac{[F_\gamma(q_{j+1})]^K - [F_\gamma(q_j)]^K}{K(F_\gamma(q_{j+1}) - F_\gamma(q_j))} \cdot f_\gamma(\gamma) d\gamma$$

The system throughput $\mathcal{R}(B_D, B_Q)$ can be lower bounded by

$$\mathcal{R}(B_D, B_Q) = \sum_{k \in \mathcal{S}} \mathcal{R}_k \geq$$

$$\sum_{k \in \mathcal{S}} \sum_{j=0}^{2^{B_Q}-1} \int_{Q_j} \log_2(1+\gamma) \frac{[F_\gamma(q_{j+1})]^K - [F_\gamma(q_j)]^K}{K(F_\gamma(q_{j+1}) - F_\gamma(q_j))} f_\gamma(\gamma) d\gamma \tag{5.38}$$

where B_D is contained both in $F_\gamma(\gamma)$ and $f_\gamma(\gamma)$.

Unfortunately, the optimization problem (5.36) does not seem to accept closed-form solution. Additionally, the solution depends on the quantization levels $q_i, 0 \leq i \leq N_Q - 1$ to be considered, thus different CQI quantization strategies will yield different solutions. To circumvent the complexity of numerical brute force optimization and the non-linearity of this optimization problem, numerical algorithms based on dynamic programming and providing a global optimum can be used [103, 104].

5.6.4 Decoupled Feedback Optimization

In this section, instead of determining jointly the optimal feedback bit split, we follow a low-complexity approach. The problem is decomposed in a two-step optimization procedure:

we first find the optimal number of CDI bits required to guarantee full multiplexing gain, implying that the feedback load allocated to CQI is $B_Q = (B_{tot} - B_D)$, and optimizing the 2^{B_Q} quantization levels by using (5.38). This approach is motivated by results showing that lack of accurate CDI feedback in the high SNR regime results in loss of multiplexing gain. As the loss of the pre-log factor of M is more detrimental on the achievable sum-rate than the loss in multiuser diversity, we believe that for such kinds of feedback rate optimizations, an efficient rule of thumb is to guarantee appropriate CDI feedback rate to achieve close-to-full spatial multiplexing gain.

To illustrate this feedback optimization technique, we apply this decoupled approach to feedback optimization of strategy 1. Based on the asymptotic growth of (5.9) for large K given in [64], we derive the scaling of CDI feedback load, which in turn determines the remaining CQI feedback bits. We define the power gap (per user) between the SINR of the above scheme, $\mathrm{SINR_I}$, and that of zero-forcing with perfect CSI, $\mathrm{SINR_{ZF}}$ as the ratio $\frac{\mathrm{SINR_I}}{\mathrm{SINR_{ZF}}} = \alpha$. Note that this power gap is translated to a rate gap. In order to achieve full multiplexing gain for finite K, the number of CDI bits B_D per receiver k should scale according to:

$$B_D = (M-1)\log_2(P/M) - (1-b)\log_2 K + c \tag{5.39}$$

where $c = \log_2(K/\mathcal{K}_i)$ is a constant capturing the multiuser diversity reduction at each step i of the greedy-SUS algorithm due to the ϵ-orthogonality constraint between scheduled users. As $b < 1$, having more users in the cell, a smaller number of feedback bits B_D per user is required in order to achieve full multiplexing gain. For example, in a system with $M = 4$ antennas, $K = 30$ users and SNR $= 10$ dB, when a 3-dB SINR gap is considered, each user needs to feed back at least $B_D = 9$ bits.

Scaling of CDI feedback bits at high SNR regime

In the high SNR regime, the role of CDI is more critical due to the effect of quantization error [10]. For $P \to \infty$ and fixed K, based on asymptotic results of [64], we can show that the feedback load should scale as

$$B_D = (M-1)\log_2 P - \log_2 K \tag{5.40}$$

For instance, for a system with $M = 4$ antennas, SNR $= 20$ dB and $K = 60$ users, $B_D = 14$ bits are required to guarantee full multiplexing gain. Expectedly, the feedback load B_D at high SNR is larger than that of (5.39). Thus, it is more beneficial to allocate more feedback bits on the quantization of channel direction (B_D) at high SNR, and assign less bits for CQI (B_Q).

5.7 Performance Evaluation

In order to assess the sum-rate performance of the proposed schemes, simulations have been performed under the following conditions: $M = 2$ transmit antennas, orthogonality constraint $\epsilon = 0.4$ and codebooks generated using random vector quantization (RVQ) [10,61]. The achieved sum rate is compared with two alternative transmission techniques for the MIMO downlink, random beamforming [9] and zero-forcing beamforming with perfect CSI (and equal power allocation).

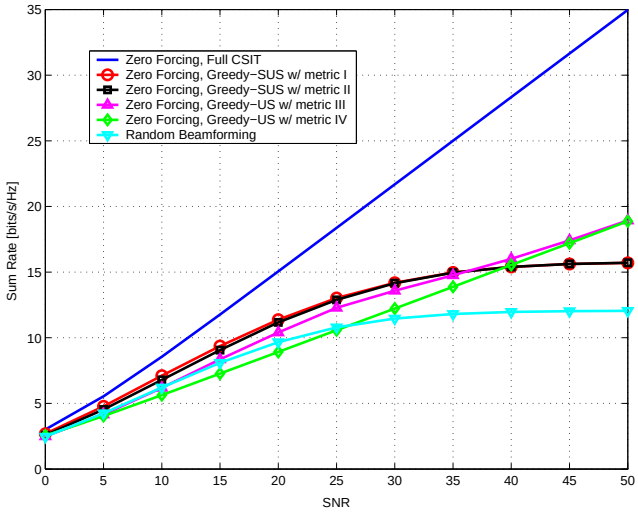

Figure 5.2: Sum rate versus the average SNR for $B_D = 4$ bits, $M = 2$ transmit antennas and $K = 30$ users.

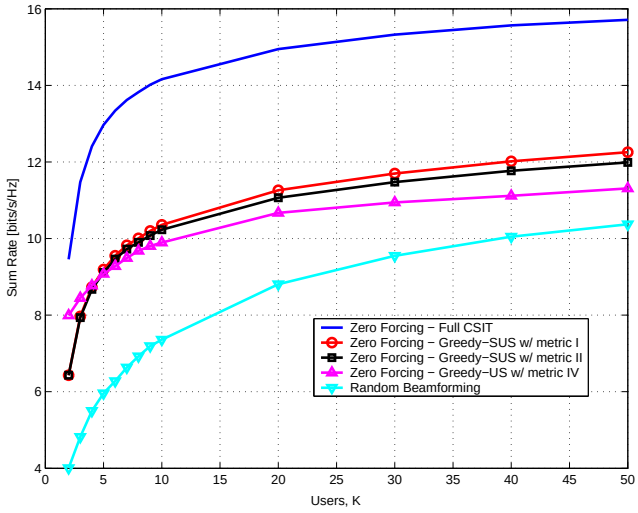

Figure 5.3: Sum rate as a function of the number of users for $B_D = 4$ bits, $M = 2$ transmit antennas and SNR $= 20$ dB.

Unquantized CQI

In Figure 5.2 we compare the sum rates of the proposed CQI metrics as a function of the average SNR, for K=30 users and $B_D = 4$ bits per user for CDI quantization. Strategy 1 (metric I) and strategy 2 (metric II) offer similar throughput, exhibiting however the same bounded behavior at high SNR, where the system capacity converges to a constant value. Given a fixed number of CDI bits B_D, the system becomes unavoidably interference-limited at high SNR and the rate curves flatten out. This is due to the fact that the accuracy of knowledge (resolution) of the quantization error remains constant for SNR increasing,

as well as due to that Greedy-SUS forces the system to schedule always M users. On the contrary, the scheme using strategy 3 (with feedback of two scalar values) provides higher flexibility by transmitting to $\mathcal{M} \leq M$ users, thus keeping a linear sum-rate growth in the interference-limited region and converging to TDMA for $P \to \infty$ (where $\mathcal{M} = 1$ is optimal).

In Figure 5.3 we plot the sum rate as a function of K for average SNR = 20dB and codebook of size $B_D = 4$ bits. It can be seen that all scalar metrics can efficiently benefit from the multiuser diversity gain. The gap with respect to the full CSIT case can be decreased by increasing the feedback load B_D. However, the slightly different scaling of strategy 4 (scheduling metric IV) is due to the fact that the user selection based on sum-rate estimates decides that $\mathcal{M} < M$ beams ought to be used. Since the calculations are performed using incomplete CSIT, erroneous or loose estimations can sometimes lead to sub-optimal decisions in terms of the number of users to be scheduled. Furthermore, in a system with fixed orthogonality factor ϵ, the accuracy of the lower bound (γ^{II}) does not improve as K increases. On the other hand, the upper bound (γ^{I}) becomes more realistic due to a higher probability of finding orthogonal quantized channels, hence yielding slightly better performance.

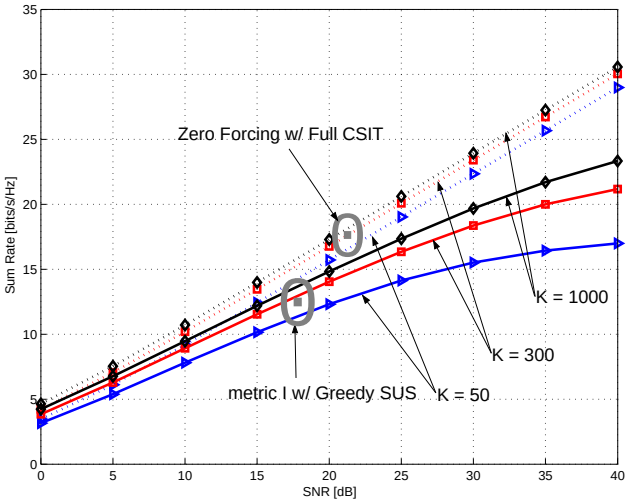

Figure 5.4: Sum rate performance as a function of the average SNR for increasing value of the number of users, with $B_D = 4$ bits of feedback per user and $M = 2$ transmit antennas.

We study now the performance of strategy 1 with different number of users and CDI feedback bits in order to obtain an insight on the CQI feedback metric design and the results of our asymptotic analysis. Figures 5.4 and 5.5 show a sum-rate comparison as a function of the average SNR, illustrating the multiplexing gain achieved by strategy 1. In both figures, it can be seen that given a fixed number of feedback bits B_D, the system becomes unavoidably interference-limited at high SNR and the rate curves flatten out. Given a fixed codebook size, Figure 5.4 shows the performance improvement of the proposed CQI metrics as the number of active users increases. Indeed, it can be seen that the performance gap between the scheme with perfect CSIT and strategy 1 with partial CSIT is narrower for K increasing. Although the scheme enters the interference limited regime for large values of P, the larger the number of users, the higher the SNR value for which the sum rate converges to a bound.

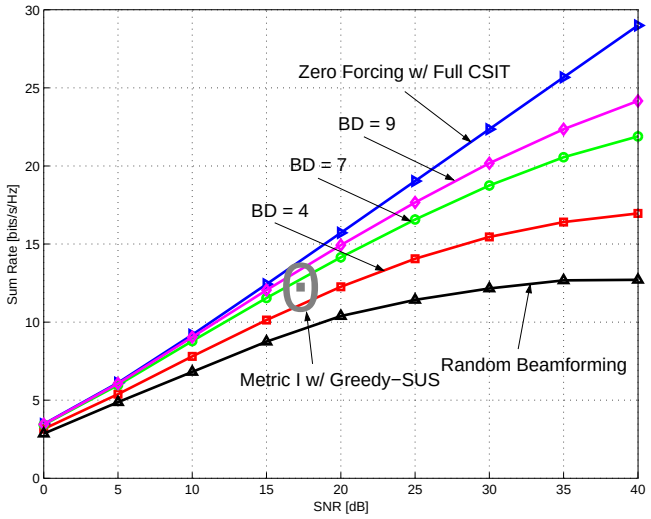

Figure 5.5: Sum rate as a function of the average SNR for increasing codebook size, $M = 2$ transmit antennas, and $K = 50$ users.

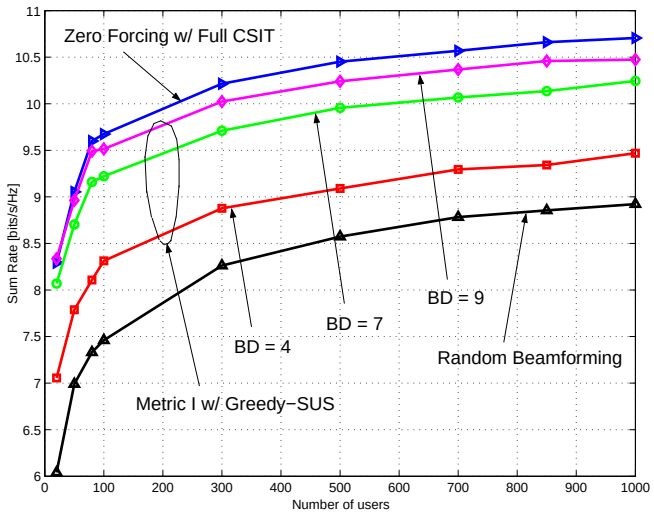

Figure 5.6: Sum rate performance as a function of the number of users for increasing codebook size, $M = 2$ transmit antennas, and SNR $= 10$ dB.

For fixed number of active users in the cell (Fig. 5.5), by increasing the number of codebook bits, strategy 1 converges to ZFBF with perfect CSIT, while providing considerable gains with respect to RBF. Note also that increasing the number of bits for channel direction quantization at high SNR is more beneficial than at low SNR. The sum rate as a function of the number of users K is shown in Figure 5.6. As the size of the codebook increases, the performance of scheme I approaches that of the scheme with perfect CSIT, showing the expected scaling with the number of users. This is due to the fact that metric I can efficiently exploit multiuser diversity.

Effect of CQI quantization

In order to evaluate the effect of CQI quantization, we consider a system in which each user has in total 10 bits available for feedback reporting. A sum-rate comparison as a function of the number of users for SNR = 20 dB is shown in Figure 5.7. We use B_D bits for feeding back the index of the quantized channel and the remaining $B_Q = (10 - B_D)$ bits for CQI quantization. For Strategy 4 (two scalar values of feedback), 2 bits are used for quantization of the channel norm ($\gamma^{(1)}$) and 3 bits for the alignment ($\gamma^{(2)}$). The random beamforming scheme uses $B_D = 1$ bits in order to specify the chosen transmitted beam ($B_D = \lceil \log_2 M \rceil$) and the remaining (9 bits) for SINR quantization. A simple quantization technique has been used that minimizes the mean squared distortion (max Lloyd algorithm). For this amount of available feedback, it can be seen that for the simulated range of K, 6 bits are enough to capture a large portion of multiuser diversity and preserve the scaling (case $B_D = 4$). Note also that the performance is similar to that of Figure 5.3, in which the CQI metrics are considered unquantized.

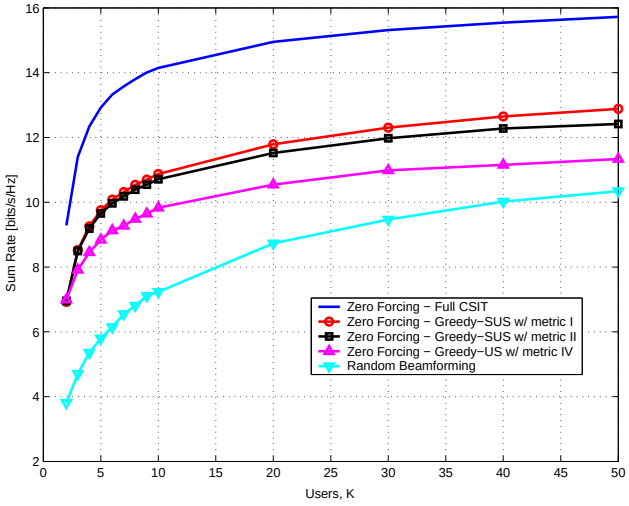

Figure 5.7: Sum rate versus the number of users for with SNR = 20 dB, $M = 2$ transmit antennas and 10-bit total feedback bits. $B_D = 5$ bits are used for codebook indexing and ($B_Q = 10 - B_D$ bits) for CQI quantization. For metric IV, 2 bits are used for quantization of the channel norm and 3 bits for the alignment.

Finite sum rate feedback constraint

We evaluate now the sum rate performance of strategy 1 under a finite sum rate feedback constraint. The total number of available feedback bits is $B_{tot} = 7$ bits. CQI quantization is performed through Max-Lloyd's algorithm. Once both the input quantization levels q_i and output representative levels γ_{q_i} are found, the quantizer sets $\gamma_{q_i} = q_i$, $0 \leq i \leq N_Q - 1$ in order to avoid information outage events.

Figures 5.8 and 5.9 show the sum rate as a function of the number of users for SNR = 10 dB and SNR = 20 dB respectively for different CDI and CQI feedback allocations. As expected, it is more beneficial to allocate more bits on channel direction quantization

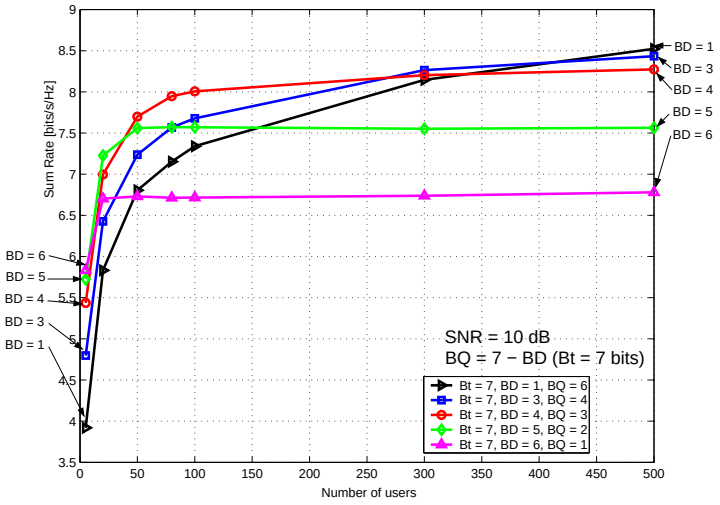

Figure 5.8: Sum rate vs. number of users for $M = 2$ and SNR $= 10$ dB.

in a system with low number of active users. On the other hand, as the number of users increases, it becomes more beneficial to allocate bits on CQI quantization instead. The black curve $B_D = 1$ bit corresponds to the RBF for $M = 2$ transmit antennas [9]. In a system with optimal quantization, i.e. matched to the PDF of the maximum CQI value among K users, the amount of necessary quantization levels is reduced as the number of users in the cell increases. Thus, fewer amounts of feedback bits are needed for CQI quantization in order to capture the multiuser diversity.

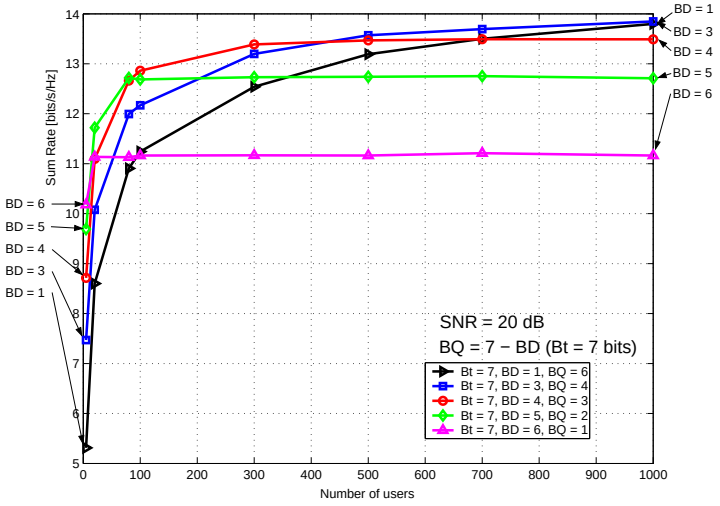

Figure 5.9: Sum rate vs. number of users for $M = 2$ and SNR $= 20$ dB.

In Figure 5.10, the envelope of the curves in the two previous figures is shown, which corresponds to a system that chooses the best B_D/B_Q balance for each average SNR and K pair. In this figure, we compare how this best pair of (B_D, B_Q) changes as the system

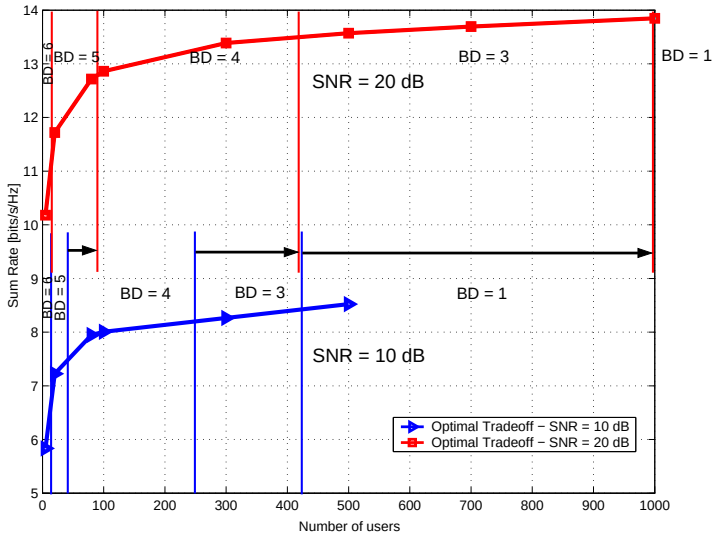

Figure 5.10: Sum rate vs. number of users in a system with optimal B_D/B_Q balancing for different SNR values.

average SNR increases. Both curves are divided in different regions, according to the optimal (B_D, B_Q) pair in each region. It can be seen that the optimal threshold for switching from $B_D \rightarrow B_D - 1$ bits (and thus $B_Q \rightarrow B_Q + 1$) is shifted to the right for higher average SNR values (upper curve). This means that as the average SNR increases, more bits should be allocated on channel direction information. Summarizing, given a pair of average SNR and K values, there exists an optimal compromise of B_D and B_Q, given that $B_{tot} = B_D + B_Q$.

5.8 Conclusion

In this chapter, we study multi-antenna broadcast channels, in which each user reports back to the BS quantized CDI and real-valued scalar CQI through a limited rate feedback channel. We proposed various scalar CQI feedback and scheduling metrics that, if combined with efficient joint scheduling and zero-forcing beamforming, can achieve a significant fraction of the capacity of the full CSI case by means of multiuser diversity. These metrics are built upon inter-user interference bounds and incorporates information on both channel gain and quantization error as a means to estimate satisfactorily the received SINR. A novel feedback strategy is also identified, which allows for adaptive switching between multiuser (SDMA) and single-user transmission (TDMA) mode was also identified as a means to compensate for the sum-rate ceiling effect at high SNR. Our scheme is shown to achieve linear sum-rate growth in the interference-limited region by dynamically adapting the number of scheduled users. Under a practically relevant fixed feedback rate constraint per user, we formulated the problem of optimal feedback balancing in order to exploit spatial multiplexing and multiuser diversity gains. A low-complexity optimization approach has been suggested in order to identify the necessary CDI and CQI feedback load scaling, revealing an interesting interplay between the number of users, the average SNR and the number of feedback bits.

Table 5.1: Greedy Semi-orthogonal User Selection with Limited Feedback

Step 0 set $\mathcal{S} = \emptyset$, $\mathcal{Q}^0 = 1, \ldots, K$

For $i = 1, 2, \ldots, M$ repeat

Step 1 $k_i = \arg \max_{k \in \mathcal{Q}^{i-1}} \mathrm{CQI}_k$

Step 2 $\mathcal{S} = \mathcal{S} \cup k_i$

Step 3 $\mathcal{Q}^i = \left\{ k \in \mathcal{Q}^{i-1} \mid |\hat{\mathbf{h}}_k \hat{\mathbf{h}}_{k_i}^H| \leq \epsilon \right\}$

Table 5.2: Greedy User Selection Algorithm with Limited Feedback

Step 0 Initialization: Set $\mathcal{S}_0 = \emptyset$, $\mathcal{R}(\mathcal{S}_0) = 0$, and $\mathcal{Q}^0 = 1, \ldots, K$

Step 1 $k_1 = \arg \max_{k \in \mathcal{Q}^0} \mathrm{CQI}_k$

Set $\mathcal{S}_1 = \mathcal{S}_0 \cup \{k_1\}$

While $i < M$ repeat

$i \leftarrow i + 1$

Step 2 $k_i = \arg \max_{k \in (\mathcal{Q}^0 - \mathcal{S}_{i-1})} \mathcal{R}(\mathcal{S}_{i-1} \cup \{k\})$

Step 3 Set $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{k_i\}$

if $\mathcal{R}(\mathcal{S}_i) \leq \mathcal{R}(\mathcal{S}_{i-1})$

Step 4 finish algorithm and $i \leftarrow i - 1$

Step 5 Set $\mathcal{S} = \mathcal{S}_i$ and $\mathcal{M} = i$

APPENDIX

5.A Proof of Theorem 5.1

Before proceeding to the proof of Theorem 5.1, we first state the following result.

Lemma 5.3: *Let* $\mathbf{U}_k \in \mathbb{C}^{M\times(M-1)}$ *be an orthonormal basis spanning the null space of* $\mathbf{w}_k$. *Then,*

$$\left\|\overline{\mathbf{h}}_k\mathbf{U}_k\right\|^2 = 1-\cos^2\theta_k \tag{5.41}$$

Proof. Define the orthonormal basis $\mathbf{Z}_k$ of $\mathbb{C}^M$ obtained by stacking the column vectors of $\mathbf{U}_k$ and $\mathbf{w}_k$: $\mathbf{Z}_k = [\mathbf{U}_k\mathbf{w}_k]$. Since $\mathbf{Z}_k\mathbf{Z}_k^H = \mathbf{I}$ and $\overline{\mathbf{h}}_k$ has unit power

$$\left\|\overline{\mathbf{h}}_k\mathbf{Z}_k\right\|^2 = \overline{\mathbf{h}}_k\mathbf{Z}_k\mathbf{Z}_k^H\overline{\mathbf{h}}_k^H = \overline{\mathbf{h}}_k\overline{\mathbf{h}}_k^H = 1 \tag{5.42}$$

Then, by definition of $\mathbf{Z}_k$ we can separate the power of $\overline{\mathbf{h}}_k$ as follows

$$\left\|\overline{\mathbf{h}}_k\mathbf{Z}_k\right\|^2 = \left\|\overline{\mathbf{h}}_k\left[\mathbf{U}_k\mathbf{w}_k\right]\right\|^2 = \left\|\overline{\mathbf{h}}_k\mathbf{U}_k\right\|^2 + \left|\overline{\mathbf{h}}_k\mathbf{w}_k\right|^2 = 1 \tag{5.43}$$

Setting $\left|\overline{\mathbf{h}}_k\mathbf{w}_k\right|^2 = \cos^2\theta_k$ and solving the above equation for $\left\|\overline{\mathbf{h}}_k\mathbf{U}_k\right\|^2$ we obtain the desired result. □

Now we can proceed to the proof of Theorem 5.1. Using the definition of $\mathbf{\Psi}_k(\mathcal{S})$ and defining $\omega_k^2 = \cos^2\theta_k$, the interference over the normalized channel for user k and index set $\mathcal{S}$, denoted as $\overline{I}_k(\mathcal{S})$, can be expressed as

$$\overline{I}_k(\mathcal{S}) = \sum_{i\in\mathcal{S},i\neq k}\left|\overline{\mathbf{h}}_k\mathbf{w}_i\right|^2 = \sum_{i\in\mathcal{S},i\neq k}\overline{\mathbf{h}}_k\mathbf{w}_i\mathbf{w}_i^H\overline{\mathbf{h}}_k^H = \overline{\mathbf{h}}_k\mathbf{\Psi}_k(\mathcal{S})\overline{\mathbf{h}}_k^H \tag{5.44}$$

The normalized channel $\overline{\mathbf{h}}_k$ can be expressed as a linear combination of orthonormal basis vectors. Using *Lemma 5.3*, all possible unit-norm $\overline{\mathbf{h}}_k$ vectors with $\left|\overline{\mathbf{h}}_k\mathbf{w}_k\right| = \omega_k$ can be written as follows

$$\overline{\mathbf{h}}_k = \omega_k e^{-j\alpha_k}\mathbf{w}_k^H + \sqrt{1-\omega_k^2}\mathbf{U}_k\mathbf{B}_k\mathbf{e}_k \tag{5.45}$$

where $\mathbf{B}_k$ is a diagonal matrix with entries $e^{j\beta_i}, i = 1,\ldots,M-1$ and $\mathbf{e}_k$ is an arbitrary unit-norm vector in $\mathbb{C}^{M-1}$. The complex phases β_i and α_k are unknown and lie in $[0, 2\pi]$. Substituting (5.45) into (5.44) we get

$$\begin{aligned}\overline{I}_k(\mathcal{S}) = \omega_k^2\ &\mathbf{w}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{w}_k\\ (a)\qquad &+\left(1-\omega_k^2\right)\mathbf{e}_k^H\mathbf{B}_k^H\mathbf{U}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{U}_k\mathbf{B}_k\mathbf{e}_k\\ (b)\qquad &+\omega_k\sqrt{1-\omega_k^2}\left[e^{-j\alpha_k}\mathbf{w}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{U}_k\mathbf{B}_k\mathbf{e}_k\right.\\ &\qquad\left.+\mathbf{e}_k^H\mathbf{B}_k^H\mathbf{U}_k^H\mathbf{\Psi}_k(\mathcal{S})\mathbf{w}_k e^{j\alpha_k}\right]\end{aligned} \tag{5.46}$$

Since the first term in (5.46) is perfectly known, the upper bound on $\overline{I}_k(\mathcal{S})$ is found by joint maximization of the summands (a) and (b) with respect to α_k, $\mathbf{B}_k$ and $\mathbf{e}_k$. We use a simpler optimization method, which consists of bounding separately each term.

(a) Defining $\mathbf{A}_k(\mathcal{S}) = \mathbf{U}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{U}_k$ for clarity of exposition, the second term can be bounded as follows

$$\max_{\mathbf{B}_k, \mathbf{e}_k} \left(1-\omega_k^2\right) \mathbf{e}_k^H \mathbf{B}_k^H \mathbf{A}_k(\mathcal{S}) \mathbf{B}_k \mathbf{e}_k = \left(1-\omega_k^2\right) \lambda_{max}\{\mathbf{A}_k(\mathcal{S})\}$$

$$s.t. \, \|\mathbf{e}_k\| = 1 \tag{5.47}$$

where the operator $\lambda_{max}\{\cdot\}$ returns the largest eigenvalue. The maximum in (5.47) is obtained when the vector $\mathbf{B}_k \mathbf{e}_k$ equals the principal eigenvector of the matrix $\mathbf{A}_k(\mathcal{S})$.
(b) Defining $\mathbf{q}_k = \mathbf{B}_k^H \mathbf{U}_k^H \mathbf{\Psi}_k(\mathcal{S}) \mathbf{w}_k e^{j\alpha_k}$ and noting that the matrix $\mathbf{\Psi}_k(\mathcal{S})$ is Hermitian by construction, the bound on the third term in (5.46) can be written as follows

$$\max_{\mathbf{q}_k, \mathbf{e}_k} \omega_k \sqrt{1-\omega_k^2} \left[\mathbf{q}_k^H \mathbf{e}_k + \mathbf{e}_k^H \mathbf{q}_k\right] = \max_{\mathbf{q}_k} 2\omega_k \sqrt{1-\omega_k^2} \, \|\mathbf{q}_k\|$$

$$s.t. \, \|\mathbf{e}_k\| = 1 \tag{5.48}$$

The left hand side is maximized for $\mathbf{e}_k = \frac{\mathbf{q}_k}{\|\mathbf{q}_k\|}$, which satisfies the unit-norm constraint, yielding the modified bound in (5.48). The solution is given by

$$\begin{aligned} \max_{\mathbf{q}_k} 2\omega_k \sqrt{1-\omega_k^2} \, \|\mathbf{q}_k\| &= \max_{\mathbf{B}_k, \alpha_k} 2\omega_k \sqrt{1-\omega_k^2} \left\| \mathbf{B}_k^H \mathbf{U}_k^H \mathbf{\Psi}_k(S) \mathbf{w}_k e^{j\alpha_k} \right\| \\ &= 2\omega_k \sqrt{1-\omega_k^2} \left\| \mathbf{U}_k^H \mathbf{\Psi}_k(S) \mathbf{w}_k \right\| \end{aligned} \tag{5.49}$$

Finally, incorporating into (5.46) the bounds obtained in (5.47) and (5.49) we obtain the desired bound.

5.B Proof of Lemma 5.1

By noting that ϵ_{ZF} corresponds to the maximum possible amplitude of the off-diagonal terms of $\left(\hat{\mathbf{H}}_k \hat{\mathbf{H}}_k^H\right)^{-1}$ and under the non restrictive assumption $\epsilon < \frac{1}{M-1}$, the bound on ϵ_{ZF} is found by bounding the amplitude of the off-diagonal terms in the Neumann series $\sum_{n=1}^{\infty} \mathit{offdiag}\left(\hat{\mathbf{H}}_k \hat{\mathbf{H}}_k^H\right)^n$, where $\mathit{offdiag}(\cdot)$ takes the off-diagonal part setting the elements in the diagonal to zero. By representing the non-normalized zero-forcing beamforming vectors as the sum of $\hat{\mathbf{h}}_k$ and its orthogonal complement $\tilde{\mathbf{w}}_k$, i.e. $\mathbf{w}_k = \hat{\mathbf{h}}_k + \tilde{\mathbf{w}}_k$ and bounding the amplitude of the diagonal terms of $\mathbf{I} + \sum_{n=1}^{\infty} \mathit{offdiag}\left(\hat{\mathbf{H}}_k \hat{\mathbf{H}}_k^H\right)^n$, we obtain the desired bound on the channel alignment $\cos\theta_k$.

5.C Proof of Theorem 5.2

By using the definition of each user's SINR_k, $\cos\theta_k$ and equal power allocation, we have that

$$\mathrm{SINR}_k = \frac{P \left|\mathbf{h}_k \mathbf{w}_k\right|^2}{\sum_{j \in \mathcal{S}, j \neq k} P |\mathbf{h}_k \mathbf{w}_j|^2 + M} = \frac{P \|\mathbf{h}_k\|^2 \cos^2 \theta_k}{P \|\mathbf{h}_k\|^2 \sum_{j \in \mathcal{S}, j \neq k} \left|\overline{\mathbf{h}}_k \overline{\mathbf{w}}_j\right|^2 + M} \tag{5.50}$$

We aim to find an upper bound on the multiuser interference given by Theorem 5.1 that takes into account the worst-case orthogonality ϵ_{ZF}. Expressing the worst-case interference

received by the k-th user in terms of $\cos\theta_k$ and ϵ_{ZF}, the following bounds can be easily derived for equation (5.16)

$$\begin{cases} \alpha_k \leq (M-1)\epsilon_{ZF}^2 \\ \beta_k \leq 1 + (M-2)\epsilon_{ZF} \\ \omega_k \leq (M-1)\epsilon_{ZF} \end{cases} \tag{5.51}$$

Hence, by substituting these values in equation (5.15), we obtain the upper bound $\overline{I}_k = \cos^2\theta_k\ (M-1)\epsilon_{ZF}^2 + \sin^2\theta_k\ [1+(M-2)\epsilon_{ZF}] + 2\sin\theta_k\cos\theta_k(M-1)\epsilon_{ZF} \leq \sin^2\theta_k$. By substituting $\epsilon_{ZF} = \vartheta$ and $\cos\theta_k = \frac{|\cos\phi_k - \sqrt{\vartheta}|}{1+\vartheta}$ (i.e. inequalities (5.17) and (5.18), respectively become equalities), where $\vartheta = \frac{\epsilon}{1-(M-1)\epsilon}$ in the previous expression, we have the upper bound given by (5.21). Using this bound on the SINR_k expression derived in (5.50), we obtain the SINR bound in equation (5.20).

5.D Proof of Lemma 5.2

Before proceeding to the proof, we first state some preliminary calculations that are useful in the derivation of the CDF of γ_k^I. To simplify the notation, we define the random variable $\nu := \|\mathbf{h}_k\|^2$ which is Gamma distributed with parameter M and mean $\mathbb{E}\{\|\mathbf{h}_i\|^2\} = M$; hence, its PDF is given by

$$f_\nu(x) = \frac{x^{M-1}}{\Gamma(M)}e^{-x} \tag{5.52}$$

where $\Gamma(M) = (M-1)!$ is the complete gamma function.

In [62], it is shown that under the ACVQ framework, the interference $Y = \|\mathbf{h}_k\|^2 \sin^2\phi_k$ follows a chi-square $\chi^2_{(2M-2)}$ distribution with $(2M-2)$ degrees of freedom weighted by δ, i.e. $Y \sim \delta\chi^2_{(2M-2)}$. Similarly, the distribution of the received signal $X = \|\mathbf{h}_k\|^2\cos^2\phi_k = \|\mathbf{h}_k\|^2(1-\sin^2\phi_k)$ is the sum of two independent weighted chi-square distributions $\chi^2_{(2)} + (1-\delta)\chi^2_{(2M-2)}$.

Define the following changes of variables

$$\begin{array}{ll} \psi := \sin^2\phi_k & u := \frac{1}{\delta}\nu(1-\psi) \\ \nu := \|\mathbf{h}_k\|^2 & v := \frac{1}{\delta}\nu\psi \end{array} \tag{5.53}$$

Then, the metric in equation (5.9) can be expressed as

$$\gamma = \frac{u}{v + \frac{M}{P\delta}} \tag{5.54}$$

The Jacobian of the transformation $u = f(\nu,\psi)$, $v = g(\nu,\psi)$ described in (5.53) is given by

$$J(\nu,\psi) = \begin{vmatrix} \frac{\partial u}{\partial \nu} & \frac{\partial u}{\partial \psi} \\ \frac{\partial v}{\partial \nu} & \frac{\partial v}{\partial \psi} \end{vmatrix} = \frac{\nu}{\delta^2} \tag{5.55}$$

Expressing ν and ψ as a function of u and v, we have $\nu = \delta(u+v)$ and $\psi = \frac{v}{u+v}$. Substituting in the Jacobian, we get $J(u,v) = \frac{(u+v)}{\delta}$. Since ν and ψ are independent random variables for i.i.d. channels, the joint PDF of u and v is obtained from $f_{uv}(u,v) = \frac{1}{J(u,v)} f_\nu\left[\delta(u+v)\right] f_\psi\left[\frac{v}{u+v}\right]$. The PDF f_ν is given by eq. (5.52) and f_ψ is given by [62]

$$f_\psi\ (x) = \begin{cases} N_D(M-1)x^{M-2} & 0 \leq x \leq \delta \\ 0 & x > \delta \end{cases} \tag{5.56}$$

Hence, we get the joint density

$$f_{uv}(u,v) = \frac{\delta}{\Gamma(M-1)} e^{-\delta(u+v)} v^{M-2} \tag{5.57}$$

The CDF of the CQI metric I is found by solving the integral

$$F_\gamma(x) = \iint_{u,v \in D_x} f_{uv}(u,v)\ du\, dv \tag{5.58}$$

The bounded region D_x in the uv-plane represents the region where the inequality $\frac{u}{v+\frac{M}{P\delta}} \leq x$ holds. In addition, since the domain of ψ is $D_\psi = [0,\delta]$, we also obtain the inequalities $\frac{v}{u+v} \geq 0$, $\frac{v}{u+v} \leq \delta$ and thus $u \geq \frac{1-\delta}{\delta} v$. Hence, $F_\gamma(x)$ is obtained by integrating $f_{uv}(u,v)$ over the first quadrant of the uv-plane, in the region defined by $u \leq x\left(v + \frac{M}{P\delta}\right)$ and $u \geq \frac{1-\delta}{\delta} v$. Depending on the slopes of these linear boundaries, the integral in (5.58) is carried out over different regions

$$F_\gamma(x) = \begin{cases} \int_0^\infty \int_{\frac{1-\delta}{\delta}v}^{x\left(v+\frac{M}{P\delta}\right)} f_{uv}(u,v)\ du\, dv & x \geq \frac{1-\delta}{\delta} \\ \int_0^{\frac{Mx}{P(1-\delta-\delta x)}} \int_{\frac{1-\delta}{\delta}v}^{x\left(v+\frac{M}{P\delta}\right)} f_{uv}(u,v)\ du\, dv & 0 \leq x < \frac{1-\delta}{\delta} \end{cases} \tag{5.59}$$

The upper integration limit along the v axis in the region $0 \leq x < \frac{1-\delta}{\delta}$, corresponds to the value of v in which the linear boundaries intersect, $v = \frac{Mx}{P(1-\delta-\delta x)}$. Solving the integrals above, we obtain the CDF of the SINR metric.

5.E Proof of Theorem 5.3

Let $\gamma_{k_i}^I$ denote the upper bound on the achieved SINR of user k_i (i.e. the user selected at the i-th iteration, for $i = 1, 2, \ldots, M$. From Theorem 1 in [64], we have that

$$\Pr\left\{ u_{\mathcal{K}_1} - \frac{P}{M}\log\log\sqrt{K} \leq \gamma_{k_1}^I \leq u_{\mathcal{K}_1} + \frac{P}{M}\log\log\sqrt{K} \right\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

with $u_{\mathcal{K}_1} = \frac{P}{M}\log(\frac{K}{\beta}) - \frac{P(M-1)}{M}\log\log(\frac{K}{\beta})$.

For $i = 2, \ldots, M$, we obtain

$$\Pr\left\{ u_{\mathcal{K}_i} - \frac{P}{M}\log\log\sqrt{K} \leq \gamma_{k_i}^I \leq u_{\mathcal{K}_i} + \frac{P}{M}\log\log\sqrt{K} \right\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

with $u_{\mathcal{K}_i} = \frac{P}{M}\log(\frac{\mathcal{K}_i}{\beta}) - \frac{P(M-1)}{M}\log\log(\frac{\mathcal{K}_i}{\beta})$.

From Greedy-SUS procedure, we have that $\gamma_{k_1}^I \geq \gamma_{k_2}^I \geq \ldots \geq \gamma_{k_M}^I$, and after some manipulations it can be shown that for large K, we have

$$\Pr\left\{ u_{\mathcal{K}_i} - \frac{P}{M}\log\log\sqrt{K} \leq \gamma_{k_i}^I \leq u_{\mathcal{K}_1} + \frac{P}{M}\log\log\sqrt{K} \right\} \geq 1 - O\left(\frac{1}{\log K}\right)$$

Since $\log(\cdot)$ is an increasing function, we have that

$$\begin{aligned} &\Pr\{\log_2\left(1 + u_{\mathcal{K}_i} - \frac{P}{M}\log\log\sqrt{K}\right) \leq \log_2\left(1+\gamma_{k_i}^I\right) \\ &\leq \log_2\left(1 + u_{\mathcal{K}_1} + \frac{P}{M}\log\log\sqrt{K}\right)\} \geq 1 - O\left(\frac{1}{\log K}\right) \end{aligned} \tag{5.60}$$

Hence,

$$\lim_{K\to\infty} \Pr\{\frac{\log_2\left(1+u_{\mathcal{K}_i}-\frac{P}{M}\log\log\sqrt{K}\right)}{\log_2(\frac{P}{M}\log K)} \leq \frac{\log_2\left(1+\gamma_{k_i}^I\right)}{\log_2(\frac{P}{M}\log K)} \leq \frac{\log_2\left(1+u_{\mathcal{K}_1}+\frac{P}{M}\log\log\sqrt{K}\right)}{\log_2(\frac{P}{M}\log K)}\} \geq 1-O\left(\frac{1}{\log K}\right) \tag{5.61}$$

By substituting $u_{\mathcal{K}_1}$ and $u_{\mathcal{K}_i}$ in the above equation, we conclude that the LHS and the RHS of the inequalities both converge to one as $K\to\infty$, therefore

$$\lim_{K\to\infty}\frac{\mathcal{R}}{\log_2\left(\frac{P}{M}\log K\right)}=1 \tag{5.62}$$

with probability one. Assuming equal power allocation and that M perfectly orthogonal users can be found, as $\Pr\{|\mathcal{S}|=M\}\overset{K\to\infty}{\to}1$, we have that the proposed scheme achieves a sum rate of $M\log_2\left(\frac{P}{M}\log K\right)$.

An upper bound on $\mathcal{R}_{opt}$ is given in [44], where

$$\Pr\left\{\frac{\mathcal{R}_{opt}}{M}\leq\log_2\left(1+\frac{P}{M}\left(\log K+O(\log\log K)\right)\right)\right\}\geq 1-O\left(\frac{1}{\log^2 K}\right)$$

Thus,

$$\Pr\Big\{\log_2(1+\gamma_{k_i}^I)-\tfrac{\mathcal{R}_{opt}}{M}\geq \log_2\left(1+u_{\mathcal{K}_i}-\tfrac{P}{M}\log\log\sqrt{K}\right)-\log_2\left(1+\tfrac{P}{M}\left(\log K+O(\log\log K)\right)\right)\Big\} \geq 1-O\left(\tfrac{1}{\log K}\right)-O\left(\tfrac{1}{\log^2 K}\right)$$

where the RHS of the inequality inside the Pr goes to zero for $K\to\infty$. As a result, for large K, we have that

$$0\leq\log_2(1+\gamma_{k_i}^I)-\frac{\mathcal{R}_{opt}}{M},\quad i=1,\ldots,M$$

with probability one, which results to (5.30) for $K\to\infty$, as $\mathcal{R}_{opt}$ is an upper bound on the sum rate of our proposed scheme.

5.F Proof of Theorem 5.4

For $P\to\infty$, we have

$$\gamma_k^{II}=\lim_{P\to\infty}\frac{\frac{P}{(1+\vartheta)^2}\left\|\mathbf{h}_k\right\|^2(\cos\phi_k-\sqrt{\vartheta})^2}{P\left\|\mathbf{h}_k\right\|^2\overline{I}_{UB_k}+M}=\frac{(\cos\phi_k-\sqrt{\vartheta})^2}{(1+\vartheta)^2\overline{I}_{UB_k}}\leq\cot^2\phi_k \tag{5.63}$$

whose PDF is given by $f_{\cot^2\phi}(x)=\frac{(M-1)N_D}{(1+x)^M}$, for $x\geq(1-\delta)/\delta$ and zero elsewhere [64].

The expected sum rate for a user set $\mathcal{S}$ (of cardinality M) is given by

$$\mathcal{R}\leq\mathbb{E}\left\{\sum_{i=1}^{M}\log_2(1+\max_{k_i\in\mathcal{K}_i}\cot^2\phi_{k_i})\right\}=\sum_{i=1}^{M}\int_0^\infty\log_2(1+x)dF_{\cot^2\phi}^{\mathcal{K}_i}(x)dx$$

$$= \sum_{i=1}^{M} \mathcal{K}_i \int_{\frac{1-\delta}{\delta}}^{\infty} \log_2(1+x) \frac{2_D^B (M-1)}{(1+x)^M} \left(1 - \frac{2_D^B}{(1+x)^{M-1}}\right)^{\mathcal{K}_i - 1} dx$$

$$\stackrel{(a)}{=} 2_D^B (M-1) \sum_{i=1}^{M} \mathcal{K}_i \sum_{k=0}^{\mathcal{K}_i - 1} \binom{\mathcal{K}_i - 1}{k} (-1)^k \int_{\frac{1-\delta}{\delta}}^{\infty} \log_2(1+x) \frac{2^{B_D k}}{(1+x)^{k(M-1)+M}} dx$$

$$= \frac{\log_2 e}{M-1} \sum_{i=1}^{M} \mathcal{K}_i \sum_{k=0}^{\mathcal{K}_i - 1} \binom{\mathcal{K}_i - 1}{k} (-1)^k \left[\frac{B_D \log 2}{k+1} + \frac{1}{(k+1)^2}\right]$$

$$\stackrel{(b)}{=} \frac{\log_2 e}{M-1} \sum_{i=1}^{M} (B_D \log 2 + H_{\mathcal{K}_i}) \tag{5.64}$$

where (a) follows from binomial expansion and to get (b) the Nörlund-Rice integral representation is applied [105]. Combining (5.64) with $\mathcal{K}_i \leq K$, we get (5.31).

Chapter 6

Feedback Reduction using Ranking-based Feedback

6.1 Introduction

In the previous chapters, we investigated several scheduling and linear beamforming techniques and tried to identify low-rate feedback measures that provide the transmitter with sufficient yet partial channel knowledge, as a means to achieve near optimal system throughput. It was shown that if some form of implicit channel knowledge (e.g. channel correlation) is exploited, it sufficient to feed back one or two scalar feedback parameters in order to achieve satisfactory performance. In this chapter, we take a different approach to the problem of feedback reduction and aim at finding a representation of feedback metrics that allows for further compression. Using the promising two-step scheduling and precoding approach proposed in Chapter 3, we point out that the channel information to be conveyed to the scheduler can be further decreased. As the first-stage channel information is mainly utilized for the purposes of user selection and not for beam design or rate allocation, we propose a new type of feedback representation, coined as *ranking-based feedback*. In this approach, each user - instead of reporting a quantized version of CSIT feedback - calculates and feeds back the ranking, an integer between 1 and $W+1$, of its instantaneous CSIT among a set of W past CSIT measurements. This representation enables the BS to select users that are on the highest peak (quantile) with respect to their own channel distribution, independently of the distribution of other users. When W is sufficiently large, the selected users are also the ones with the most favorable channel conditions. An interesting property of this method is that temporal fairness is restored in heterogeneous networks, i.e. systems in which users' channels are not identically distributed and mobile terminals experience different average SNRs.

Feedback reduction in SDMA systems has in fact evolved in a topic of research in its own right and many possible strategies can be pointed out. Apart from the approaches already presented and proposed in the previous chapters, a few selected additional ones are briefly exposed here. A popular approach, referred to as selective or threshold-based feedback, allows a user to send back information depending on whether its current channel conditions exceed a certain threshold or not. This concept was first proposed in [106] for a downlink single-input, single-output (SISO) system and SNR-dependent thresholds, and is shown to reduce statistically the required total amount of feedback by means of multiuser diversity. The feedback rate can be further reduced, at the cost of feedback delay, by using an adaptive threshold [107]. The selective feedback idea was extended for MISO systems in [108]. In [109], a scheme based on [53] and one bit feedback was shown to achieve the optimal capacity growth rate when $K \to \infty$. A scheme based on multi-beam random beamforming was proposed in [110,111], where it was shown that a deterministic feedback of $\log_2(1+M)$ bits per user is enough to guarantee the optimal capacity scaling law for fixed M and single-antenna receivers. A common limitation of the above feedback reduction techniques is that the total feedback rate grows linearly with the number of users, thus reducing the effective system throughput when the number of users is large. SDMA under a sum feedback rate constraint is considered in [112], in which threshold-based feedback on the channel quality and the channel direction is used for feedback reduction in order to satisfy a sum feedback rate constraint. Differently from the previous approaches in which users are assumed to send feedback through dedicated channels, the authors in [113] consider a contention-based feedback protocol, in which users compete to gain access in a shared medium. In this system, the feedback resources are fixed random access minislots, and active users attempt to convey feedback messages only if their channel gain is above a threshold.

In this chapter, we adopt a two-stage SDMA downlink technique. During the scheduling phase, all active users K are allowed to feedback some kind of finite rate CQI, whereas in the second step, information on the transmission rate is requested only from the $M \ll K$ selected users. Our work builds upon recently proposed ideas in the context of scheduling [114]. Therein, a so-called 'score-based' opportunistic scheduler was proposed for realistic scenarios with asymmetric fading statistics and data rate constraints. Similar channel distribution-based schedulers have also been proposed in [115–117] as a means to schedule a user whose instantaneous rate is in the highest quantile of its distribution. Interestingly, these works were solely focused on scheduling at the transmitter side, and neither in the context of feedback reduction nor that of MIMO systems. The contributions of this chapter are the following:

- We propose a new concept of CSIT representation, coined as 'ranking-based feedback', for the sole purpose of user selection as a means to reduce the required feedback load. The ranking-based CSIT consists of an integer value that represents the rank of each user's instantaneous CQI among a number of stored CQI values observed over the W past slots.

- The key advantage of the proposed method is two fold: 1) ranking-based feedback is already in digital form, which helps for further compression and simple scalar quantization, 2) ranking-based feedback provides not only information about the channel

quality but also about the relative quality level, in a way that is independent of the users' fading statistics, hence providing inherent fairness. This type of limited feedback representation enables the BS to select users that are on the peak of their own channel distribution, independently of the channel conditions of other users.

- We analyze the sum-rate performance of a modified MISO downlink system with random orthogonal beams as in [9, 53], in which users are selected based on ranking-based CSIT. Furthermore, we provide analytic expressions for the sum rate when W is finite. We quantify the effect of finite W and the error introduced in the scheduling decisions compared to the optimal case of $W \to \infty$.

- We study the additional merit of ranking-based CSIT in heterogeneous networks by showing that such form of feedback information is able to offer temporal fairness among users, since the probability of a user to be selected is $1/K$, independently of the other users' channel distributions and its own average SNR.

6.2 Ranking-based Feedback Framework

We present here the concept of *ranking-based feedback* and show its intrinsic advantages when it is used as a user selection metric during the scheduling stage in a broadcast channel with M transmit antennas and $K \geq M$ single-antenna users.

6.2.1 Two-stage approach

We assume a two-stage feedback approach by splitting the feedback resource into two stages (scheduling followed by transmission). In the scheduling stage, all K active users compete for medium access and each user k is allowed to report instantaneous CQI, denoted as γ_k, which is a certain function of its channel, i.e. $\gamma_k = f(\mathbf{h}_k)$. This CQI metric can generally take on any form of channel information representation. For instance, in a TDMA context, γ_k may represent the SNR or the transmission rate of user k, whereas in a SDMA setting, the CQI may be the channel norm or the received SINR (achievable or estimated). Actually, all the feedback metrics that we presented in the previous chapters can be used here as CQI feedback. However, as in this chapter the CQI is used *solely* for purposes of user selection, coarser channel information can possibly be used. Given a set of selected users $\mathcal{S}$ with cardinality $|\mathcal{S}| = \mathcal{M} \leq M$, a second step exploiting precoding is applied to serve the selected users. During the second step, the transmitter may request for variable levels of additional CSIT feedback from the $\mathcal{M} << K$ pre-selected users. The second-step precoding matrix may require variable levels of additional CSIT feedback to be computed, depending on design. The second-stage CSIT feedback can be used for precoding design as well as for link adaptation. For simplicity of exposition, we consider a system where a random, unitary precoder is generated at each time slot during the first stage. Moreover, the second-step precoder is the same as the one used in the scheduling step and the selected users feed back their transmission rates for the purposes of link adaptation. Alternatively, the need for a second stage in order to inform the BS on the transmission rate can potentially be circumvented by assuming that the CDFs of different users' channels are known a priori at the transmitter. This assumption can be justified in systems, where the statistical reciprocity

between the downlink and uplink channels allows the BS to estimate the distributions by aggregating each user's CQI feedback.

6.2.2 Ranking-based CQI Representation

At time instant t, each user measures its CQI on each of $\mathcal{B}$ randomly generated beams (columns of the first-stage precoding matrix). In addition to the instantaneous CQI value on each beam m, $\{\gamma_{k,m}(t)\}_{m=1}^{\mathcal{B}}$, each user also keeps record of a set of past CQI values, denoted as $\mathcal{W}_{k,m}$, observed over a window of size W, i.e.

$$\mathcal{W}_{k,m} = \{\gamma_{k,m}(t-1), \gamma_{k,m}(t-2), \ldots, \gamma_{k,m}(t-W+1)\}$$

Then, each user, say the k-th, calculates the ranking (order) $r_{k,m}(t) \in \{1, \ldots, W+1\}$ of its current CQI metric $\gamma_{k,m}(t)$ on beam m among the W past values contained in the set $\mathcal{W}_{k,m}$. In other words, if $\gamma_{k,m}(t)$ is the third largest value within the set of W latest measured values, $r_{k,m}(t) = 3$. The rank value of user k at slot t on beam m is mathematically given by [114]

$$r_{k,m}(t) = 1 + \sum_{w=1}^{W-1} 1\{\gamma_{k,m}(t) < \gamma_{k,m}(t-w)\} + \sum_{w=1}^{W-1} 1\{\gamma_{k,m}(t) = \gamma_{k,m}(t-w)\} Z_w \quad (6.1)$$

where Z_w are i.i.d. random variables on $\{0,1\}$ with $\Pr\{Z_w = 0\} = 1/2$ corresponding to the case where the instantaneous CQI is equal to one or several of the past values, in which either rank value is randomly chosen with equal probability.

The key ideas are as follows:
1) each user selects its minimum rank value over the beams, i.e.,

$$r_k(t) = \min_{m=1,\ldots,\mathcal{B}} r_{k,m}(t) \quad (6.2)$$

2) each user, instead of reporting directly its maximum CQI value over the beams, feeds back a quantized value $\hat{r}_k(t)$ of the integer $r_k(t)$, along with the beam index m in which the ranking value is minimum, i.e.

$$\hat{r}_k(t) = \mathcal{Q}(r_k(t)) \quad (6.3)$$

where $\mathcal{Q}(\cdot)$ represents a $N = 2^B$-level quantizer. Thus, the feedback load per user is $\lceil \log_2 N \rceil$ bits for the ranking and $\lceil \log_2 M \rceil$ bits for the index of its preferred beam.

At the transmitter side, the scheduler assigns each beam m to the user k_m^* with the minimum reported ranking value, that is,

$$k_m^*(t) = \arg \min_{1 \le k \le K} \hat{r}_k(t) \quad (6.4)$$

As stated before, once the users $\{k_m^*(t)\}_{m=1}^{\mathcal{B}}$ are selected based on ranking-based CSIT, they are polled and requested to report the transmission rate that can be supported by their instantaneous channel conditionss.

The W past CQI measurements are samples of each user's CQI empirical process. Therefore, the length of the observation window provides a measure of how accurately the CQI distribution is monitored by the user. The larger the W, the better a user can track the distribution of its CQI process, thus identifying more accurately the peaks with respect to

its own distribution. In other words, ranking-based CSIT enables each user to have an estimate of the quantile of its CQI using W previous CQI samples, where the sample quantile[1] of order p is defined as the statistical functional $\hat{F}_W^{-1}(p) = \inf\left\{x : \hat{F}_W(x) \geq p\right\}$ for $p \in (0,1)$ and $\hat{F}_W(\cdot)$ denoting the empirical distribution function of W samples. In the asymptotic case of $W \to \infty$, the observation window captures the entire CQI distribution and corresponds to the case in which ranking-based CSIT provides exact information on the CDF of the CQI process. In this case, the user with the minimum ranking-based CQI value is the one whose instantaneous CQI is in the highest quantile.

6.3 Performance analysis

We evaluate the average rate of a system employing random opportunistic beamforming in which ranking-based feedback is used as user selection metric. We assume that the CQI takes on the form of user rate, i.e., $\gamma_{k,m} = \log_2(1 + \mathrm{SINR}_{k,m})$. Let $X_{k,m}$ denote the rate process of the k-th user rate on the m-th beam with CDF denoted as $F_{X_{k,m}}(\cdot)$. The distribution function is assumed to be strictly increasing and continuous, such that its inverse $F^{-1}_{X_{k,m}}(\cdot)$ exists. Unless otherwise stated, we assume a homogenous network where all users have identical average SNR (i.i.d. channel statistics).

6.3.1 Asymptotic optimality of ranking-based feedback for large window size W

For finite window size W, ranking-based CSIT enables each user to estimate the quantile of its instantaneous CQI based on W samples of its empirical CQI process. For fixed x the number of r.v.s X_i such that $X_i \leq x$ follows a binomial distribution with probability of 'success' $p = F(x)$, hence the r.v. $\hat{F}_X^W(x)$ follows a binomial distribution with possible values $0, 1/W, \ldots, 1$. We examine here the behavior of the empirical function $\hat{F}_X^W(x)$ for W increasing and show how likely is $\hat{F}_X^W(x)$ to be close to $F(x)$ for arbitrary large W and x fixed.

Let the collection of r.v. $\mathcal{X} = \{X_t : t \in \mathbb{N}^+\}$ be a discrete-time stochastic process for each user defined on the same probability space. $\mathcal{X}$ is assumed stationary and ergodic and for exposition convenience we omitted the user index k from the stochastic process. The random sample of i.i.d. r.v. $X_1, X_2, \ldots, X_W$ is an empirical process, whose empirical distribution $\hat{F}_X^W(\cdot)$ is defined as the CDF that puts mass $1/W$ at each sample point X_i, i.e.

$$\hat{F}_X^W(x) = \frac{1}{W}\sum_{i=1}^{W} \mathbb{I}\{X_i \leq x\} \tag{6.5}$$

where $\mathbb{I}\{X_i \leq x\}$ is an indicator function defined as

$$\mathbb{I}\{X_i \leq x\} = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases} \tag{6.6}$$

We can show that for $W \to \infty$, the empirical CDF converges to the CDF of the CQI distribution, which implies that the user with minimum ranking feedback value is the user

[1] More formally, for a process $(Y(t), t \geq 0)$ with stationary and independent increments with $Y(0) = 0$, the p-quantile of $(Y(s), 0 \leq s \leq t)$ for $0 < p < 1$ is defined by $M(p,t) = inf\left\{x : \int_0^t \mathbf{1}(Y(s) \leq s)ds > pt\right\}$.

with the maximum CQI value.

Proposition 6.1: *In a system where users have i.i.d. channel statitics, user selection based on ranking-based feedback converges to the capacity-optimal max-rate scheduling for* $W \to \infty$.

Proof. The proof is given in Appendix 6.A. □

6.3.2 Throughput for infite observation window size W

In this section, we study the average sum rate in the large W regime. Assuming W to be infinitely large, we can easily see that user selection based on ranking-based CSIT is equivalent to minimum complementary CDF (CCDF) scheduling. This means that if $r_{k,m}$ captures the distribution of received SINR process, denoted as $\Gamma_{k,m}$, then $\lim_{W\to\infty} \frac{r_{k,m}}{W} = \bar{F}_{\Gamma_{k,m}}(\gamma_{k,m})$, where $\bar{F}_{\Gamma_{k,m}}(\gamma_{k,m}) = 1 - F_{\Gamma_{k,m}}(\gamma_{k,m})$ is the CCDF of the CQI metric $\gamma_{k,m}$. As shown in Proposition 6.1, selecting on each beam m the user k_m^* with the minimum ranking value is equivalent to selecting the user with the minimum tail of CDF, i.e.

$$\begin{aligned} k_m^* &= \arg\min_{1\le k\le K} \bar{r}_{k,m}(t) = \arg\min_{1\le k\le K} 1 - F_{\Gamma_{k,m}}(\gamma_{k,m}(t)) \\ &= \arg\max_{1\le k\le K} F_{\Gamma_{k,m}}(\gamma_{k,m}(t)) \qquad m = 1,\dots,\mathcal{B} \end{aligned} \tag{6.7}$$

where $\bar{r}_{k,m}(t) = r_{k,m}(t)/W$ is the normalized ranking value and $\gamma_{k,m}(t)$ is the realization of $\Gamma_{k,m}$ at slot t.
The rate of user k on beam m, prior to channel-aware scheduling, is given by

$$\mathcal{R}_{k,m} = \int_0^\infty \log_2(1+\gamma_{k,m})\, f_\Gamma(\gamma) d\gamma = \int_0^1 \log_2\left(1 + F^{-1}_{\Gamma_{k,m}}(\bar{r})\right) d\bar{r} \tag{6.8}$$

where $f_{\Gamma_{k,m}}(\cdot)$ is the PDF of CQI metric γ. If we assume i.i.d. channel statistics and that the user on the highest quantile is scheduled on each beam m, then the average sum rate is given by the following proposition:

Proposition 6.2: *The average sum rate* $\mathcal{R}$ *of a symmetric network (i.i.d. users) where user selection is performed based on ranking-based feedback is given by*

$$\mathcal{R} = \mathcal{B}K \int_0^1 \log_2\left(1 + F_\Gamma^{-1}(z)\right) z^{K-1} dz \tag{6.9}$$

Proof. The proof is straightforward by changing the variable $F_\Gamma(\gamma) = z$ in the sum rate given by $\mathcal{R} = \mathcal{B}\int_0^\infty \log_2(1+\gamma) dF_\Gamma^K$, where F_Γ^K is the CDF of the best user selected among K i.i.d. users with common parent distribution $F_\Gamma(\gamma)$. □

Note that similar result has been derived in [115]. Therein, the authors derive the average user rate for the general case where the channel distributions are not necessarily identically distributed and $\mathcal{B} = 1$. Proving that the probability that user k is selected at time slot t given that the user rate $X_k(t) = x_k$ is $\Pr\{k^*(t) = k | X_k(t) = x_k\} = F_{X_k}^{K-1}(x_k)$, they showed that the average rate of a user is given by $R_k = \int_0^1 u^{K-1} F_{X_k}^{-1}(u) du$.

Unfortunately, equation (6.9) does not always result in closed-form expressions. For instance, the sum rate of multi-beam RBF given by $\mathcal{R}_{RBF} = \mathcal{B}K\int_0^1 F_{X_k}^{-1}(u) u^{K-1} du$, where $F_{X_k}^{-1}(u)$ is the inverse of $F_{X_k}(u) = 1 - \frac{e^{-\mathcal{B}/P} e^{-2^u \mathcal{B}/P}}{2^{(\mathcal{B}-1)u}}$ requires numerical calculation. Analytic expressions can be derived in specific regimes, such as the high and low power regions.

6.3.3 Throughput for finite observation window size W

Let $X_{k_m^*}(t)$ denote the rate process of the user k selected on beam m with distribution function $F_{X_{k_m^*}}(x) = [\Pr\{X_{k,m} \leq x\}]^K$. The expected rate $\mathcal{R}_{k,m}$ of k-th user when scheduled on beam m is given by

$$\mathcal{R}_{k,m} = \mathbb{E}\left\{X_{k_m^*}(t)\right\} = \int_0^\infty \Pr\left\{\max_{1\leq k\leq K} X_{k,m}(t) > x\right\} dx \tag{6.10}$$

Proposition 6.3: *The average sum rate $\mathcal{R}$ of a system generating $\mathcal{B}$ random orthonormal beams and scheduling $\mathcal{B}$ users among K active users based on ranking-based feedback with observation window W is given by*

$$\mathcal{R} = \sum_{m=1}^{\mathcal{B}} \left(\int_0^\infty (1 - (F_{X_{k_m^*}}(x))^W) dx - \sum_{w=1}^{W} \left(\frac{W-w}{W}\right)^K \int_0^\infty F_{w,m}(x) dx \right) \tag{6.11}$$

where $F_{w,m}(x) = \binom{W}{w} \left(F_{X_{k_m^*}}(x)\right)^{W-w} \left(1 - F_{X_{k_m^*}}(x)\right)^w$.

Proof. The proof is given in Appendix 6.B. □

Using the above proposition, we can show that the throughput $\mathcal{R}_{\text{TDMA}}$ of single-beam RBF [53] is given by

$$\mathcal{R}_{\text{TDMA}} = \sum_{w=0}^{W} \left[1 - \left(\frac{W-w}{W}\right)^K\right] \binom{W}{w} \int_0^\infty (F_{X_{k_m^*}}(x))^{W-w} (1 - F_{X_{k_m^*}}(x))^w dx \tag{6.12}$$

with $F_{X_{k_m^*}}(x) = \left(1 - e^{-\frac{2^x - 1}{P}}\right)^K$. The constant term $\mathcal{G} = \sum_{w=0}^{W} \left[1 - \left(\frac{W-w}{W}\right)^K\right]$ can be evaluated analytically as $\mathcal{G} = 1 + W + (-1)^K W^{-K}(\zeta(-K) - \zeta(-K, -W))$, where $\zeta(s)$ and $\zeta(s, a)$ are the Riemann zeta function and Hurwitz zeta function, respectively. Equation (6.12) does not seem to have closed-form representation for exponentially distributed channel gains. However, in the high power regime the following series representation can be obtained:

Corollary 6.1: *At high SNR, the average sum rate R_{high}^W of multi-beam RBF with $\mathcal{B} = 2$ beams, finite W and ranking-based user selection is given by*

$$\mathcal{R}_{high}^W = 2 \sum_{w=1}^{W} \binom{W}{w} \left[1 - \left(\frac{W-w}{W}\right)^K\right] \frac{\Gamma(Kw-1)\Gamma(KW - Kw + 1)}{\Gamma(KW)} \tag{6.13}$$

For large enough W, a good approximation of the binomial distribution is given by the normal distribution (De Moivre-Laplace Theorem). Let $q = F_{X_{k_m^*}}(x)$ and $p = 1 - F_{X_{k_m^*}}(x)$, then $F_{w,m}(x)$ can be approximated by

$$F_{w,m}(x) \approx \frac{1}{\sqrt{2\pi W p q}} e^{-\frac{(w-Wp)^2}{2Wpq}} \tag{6.14}$$

which simplifies the calculation of the integral in (6.11) as $\int_0^\infty F_{w,m}(x) dx = Q\left(\sqrt{2Wp/q}\right)$, where $Q(\cdot)$ is the standard normal CDF.

6.3.4 Performance reduction bound for finite window size W

In the previous two sections we evaluated the throughput performance for finite and infinite observation window size W. In order to quantify the system throughput reduction due to finite values of W, a bound on the difference between the rate when each user knows perfectly its CDF and the throughput when ranking-based feedback is based on the empirical distribution of each user's channel distribution over W is of interest. Intuitively, the sum rate performance is a monotonically decreasing function with W, thus for W decreasing, the performance degradation is increased. However, a bound on the difference does not seem tractable. The main difficulty is that the user rate distribution, as $F_{X_{k,m}}(x)$ is not a linear function of the CQI distribution, i.e. $F_{X_{k,m}}(x) = F_{\Gamma_{k,m}}(2^x - 1)$. Nevertheless, a bound on the the ratio $\mathcal{F}(W,K) = \hat{F}^W_{X_{k^*_m}}(x)/F_{X_{k^*_m}}(x)$, where $\hat{F}^W_{X_{k^*_m}}(\cdot)$ is rate distribution seen by user k when is scheduled based on ranking-based feedback estimated using W samples is derived in [117].

Proposition 6.4: *For a system with K active users employing ranking-based CSIT observed over W past values, the ratio $\mathcal{F}(W,K)$ is lower bounded as*

$$\mathcal{F}(W,K) \geq \left(1 - \left(\frac{W}{W+1}\right)^K\right)\frac{W+1}{K} \leq (1 - e^{-K/W})\frac{W+1}{K} \tag{6.15}$$

where the Bernoulli inequality is used for bounding $\left(\frac{W}{W+1}\right)^K$.

Expanding $e^{-K/W}$ in Taylor series, we have that $(1 - e^{-K/W})\frac{W+1}{K} = (1 - \frac{K}{2W})\frac{W+1}{W} \gtrsim 1 - \frac{K}{W+2}$. Hence, for fixed throughput reduction, the number of samples W required to be stored in memory has to scale almost linearly with the number of active users K in the system.

In addition to the previous bound, a sharp non-asymptotic bound can be derived based on the Dvoretsky-Kiefer-Wolfowitz (DKW) inequality [118, 119]:

Theorem 6.1: *Let $X_1, X_2, \ldots, X_W \sim F_{X_{k,m}}$, then for any $\epsilon > 0$*

$$\Pr\left\{\sup_x \left|\hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x)\right| > \epsilon\right\} \leq 2e^{-2W\epsilon^2} \tag{6.16}$$

Based on Theorem 6.1, we can construct a confidence set that gives us a measure of the required window size W. Given $\alpha \in (0,1)$, say that a random set $S(x)$ is a $(1-\alpha)$ confidence set for the parameter θ if

$$\Pr\{\theta \in S(x)\} \geq 1 - \alpha \tag{6.17}$$

Then, for any F, we have that

$$\Pr\left\{\ell_1(x) \leq F_{X_{k,m}}(x) \leq \ell_2(x), \quad \forall x\right\} \geq 1 - \alpha \tag{6.18}$$

where the two sequences $\ell_1(x) = \max\left\{\hat{F}^W_{X_{k,m}}(x) - \epsilon_W, 0\right\}$, $\ell_2(x) = \min\left\{\hat{F}^W_{X_{k,m}}(x) + \epsilon_W, 1\right\}$ and $\epsilon_W = \sqrt{\frac{1}{2W}\log(2/\alpha)}$. This implies that if one wishes to draw a large enough sample to ensure that the deviation between the empirical distribution and the actual CDF is less than or equal to 10%, with 90% confidence, then for $\epsilon = 0.1$ in (6.16), a sample size of approximately W=150 samples is needed.

6.3.5 Window size versus feedback reduction tradeoff

In the previous section, it has been shown that the performance difference between ranking-based user selection and max-rate scheduling is decreased for W increasing. In practical systems, the feedback channel shared by all users has a fixed bandwidth and thus the rate of reporting $\hat{r}_k(t)$ is finite and generally fixed. As a result, under a fixed feedback rate constraint of $B = \lceil \log_2 N \rceil$ bits, when W is increased, the accuracy of $\hat{r}_k(t)$ is decreased as the distortion of the quantizer $\mathcal{Q}(\cdot)$ is increased. This is evidently due to the fact that the dynamic range of the integer values $r_k(t) \in (0, W+1]$ to be quantized by B bits is increased. In order to guarantee the same throughput performance for increasing W, the number of feedback bits B should scale accordingly so that the quantization error is fixed. This results in an interesting tradeoff between:

- the capacity performance
- the window size W
- the number of feedback bits B

Consider that uniform scalar quantization is used to quantize a source R that is uniformly distributed over [0,1]. The error variance (distortion) is given by:

$$\sigma_Q^2 = \mathbb{E}\left\{(R - \mathcal{Q}(R))^2\right\} = \int_{-\infty}^{+\infty} (r - \mathcal{Q}(r))^2 f_R(r) dr = \frac{(r_{max} - r_{min})^2}{12N^2} \tag{6.19}$$

where $f_R(r)$ is the PDF of the uniform source R, and r_{max} and r_{min} are the maximum and minimum value of ranking-based feedback, respectively. For fixed variance of the quantization error $\sigma_Q^2 = \delta^2$, $r_{min} = 1$ and $r_{max} = W + 1$, the number of bits B should scale proportionally to $B \sim (\log_2 (W/\delta) - 1.8)$ bits. This feedback requirement can be decreased if non-uniform quantization (e.g. optimal entropy-constrained) is employed. The problem of optimum quantization design for ranking-based feedback has not been investigated in the thesis.

6.4 Ranking-based CDI Model

The concept of ranking-based feedback, as presented above, is not restrictive to RBF schemes; it can be generalized to other downlink precoding configurations. The ranking-based concept can indeed be applied to any kind of feedback information of interest utilized for user selection purposes. In codebook-based SDMA downlink systems, for instance, it can be additionally used to represent some kind of CDI as a means to select near orthogonal user with large channel gains. Consider a system in which each user can report CDI feedback based on a predefined codebook in addition to the CQI value that can take on the form of channel norm or estimate of SINR [64, 94]. If we assume that the quantization codebook $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_D}\}$ containing N_D unit norm vectors $\mathbf{v}_i \in \mathbb{C}^M$, for $i = 1, \ldots, N_D$ is known to both the transmitter and receivers, each receiver k quantizes its channel to the codevector that maximizes the following inner product:

$$\hat{\mathbf{h}}_k = \arg\max_{\mathbf{v}_i \in \mathcal{V}} \cos^2(\angle(\bar{\mathbf{h}}_k, \mathbf{v}_i)) \tag{6.20}$$

where the normalized channel vector $\bar{\mathbf{h}}_k = \mathbf{h}_k / \|\mathbf{h}_k\|$ corresponds to the channel vector direction, and $\hat{\mathbf{h}}_k$ is the k-th user channel quantization.

Denote $r_{g,k}$ as the k-th user ranking of its CQI among W past values, and let $r_{d,k}$ be the ranking-based CDI given by the alignment between the directions of the actual channel and the quantized one, i.e. $\cos^2(\angle(\bar{\mathbf{h}}_k, \hat{\mathbf{h}}_k))$. In a centralized approach, each user reports back to the transmitter both $r_{g,k}$ and $r_{d,k}$ and the scheduler selects the user set with minimum ranking values in both CQI and CDI, thus selects the users with high instantaneous channel gain and small quantization error. In a distributed protocol, the set of scheduled users can be constructed such that only the subset $\mathcal{L}$ of users whose ranking values are below a threshold is allowed to report their CSIT to the BS. This pre-selection protocol is given by

$$\mathcal{L} = \{1 \leq k \leq K : r_{g,k} \leq \tau_g \text{ and } r_{d,k} \leq \tau_d\} \tag{6.21}$$

where τ_g, τ_d are thresholds for the channel norm and channel alignment, respectively. The fact that $r_{g,k}$, $r_{d,k}$ are uniformly distributed facilitates the calculation of optimal threshold values.

6.5 Scheduling with Heterogeneous Users

Up to this point, we considered a system with statistically identical users and studied the system throughput when all users exhibit identical average SNRs. However, in a typical wireless network, user channels are not necessarily i.i.d. and mobile terminals experience unequal average SNRs due to different distances from the BS and the corresponding different path losses (near-far effects). Hence, if a max-rate scheduler is used, the throughput will be maximized by transmitting to the users with the strongest channels. As the selected users are highly likely to be the ones closest to the BS, the issue of fairness arises. Restoring fairness requires considering a different scheduling policy that sacrifices capacity for the sake of equalizing the probability that a user is scheduled.

In heterogeneous system configurations, the sum rate is no longer an appropriate performance metric, as it cannot guarantee any fairness constraints and rate balancing among users with non-symmetric average SNRs. We focus on the problem of maximizing the weighted sum rate, in order to reflect the potential fairness issues that arise. Assume that the channel vector of each user can be written as $\mathbf{h}_k = \sqrt{\rho_k}\tilde{\mathbf{h}}_k$, where ρ_k denotes the k-th user average SNR and $\tilde{\mathbf{h}}_k \sim \mathcal{CN}(0,1)$. The equivalent channel model becomes

$$y_k = \sqrt{\rho_k}\tilde{\mathbf{h}}_k \mathbf{x} + n_k, \qquad k = 1, \ldots, K \tag{6.22}$$

We consider a weighted sum-rate maximization criterion, which results in the optimization problem

$$\begin{aligned} &\max_{\mathcal{S}\in\mathcal{G}} \sum_{k\in\mathcal{S}} w_k \mathcal{R}_k \\ &\text{s.t} \sum_{k\in\mathcal{S}} w_k = 1 \\ &\quad w_k \geq 0 \ \forall k \end{aligned} \tag{6.23}$$

where $\mathcal{R}_k$ and w_k are the rate and weighting factor of the k-th user, respectively. Let φ_k be the fraction of time slots allocated to user k, with $\sum_{k=1}^{K} \varphi_k = 1$. A general CCDF-based

user selection policy on m-th beam is defined as:

$$k_m^* = \arg \max_{1 \leq k \leq K} (1 - F_{X_{k,m}}(x_{k,m}))^{1/\varphi_k} \tag{6.24}$$

In other words, using the minimum tail scheduler, user k can gain access to the channel with probability φ_k. In [115], it has been shown that this scheduling policy can guarantee equal access to the channel for heterogeneous users. This can be also achieved if ranking-based feedback is employed during the scheduling stage. More formally, let $\mathcal{A}_{k,m}$ be the event that user k is selected on beam m based on ranking-based feedback. If all users have the same time fraction, i.e. $\varphi_k = 1/K$, then following the proof in [115] we have

$$\begin{aligned} \Pr\{\mathcal{A}_{k,m}\} &= \int_0^\infty \Pr\{\mathcal{A}_{k,m} | X_{k,m} = x\} f_{X_{k,m}}(x) dx \\ &= -\int_0^\infty \left(1 - F_{X_{k,m}(x)}\right)^{\frac{1-K}{K}} dF_{X_{k,m}(x)} = 1/K \end{aligned} \tag{6.25}$$

Interestingly, the probability that the k-th user is selected $\Pr\{\mathcal{A}_{k,m} = 1\}$ does not depend on the distribution of the other users, even if the users' channels are independent but not necessarily identically distributed. The independence of the selection probability from the other users' statistics can be inferred from the fact that the ranking of each user's CQI follows a uniform distribution independently of the other users' fading characteristics. Thus, in addition to its feedback reduction merits, ranking-based metric can also restore temporal fairness by sharing the scheduling time slots in a fair manner among users. The average user throughput for independent non-identically distributed (i.ni.d) channel statistics with $\mathcal{B} = 1$ and max-CDF scheduling is studied in [115]. In the appendix, we provide an additional proof of following result [115]:

Proposition 6.5: *The average sum rate, $\mathcal{R}$, of a heterogeneous system in which ranking-based feedback is used for the purposes of user selection is given by*

$$\mathcal{R} = \sum_{m=1}^{\mathcal{B}} K \int_0^1 F_{X_{k,m}}^{-1}(z) z^{K-1} dz \tag{6.26}$$

Proof. The proof is given in Appendix 6.C. □

6.6 Performance Evaluation

In this section, we compare the performance of following schemes:

- Scheme I: RBF employing quantized ranking-based CQI for user selection in the scheduling stage.
- Scheme II: RBF in which users are selected based on quantized SNR/SINR feedback in the scheduling stage.

Using two-stage approach, the proposed CSIT representation is used solely for selecting the group of scheduled users. Thus, in both schemes, once the group of users (among all active K ones) is identified in the first stage, the BS requests the transmission rate of the $\mathcal{M}$ selected users in order to perform link adaptation.

In the first set of simulations, we consider single-beam RBF [53] as downlink transmission scheme with $M = 2$ transmit antennas and SNR $= 10$ dB. In Figure 6.1 the throughput difference between scheme I and II is plotted as a function of observation window size W. Expectedly, for small values of W, ranking-based feedback cannot capture sufficiently the CQI distribution, failing to select the users that are on their highest quantile of their distribution. This results in a rate reduction penalty as the system does not exploit multiuser diversity and does not schedule users with large channel gains. For W increasing, the performance of ranking-based system converges to that of max-rate scheduler (for $W \to \infty$), as stated in Proposition 6.1.

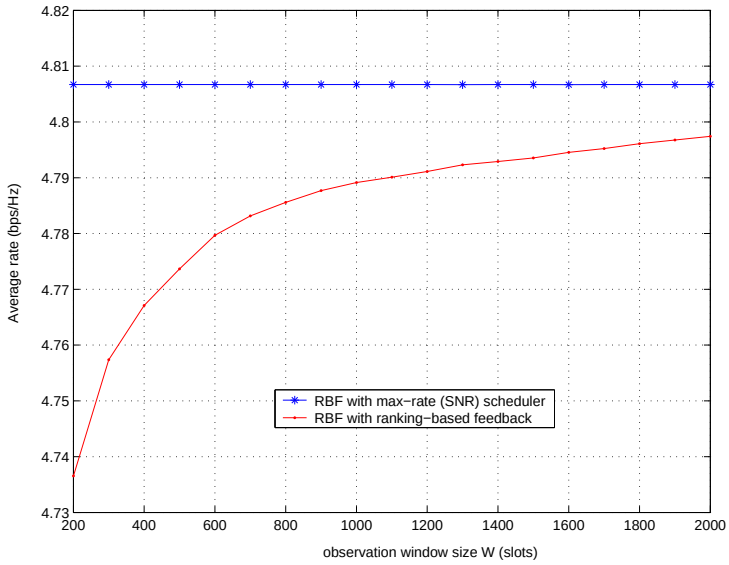

Figure 6.1: Throughput comparison as a function of window size W for single-beam RBF with $M = 2$ antennas, SNR $= 10$ dB and $K = 10$ active users.

Figures 6.2 and 6.3 show the effect of feedback quantization on the system throughput. In Figure 6.2 the SNR feedback value is quantized with $B = 5$ bits using the optimal Max-Lloyd algorithm, whereas the ranking-based CQI is quantized using $B = 3$ bits. For different values of W, the proposed feedback representation is able to identify correctly the users with the highest instantaneous rate as compared to the quantized SNR feedback, resulting in capacity gain even with a feedback load reduction of 40%. This is mainly due to the inherent digital form of ranking-based CQI and its dynamic range, which allows for efficient compression. In Figure 6.3 the performance of ranking-based user selection for different quantization bit rates is compared with that of SNR-based CQI for fixed observation window size. The feedback load can be reduced up to 40% with negligible capacity reduction (~ 0.1bps/Hz).

In the second set of simulations, the multi-beam variant of RBF [9] is used as transmission scheme. The SINR feedback is quantized using $B = 5$ bits, whereas only 3 bits are used for ranking-based CQI quantization. As shown in Figure 6.4, the proposed feedback representation in an SDMA downlink with $M = 2$ antennas provides similar results as in the single-beam case by representing more efficiently the user selection metric, thus reducing the uplink channel rate with no compromise on the system throughput.

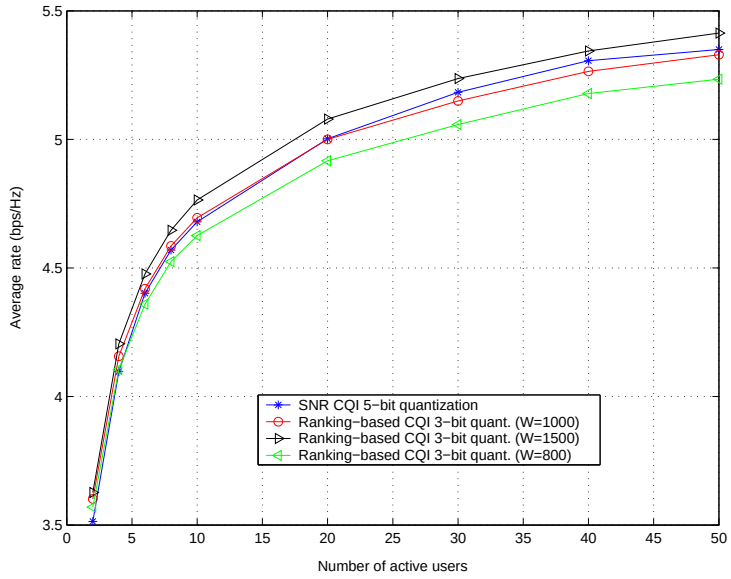

Figure 6.2: Average rate as a function of the number of users for single-beam RBF with M=2 antennas, SNR = 10 dB and different values of window size W.

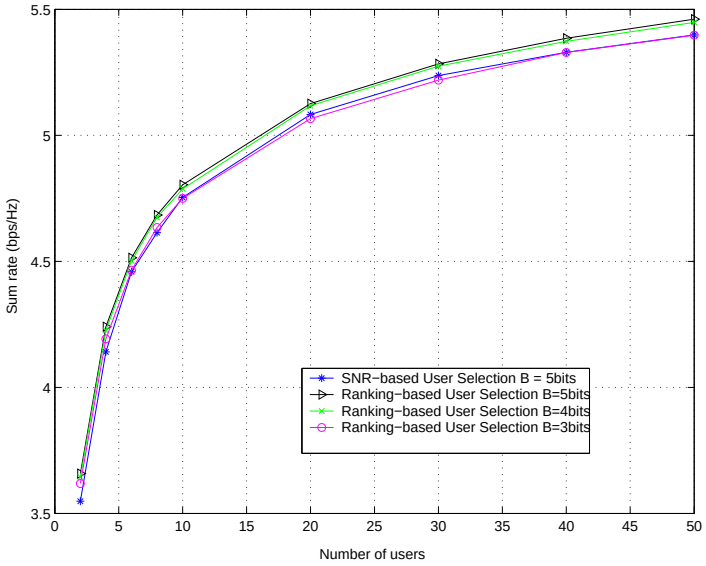

Figure 6.3: Average rate as a function of the number of users for single-beam RBF with $M = 2$ antennas, SNR = 10 dB, W=1000 slots, and ranking-based CQI metric quantized with different resolutions.

In the last part of numerical results, we study a multi-beam RBF system with $M = 4$ antennas and users with i.ni.d. channels, whose average SNRs are uniformly distributed from -10 to 30 dB. The loss in sum rate observed in Figure 6.5 is expected since in the het-

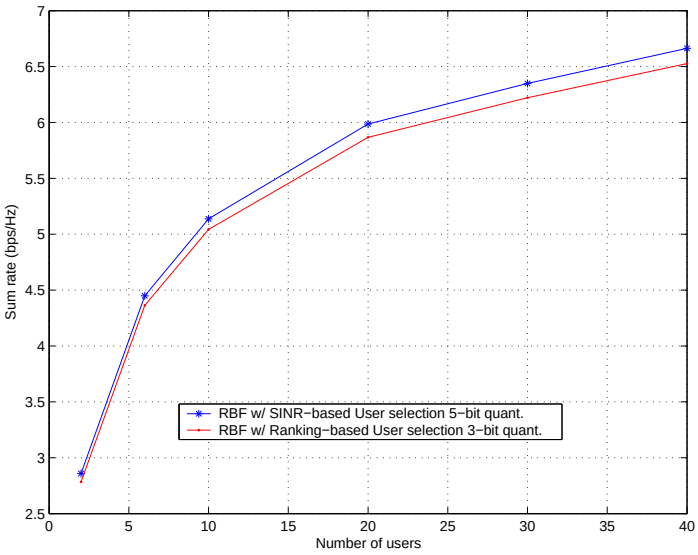

Figure 6.4: Sum rate as a function of the number of users for multi-beam RBF with $M = 2$ antennas, SNR $= 10$ dB and $W = 1000$ slots.

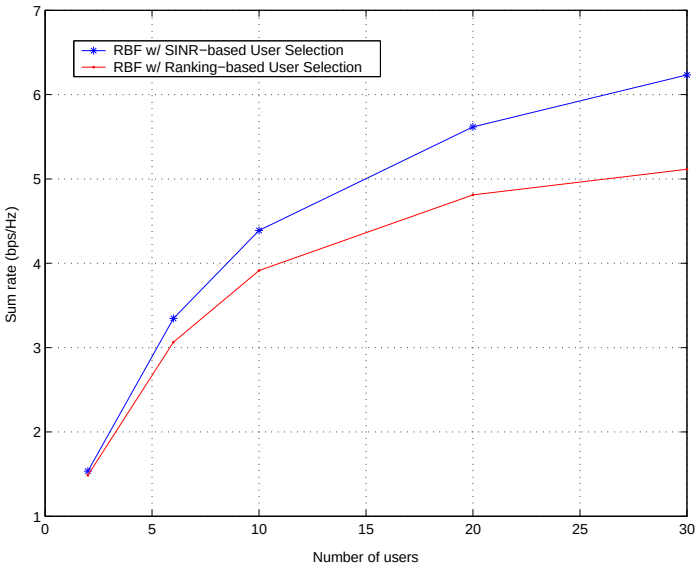

Figure 6.5: Sum rate as a function of users for multi-beam RBF in a heterogeneous network in which users' average SNRs range from -10 dB to 30 dB, $M = 4$ antennas and $W = 1000$ slots.

erogeneous network case, the users with the minimum ranking-based CQI are not generally the ones with the highest absolute instantaneous CQI values, but those whose instantaneous CQI values are near to a peak with respect to their own distribution. Nevertheless, cell-edge users that enjoy lower average SNRs have equal probability of being selected, if their channels are instantaneously on the highest quantile. Selecting users with higher pathloss

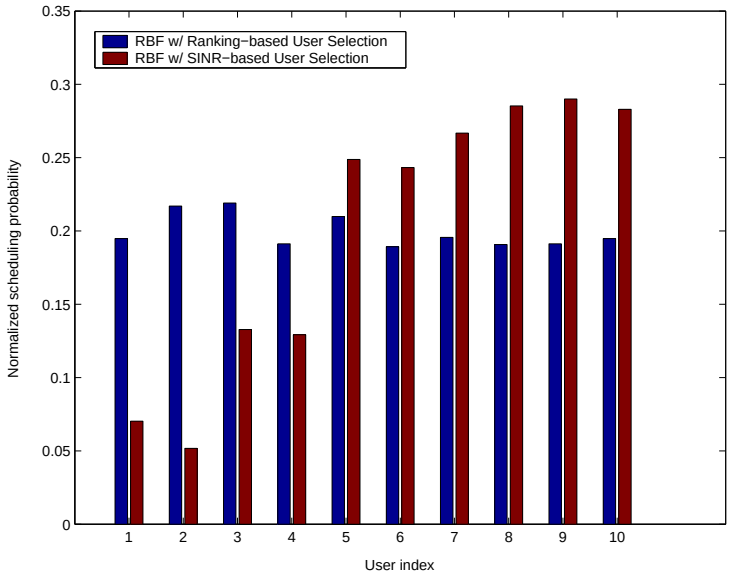

Figure 6.6: Normalized scheduling probability vs. user index for multi-beam RBF with $M = 4$ antennas and $K = 10$ users. The users are sorted from the lowest to the highest average SNR and the SNR range is from -10 dB to 30 dB.

(lower average SNR) results in system throughput reduction, however temporal fairness is restored as the access time per user is equalized independently as shown in Figure 6.6.

6.7 Conclusion

In this chapter, the problem of feedback reduction is addressed under a different perspective. We proposed a novel type of CSIT representation, coined as ranking-based feedback, as a means to further reduce the required feedback load during the scheduling stage in multi-antenna broadcast channels. Based on a two-stage scheduling/random beamforming approach, we analyzed the performance of a system in which users are preselected based on ranking-based feedback. When users exhibit i.i.d. channel statistics, it was shown that ranking-based user selection can reduce substantially (up to 40%) the uplink feedback load with negligible or no decrease in multiuser diversity gain and system throughput. In heterogeneous networks (i.ni.d. channels), temporal fairness can be achieved at little expense of throughput due to the fact that users have equal access probability to the channel medium, irrespective to the distribution of other users. In other words, users at cell edges or in deep fades (i.e. in poor channel conditions) have the same chances of being served as users that enjoy favorable channel conditions.

APPENDIX

6.A Proof of Proposition 6.1

The ranking $r_{k,m}(t)$, measured over W past samples, provides information about the empirical distribution of the rate process, i.e. $\frac{r_k(t)}{W} \approx 1 - \hat{F}^W_{X_{k,m}}(x)$. We want to show that the difference between $\hat{F}^W_{X_{k,m}}(x)$ and the actual cdf $F_{X_{k,m}}(x)$ vanishes to zero when $W \to \infty$. A measure of closeness of the two functionals, called *maximum discrepancy* (Kolmogorov-Smirnov statistic), is given by

$$D_W = \sup_{-\infty<x<\infty} \left| \hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x) \right| \tag{6.27}$$

whose probability density function is independent of $F(\cdot)$ provided that $F(\cdot)$ is continuous. Considering the above distance metric as a measure of the difference, Proposition 6.1 is a direct consequence of the following theorem:

Theorem 6.2 (Glivenko-Cantelli [120]): *Let $X_1, X_2, \ldots, X_W \sim F_{X_{k,m}}(x)$, then the sample paths of $\hat{F}^W_{X_{k,m}}$ get uniformly closer to $F_{X_{k,m}}$ as $W \to \infty$, i.e.*

$$\left\| \hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x) \right\|_\infty = \sup_x \left| \hat{F}^W_{X_{k,m}}(x) - F_{X_{k,m}}(x) \right| \overset{as}{\to} 0 \tag{6.28}$$

The above theorem implies that for large W the empirical distribution converges to the distribution function almost surely (as). Hence $\hat{F}^W_{X_{k,m}}$, which is observed over a window of size W, is almost surely a good approximation for $F_{X_{k,m}}$, and the approximation becomes better as the number of observations increases. In this case, user selection based on ranking-based CSIT becomes equivalent to max-CDF scheduling, which in turn is equivalent to max-rate scheduling for large W and i.i.d. channel distributions, i.e.

$$k^*_m(t) = \arg \min_{1\le k\le K} r_k(t) = \arg \min_{1\le k\le K} \left(1 - F_{X_{k,m}}(x_{k,m})(t)\right) = \arg \max_{1\le k\le K} x_{k,m}(t) \tag{6.29}$$

6.B Proof of Proposition 6.3

Let $F_{X_{k^*_m}}(x) = \Pr\left\{X_{k^*_m}(t) \le x\right\}$ be the rate distribution of the selected user k over beam m and $F_{w,m}(x)$ be the probability that in beam m, the w largest values among W are greater than x, then for a selected user k^*_m over beam m conditioning on $F_{w,m}(x)$ we have

$$\Pr\left\{X_{k^*_m}(t) \le x\right\} = \sum_{w=0}^{W-1} \Pr\left\{r_{k^*_m}(t) > w\right\} F_{w,m}(x) = \sum_{w=0}^{W-1} \left(\frac{W-w}{W}\right)^K F_{w,m}(x) \tag{6.30}$$

where $\Pr\left\{r_{k^*_m}(t) > w\right\} = \Pr\left\{\min_{1\le k\le K} r_{k,m}(t) > w\right\} = [1 - F_r(w)]^K = \left(\frac{W-w}{W}\right)^K$ as the ranking-based CSIT is uniformly distributed with CDF $F_r(w)$ over the set of W past values. Using results from order statistics [121], we have that

$$F_{w,m}(x) = \binom{W}{w} (F_{X_{k^*_m}}(x))^{W-w} (1 - F_{X_{k^*_m}}(x))^w \tag{6.31}$$

Therefore, the expected sum rate $\mathcal{R}$ is given by

$$\mathcal{R} = \sum_{m=1}^{\mathcal{B}} \int_0^{\infty} \Pr\left\{X_{k_m^*}(t) > x\right\} dx = \sum_{m=1}^{\mathcal{B}} \int_0^{\infty} \left(1 - \Pr\left\{X_{k_m^*}(t) \leq x\right\}\right) dx \tag{6.32}$$

$$= \sum_{m=1}^{\mathcal{B}} \int_0^{\infty} 1 - \sum_{w=0}^{W-1} \left(\frac{W-w}{W}\right)^K F_{w,m}(x) dx \tag{6.33}$$

which gives (6.11) as $F_{0,m}(z) = (F_{X_{k_m^*}}(x))^W$.

6.C Proof of Proposition 6.5

Before proceeding to the proof, we state the following result:

Lemma 6.1: *The random variable $U_{k,m} = F_{X_{k,m}}(X_{k,m})$ is uniformly distributed on the interval [0,1].*

Proof. In the lines of [115], suppose that x is an arbitrary number and $u = F_{X_{k,m}}(x)$, with $0 \leq u \leq 1$. The distribution function (CDF) of $U_{k,m}$ is given as

$$\begin{aligned} F_{U_{k,m}}(u) &= \Pr\{U_{k,m} \leq u\} = \Pr\left\{F_{X_{k,m}}(X_{k,m}) \leq u\right\} \\ &= \Pr\left\{X_{k,m} \leq F^{-1}_{X_{k,m}}(u)\right\} = u, \quad 0 \leq u \leq 1 \end{aligned} \tag{6.34}$$

which implies that $U_{k,m}$ is uniformly distributed on [0,1]. □

The average sum rate of RBF is given by

$$\mathcal{R} = \sum_{m=1}^{\mathcal{B}} \mathcal{R}_{k,m} \tag{6.35}$$

where $\mathcal{R}_{k,m}$ is the average rate of the selected user k on beam m given by

$$\mathcal{R}_{k,m} = \mathbb{E}\{X_{k,m}^{(K)}\} \tag{6.36}$$

where $X_{k,m}^{(K)} = \max\left\{X_{k,m}^1, X_{k,m}^2, \ldots, X_{k,m}^K\right\}$ (maximum over K i.i.d. random variables) with $X_{k,m}^i \sim X_{k,m}$. Since $\mathbb{E}\{X_{k,m}^{(K)}\} = \mathbb{E}\{F^{-1}_{X_k,m}(U_{k,m}^{(K)})\}$ with $U_{k,m}^{(K)} = \max\left\{U_{k,m}^1, U_{k,m}^2, \ldots, U_{k,m}^K\right\}$, from order statistics [121] (eq. 3.1.1) we have that

$$\mathbb{E}\{F^{-1}_{X_{k,m}}(U_{k,m}^{(K)})\} = K \int_0^1 F^{-1}_{X_{k,m}}(z) z^{K-1} dz \tag{6.37}$$

Inserting (6.37) into (6.35) results in (6.26).

Chapter 7

System Aspects in Multiuser MIMO Systems

7.1 Introduction

MIMO techniques have been widely recognized as a key technology in the evolution of next-generation broadband wireless access systems. Their potential for high spectral efficiency, increased diversity, and interference suppression has motivated significant amount of work and research, not only from academia but also from numerous companies that try to implement and commercialize multiuser MIMO technology. The scarce bandwidth resources, the introduction of data services and best effort applications, the transition from circuit-switched to packet-switched networks, as well as the need for enhanced quality of service (QoS) are some of the motivating factors that made MIMO technology key element of forthcoming wireless systems. Multiuser multi-antenna techniques are currently envisioned in 3GPP long term evolution (LTE), WCDMA/HSDPA, IEEE 802.16e (WiMAX), and IEEE 802.11n.

For applications such as wireless LANs, broadband wireless MANs and cellular telephony, MIMO systems will likely be deployed in environments where a single base station communicates and delivers information to multiple users sharing the same spatial channel. In such network deployments, the spatial degrees of freedom offered by multiple antennas can be advantageously exploited to enhance system throughput, by scheduling simultaneously multiple users. The design of multiuser MIMO systems hinges on the problem of the joint design of a good antenna combining technique (e.g. beamforming, space-time coding) with a properly matched channel access protocol that may include some degree of SDMA. At the heart of this problem lies that of CSIT acquisition. Information-theoretic results and throughput gain promises may often become questionable if a constraint of reasonably low-rate CSIT feedback and complexity is taken into account.

This chapter focuses on several system issues and design challenges that arise in real-world wireless system design. We discuss the main practical challenges that we should consider when deploying techniques as those proposed in Chapters 3-6. We also propose a generalization of the proportional fair scheduler (PFS) for multiuser contexts (e.g. SDMA, OFDMA, etc.).

7.2 Channel State Information Acquisition

In FDD systems and TDD systems without calibration, the only way probably to acquire channel state information at the BS from each user is through a feedback control channel, similarly to the control channels used for power control or adaptive modulation. Since the bandwidth required for those feedback control channels is considered as overhead that reduces the overall system spectral efficiency, and which grows in proportion to the number of active users, there is a substantial interest in compressing the required amount of information. The issue of feedback reduction becomes imperative in systems with wideband (e.g. OFDM) communication or high mobility (such as 3GPP-LTE and WiMAX).

7.2.1 CSI at the Receiver

Channel acquisition at the receiver is usually acquired through transmission of training sequences (pilot symbols) by the transmitter that enable the mobile terminals to perform channel estimation. It is also possible to use blind methods that do not require any training symbols but exploit knowledge of the structure of the transmitted signal or the channel. The assumption that the receiver enjoys accurate channel state information is often reasonable, especially in the downlink, where pilot-symbol-based channel estimation is more efficient since the terminals can share a common pilot channel. Note however that in practical systems, there is a tradeoff between the accuracy of CSIR and the achievable throughput, since in order to estimate the channel, a portion of the transmission time and a fraction of the power is spent to the training phase. Clearly the longer the training interval, the more accurate the channel estimate, and the higher the achievable rate; however the longer the training phase, the less time the BS disposes to transmit data to the users.

7.2.2 CSI at the Transmitter

Channel acquisition at the transmitter can be performed either implicitly or in an explicit way by relying on channel measurements at the receiver side. The methods available to gather CSI at the transmitter can be classified into two categories, relying either on reciprocity or feedback.

Implicit CSIT: reciprocity-based acquisition

The reciprocity principle is based on the property that electromagnetic waves propagating in both directions will undergo the same propagation phenomena, thus in systems operating at the same frequency band in both uplink and downlink (TDD systems), the instantaneous forward channel is identical to the transpose of the reverse channel. Therefore, the BS can

estimate the downlink channel from the uplink as long as the downlink-uplink switching time is much smaller than the channel coherence time.

Ideally, reciprocity requires the forward and reverse channels to operate at the same frequency, the same time, and the same antenna array. Although this assumption may not always hold in practice, reciprocity still hold if any time lag between the forward and reverse transmissions is much smaller than the channel coherence time T_c. Similarly, any frequency offset must be much smaller than the channel coherence bandwidth B_c, and the antenna location differences between uplink and downlink must be much smaller than the channel coherence distance d_c.

Reciprocity-based channel acquisition is usually applied in TDD systems, whereas it is normally not applicable in FDD systems where the temporal and spatial dimensions may be identical, but the frequency offset between uplink and downlink is generally much larger than the channel coherence bandwidth. However, the users' spatial signatures vary more slowly that fast fading. Therefore, depending on the angle spread, channel directional information can be extracted from the uplink even in FDD systems. Note also that the reciprocity principle requires RF hardware chains with identical frequency transfer function characteristics. Therefore, accurate RF chain calibration must be performed periodically to track the slow time variations of the RF chains and adjust the difference in the frequency response.

Explicit CSIT: feedback-based acquisition

Feedback of CSI through an uplink channel is employed in system settings where the uplink and downlink utilize different frequency bands (e.g. FDD systems), or when the reciprocity-based approach in a TDD system is not reliable due to temporal variation of the channel. In this approach, the channel is first estimated at the receiver side and then conveyed to the transmitter using a feedback link.

In the previous chapter, we used the idealized assumption of infinite rate and zero-delay feedback channel. However, for channels with relatively small coherence time, e.g. multiuser outdoor systems with high mobility users, the zero-delay and error-free assumptions are often unreasonable. The feedback delay between the measured channel and the one employed by the transmitter may result in outdated CSI which can be a significant source of error.

Channel acquisition using feedback can be applied in both TDD and FDD systems; however it is more common in FDD scenarios. Although feedback-based channel acquisition has been successfully applied in simple systems, the requirement on uplink bandwidth can become prohibitively large for complex system settings such as frequency-selective MIMO channels. Moreover, in wideband systems, obtaining CSIT and CSIR per each subcarrier can be costly in terms of training overhead. However, the performance of feedback and channel estimation can be improved by exploiting the high degree of correlation between channels of adjacent subcarriers.

7.3 Codebook-based Precoding

Codebook-based downlink precoding has been already considered as transmission scheme for next-generation wireless standards (e.g. HSDPA) and has sparked a vivid debate in

3GPP-LTE standardization activities. Current scenarios envisaged that mobile terminals use a codebook of size 2^{B_D} and are allowed to convey back to the transmitter a quantization index and a real-valued CQI via an uplink feedback channel. The candidate schemes that are envisioned can be divided in two categories:

- In the first group, often referred to as unitary precoding, the codebook contains a set of $L = 2^{B_D}/M$ pre-determined unitary beamforming matrices of size $M \times M$. Each terminal selects from the codebook the beamforming matrix that offers the highest SINR for each of its M beamforming vectors, assuming that the other $M-1$ vectors are used for transmission to interfering users. The most popular scheme of this category is per-user unitary and rate control (PU^{2}RC) [122].

- In the second group, often referred to as non-unitary precoding, the codebook contains 2^{B_D} unit-norm quantization vectors and is used by each terminal to quantize its channel vector direction (of dimension M). As the employed channel vector is normalized, this feedback value captures information regarding only the spatial direction of the channel vector. Since the terminal does not know a priori the beamforming vectors of the interfering users, the reported CQI contains an estimate (lower or upper bound) of the achievable SINR. Zero-forcing precoding is usually utilized to design the beamforming matrix.

According to the taxonomy we provided in Section 2.8, unitary precoding can be viewed as projection-based technique, while non-unitary precoding as quantization-based technique. We should note that only unitary precoding is employed in 3GPP-LTE standard. Zero-forcing beamforming, despite being proposed by several companies, has not been considered as a candidate multiuser MIMO scheme.

Several codebook design challenges arise in practice, especially since defining near-optimal quantization regions depend on various system parameters, including the channel properties and statistics and the antenna configuration and correlation. In Chapter 5 we studied quantization-based approaches considering for simplicity unstructured codebooks that contain M-dimensional random vectors. Such codebooks are designed specifically for uncorrelated channels whose direction is isotropically distributed in the unit sphere. Therefore, in practically relevant correlated channels, structured codebooks are expected to perform significantly better [123]. A practical codebook design offering good performance with line-of-sight channels or channels with a small angle spread is the Fourier codebook [124]. This codebook is simply constructed by extracting the top M rows of the discrete Fourier transform (DFT) matrix of size N_D.

Another design challenge is related to whether quantization codebooks should be common or user-specific. Clearly, the performance is increased by considering that each receiver uses a different and independently generated codebook, especially in networks with low number of users served using non-unitary precoding. If both the codebook size and the number of users to select from are small, it is highly likely that several users may quantize their channels to the same quantization vectors. Therefore, if ZFBF is applied on the channel quantizations, the probability that M near-orthogonal users are found by the scheduler is decreased (reduction in the spatial dimensions available). The complexity of generating a different codebook for each user can be reduced by generating a common, general codebook

$\mathcal{V}_g$ known at both ends of the link, and afterwards each user obtains its specific codebook through random unitary rotation of $\mathcal{V}_g$. In that case, each codevector is independent from user to user.

Finally, two questions that often arise in practice are related to the codebook size and how often it should be updated depending on the channel coherence time. For instance, although the performance of ZFBF-based codebook techniques is increased for N_D increasing, unitary precoding performs better for small codebook sizes. Actually, the multiplexing gain of unitary precoding based schemes vanishes to one for large codebooks, due to the fact that the average number of users selecting the same beamforming matrix decreases exponentially with the number of quantization bits B_D.

7.4 CQI feedback metrics and Link Adaptation

The utility of CQI feedback is two-fold: on one hand, it is employed by the SDMA scheduler as a means to select users with favorable channel conditions and separable spatial signatures. On the other hand, it is used from the link adaptation protocol to select the appropriate coding and modulation schemes and to adapt the rate of the link.

The information encapsulated in the CQI feedback parameters limits the decision and the degrees of freedom available at the transmitter. For instance, if the CQI contains information on the channel norm, the scheduler can easily identify the users with the strongest channels, but fails to derive any information on their spatial separability and the interference they cause to each other. If more than one user access simultaneously the channel, such CQI metric cannot be generally utilized by the link adaptation protocol, since the instantaneous rate allocated ignoring the inter-user interference may fall above the instantaneous mutual information of the fading channel. However, in quantization-based systems, one challenge when designing feedback metrics is that information on received SINR is in principle not available to the individual users who only have knowledge of their own channels. The SINR measurement depends, among others, on the channel as well as on the number of other mobiles being simultaneously scheduled along with the user making the measurement and their respective beamforming vectors. As user cooperation is not allowed, the number of simultaneous users and the available power for each of them will generally be unknown at the mobile. In Chapter 5, we show that in the large number of user case, simplifications arise which give the user the possibility of estimating with satisfactory accuracy the received SINR. SINR-like metrics that rely on statistical bounds can be efficient scheduling decision metrics, however they cannot guarantee QoS and information outage-free rate adaptation. Note also that in practical systems, such as HSDPA, CQIs take discrete values representing one of the possible modulation and coding schemes (MCS).

7.5 Opportunistic Scheduling: System Issues

Opportunistic scheduling protocols are designed towards a better utilization of the spectrum by granting channel access to users that experience favorable channel conditions (multiuser diversity). However, the promised throughput gains can be realized only if dynamic link adaptation techniques are available to take advantage of the improvement in channel condi-

tions. In other words, the BS should have access to channel quality measurements and the ability to adapt the rate as a function of the instantaneous CQI. Apart from the problem of feedback overhead and the requirement for channels with fast fluctuations, multiuser diversity is gained at the expense of throughput fairness and delay. In an idealized scenario where users' fading statistics are the same, the strategy of communicating with the user that exhibits the best channel maximizes not only the total capacity of the system but also the throughput of individual users. However, in practice, the statistics are not symmetric and identically distributed: there are users who are closer to the BS with higher average SNR or users at the cell edge with poor SNR; there are users who are stationary and some that are moving; there are users who are in a rich scattering environment and some with no scatterers around them. In these scenarios, opportunism may lead to unfair resource allocation since the users with poor channel conditions may get negligible or zero throughput. Due to its particular importance from a user-centric point of view, fairness is analyzed in detail in the following section.

7.6 Fairness

The concept of fairness has been extensively studied in the literature of resource allocation for wireline and computer networks [125], whereas most theoretical approaches arose from the field of political economics. In this field, the concept of utility and welfare functions were developed in order to define fairness. In order to express user's satisfaction with the service delivered by the network, utility functions are defined to formalize a notion of network performance evaluated in terms of the degree to which the network satisfies the service requirements of each user's applications. Let r_k denote the resource (service) assigned to user k, and which may contain all the relevant QoS measures (delay, throughput, packet loss, etc.). The utility function $U_k(r_k)$ maps the resource into the performance of the service. For elastic traffic, such as file transfer, email and remote terminal, which are delay tolerant and their satisfaction is generally measured in terms of bandwidth, the utility function is commonly defined as $U_k(r_k) = \log r_k$. The welfare function $W(U_1, \ldots, U_K)$ is defined as the one that aggregates the individual utility functions U_k. A fair resource allocation is the one that maximizes the welfare function $\max\{W(U_1, \ldots, U_K)\}$.

7.6.1 Definition of Fairness in Scheduling

There is no unique or general definition of fairness and one can find at least three main definitions of fairness in the resource allocation literature:

- **Max-Min Fairness**: The idea behind max-min fairness is to allocate resources as equally as possible among the competing users, thus this criterion might be the preferred option for the terminals in a bad condition, since it assures that all users receive the same resource sharing. Formally, max-min fairness is expressed

$$\max_{r_k} \min_{k} U_k(r_k) \tag{7.1}$$

  for concave utility functions. This corresponds to the welfare function $W(U_1, \ldots, U_K) = \min_{U_k}(U_1, \ldots, U_K)$. It has the property that for a feasible resource allocation vector

$\mathbf{r} = (r_1, \ldots, r_K)$, an increase of any rate within the domain of feasible rate allocations must be at the cost of a decrease of some already smaller rate, i.e. the utility $U_k(r_k)$ cannot be increased without simultaneously decreasing $U_j(r_j)$ for some j with $U_j(r_j) \leq U_k(r_k)$. Depending on the resource allocation problem, a max-min fair allocation does not always exist; however existence results in uniqueness.

- **Proportional Fairness**: The idea behind proportional fairness is to maximize the global performance, meaning that a user with bad conditions may see its utility decreased if this allows a large enough increase to a user with already good conditions (for the sake of the overall throughput). The welfare function of proportional fair allocation is $W(U_1, \ldots, U_K) = \sum_k U_k$. A rate allocation vector $\mathbf{r}$ is proportionally fair if it is feasible, and if for any other feasible allocation $\mathbf{r}^{'} = (r_1^{'}, \ldots, r_K^{'})$, the aggregate of proportional changes is zero or negative:

$$\sum_{k=1}^{K} \frac{r_k^{'} - r_k}{r_k} \leq 0 \tag{7.2}$$

- **Weighted Fairness**: If weights w_k are associated with the relative importance of each user for the system, both max-min and proportional fairness can be generalized. The welfare function for the weighted max-min fairness is then given by $W(U_1, \ldots, U_K) = \min\{U_k(r_k/w_k)\}$ and for the weighted proportional fairness is $W(U_1, \ldots, U_K) = \sum_k w_k U_k$. Under weighted fairness, each utility function is increased according to its associated weight w_k.

As the concept of fairness is generally subjective, it is not clear which definition is the best one. Normally, the scheduler selects the appropriate fairness measure for the system, depending on the burstiness of the traffic, the number of users, the price that users are willing to pay, the system time scale, etc. Two commonly used measure of fairness are: the Jain index [126] and the Gini index [127].

7.6.2 Proportional Fair Scheduler (PFS)

Proportional fair scheduler was used for the downlink scheduling in IS-856 (also known as 1xEV-DO or HDR) and was adopted in [53] as a means to meet the challenges of delay and fairness constraints while harnessing multiuser diversity. PFS maintains resource fairness by providing a fair sharing of transmission time proportional to past user throughputs over a fixed window length. On a time-slotted transmission, let $\mathcal{R}_k(t)$ be the date rate requested by user k at time slot t and supported by its instantaneous channel quality. The scheduler selects at each scheduling slot the user k^* with:

$$k^* = \arg \max_{1 \leq k \leq K} \frac{\mathcal{R}_k(t)}{\bar{\mathcal{R}}_k(t)} \tag{7.3}$$

among all active users K for which the base station has data to send. The rate $\bar{\mathcal{R}}_k(t)$ is the k-th user's average throughput in a past window of length t_c, and is updated slot-wise using an exponential filter as follows:

$$\bar{\mathcal{R}}_k(t+1) = \begin{cases} (1 - \frac{1}{t_c})\bar{\mathcal{R}}_k(t) + \frac{1}{t_c}\mathcal{R}_k(t), & k = k^* \\ (1 - \frac{1}{t_c})\bar{\mathcal{R}}_k(t), & k \neq k^* \end{cases} \tag{7.4}$$

The parameter t_c defines the time horizon in which we want to achieve fairness and is constrained by the maximum delay tolerance. Obviously, the larger t_c, the less stringent the fairness constraint, and thus longer delays start appearing between successive transmissions to the same user. For instance, in IS-856 $t_c \approx 1.67$ seconds.

Note that the above PFS rule computes the proportionally fair allocation based on the following practical result (theorem): there exists one unique PF allocation and is obtained by maximizing $\sum_k \log r_k$ over the set of feasible resource allocations. In [53] it was shown that PFS maximizes the sum of the logarithm of the average throughput $\sum_k \log \bar{\mathcal{R}}_k$ almost surely among the class of all schedulers when $t_c \to \infty$. In other words, PFS maximizes the product of user long-term average throughputs, rather than the sum throughput. Therefore, when users are charged equally in terms of price per unit share, PFS brings the maximum revenue to the network operator according to [128].

In [53], PFS exploits the multiuser diversity by assigning the radio resource to a user when its SNR is at or near its peak. In this sense, PFS can be thought as an approximation of greedy scheduling under resource fairness constraint. Its performance is affected by both the user fading statistics and the number of active users, and the optimum multiuser diversity can be obtained when each user has the same i.i.d. small-scale fading over time. Note that users with higher SNR and greater fading variations get higher throughput than those with the opposite condition. However, regardless of the user average SNR, the PFS algorithm provides equal opportunity of transmission to users with the i.i.d. fading statistics, and only slightly better chances of transmission to those with smaller channel variations over the long term [129]. Detailed theoretical analysis of the properties of PFS can be found in [130,131].

7.6.3 Multiuser Proportional Fair Scheduler (M-PFS)

PFS was originally proposed for systems that serve only one user at each scheduling window. In this section, we generalize the PFS policy for any multiuser transmission system. Let $\mathcal{G}$ be the set of all possible subsets of cardinality $|\mathcal{G}| = M$ of disjoint indices among the complete set of user indices $\{1, \cdots, K\}$. Let $\mathcal{S}_t \in \mathcal{G}$, be one such group of M users selected for transmission at a given time slot t.

Proposition 7.1: *The multiuser proportional fair scheduling policy (M-PFS) is such that the users are selected as*

$$\mathcal{S}_t^* = arg \max_{\mathcal{S} \in \mathcal{G}} \prod_{k \in \mathcal{S}} \left(1 + \frac{\mathcal{R}_{k|\mathcal{S}}(t)}{(t_c - 1)\bar{\mathcal{R}}_k(t)} \right) \tag{7.5}$$

where $\mathcal{R}_{k|\mathcal{S}}(t)$ is the rate of user $k \in \mathcal{S}$ conditioned to the scheduling set $\mathcal{S}$.

Proof. In order to show that (7.5) is a proportional fair scheduler, we need to show that it maximizes the sum of the logarithms of the average throughputs, i.e. $\sum_k \log \bar{\mathcal{R}}_k(t)$. Consider the objective function $\mathcal{J} = \sum_k \log \bar{\mathcal{R}}_k(t+1)$. Then we have:

$$\begin{aligned} \mathcal{J} &= \sum_{k \notin \mathcal{S}} \log \left(\left(1 - \frac{1}{t_c} \right) \bar{\mathcal{R}}_k(t) \right) + \sum_{k \in \mathcal{S}} \log \left(\left(1 - \frac{1}{t_c} \right) \bar{\mathcal{R}}_k(t) + \frac{1}{t_c} \mathcal{R}_{k|\mathcal{S}}(t) \right) \\ &= \sum_{k} \log \left(\left(1 - \frac{1}{t_c} \right) \bar{\mathcal{R}}_k(t) \right) + \sum_{k \in \mathcal{S}} \log \left(1 + \frac{\mathcal{R}_{k|\mathcal{S}}(t)}{(1 - t_c)\bar{\mathcal{R}}_k(t)} \right) \end{aligned} \tag{7.6}$$

The first term in (7.6) can be omitted since it does not depend on the particular choice of the scheduling set $\mathcal{S}$, hence selecting the users that maximize the objective function results in the following optimization problem:

$$\mathcal{S}_t^* = \arg\max_{\mathcal{S}\in\mathcal{G}} \mathcal{J} = \arg\max_{\mathcal{S}\in\mathcal{G}} \log \prod_{k\in\mathcal{S}} \left(1 + \frac{\mathcal{R}_{k|\mathcal{S}}(t)}{(1-t_c)\overline{\mathcal{R}}_k(t)}\right) \tag{7.7}$$

which results in (7.5) since the logarithm is a monotonically increasing function. □

By developing the above expression we have

$$\mathcal{S}_t^* = \arg\max_{\mathcal{S}\in\mathcal{G}} \left(1 + \sum_{k\in\mathcal{S}} \frac{\mathcal{R}_{k|\mathcal{S}}(t)}{(1-t_c)\overline{\mathcal{R}}_k(t)} + b\right)$$

where b is the by-products from the multiplication. If we consider a system with parallel channels, in which the rate provided to user k does not depend on the rate of users $j, j \in \mathcal{S}, j \neq k$, then b can be omitted resulting in the following M-PFS expression

$$\mathcal{S}_t^* = \arg\max_{\mathcal{S}\in\mathcal{G}} \sum_{k\in\mathcal{S}} \left(\frac{\mathcal{R}_{k|\mathcal{S}}(t)}{\overline{\mathcal{R}}_k(t)}\right) \tag{7.8}$$

We remark that (7.5) can be directly applied as the PFS policy for multiuser SDMA downlink systems, multi-carrier (e.g. OFDMA), and multi-cell networks.

Chapter 8

Conclusions and Perspectives

In this dissertation, we have focused on resource allocation and performance optimization for multiuser multi-antennas systems with incomplete CSIT. Limited feedback techniques that allow the transmitter to live well with partial channel knowledge and still achieve a significant fraction of the optimal capacity achieved under perfect CSIT is the leitmotiv of this thesis.

One first key idea is based on splitting the feedback information between the scheduling and the final beam design (or "user serving") stage, thus taking profit from the fact the numbers users to be served at each scheduling slot is much less than the number of users simultaneously requesting data packets during one given scheduling window. We introduced a two-stage framework that decouples the scheduling and beamforming problems, showing that user selection can be performed well using rough channel estimates, while the stage of serving the selected users is better accomplished with more accurate feedback. In one proposed setting, random beamforming is exploited to identify good, spatially separable, users in a first stage. In the second stage, the initial random beams of the selected users are refined based on the available feedback as a means to offer improved performance and robustness. Several refinement strategies, including beam power control and beam selection, are proposed, offering various feedback reduction and performance tradeoff. The common features of the above schemes is to restore robustness of RBF with respect to sparse network settings (low to moderate number of active users), at the cost of a moderate complexity increase. The established framework is suitable for resource allocation in slow varying multi-antenna networks with best effort, elastic traffic.

Furthermore, we have studied the problem of user selection and precoding with partial CSIT in more realistic channel scenarios. We showed that useful information that lies hidden in the second-order statistics of the channel - either in the temporal or in the spatial domain - can be exploited by the SDMA scheduler. In time-correlated channels, the redundancy (memory), which appears due to the channel structure, is exploited in order to successively refine over time the random beams of RBF. A framework, coined as memory-based oppor-

tunistic beamforming, has been established, which allows to fill the capacity gap between a purely opportunistic RBF and a channel-aware precoding and scheduling scheme with full CSIT. Our approach is suitable for low mobility (indoor) settings (i.e. limited Doppler spread), while is shown to approach the capacity of optimal unitary precoding with full CSIT for channels with large coherence time.

In spatially-correlated MIMO channels, long-term statistical channel knowledge can reveal information about the mean spatial separability of users, which is instrumental to a proper beamforming design. The merit of combining statistical and instantaneous channel information has been highlighted through several approaches. A maximum-likelihood (ML) channel estimation framework is established, which effectively combines slowly varying statistical CSIT, assumed available at the transmitter, with instantaneous low-rate CSIT. In particular, we considered both channel norm and effective channel gain (beam gain information) as scalar CQI feedback. Efficient algorithms were developed for computing the coarse ML estimates, which enable the SDMA scheduler to identify users with large gains and separable spatial signatures. A greedy user selection scheme and a low-complexity, SDMA eigenbeamforming technique based on multiuser interference bounds were also proposed and evaluated. It was demonstrated that, in systems with reasonably limited angle spread at the transmitter, such as wide-area cellular networks with elevated base stations, it is sufficient to feed back a single scalar but properly designed CQI parameter and combine it with long-term statistical CSIT in order to achieve near-optimal throughput performance.

Limited feedback strategies utilizing quantization codebooks were also investigated in the thesis. In particular, the problem of efficient, sum-rate maximizing CQI metric design is addressed. We identified several scalar feedback metrics that incorporate information on the channel gain, the channel direction, and the quantization error, and can be interpreted as reliable estimates of the received SINR. For that, bounds on the instantaneous inter-user interference when ZFBF is employed were derived. Although the exact SINR is in principle not available to the individual users, the use of interference bounds and approximate expressions results in simplifications that give users the possibility of estimating a priori their individual received SINR. It was demonstrated that scalar CQI feedback combined with CDI and efficient user selection and ZFBF can achieve a significant fraction of the capacity of the full CSIT case by means of multiuser diversity. However, a major limitation of SDMA systems relying on quantized CSIT is that they become interference dominated and their multiplexing gain is reduced at high SNR under fixed feedback load rate. Motivated by the fact that SDMA does not always outperform TDMA when the transmitter relies on incomplete CSIT, we showed the importance of dynamic SDMA/TDMA transition algorithms. Properly designed scheduling metrics allowing a soft, adaptive switching from multiuser to single-user transmission mode are shown to be a promising means to circumvent this problem, guaranteeing a linear sum-rate growth at any SNR range. Moreover, we considered a practically relevant system in which each user has a sum feedback rate constraint. A tradeoff between multiuser diversity and spatial multiplexing has been identified, since the available feedback bits ought to be shared between CDI and CQI information. The problem of optimizing the feedback bit split has been studied, revealing an interesting interplay between the number of active users, the average SNR and the feedback load.

Finally, a low-rate representation of CSIT feedback parameters, referred to as ranking-based feedback, was identified as a means to further compress the reported channel feedback

information. Each user calculates and reports to the BS the integer-valued ranking of its instantaneous CSIT among a set of stored past CSIT measurements. This alternative representation enables the scheduler to identify users that are instantaneously on the highest peak (quantile) with respect to their own channel distribution, independently of the distribution of other users. Interestingly, in non-symmetric networks, with i.ni.d. channel statistics among users, the proposed ranking-based feedback allows to restore temporal fairness since it equalizes the probability that a user will be selected, independently of its average SNR.

Future Research

The results of this dissertation shed some light on how to achieve a significant fraction of the multi-antenna broadcast capacity as promised by information-theoreric results, even when the transmitter relies on limited and incomplete channel knowledge. In parallel, the thesis brought up several interesting open issues and topics for further research, as briefly discussed in what follows.

Our work in Chapters 3 to 5 have identified linear precoding combined with efficient user selection and limited as a promising technique to achieve the sum rate of MIMO broadcast channels. Nevertheless, the results rely on several simplifying assumptions on the behavior of the feedback channel. Since the uplink channel is not instantaneous and error-free in practice, a natural extension to these results can be studying the effect of feedback channel noise, delays and CSIT estimation on the system performance. This investigation is of primary importance in high mobility networks with large Doppler spread channels where delays are more prominent. Clearly, the feedback delay would affect the validity of the feedback and would cause the scheduler to mistakenly choose users that do not have the most favorable channel conditions. One simple method would be to back off the reported CQI; however understanding the amount of back off and the effect of estimation error variance on the throughput are challenging open problems.

In all our work, except in Chapter 6, we study network settings with i.i.d. channel fading statistics. It is of particular interest to assess the real throughput gain of the proposed methods in channels with shadowing and path loss, in which the users exhibit unequal average SNRs. Such scenarios would certainly impact the multiuser diversity gains as well as the system overall sum-rate and fairness performance. Additionally, if we consider the impact of realistic traffic models and system loads, the available degrees of freedom at the disposal of the scheduler can be severely reduced. It might be of interest to identify how many effective active users are available for selection by the scheduler at each time and how to take advantage of the different degrees of freedom to satisfy the QoS constraints for different types of traffic. Fairness issues, which have not been taken into account in our work here presented, need to be incorporated, in order to provide high throughput while satisfying certain QoS constraints.

Extensions of the problem of resource allocation for multiuser multi-antenna downlink channels with limited feedback to wideband systems and multicell settings are also problems of timely relevance that require further research.

Finally, we have investigated techniques matched to a quantized (digital) channel feedback where each user sends back a suitably encoded and modulated quantization index. Nevertheless, recent findings have started considering analog feedback schemes. Although digital feedback is shown to be superior in most cases [132], such practically relevant frame-

work may give rise to hybrid digital/analog feedback approaches. For instance, the feedback link design can be modeled as a Wyner-Ziv coding problem, where the transmitter combines the digital, quantized CSIT information that combines with analog side information.

In order to conclude, we might say that the theoretical limits of multiuser multi-antenna systems are relatively well understood nowadays. However, the gap between the current practical schemes and the theoretical limits is still significant, making the optimal design of limited feedback multiuser MIMO transmission an open and exciting problem.

Bibliography

[1] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Seattle, June 1995.

[2] 3GPP, "Long Term Evolution, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; General description," *TS 36.201 v1.0.0*, March 2007.

[3] IEEE, "Air interface for fixed and mobile broadband wireless access systems amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands," *IEEE Std 802.16e-2005*, Febr. 2006.

[4] "NGMN: 'Next Generation Mobile Networks Beyond HSPA and EVDO - A white paper', V3.0," *Available at http://www.ngmn-cooperation.com*, December 2006.

[5] J.-C. Belfiore, G. Rekaya, and E. Viterbo, "The Golden code: a 2×2 full-rate space-time code with non-vanishing determinants," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1432–1436, Apr. 2005.

[6] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[7] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

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

[9] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channel with partial side information," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 506–522, Febr. 2005.

[10] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.

[11] G. Dimić and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm," *IEEE Trans. Sig. Processing*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.

[12] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Jour. on Sel. Areas in Commun. (JSAC)*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[13] T. S. Rappaport, *Wireless Communications, 2nd ed.*, Prentice Hall, NJ, 2002.

[14] T. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[15] A. El Gamal, "The capacity of a class of broadcast channels," *IEEE Trans. Inform. Theory*, vol. 25, no. 2, pp. 166–169, Mar. 1979.

[16] W. Yu and J. Cioffi, "The sum capacity of a Gaussian vector broadcast channel," *IEEE Trans. Inform. Theory*, vol. 50, no. 9, pp. 1875–1892, Sept. 2004.

[17] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[18] P. Viswanath and D. N. Tse, "Sum capacity of the vector Gaussian channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[19] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple access and broadcast channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

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

[21] W. Yu and T. Lan, "Minimax duality of Gaussian vector broadcast channels," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Chicago, IL, USA, June 2004.

[22] W. Yu and W. Rhee, "Degrees of freedom in wireless multiuser spatial multiplex systems with multiple antennas," *IEEE Trans. Commun.*, vol. 54, no. 10, pp. 1744–1753, Oct. 2006.

[23] D. J. Mazzarese and W. A. Krzymien, "Throughput maximization and optimal number of active users on the two transmit antenna downlink of a cellular system," in *Proc. IEEE Pacific Rim Conf. on Commun., Comp. and Sig. Processing (PACRIM)*, Victoria, BC, Canada, Aug. 2003.

[24] P. Bergman, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. 19, no. 3, pp. 197–207, Mar. 1973.

[25] S. A. Jafar and A. Goldsmith, "Isotropic fading vector broadcast channels: The scalar upper bound and loss in degrees of freedom," *IEEE Trans. Inform. Theory*, vol. 51, no. 3, pp. 848–857, Mar. 2005.

[26] A. Lapidoth, S. Shamai (Shitz), and M. Wigger, "On the capacity of fading MIMO broadcast channels with imperfect transmitter side-information," in *Proc. of 43rd Allerton Conf. on Commun., Control and Comput.*, Monticello, IL, USA, Sept. 2005.

[27] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1250–1276, June 2002.

[28] T. Philosof, U. Erez, and R. Zamir, "Combined shaping and precoding for interference cancellation at low SNR," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Yokohama, Japan, June 2003.

[29] U. Erez and S. ten Brink, "Approaching the dirty paper limit for canceling known interference," in *Proc. 41st. Allerton Conf. on Com., Cont. and Comp.*, Monticello IL, USA, Oct. 2003.

[30] A. Bennatan, D. Burstein, G. Caire, and S. Shamai (Shitz), "Superposition coding for side information channels," in *Proc. of Int. Symp. on Inform. Theory and Its Appl. (ISITA)*, Parma, Italy, Oct. 2004.

[31] F. Boccardi, F. Tosato, and G. Caire, "Precoding Schemes for the MIMO-GBC," in *Proc. of Int. Zurich Sem. on Comm. (IZS'06)*, Zurich, Switzerland, Febr. 2006.

[32] R. F. H. Fischer and C. H. Windpassinger, "Improved MIMO precoding for decentralized receivers resembling concepts from lattice reduction," in *Proc. IEEE Glob. Telecom. Conf. (Globecom)*, San Francisco, CA, USA, Dec. 2003.

[33] W. Yu and J. Cioffi, "Trellis precoding for the broadcast channel," in *Proc. IEEE Glob. Telecom. Conf. (Globecom)*, San Antonio, TX, USA, Nov. 2001.

[34] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multi-antenna multi-user communication - part I: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[35] M. Airy, A. Forenza, R. W. Heath Jr., and S. Shakkottai, "Practical Costa pre-coding for the multiple antenna broadcast channel," in *Proc. IEEE Glob. Telecom. Conf. (Globecom)*, Dallas, TX, USA, Nov. 2004.

[36] C. Windpassinger, R. F. H. Fischer, and J. B. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, pp. 2057–2060, Dec. 2004.

[37] R. F. H. Fischer and C. H. Windpassinger, "Even-integer precoding for broadcast channels," in *Proc. of 5th Int. ITG Conf. on Source and Ch. Coding (SCC)*, Munich, Germany, Jan. 2004.

[38] 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, pp. 537–544, Mar. 2005.

[39] M. Tomlinson, "New automatic equalizer employing modulo arithmetic," *Electronics Letters*, vol. 7, no. 5/6, pp. 138–139, Mar. 1971.

[40] M. Miyakawa and H. Harashima, "A method of code conversion for a digital communication channel with intersymbol interference," *Trans. Inst. Electron. Commun. Eng. (IEICE) Japan*, vol. 52-A, pp. 272–273, June 1969.

[41] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Trans. Vehic. Tech.*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[42] M. Stojnic, H. Vikalo, and B. Hassibi, "Maximizing the sum-rate of multi-antenna broadcast channels using linear preprocessing," *IEEE Trans. Wireless Comm.*, vol. 5, no. 9, pp. 2338–2342, Sept. 2006.

[43] Z. Tu and R. S. Blum, "Multi-user diversity for a dirty paper approach," *IEEE Comm. Lett.*, vol. 7, no. 8, pp. 370–372, Aug. 2003.

[44] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 11–15, Jan. 2007.

[45] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Sig. Processing*, vol. 52, no. 2, pp. 461– 471, Febr. 2004.

[46] B. Hochwald and S. Viswanath, "Space time multiple access: linear growth in the sum rate," in *Proc. of 40th Allerton Conf. on Commun., Control and Comput.*, Monticello, IL, USA, Oct. 2002.

[47] A. Lapidoth, "On the high-SNR capacity of non-coherent networks," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3025–3036, Sept. 2005.

[48] W. Choi J. G. Andrews, "The capacity gain from base station cooperative scheduling in a MIMO DPC cellular system," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Seattle, WA, USA, July 2006.

[49] L. Breiman, "On some limit theorems similar to the arc-sin law," *Theory of Probability and its Appl.*, vol. 10, 1965.

[50] M. R. Leadbetter and H. Rootzen, "Extremal theory for stochastic processes," *Ann. Probab.*, vol. 16, pp. 431–478, 1988.

[51] D. J. Love, R. W. Heath Jr., W. Santipach, and M. L. Honig, "What is the value of limited feedback for MIMO channels?," *IEEE Comm. Mag.*, vol. 42, no. 10, pp. 54–59, Oct. 2003.

[52] P. Ding, D. J.Love, and M. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. Sig. Processing*, vol. 55, no. 7, pp. 3417–3428, July 2007.

[53] P. Viswanath, D. N. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[54] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE Jour. on Sel. Areas in Commun. (JSAC)*, vol. 16, no. 8, pp. 1423–1436, Oct. 1998.

[55] S. Zhou, Z. Wang, and G. B. Giannakis, "Quantifying the power loss when transmit beamforming relies on finite rate feedback," *IEEE Trans. Wireless Comm.*, vol. 4, no. 4, pp. 1948–1957, July 2005.

[56] K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.

[57] W. Santipach and M. Honig, "Asymptotic capacity of beamforming with limited feedback," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Chicago, IL, USA, July 2004.

[58] D. Love, R. W. Heath Jr. and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[59] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing Lines, Planes, etc.: Packings in Grassmannian Spaces," *Journal of Exper. Math.*, vol. 5, pp. 139–159, 1996.

[60] W. Santipach and M. Honig, "Signature optimization for CDMA with limited feedback," *IEEE Trans. Inform. Theory*, vol. 51, no. 10, pp. 3475–3492, Oct. 2005.

[61] C. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. Wireless Comm.*, vol. 6, no. 2, pp. 458–462, Febr. 2007.

[62] T. Yoo, N. Jindal, and A. Goldsmith, "Finite-rate feedback MIMO broadcast channels with a large number of users," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Seattle, WA, USA, July 2006.

[63] J. C. Roh and B. D. Rao, "Transmit beamforming in multiple-antenna systems with finite rate feedback: A VQ-based approach," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1101–1112, Mar. 2006.

[64] T.Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna broadcast channels with limited feedback and user selection," *IEEE Jour. on Sel. Areas in Commun. (JSAC)*, vol. 25, no. 7, pp. 1478–1491, Sept. 2007.

[65] J. Wagner, Y.-C. Liang, and R. Zhang, "On the balance of multiuser diversity and spatial multiplexing gain in random beamforming," *to appear in IEEE Trans. Wireless Comm.*, 2007.

[66] N.Zorba and A. I. Pérez-Neira, "Robust multibeam opportunistic schemes under quality of service constraints," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Glasgow, Scotland, June 2007.

[67] N.Zorba and A. I. Pérez-Neira, "A multiple user opportunistic scheme: the Grassmannian approach," in *Proc. of Int. Zurich Sem. on Comm. (IZS'06)*, Zurich, Switzerland, Febr. 2006.

[68] R. Bosisio, J. L. Vicario, C. Anton-Haro, and U. Spagnolini, "Diversity-multiplexing tradeoff in multi-user scenario with selective feedback," in *Proc. of 15th IST Mob. & Wir.Commun. Summit,*, Mykonos, Greece, June 2006.

[69] N. V. Smirnov, "Limit distributions for the terms of a variational series," *Trudy Math. Inst. Steklov (trans. by Amer. Math. Soc. Tran., vol. 67, no. 16, 1952)*, vol. 25, 1949.

[70] M. Maddah-Ali, M. Ansari, and A. Khandani, "An efficient signaling scheme for MIMO broadcast systems: design and performance evaluation," *submitted to IEEE Trans. Inform. Theory*, July 2005.

[71] R. Zakhour and D. Gesbert, "A two-stage approach to feedback design in multi-user MIMO channels with limited channel state information," in *Proc. IEEE Int. Symp. on Pers., Indoor and Mobile Radio Comm. (PIMRC)*, Athens, Greece, Sept. 2007.

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

[73] N. Jindal, W. Rhee, S. Vishwanath, S.A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[74] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Jour. on Sel. Areas in Commun. (JSAC)*, vol. 13, no. 7, pp. 1341–1347, Sept. 1995.

[75] K. W. Shum, K. K. Leung, and C. W. Sung, "Convergence of iterative waterfilling algorithm for Gaussian interference channels," *IEEE Jour. on Sel. Areas in Commun. (JSAC)*, vol. 25, no. 6, pp. 1091–1100, Aug. 2007.

[76] J. Papandriopoulos and J. S. Evans, "Low-complexity distributed algorithms for spectrum balancing in multi-user DSL networks," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Istanbul, Turkey, June 2006.

[77] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, 1952.

[78] T. Furuta, "Specht ratio S(1) can be expressed by Kantorovich constant K(p): S(1)=exp[K'(1)] and its application," *Math. Inequalities & Applications*, vol. 6, no. 3, pp. 521–530, 2003.

[79] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press Inc., New York, 1994.

[80] W. Dziubdziela, "On convergence rates in the limit laws of extreme order statistics," in *Trans. 7th Prague Conf. and 1974 Europ. Meeting of Statisticians*, 1974.

[81] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input queued switch," in *Proc. of IEEE Conf. on Computer Commun. (INFOCOM)*, San Francisco, CA, USA, Mar. 1996.

[82] P. Giaccone, B. Prabhakar, and D. Shah, "Towards simple, high-performance schedulers for high-aggregate bandwidth switches," in *Proc. of IEEE Conf. on Computer Commun. (INFOCOM)*, New York, NY, USA, June 2002.

[83] C.-S. Hwang, W. Lee, and J. Cioffi, "Randomized Scheduler for Temporally-Correlated Channels," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP)*, Hawaii, HI, USA, Apr. 2007.

[84] D. Avidor, J. Ling, and C. B. Papadias, "Jointly opportunistic beamforming and scheduling (JOBS) for downlink packet access," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Paris, France, June 2004.

[85] P. Svedman, L. J. Cimini, M. Bengtsson, S. K. Wilson, and B. Ottersten, "Exploiting temporal channel correlation in opportunistic SD-OFDMA," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Istanbul, Turkey, June 2006.

[86] M. Bengtsson D. Samuelsson and B. Ottersten, "Improved multiuser diversity using smart antennas with limited feedback," in *Proc. of Europ. Sig. Proces. Conf. (EUSIPCO)*, Antalya, Turkey, Sept. 2006.

[87] D. Hammarwall and B. Ottersten, "Exploiting the spatial information provided by channel statistics and SNR feedback," in *Proc. IEEE Workshop on Sign. Proc. Adv. in Wireless Comm. (SPAWC)*, Cannes, France, July 2006.

[88] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Beamforming and user selection in SDMA systems utilizing channel statistics and instantaneous SNR feedback," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP)*, Hawaii, HI, USA, Apr. 2007.

[89] R. B. Ertel, P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Pers. Commun.*, vol. 5, no. 1, pp. 10–22, Febr. 1998.

[90] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Sälzer, "Low complexity scheduling and beamforming for multiuser MIMO systems," in *Proc. IEEE Sig. Proc. Adv. on Wir. Commun. (SPAWC'06)*, Cannes, France, July 2006.

[91] S. Jafar, S. Viswanath, and A. Goldsmith, "Channel capacity and beamforming for multiple transmit and multiple receive antennas with covariance feedback," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Helsinki, Finland, June 2001.

[92] K. Huang, J. G. Andrews and R. W. Heath Jr., "Orthogonal beamforming for SDMA downlink with limited feedback," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP)*, Hawaii, HI, USA, Apr. 2007.

[93] C. Swannack, G. Wornell, and E. Uysal-Biyikoglu, "MIMO Broadcast Scheduling with Quantized Channel State Information," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Seattle, WA, USA, July 2006.

[94] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Sälzer, "Efficient metrics for scheduling in MIMO broadcast channels with limited feedback," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP)*, Hawaii, HI, USA, Apr. 2007.

[95] M. Trivellato, F. Boccardi, and F. Tosato, "User selection schemes for MIMO broadcast channels with limited feedback," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Dublin, Ireland, Apr. 2007.

[96] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Sälzer, "Efficient metric for Scheduling in MIMO Broadcast Channels with Limited Feedback," FT060303 - *France Telecom R&D internal report*, Mar. 2006.

[97] N. Jindal, "Finite Rate Feedback MIMO Broadcast Channels," in *Workshop on Inform. Theory and its Applications (ITA)*, UC San Diego, USA (invited), Febr. 2006.

[98] T. Yoo and A. Goldsmith, "Sum-rate optimal multi-antenna downlink beamforming strategy based on clique search," in *Proc. IEEE Glob. Telecom. Conf. (Globecom)*, St. Louis, MO, USA, Dec. 2005.

[99] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "Finding NEMO: Near Mutually Orthogonal sets and applications to MIMO broadcast scheduling," in *Proc. Int. Conf. on Wir. Networks, Commun. and Mob. Computing*, June 2005.

[100] C.-P. Chen and F. Qi, "The best bounds of the n-th harmonic number," *Global Jour. of Math. and Math. Sciences 2 (accepted)*, 2006.

[101] L. Zheng and D. N. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[102] F. Floren, O. Edfors, and B.-A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. IEEE Glob. Telecom. Conf. (Globecom)*, San Francisco, CA, USA., Dec. 2003.

[103] D. K. Sharma, "Design of absolutely optimal quantizers for a wide class of distortion measures," *IEEE Trans. Inform. Theory*, vol. 24, no. 6, pp. 693–702, Nov. 1978.

[104] X. Wu and K. Zhang, "Quantizer monotonicities and globally optimal scalar quantizer design," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 1049–1053, May 1993.

[105] N. E. Nörlund, *Vorlesungen über Differenzenrechnung*, Chelsea Publishing Company, New York, 1954.

[106] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Paris, France, June 2004.

[107] V. Hassel, M.-S. Alouini, D. Gesbert, and G. Oien, "Exploiting multiuser diversity using multiple feedback thresholds," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Stockholm, Sweden, May-June 2005.

[108] S. Sanayei and A. Nosratinia, "Opportunistic beamforming with limited feedback," in *Proc. Asilomar Conf. on Sign., Syst. and Computers*, Pacific Grove, CA, USA, Nov. 2005.

[109] S. Sanayei and A. Nosratinia, "Exploiting multiuser diversity with only 1-bit feedback," in *Proc. IEEE Wireless Comm. and Net. Conf.*, New Orleans, LA, USA, Mar. 2005.

[110] J. Diaz, O. Simeone and Y. Bar-Ness, "How many bits of feedback is multiuser diversity worth in MIMO downlink?," in *Proc. of IEEE Int. Symp. on Spread Spect. Techn. and Appl. (ISSSTA)*, Manaus, Brazil, Aug. 2006.

[111] J. Diaz, O. Simeone and Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Seattle, WA, USA, July 2006.

[112] K. Huang, R. W. Heath Jr., and J. G. Andrews, "Space division multiple access with a sum feedback rate constraint," *IEEE Trans. Sig. Processing*, vol. 55, no. 7, pp. 3879–3891, July 2007.

[113] T. Tang and R. W. Heath Jr., "Opportunistic feedback for downlink multiuser diversity," *IEEE Comm. Lett.*, vol. 9, no. 10, pp. 948–950, Oct. 2005.

[114] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," in *Proc. of European Wireless*, Barcelona, Spain, Febr. 2004.

[115] D. Park, H. Seo, H. Kwon, and B. G. Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Trans. Commun.*, vol. 53, no. 11, pp. 1919–1929, Nov. 2005.

[116] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks with heterogeneous users," in *Proc. Conf. on Inform. Sciences and Systems (CISS)*, Princeton, NJ, USA, Mar. 2004.

[117] S. Patil and G. de Veciana, "Measurement-based opportunistic scheduling for heterogeneous wireless systems.," submitted to IEEE/ACM Transactions on Networking, Jan. 2006.

[118] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Statistics*, vol. 27, pp. 642–669, 1956.

[119] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Ann. Probab.*, vol. 18, no. 3, pp. 1269–1283, 1990.

[120] P. Billingsley, *Probability and Measure, 3rd edition*, J. Wiley and Sons, Inc., New York, 1995.

[121] H. A. David and H. N. Nagaraja, *Order Statistics, 3rd edition*, J. Wiley and Sons, Inc, New York, 2003.

[122] Samsung, "Downlink MIMO for EUTRA," *3GPP TGS RAN WG1, R1-063028*, Febr. 2006.

[123] Texas Instruments, "Evaluation of codebook-based precoding for LTE MIMO systems," *3GPP TGS RAN WG1, R1-061439*, May 2006.

[124] Philips, "System-level simulation results for channel vector quantisation feedback for MU-MIMO," *3GPP TGS RAN WG1, R1-063028*, Nov. 2006.

[125] R. Denda, A. Banchs, and W. Effelsberg, "The fairness challenge in computer networks," in *Proc. of 1st Int. Work. on Quality for Fut.Internet Services (QofIS)*, Berlin, Germany, Sept. 2000.

[126] R. Jain, D. M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," in *DEC Research Report TR-301*, Sept. 1984.

[127] C. Gini, "Measurement of inequality and incomes," *The Economic Journal*, vol. 31, no. 121, pp. 124–126, Mar. 1921.

[128] F. Kelly, "Charging and rate control for elastic traffic," *Europ. Trans. on Telecom.*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[129] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proc. IEEE Int. Symp. on Pers., Indoor and Mobile Radio Comm. (PIMRC)*, San Diego, CA, USA, Sept. 2001.

[130] S. C. Borst and P. A. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proc. of IEEE Conf. on Computer Commun. (INFOCOM)*, Anchorage, AK, USA, Apr. 2001.

[131] D. Avidor, S. Mukherjee, J. Ling, and C. B. Papadias, "On some properties of the proportional fair scheduling policy," in *Proc. IEEE Int. Symp. on Pers., Indoor and Mobile Radio Comm. (PIMRC)*, Barcelona, Spain, Sept. 2004.

[132] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Quantized vs. analog feedback for the MIMO downlink: A comparison between zero-forcing based achievable rates," in *Proc. IEEE Int. Symp. Info. Th. (ISIT)*, Nice, France, 2007.