Capacity-Approaching Data Transmission in MIMO Broadcast Channels

—A Cross-Layer Approach

Jing Jiang

Dissertation submitted to the Faculty of
the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

R. Michael Buehrer (Co-Chair)
William H. Tranter (Co-Chair)
Jeffrey H. Reed
Ira Jacobs
Scott F. Midkiff
Naren Ramakrishnan

July 6, 2004
Blacksburg, Virginia

Keywords: Multi-Antenna Broadcast Channel, Capacity, Diversity, Spatial Multiplexing, Precoding

Capacity-Approaching Data Transmission in MIMO Broadcast Channels —A Cross-Layer Approach

Jing Jiang

Abstract

This dissertation focuses on downlink multi-antenna transmission with packet scheduling in a wireless packet data network. The topic is viewed as a critical system design problem for future high-speed packet networks requiring extremely high spectral efficiency. Our aim is to illustrate the interaction between transmission schemes at the physical layer and scheduling algorithms at the medium access control (MAC) layer from a sum-capacity perspective. Various roles of multiple antennas are studied under channel-aware scheduling, including diversity, beamforming and spatial multiplexing. At a system performance level, our work shows that downlink throughput can be optimized by joint precoding across multiple transmit antennas and exploiting small-scale fading of distributed multiple input and multiple output (MIMO) channels.

There are three major results in this dissertation. First, it is shown that over a MIMO Gaussian broadcast channel, and under channel-aware scheduling, open-loop transmit antenna diversity actually reduces the achievable sum rate. This reveals a negative interaction between open-loop antenna diversity and the closed-loop multiuser diversity through scheduling. Second, a suboptimal dirty paper coding (DPC) approach benefits greatly from multiuser diversity by an efficient packet scheduling algorithm. Performance analysis of a suboptimal greedy scheduling algorithm indicates that, compared with the receiver-centric V-BLAST method, it can achieve a much larger scheduling gain over a distributed MIMO channel. Further, pre-interference cancellation allows for transmissions free of error propagation. A practical solution, termed Tomlinson-Harashima precoding (THP), is studied under this suboptimal scheduling algorithm. Similar to V-BLAST, a reordering is applied to minimize the average error rate, which introduces only a negligible sum-rate loss in the scenarios investigated. Third, for an orthogonal frequency division multiplexing (OFDM) system using MIMO precoding, it is shown that a DPC-based approach is readily applicable and can be easily generalized to reduce the peak-to-average power ratio (PAR) up to 5 dB without affecting the receiver design. Simulations show that in an interference-limited multi-cell scenario, greater performance improvement can be achieved by interference avoidance through adaptive packet scheduling, rather than by interference diversity or averaging alone. These findings suggest that, coordinated with channel-aware scheduling, adaptive multiplexing in both spatial and frequency domains provides an attractive downlink solution from a total capacity point of view.

Acknowledgment

I feel very blessed to have learned from and have worked with so many incredible individuals during my Ph.D. study. Foremost, I thank my advisor Professor William H. Tranter for his guidance and support throughout the past five years. I have benefited tremendously from his unique blend of wisdom, vision and technical insights. I thank my co-advisor R. Michael Buehrer for being a great teacher and a friend. Professor Buehrer's practical sensibility, generosity and integrity will be an inspiring role model for my future career. I thank Professor William J. Ebel for introducing me to space-time codes and for taking me into such an amazing area in wireless communication. I am very much encouraged by his enthusiasm. I also thank my committee for always being available and very supportive.

I thank many of my previous colleagues for their constant encouragement and guidance in my career. I owe them a great deal for their friendship. The discussion with Richard Goodson on precoding and shaping techniques provided inspiration for my PhD research. I have benefited greatly from his leadership at Adtran and profound thinking. I thank Keith Conner for teaching me equalization techniques in GSM systems and for good career advice. I also thank Shupeng Li for introducing me to the network simulator, which helped me grow a strong interest in network simulation. I am very grateful for the friendship from some past MPRG alumni, in particular Hao Xu and Zhigang Rong, with whom I have had opportunities to collaborate on various subjects. I thank all my colleagues at MPRG for many stimulating discussions on simulation methodologies, multi-antenna techniques, new commercial products and industrial trends. I am deeply indebted to them for their ideas and insights, and will miss this fun and intellectually stimulating environment.

I thank my family for always being a great comfort and encouragement, and to them I devote this work. Above all, I give my gratitude and praise to God the Father for his fathomless mercy and my Lord Jesus Christ for his sacrificial love.

Finally, I acknowledge the Virginia Tech Electrical and Computer Engineering Department and the National Science Foundation for providing funding for my Ph.D. study, and the IEEE/Motorola Daniel E. Noble Fellowship for additional support.

Contents

List of Symbols[1]

$\mathbb{R}^{m\times n}$:	set of $m\times n$ real matrices
$\mathbb{C}^{m\times n}$:	set of $m\times n$ complex matrices
$\mathbf{A}_{ij}$:	or $[\mathbf{A}]_{ij}$, the (i,j)-th entry of matrix $\mathbf{A}$
$(\mathbf{A})^T$:	transposition of matrix $\mathbf{A}$
$(\mathbf{A})^H$:	Hermitian transposition of matrix $\mathbf{A}$
$\det(\mathbf{A})$:	determinant of a nonsingular square matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$:	rank of a matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$:	trace of a square matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	diagonal matrix formed by diagonal entries of matrix $\mathbf{A}$
$\mathrm{E}(\bullet)$:	expectation of a random scalar/vector/matrix
$\mathbf{A}^{-1}$:	inverse of a nonsingular square matrix $\mathbf{A}$
$\mathbf{A}^{+}$:	pseudoinverse of matrix $\mathbf{A}$
$\|\mathbf{A}\|_F$:	Frobenius (or Euclidean) norm of matrix $\mathbf{A}$
$\mathbf{I}_n$:	identity matrix of dimension n
$\|\mathbf{x}\|_2$:	or $\|\mathbf{x}\|$, L_2 (or Euclidean) norm of vector $\mathbf{x}$
$\mathrm{Re}(\bullet)$:	real part of a complex scalar/vector/matrix

[1] Bold-face and lower-case letters are used to denote vectors. Unless specified otherwise, all vectors are column vectors.

$\sim$:	is distributed as
$\triangleq$:	is defined as
$\approx$:	is approximately equal to
$\mathcal{CN}(\mu,\Sigma)$:	complex circular symmetric Gaussian distribution with mean μ and covariance matrix Σ
$(x)_+$:	$\max(0,x)$
$\lfloor x \rfloor$:	floor function; the largest integer not greater than x
$\lceil x \rceil$:	ceiling function; the smallest integer not less than x
ch($\mathbf{A}$)	the lower triangular matrix from Cholesky factorization of matrix $\mathbf{A}$
$\mathcal{A}\backslash\mathcal{B}$	set $\mathcal{A}$ excluding set $\mathcal{B}$
$\mathcal{A}\bigcup\mathcal{B}$	union of set $\mathcal{A}$ and set $\mathcal{B}$
$\mathcal{A}\bigcap\mathcal{B}$	intersection of set $\mathcal{A}$ and set $\mathcal{B}$
$\lvert\mathcal{A}\rvert$	cardinality of set $\mathcal{A}$
$\{\mathcal{A}\}^c$	complement of set $\mathcal{A}$
$\mathcal{P}(\mathbf{A})$	projection onto the orthogonal complement of the subspace spanned by the rows of matrix $\mathbf{A}$
C_n^m:	the number of m combinations out of n
P_n^m:	the number of m permutations out of n
$N!$:	the factorial of a non-negative integer N with $0!\triangleq 1$
vec($\mathbf{A}$):	vectorized matrix $\mathbf{A}$ formed by stacking the columns of $\mathbf{A}$ over each other
$\mathbf{A}\otimes\mathbf{B}$:	Kronecker product (direct product) of the matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A}>0$:	matrix $\mathbf{A}$ is positive definite

$\delta[i-j]$: Kronecker delta function, $\delta[i-j] \triangleq \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$

$u(x)$: unit step function, $u(x) \triangleq \begin{cases} 1, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$

$Q(x)$: Q function, $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$

$\mathcal{X}_n^2$: chi-square distribution with n degrees of freedom

List of Acronyms

cdf	cumulative distribution function
i.i.d.	independent and identically distributed
pdf.	probability density function
pmf	probability mass function
AWGN	additive white Gaussian noise
BC	broadcast channel
BLER	block error rate
BSC	base station controller
BTS	base station transceiver
CDMA	code division multiple access
CIR	carrier to interference ratio
COOP	cooperative (receivers)
CP	cyclic prefix
CQI	channel quality information
CSI	channel state information
CSIR	CSI at receiver only
CSP	complex sphere precoding
DPC	dirty paper coding
DSL	digital subscriber loop
DSM	degraded same marginal
FDD	frequency division duplex
FH-OFDM	frequency-hopped OFDM
GDFE	generalized decision-feedback equalizer
H-ARQ	hybrid automatic repeat request
HDR	high data rate
IP	Internet protocol
LAN	local area networks

MAC	medium access control or multiaccess channel
MAN	metropolitan area network
MIMO	multiple input and multiple output
MISO	multiple input and single output
MMSE	minimum mean square error
MMSE-DPC	minimum mean square error dirty paper coding
MOLS	mutually orthogonal latin square
MRC	maximal ratio combining
MSE	mean square error
MRC-BF	MRC beamforming
MSC	mobile switching center
MTSO	mobile telephone switching office
M-PSK	M-nary phase shift keying
M-QAM	M-nary quadrature amplitude modulation
OFDM	orthogonal frequency division multiplexing
OIN-OFDM	OFDM system with opportunistic interference nulling
PAR	peak-to-average power ratio
PDP	power delay profile
PF	proportional fairness or proportional fair
PSTN	public switched telephone network
P/S	parallel to serial conversion
QAM	quadrature amplitude modulation
QoS	quality of service
QPSK	quadrature phase shift keying
RZF-DPC	regularized ZF-DPC
SER	symbol error rate
SIC	successive interference cancellation
SINR	signal to interference and noise ratio
SISO	single input and single output
SNR	signal to noise ratio

S/P	serial to parallel conversion
TCP	transmission control protocal
TDMA	time division multiple access
TDD	time division duplex
THP	Tomlinson-Harashima precoding
V-BLAST	vertical Bell labs layered space-time
VLSI	very large scale integration
ZF-DPC	zero-forcing DPC
ZF-LBF	zero-forcing linear beamforming
1xEV-DO	CDMA20001x evolution – data only
2/3G	second/third generation

List of Tables

List of Figures

Chapter 1
Introduction

Wireless communication dates back to 1897 when Guglielmo Marconi successfully demonstrated wireless telegraphy. In the following century, various wireless systems have flourished, and often disappeared later. As we look back today, the past decade has witnessed the success of the second-generation (2G) digital wireless cellular standards, the dramatic progress in VLSI technology, which has enabled the small-area and low-power implementation of advanced signal processing algorithms and coding techniques, and a research shift from voice to data networks. This dissertation is motivated by the recent advancements in wireless, which is still being driven by the ever increasing demand for high-speed tetherless connections to realize "anytime and anywhere" communications.

1.1 Motivation and Scope

Information theory started with Shannon's paper "A Mathematical Theory of Communication" published in 1948 [1], which also founded the modern theory of communication. Before 1995, the mainstream of wireless communication research focused on improving radio link reliability and capacity motivated by Shannon's work. The field changed around 1995, when the seminal papers by Foschini [2] and Teletar [3] independently showed that in a rich-scattering environment, the capacity of a point-to-point wireless link can be greatly increased by employing multi-element antenna arrays at both ends of the link. The underlying wireless link with multiple inputs and multiple outputs (MIMO), possibly

correlated, is therefore named a *MIMO channel*. Generally, multiple antennas can play multiple roles, such as diversity, beamforming, spatial multiplexing, or interference cancellation. This dissertation focuses on the role of multiple antennas in a multiuser downlink channel, where multiple antennas at geographically isolated users constitute one end of a MIMO link in a distributed sense.

Practical communication systems are often multiuser in nature. For example, the mobile wireless system is inherently a multiuser system because signals of mobile users share a common air interface. Compared with single-user communication, multiuser communication is substantially more complicated. This complexity results mainly because real systems are often constrained by limited physical resources such as power and bandwidth, and economical designs are needed to allow multiple users to share scarce resources efficiently. The major purpose of this dissertation is to study such system designs so that the total throughput is maximized under certain *quality-of-service* (QoS) constraints. This throughput-oriented design is also motivated by 3G/4G wireless systems that target delay-tolerant, high-speed data transmission, in particular Internet access.

In future wireless systems, packet data services based on TCP/IP are expected to be a dominant application at data rates that are orders-of-magnitude higher than are supportable today. In the context of high-speed packet transmission, the design philosophy of 2G wireless systems is seriously challenged by new design requirements in terms of both multiple asymmetric links and various QoS provisions. While the traditional design philosophy seeks to isolate functionality of the network into distinct layers, recent development of wireless communications has led to a broader network perspective, and pointed towards cross-layer design. A second goal of this dissertation is therefore to explore the methodology of cross-layer design. To this end, a joint-design approach is taken across the multi-antenna transmission schemes at the *physical layer* and the channel-aware

scheduling algorithms at *medium access control* (MAC) layer.

While the main objective of this dissertation is to explore the performance benefits of various data transmission strategies over MIMO broadcast channels, which have been identified as the bottleneck of future wireless communication systems with Internet access, our focus is on the interaction between multi-antenna transmission and multiuser scheduling. Motivated by recent progress in information theory and the success of 2G wireless digital standards, we take an information-theoretic perspective in the context of cellular systems throughout this dissertation. We consider the case of a distributed MIMO channel in which users do not cooperate at reception. In such a case, we try to achieve a system-wide benefit by explicitly leveraging so called “dirty paper coding” [4-5] for interference cancellation at transmitter and the intelligent packet scheduling for multiuser diversity. From a cross-layer point of view, we concentrate on the coordination between the physical-layer data transmission and the MAC-layer packet scheduling, and evaluate the throughput benefit from the information sharing across these two layers. Our major interest is on the management of fading and interference in a multiuser data network and its ramification across multiple layers, rather than the network-layer issues of routing, dissemination of control information, and so forth, which are of primary concern in most ad hoc networks. Although only the interaction between the physical layer and MAC layer is considered, we believe the lessons learned will shed light on future wireless communication design and provide insight into joint optimization.

1.2 Overview of Cellular Systems

This section presents a brief overview of cellular systems. Throughout our work, we present our analysis in the context of a cellular architecture for the sake of example. However,

it should be noted that both the radio transmission and the packet scheduling techniques discussed in this dissertation are not necessarily confined to a cellular network. Rather, the entire work would be applicable to any network with centralized control of specific geographic areas, e.g., wireless local area network (WLAN).

1.2.1 Cellular Architecture

A cellular network consists of a large number of wireless subscribers (mobile users), and a number of fixed base stations arranged to provide coverage to the subscribers. The area covered by a base station is called a cell, which is often represented by a hexagonal region with the base station at the cell center. Theoretically, a hexagonal lattice of cells will lead to the maximum covered area given a fixed number of base stations (or investment). An idealized hexagonal cell coverage is illustrated in Figure 1.1 with one tier of cells around the center cell.

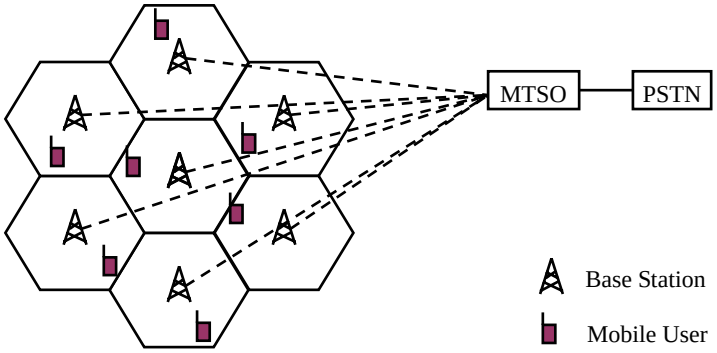


Figure 1.1. An idealized hexagonal cell layout with centered base stations.

In reality, however, the base station placements are somewhat irregular, depending on the locations where base stations can be established. Correspondingly, the actual mobile users connected to a base station are chosen by good communication paths rather than geographic distance. In Figure 1.1, the *mobile telephone switching office* (MTSO, also called a *mobile switching center* MSC) connects base stations in a given area by high speed wire

connections or microwave links. The MTSO is then connected to the *public switched telephone network* (PSTN) to establish the wireless-to-wired network connection. In a cellular network, the MTSO plays a major role in coordinating which base station will handle a call to or from a mobile user and when to handoff a mobile user from one base station to another. From a system perspective, we see that a cellular network is not an independent network, but rather an appendage to the wired network.

We differentiate two types of channels in a cellular network. The wireless link from a base station to a mobile user is interchangeably called the *downlink* or *forward channel*, and the wireless link from a mobile user to a base station is called the *uplink* or *reverse channel*. In a cellular network, there are usually many users connected to a single base station, and thus, for the forward channels, the base station must multiplex together the signals to the various connected users and broadcast one waveform from which each user can extract its own signal. The combined channel from the one base station to the multiple users is called a *broadcast channel*. For the reverse channels, each user connected to a given base station transmits his own waveform, and the base station receives the sum of the waveforms from the various users plus noise. The base station must then separate the signals from each user and forward these signals to the MTSO. The combined channel from each user to the base station is called a *multiaccess channel*.

1.2.2 Wireless Challenges

There are several aspects of wireless communication that make the problem challenging and interesting. These aspects are by and large not as significant in wireline communication. The first is the phenomenon of *fading*: the time-variation of the signal strength due to both the small-scale effect of multipath propagation, as well as the larger scale effects such as shadowing by obstacles and path loss via distance attenuation. Among

the three factors contributing to signal fading, path loss is the factor which varies the slowest. This factor normally takes time on the order of minutes or hours for noticeable change depending on the mobility of the user. Empirically, the path loss is often modeled as an exponential loss with distance, with the attenuation rate specified by a path loss exponent that can be obtained by measurement. The shadowing attenuation varies at a rate faster than the path loss, and the duration of a shadow fade can last for several seconds or minutes. Also the attenuation due to shadowing is exponential in the width of the barrier that must be passed through. In reality, the shadowing attenuation is often modeled by a log-normal random variable with a variance depending on the scattering structures [6].

The small-scale fading factor due to multipath propagation occurs on a time scale of milliseconds (depending on the speed of mobility), which can significantly affect the performance of the modulation and coding techniques over the link. This fast fluctuation of the signal envelope is due to the constructive and destructive combination of multiple reflected radio wave paths. In a dense multipath environment with sufficient scatterers (rich scattering), the multipath attenuation can be effectively modeled by a complex Gaussian random variable at baseband. This attenuation is known as *Rayleigh fading*. If there is also a non-negligible line-of-sight signal component, the so-called *Rician fading* model with a deterministic non-zero mean is more appropriate. In the evaluation of wireless link performance, the signal fading rate relative to the data rate is an important parameter. A relevant concept is the *coherence time* [6] which is a measure of the time duration over which the channel gain stays almost constant, or highly correlated with a correlation coefficient above 0.5. A fading process is called a *fast fading* channel if the signal symbol duration is greater than the fading coherence time. Otherwise it is called a *slow fading* channel. The transmitted symbol duration is also used as a reference to measure the frequency selectivity of channel fading. The channel fading is *frequency selective* if the delay spread of the

channel power delay profile (PDP) [6] is greater than one-symbol duration; otherwise, it is *flat*. Therefore, channel fading can affect both the temporal statistics of the signal, and the signal frequency response with its frequency selectivity.

Second, since the wireless medium is shared among multiple users, there is significant *interference* between them. Therefore unlike wired networks, wireless communication is interference-limited rather than noise-limited. In a cellular network, the interference can be between the uplinks or downlinks, or between users in different cells. In the multiaccess channel based on CDMA with frequency reuse factor of 1, the out-of-cell interference can be usually modeled as additive white Gaussian noise (AWGN) from the central limit theorem. However, the assumption may not be valid in a broadcast channel where the out-of-cell interference is only from a few close base stations [7]. Furthermore, the interference pattern in wireless communication is also contingent on the mode of duplex, for example, frequency division duplex (FDD) or time division duplex (TDD). Finally, interference is also influenced by whether the network is operated synchronously or asynchronously.

In addition to fading and interference, mobility is an issue in wireless communication. High mobility makes channel estimation and prediction hard to implement. Mobility can also make resource allocation difficult, increasing the handoff overhead.

Finally, in wireless communications, the signal processing at the mobile user set is constrained by the limited battery energy and cost. Low complexity often dominates the mobile transceiver design.

1.3 Contributions

The main contributions of this work are:

- A study of the interaction between antenna diversity and scheduling in wireless packet data networks;
- High-speed transmission schemes for downlink packet data using spatial multiplexing and scheduling;
- Performance evaluation of a joint design using multiple antennas and multiuser scheduling in a MIMO broadcast channel;
- A new algorithm for joint scheduling and reordering in a suboptimal dirty paper coded system and performance evaluation via simulation;
- An original algorithm for peak to average power ratio reduction in a MIMO-OFDM system using nonlinear precoding;
- A system-level performance study of a multi-cell MIMO-OFDM system using multiple transmit antennas and intelligent scheduling.

Paper submissions and publications:

- J. Jiang, R. M. Buehrer, and W. H. Tranter, "Peak to average power ratio reduction with power constraint for MIMO-OFDM wireless system using nonlinear precoding," accepted for presentation at IEEE Globecom 2004, Dallas, TX, Nov/Dec 2004.
- J. Jiang, R. M. Buehrer, and W. H. Tranter, "Multi-antenna downlink data transmission with precoding and scheduling," to be submitted to IEEE Trans. Vehicular Tech., July 2004.
- J. Jiang, R. M. Buehrer, and W. H. Tranter, "Adaptive spatial multiplexing using nonlinear precoding in multiuser MIMO OFDM systems," in preparation.
- J. Jiang, R. M. Buehrer, and W. H. Tranter, "Antenna diversity in multiuser data networks," IEEE Trans. on Commun., vol. 52, pp. 490-497, Mar 2004.
- J. Jiang, R. M. Buehrer, and W. H. Tranter, "High-speed downlink packet

transmission with spatial multiplexing and scheduling," in Proc. IEEE Wireless Commun. and Networking Conf., WCNC 2004, Atlanta, GA, USA, Mar 2004.

- J. Jiang, R. M. Buehrer, and W. H. Tranter, "Greedy scheduling performance for a dirty paper coded system," submitted to IEEE Trans. on Commun., Sept 2003.
- J. Jiang, R. M. Buehrer, and W. H. Tranter, "Spatial T-H precoding for packet data systems with scheduling," in Proc. IEEE VTC2003-Fall, Orlando, Florida, USA, Oct. 2003.
- J. Jiang, R. M. Buehrer, and W. H. Tranter, "A network view of radio links," in Propagator, Spring/Summer, 2003.
- K. K. Bae, J. Jiang, and W. H. Tranter, "Downlink WCDMA performance analysis with diversity techniques combined with beamforming," in Proc. IEEE Wireless Commun. and Networking Conf., WCNC 2003, vol. 1, pp. 202-206, Mar 2003.
- J. Jiang, K. K. Bae, and W. H. Tranter, "Downlink WCDMA bit error rate performance in indoor system applications," in Proc. IEEE 56th Vehicular Technology Conf., VTC2002-Fall, vol. 2, pp. 825–828, Vancouver, Canada, Sept. 2002.
- K. K. Bae, J. Jiang, and W. H. Tranter, et al., "WCDMA STTD Performance with Transmitter Location Optimization in Indoor Systems using Ray Tracing Technique," in Proc. IEEE Radio and Wireless Conf., RAWCON2002, pp. 123-127, Sept. 2002.

Joint work with VT Computer Science Department:

- J. He, A. Verstak, L. T. Watson, C. A. Stinson, N. Ramakrishnan, C. A. Shaffer, T. S. Rappaport, C. R. Anderson, K. Bae, J. Jiang, W. H. Tranter, "Globally optimal transmitter placement for indoor wireless communication systems", to

appear in IEEE Trans. on Wireless Commun., 2004.

- A. Verstak, N. Ramakrishnan, L. T. Watson, J. He, C. A. Shaffer, K. Bae, J. Jiang, W. H. Tranter, and T. S. Rappaport, “BSML: A binding schema markup language for data interchange in PSEs”, Scientific Programming, 10(4), Dec 2002.
- J. He, L. T. Watson, N. Ramakrishnan, C. A. Shaffer, A. Verstak, J. Jiang, K. Bae, and W. H. Tranter, “Dynamic data structures for a direct search algorithm”, Computational Optimization and Applications, 23(1), pages 5-25, Oct 2002.
- A. Verstak, J. He, L. T. Watson, N. Ramakrishnan, C. A. Shaffer, T. S. Rappaport, C. R. Anderson, K. Bae, J. Jiang, and W. H. Tranter, “S4W: Globally optimized design of wireless communication systems”, in Proc. of the Next Generation Software Workshop, 16th Intl. Parallel and Distributed Processing Symposium (IPDPS'02), Fort Lauderdale, FL, April 2002.
- J. He, A. Verstak, L. T. Watson, T. S. Rappaport, C. R. Anderson, N. Ramakrishnan, C. A. Shaffer, W. H. Tranter, K. Bae, J. Jiang, “Global optimization of transmitter placement in wireless communication systems”, in Proceedings of the High Performance Computing Symposium, Advanced Simulation Technologies Conference, A. Tentner (Ed.), Society for Modeling and Simulation International, San Diego, CA, pages 328-333, April 2002.

1.4 Outline of Dissertation

The results of this dissertation are built on multiuser information theory. Chapter 2 presents an overview of information theory with a specific focus on the capacity region of

broadcast channels. Starting with the capacity region of a scalar broadcast channel, it includes recent developments on the MIMO broadcast channel. Throughout this work, it is assumed that the channel state information (CSI), partially or fully, is available at both the transmitter and receivers. The concepts of capacity outer bounds and sum capacity are introduced. Chapter 2 also describes some signal processing and coding schemes which achieve or approach the channel capacity region or sum capacity from an information-theoretic point of view. Chapter 2 is thus entirely background material.

Chapter 3 begins the original work presented in this dissertation and studies the interaction between antenna diversity and multiuser diversity from a sum-capacity point of view. When multiple antennas are used for pure diversity purposes, a sum-capacity analysis shows that the orthogonal-design based space-time codes [8] actually reduces the average sum capacity in the presence of channel-aware scheduling. This result is contrary to that commonly observed over a single wireless link, where open-loop transmit diversity always improves link performance in fading. The implication of the proportional fair (PF) scheduling [9] on receive antenna diversity is also investigated.

Chapter 4 focuses on a suboptimal dirty paper coding technique termed “zero-forcing dirty paper coding” (ZF-DPC) [10]. The chapter begins with review of the theory of ZF-DPC. The maximum sum rate of ZF-DPC is presented and its sum-rate optimality at both high and low SNRs is shown. Also revealed is its duality to V-BLAST [11], [12] strategy. Chapter 4 then proposes original suboptimal scheduling algorithms for ZF-DPC based transmission and presents performance results. To demonstrate the performance advantage of ZF-DPC in a distributed-MIMO broadcast channel with transmitter knowledge of CSI, ZF-DPC is compared with that of V-BLAST under suboptimal scheduling. Finally, it is shown that analogous to V-BLAST, an optimal reordering can be imposed on top of scheduling to minimize the average error rate, while the associated sum-rate loss is negligible.

Chapter 5 starts with a practical precoding scheme named Tomlinson-Harashima precoding (THP) [13], [14]. Essentially a generalized decision-feedback equalizer (GDFE), THP has roots in digital subscriber loop (DSL) systems. The achievable rate of THP is analyzed and its connection with ZF-DPC is established in the chapter. Under the proposed suboptimal scheduling, Chapter 5 provides the original performance analysis of THP for symmetric users in both i.i.d. and correlated Raleigh flat fading channels. THP simulations corroborate substantial average error-rate improvement with reordering.

In Chapter 6, the concept of ZF-DPC is extended to an OFDM system on a per-tone basis. A generalized THP technique, termed complex sphere precoding, is introduced to reduce the peak-to-average power ratio (PAR) of the MIMO-precoded OFDM signals. An efficient implementation of a modified complex Fincke-Pohst algorithm is illustrated and its application for PAR reduction is introduced. Simulations show that a significant PAR reduction can be achieved by a sequential search over reduced and reordered individual tones. Multiuser scheduling across multiple tones also indicates substantial scheduling gain through multiuser diversity.

Chapter 7 compares two design philosophies based on interference avoidance and interference averaging, respectively, in a multi-cell MIMO-OFDM system using the previous nonlinear precoding. In particular, we demonstrate that the total system throughput gain is dominated by multiuser scheduling over all tones, and that the scheduling-based interference avoidance outperforms the orthogonal frequency-hopping based interference averaging. This is particular advantageous for packet transmission where effective power control is problematic due to the rapid arrival and departure of interfering packets. This performance advantage is still apparent even under the proportional fairness constraint, which is used as a tradeoff of total throughput for user delay and resource fairness.

Chapter 8 summarizes the main points of the dissertation.

Chapter 2
Review of Broadcast Channel Capacity

This chapter gives an overview of Gaussian broadcast channel capacity. The goal is to characterize the capacity region and to provide insights on optimal transmission strategies over MIMO broadcast channels. We start with known results and proceed to the recent developments in the area.

2.1 Capacity Region and Sum Capacity

A broadcast channel with K users is defined by an input alphabet $\mathcal{X}$, K output alphabets $\mathcal{Y}_k$, $k = 1,...,K$, and a conditional probability distribution

$$P(y_1, y_2, ..., y_K \mid x)$$

where $x \in \mathcal{X}$ and $y_k \in \mathcal{Y}_k$, $k = 1,...,K$. The *capacity region* $\mathcal{C}$ of this broadcast channel is the closure of the set of K-tuple rate (R_1,R_2,…,R_K), at which reliable communications can be made simultaneously. The *sum capacity* is the maximum sum rate $\sum_{k=1}^{K} R_k$ over the capacity region. An important feature of broadcast channel is that broadcast channels that have the same marginal distributions $P(y_k \in \mathcal{Y}_k \mid x)$, for $k = 1,...,K$, have the same capacity region $\mathcal{C}$ [15]. Therefore, we can consider any joint distribution with the right conditional marginals for a broadcast channel with capacity region $\mathcal{C}$. In the remainder of this section, we review some fundamental results on the capacity characterization for broadcast channel.

2.1.1 Degraded Broadcast Channel

A broadcast channel is *physically degraded* if its transition probability can be factored as

$$P(y_1,...,y_K \mid x) = P(y_{\pi(1)} \mid x)P(y_{\pi(2)} \mid y_{\pi(1)}) \cdots P(y_{\pi(K)} \mid y_{\pi(K-1)}) \qquad (2.1)$$

for some permutation $\pi(\cdot)$ on the user set $\{1,\ldots,K\}$. Or equivalently, a broadcast channel is physically degraded if the input and outputs

$$X \to Y_{\pi(1)} \to Y_{\pi(2)} \to ... \to Y_{\pi(K)}$$

form a Markov chain. A broadcast channel can also be *stochastically degraded* if its conditional marginal distributions are the same as those of a physically degraded broadcast channel. From the broadcast channel property above, we know that the capacity region of a stochastically degraded broadcast channel is the same as that of the corresponding physically degraded broadcast channel. Hence, from the capacity region point of view, we will not differentiate these two classes of degraded broadcast channels, and call both degraded from the capacity perspective.

The capacity region of a degraded broadcast channel is known to be the convex hull of the closure of rate tuple $(R_1,R_2,\ldots,R_K)$ satisfying [16]

$$0 \le R_k \le I(U_k; Y_k \mid U_{k+1}U_{k+1}...U_K),\ k = 1,...,K, \qquad (2.2)$$

where $P(u_1,...,u_K, x, y_1,..., y_K)$ factors as

$$P(u_1,...,u_K, x)P(y_1 \mid x)\prod_{k=2}^{K} P(y_k \mid y_{k-1})$$

with U_k representing the message for receiver k, and $I(X;Y)$ denoting the mutual information between X and Y. The corresponding coding theorem [16] and the converse [17] were first proved in early 70's, which gave only an information-theoretical expression for the capacity

region. No information on the optimal input distribution was given for a general degraded broadcast channel, nor the practical code construction. However, for a degraded broadcast channel, it is obvious from (2.1) that maximizing the sum rate is the same as sending to the best user, since all the other received signals $Y_{\pi(k)}$, k>1, are just degraded versions of $Y_{\pi(1)}$.

However, for a special degraded broadcast channel, scalar Gaussian broadcast channel, Bergmans first showed that Gaussian inputs are optimal [18]. A scalar Gaussian broadcast channel has

$$Y_k = X + Z_k, \quad k = 1, ..., K \tag{2.3}$$

where the complex input random variable is subject to a power constraint $\mathrm{E}\left[| X |^2 \right] \le P$, and Z_k is independent and identically distributed (i.i.d.) circular symmetric complex Gaussian noise with variance N_k, which is denoted as $Z_k \sim \mathcal{CN}(0, N_k)$. Figure 2.1 illustrates an equivalent model [15] for a degraded scalar Gaussian broadcast channel of two users.

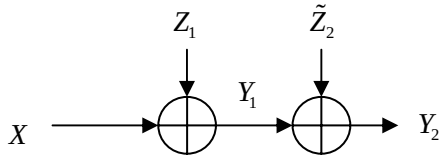


Figure 2.1 A degraded scalar Gaussian broadcast channel for K=2.

For this channel model, we have

$$Y_1 = X + Z_1$$

and

$$Y_2 = X + Z_2 = Y_1 + \tilde{Z}_2$$

for $Z_1 \sim \mathcal{CN}(0, N_1)$ and $\tilde{Z}_2 \sim \mathcal{CN}(0, N_2 - N_1)$. We see that one output can be expressed as a degraded version of the other output. The model can be easily generalized for K users.

Consider the capacity region for a scalar Gaussian broadcast channel with K users and choose

$$X = \sum_{k=1}^{K} U_k$$

where U_k are independent, circular symmetric complex Gaussian random variables distributed as $U_k \sim \mathcal{CN}(0, Q_k)$ with a total power constraint $\sum_{k=1}^{K} Q_k \le P$, then the achievable rate region can be defined by [19]

$$0 \le R_k \le \log_2(N_k + \sum_{i=1}^{k} Q_i) - \log_2(N_k + \sum_{i=1}^{k-1} Q_i),\ k = 1, ..., K \qquad (2.4)$$

It was shown by Bergmans [18] that capacity is achieved by optimizing over all Q_k, $k = 1, ..., K$, with $\sum_{k=1}^{K} Q_k \le P$, and Gaussian input distributions are optimal in this case.

2.1.2 Vector Broadcast Channel

A vector broadcast channel has vectors as inputs and outputs. For a vector Gaussian broadcast channel, we have the output $\mathbf{Y}_k \in \mathbb{C}^{m_k \times 1}$ at receiver k as

$$\mathbf{Y}_k = \mathbf{H}_k \mathbf{X} + \mathbf{Z}_k$$

for $k = 1, ..., K$. The channel input $\mathbf{X} \in \mathbb{C}^{n \times 1}$ has the covariance matrix Q_X and a power constraint $\mathrm{tr}(Q_X) \le P$, where $\mathrm{tr}(\cdot)$ is the trace operator. The channel matrix of user k is represented by $\mathbf{H}_k \in \mathbb{C}^{m_k \times n}$, and the noise vector $\mathbf{Z}_k \in \mathbb{C}^{m_k \times 1}$ at user k is i.i.d. and circular symmetric complex Gaussian. The vector Gaussian broadcast channel is of practical importance to the multi-antenna downlink transmission in wireless systems, which is the focus of this dissertation. But unfortunately, the capacity region of a general vector Gaussian broadcast still remains unknown, and moreover, a vector Gaussian broadcast channel is not degraded in general.

Despite these difficulties, recent progress in the characterization of broadcast channel capacity has shown that a coding technique named dirty paper coding [4] achieves the sum capacity of a MIMO broadcast channel [20]-[22]. Further, if Gaussian inputs are optimal, it has been proved that dirty paper coding actually achieves the capacity region of a vector broadcast channel [23], [24]. In the following, the achievable rate region of dirty paper coding (dirty paper region in [20]) is described for a general MIMO broadcast channel. Before that, we first introduce some known capacity bounds for the broadcast channel in general, and the vector broadcast channel in specific.

2.2 Capacity Bounds of the Broadcast Channel

We start with a general bound given by Cover [15] and Bergmans [16], which shows that the capacity region of a broadcast channel $P(y_1,...,y_K \mid x)$ satisfies, for all k,

$$R_k \le \max_{P_X} I(X;Y_k) \tag{2.5}$$

where the maximization is over all input distributions with power constraint P_X . This result was extended for degraded broadcast channels by Wyner [16, Ack.] as

$$\sum_{i=k}^{K} R_i \le \max_{P_X} I(X;Y_k), \tag{2.6}$$

which is essentially an upper bound on the achievable sum rate. The upper bound in (2.6) was soon superseded by Gallager's capacity characterization in [17].

A capacity outer bound for a general broadcast channel, named Korner-Marton outer bound, was given in [25] as follows. The capacity region of two user broadcast channel $P(y_1, y_2 \mid x)$ is outer bounded by

$$R_1 \le I(X;Y_1) \tag{2.7a}$$

and

$$R_2 \le I(U;Y_2), \tag{2.7b}$$

also

$$R_1 + R_2 \le I(X;Y_1|U) + I(U;Y_2) \tag{2.7c}$$

for some $P(u,x)$ where $P(u,x,y_1,y_2) = P(u,x)P(y_1,y_2|x)$. The bounds in (2.7) can be supplemented with bounds obtained by swapping Y_1 and Y_2. The Korner-Marton outer bound has been the best known outer bound so far. However, due to its high complexity of evaluation, simpler bounds were derived based on it and other important features of the broadcast channel, among which are Sato bound [26] and DSM bound [23], [24].

The Sato bound provides an upper bound on the achievable sum rate by exploiting the fact that broadcast channels that have the same marginal distributions have the same capacity region, and by letting the receivers co-operate to obtain a point-to-point channel. Basically, it states that the capacity region $\mathcal{C}$ of a broadcast channel $P(y_1,...,y_K|x)$ is bounded by

$$\sum_{k=1}^{K} R_k \le \min_{P_{\tilde{Y}_1...\tilde{Y}_K|X} \in \mathcal{P}} \max_{P_X} I(X;\tilde{Y}_1\tilde{Y}_2..\tilde{Y}_K) \tag{2.8}$$

in terms of the achievable sum rate, where $\mathcal{P}$ is the set of channels $P(\tilde{y}_1...\tilde{y}_k|x)$ having the same marginals as $P(y_1,...,y_K|x)$, i.e., $P_{\tilde{Y}_k|X}(y|x) = P_{Y_k|X}(y|x)$ for all k, x and y. As noted above, the Sato bound is looser than the Korner-Marton bound. However, as shown by Sato [26] in a scalar Gaussian broadcast channel, this bound is tight for the sum rate. What may be more surprising is that recently, the Sato upper bound has been shown to be quite tight for some non-degraded Gaussian broadcast channel by Caire and Shamai [10], and the discovery that Sato upper bound is indeed tight in a general MIMO Gaussian broadcast channel [20]-[22]. The significance of this discovery leads to a renewed interest in a coding technique

termed dirty paper coding, which has been shown to achieve the sum capacity of a MIMO Gaussian broadcast channel and provides a theoretical foundation for this dissertation.

Another simplified general capacity bound obtained from the Korner-Marton bound is the so-called DSM outer bound, standing for degraded, same marginal outer bound. Unlike the Sato bound, which allows for the cooperation among all receivers, the DSM bound tightens the Sato bound by providing receiver $\pi(k)$ with outputs $Y_{\pi(k+1)},...,Y_{\pi(K)}$ of receivers $\pi(k+1),...,\pi(K)$, and bounds the capacity region by

$$\mathcal{C} \subseteq \bigcap_{\pi} \left\{ \bigcap_{P_{\tilde{Y}_1...\tilde{Y}_K|X} \in \mathcal{P}} \mathcal{R}_D(\pi, P_{\tilde{Y}_1...\tilde{Y}_K|X}) \right\} \tag{2.9}$$

where $\mathcal{R}_D(\pi, P_{\tilde{Y}_1...\tilde{Y}_K|X})$ is the capacity region of the degraded broadcast channel created by giving receiver $\pi(k)$ the outputs of receivers $\pi(k+1),...,\pi(K)$, and $\mathcal{P}$ is the set of channels $P(\tilde{y}_1...\tilde{y}_k \mid x)$ having the same marginals as $P(y_1,...,y_K \mid x)$, i.e., $P_{\tilde{Y}_k|X}(y \mid x) = P_{Y_k|X}(y \mid x)$ for all *k*, *x* and *y*. In [23] and [24], it was shown that the optimization of the DSM outer bound over all $P(\tilde{y}_1...\tilde{y}_k \mid x)$ with the same marginals leads to the achievable rate region of dirty paper coding, which we describe next for a MIMO Gaussian broadcast channel.

2.3 Achievable Rate Region of Dirty Paper Coding

Recent development on the achievable rate region of a MIMO broadcast channel has basically been based on dirty paper coding. Here we present some relevant results.

2.3.1 Dirty Paper Coding and Its Generalization

Dirty paper coding (DPC) is a coding method of Gel'fand and Pinsker [27] using

non-causally known interference cancellation. The name dirty paper coding is attributed to Costa [4], who showed that for a scalar discrete-time point-to-point memoryless channel,

$$y_i = x_i + s_i + n_i, \quad i = 1, ..., T, \tag{2.10}$$

where x_i and y_i are the transmitted and the received signals respectively, the interfering signal s_i is known to the transmitter but not to the receiver, and n_i is the unknown noise. If both s_i and n_i are i.i.d. Gaussian, and if the entire non-causal realization[1] of s_i is known to the transmitter prior to transmission, the channel capacity of (2.10) is the same as that of the AWGN channel $y_i=x_i+n_i$, i.e., as if the interference s_i were not present. In addition, the optimal transmit signal x_i is statistically independent of s_i. This tells us that knowing the interference non-causally at the transmitter is as powerful as knowing it at both the transmitter and the receiver. Costa's result was generalized to a vector point-to-point memoryless channel in [28].

The achievability of Costa's result relies on the fact that both the noise and the interference are i.i.d. Gaussian. This result was generalized in [29] and was shown that the same rate can be achieved for arbitrary noise distribution provided that the interference is i.i.d. Gaussian, or for arbitrary interference distribution provided that the noise is Gaussian (possibly colored). In [30], [31], it was further generalized to that the same result holds for arbitrary interference, provided that the transmitter and the receiver share a common random dither signal.

2.3.2 Dirty Paper Region of MIMO Broadcast Channel

This section shows the DPC achievable rate region for a MIMO broadcast channel. Consider a general memoryless MIMO broadcast channel of K users, with n_t transmit

[1] The entire non-causal realization here means the entire coded signal block to be transmitted.

antennas at the base station and m_k receive antennas at user k, k=1,…,K. Assume an i.i.d. block-wise flat fading model and over a coherence interval of T symbols, the received signals at K users are

$$\mathbf{Y}(t) \triangleq \begin{bmatrix} \mathbf{Y}_1(t) \\ \vdots \\ \mathbf{Y}_K(t) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1(t) \\ \vdots \\ \mathbf{H}_K(t) \end{bmatrix} \sum_{k=1}^{K} \mathbf{X}_k(t) + \begin{bmatrix} \mathbf{W}_1(t) \\ \vdots \\ \mathbf{W}_K(t) \end{bmatrix} \triangleq \mathbf{H}(t)\mathbf{X}(t) + \mathbf{W}(t), \qquad (2.11)$$

where $\mathbf{X}_k(t) \in \mathbb{C}^{n_t \times T}$, $\mathbf{Y}_k(t) \in \mathbb{C}^{m_k \times T}$ and $\mathbf{W}_k(t) \in \mathbb{C}^{m_k \times T}$ are transmitted signal, received signal, and noise matrix of user k at slot t, respectively, and the additive noise $\mathbf{W}_k(t)$ is assumed zero-mean, circular symmetric complex Gaussian with $\mathrm{vec}(\mathbf{W}_k(t)) \sim \mathcal{CN}(0, \mathbf{I}_T \otimes \mathbf{I}_{m_k})$, for $k = 1,...,K$, where $\mathrm{vec}(\mathbf{W}_k(t))$ is a column vector of size $m_k T \times 1$ by stacking the columns of $\mathbf{W}_k(t)$ under each other, and $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product (or direct product) of matrices $\mathbf{A}$ and $\mathbf{B}$. The broadcast channel capacity does not depend on the covariance (or correlation if zero mean) among the noise matrices $\mathbf{W}_k(t)$, and it depends only on the marginal covariance matrices of $\mathbf{W}_k(t)$, $k = 1,...,K$. We assume that the transmitted signal is subject to a total average transmit power P, i.e., $\mathrm{tr}\left[\sum_{k=1}^{K} \mathrm{E}(\mathbf{X}_k(t)\mathbf{X}_k^H(t))\right] = \mathrm{E}\left[\|\mathbf{X}(t)\|_F^2\right] \le PT$, with $\|\mathbf{X}(t)\|_F$ being the Frobenius norm of the matrix $\mathbf{X}(t)$ and the expectation $\mathrm{E}(\cdot)$ is over channel variations. With the unit noise sample variance, P is also the total average transmit SNR. For notational brevity, we omit the slot index t due to the i.i.d. block-fading assumption. Let $\mathbf{H}_k \in \mathbb{C}^{m_k \times n_t}$ denote the user-k channel matrix from n_t transmit to m_k receive antennas, which is assumed known to the transmitter and user-k receiver. Throughout this section, we assume this knowledge of CSI at the transmitter and that the receivers know only their own CSI, since this is a standard assumption of DPC.

Without loss of generality, we consider the natural ordering of users $k = 1,...,K$, and the *k*-th user's received signal can be written as

$$\mathbf{Y}_k = \mathbf{H}_k \mathbf{X}_k + \mathbf{H}_k \sum_{j<k} \mathbf{X}_j + \mathbf{H}_k \sum_{j>k} \mathbf{X}_j + \mathbf{W}_k, \quad k = 1,...,K. \tag{2.12}$$

If we apply the dirty-paper encoding at the transmitter by generating the auxiliary input matrix $\mathbf{U} = [\mathbf{U}_1^H,...,\mathbf{U}_K^H]^H$ with $\mathbf{U}_k \in \mathbb{C}^{m_k \times T}$, and the precoding matrix $\mathbf{B} = [\mathbf{B}_1,...,\mathbf{B}_K]$ with $\mathbf{B}_k \in \mathbb{C}^{n_t \times m_k}$, and encode each data stream as $\mathbf{X}_k = \mathbf{B}_k \mathbf{U}_k$ so that (2.12) can be rewritten as

$$\mathbf{Y}_k = \mathbf{G}_{k,k} \mathbf{U}_k + \sum_{j<k} \mathbf{G}_{k,j} \mathbf{U}_j + \sum_{j>k} \mathbf{G}_{k,j} \mathbf{U}_j + \mathbf{W}_k, \quad k = 1,...,K, \tag{2.13}$$

where $\mathbf{G} = \mathbf{HB}$ has block entries $\mathbf{G}_{i,j} = \mathbf{H}_i \mathbf{B}_j \in \mathbb{C}^{m_i \times m_j}$, $i, j = 1,...,K$. The components of $\mathbf{U}$ are generated by successive DPC with Gaussian codebooks and satisfy $\mathrm{E}[\mathbf{U}\mathbf{U}^H] = \mathbf{I}_m$, where $m \triangleq \sum_{k=1}^{K} m_k$, and the precoding matrix $\mathbf{B}$ is constrained by $\mathrm{tr}[\mathrm{E}(\mathbf{B}\mathbf{B}^H)] \le \mathrm{P}$, so that the average input power is constrained to P. Now the encoder considers the inference in (2.13) caused by users $j < k$ as known non-causally and the *k*-th decoder only sees the interference from users $j > k$, and treats them as additional noise. The SINR seen by the user *k* is then

$$SINR_k = \frac{\frac{1}{m_k}\mathrm{tr}(\mathbf{G}_{k,k}\mathbf{G}_{k,k}^H)}{1 + \sum_{j>k} \frac{1}{m_j}\mathrm{tr}(\mathbf{G}_{k,j}\mathbf{G}_{k,j}^H)}, \quad k = 1,...,K. \tag{2.14}$$

Therefore, through DPC and minimum Euclidean distance decoding, the joint reliable communication of data streams is possible at rates

$$R_k^{\mathrm{dpc}} = \log_2(1 + SINR_k), \quad k = 1,...,K. \tag{2.15}$$

The achievable rate region through DPC is then the convex closure as [10]

$$\text{co} \bigcup_{\pi} \bigcup_{\mathbf{B}:\, \text{tr}[\text{E}(\mathbf{BB}^H)] \le P} \left\{ 0 \le R_{\pi(k)} \le \log_2(1 + SINR_{\pi(k)}),\ \ k = 1, ..., K \right\}, \qquad (2.16)$$

where "co" denotes convex closure and π runs over all permutations of K users. This is the dirty-paper region characterized in [20]-[22]. Further, it was pointed out in [28] that time sharing may be necessary to achieve the convex hull in (2.16).

2.4 Summary

This chapter provides a review on the capacity of broadcast channel. The goal is to shed light on the optimal transmission schemes in a MIMO broadcast channel from a capacity point of view. Starting with a degraded broadcast channel, we introduced the capacity region of a scalar Gaussian broadcast channel and some known capacity bounds for broadcast channels. For the MIMO Gaussian broadcast channel, we gave the achievable rate region of dirty paper coding. We emphasize that the dirty paper region is not a new achievability result for the broadcast channel. Rather, it is a subset of Marton's region. Most recently, it has been shown that the dirty paper region is the actual capacity region of the broadcast channel [94], which has finally solved a long-standing open problem in multiuser information theory. To our interest, the fact that the dirty paper region of a MIMO broadcast channel is equal to the capacity region of a dual MIMO multiaccess channel[2] with an equal average sum-power constraint, as well as the earlier discovery that DPC achieves the sum capacity of a general MIMO Gaussian broadcast channel, make it readily applicable to multiple antenna broadcast channels.

---

[2] Here a dual MIMO multiaccess channel is defined as that in [20], with the average sum-power constraint equal to that of the broadcast channel, and the channel matrix the Hermitian of the broadcast channel matrix.

Chapter 3
Interaction of Antenna and Multiuser Diversity

This chapter provides an interaction study of the antenna diversity and multiuser diversity. We consider the use of multiple antennas at the transmitter and/or receivers to provide open-loop diversity, rather than spatial multiplexing. When user channel quality information is available to the transmitter, our results reveal a negative interaction between these two forms of spatial diversity in terms of total throughput.

3.1 Asymptotic Average Sum Rate

We first derive the asymptotically achievable sum rate in the limit of a large number of users, with multiple-antenna diversity in i.i.d. flat Rayleigh and Rician fading.

3.1.1 Multiuser Channel Model

We use a simple downlink model of a cellular packet data system, in which the base station transmitter sends packets to K mobile terminals. There are n_T transmit antennas at the base station and n_R receive antennas at each mobile user for antenna diversity. We assume that the transmission time is divided into consecutive and equal time slots, with the duration of each slot being less than the fading coherence time and much less than the possible delay constraint of data services, but sufficiently long so that the information-theoretic assumption of infinitely long code block length is meaningful. At each slot t, the packet scheduler at the base station decides to send a packet to user k^* with the largest effective SNR $\gamma_k(t)$ at the

receive antenna combiner output

$$k^* = \arg \max_{k=1,\cdots,K} \gamma_k(t) \tag{3.1}$$

with

$$\gamma_k(t) = \frac{\gamma_s}{n_T} \sum_{j=1}^{n_R} \sum_{i=1}^{n_T} \left| \alpha_{i,j}^{(k)}(t) \right|^2 \tag{3.2}$$

where γ_s is the expected SNR at each receive antenna branch for each user, and is assumed a constant independent of the number of transmit antennas. In (3.2), $\alpha_{i,j}^{(k)}(t)$ represent the ergodic channel fading processes from transmit antenna *i* to receive antenna *j* of user *k* over slot *t*. Unless specified otherwise, we assume a block-fading channel model with the fading coefficients $\alpha_{i,j}^{(k)}(t)$ being fixed over slot *t*, and varying independently from slot to slot over time. All $\alpha_{i,j}^{(k)}(t)$ terms are i.i.d. between each pair of transmit and receive antennas, which assumes that antenna elements are spaced sufficiently with respect to the angle spread. This model presupposes that the the channel fading process can be assumed flat in frequency. It also assumes that the average total transit power is fixed and equally split over n_T transmit antennas. For simplification, we assume that perfect and instantaneous user channel SNR feedback exists between each transmitter-receiver pair. Also for antenna diversity, we assume that the effective user SNRs, after coherent combing, conform to (3.2). However, from the orthogonal design for the space-time codes [8], no "full-rate" complex orthogonal design exists for n_T greater than two, for which only fractional rates are achievable (see [8]).

At slot *t*, we define the maximum SNR among the *K* users as

$$\gamma(t) \triangleq \max_{k=1,\cdots,K} \gamma_k(t) . \tag{3.3}$$

Under the assumption of symmetric user channels with i.i.d. SNRs over each slot, and according to the order statistics for i.i.d. continuous random variables [32], the cumulative

distribution function (cdf) of γ at any slot is

$$Y(\gamma) = \left(F(\gamma)\right)^K \tag{3.4}$$

and the probability density function (pdf) of γ is

$$y(\gamma) = \begin{cases} K \cdot f(\gamma) \cdot F(\gamma)^{K-1}, & \gamma > 0 \\ 0, \quad \text{otherwise} \end{cases} \tag{3.5}$$

where $f(\bullet)$ and $F(\bullet)$ are the pdf and cdf of the i.i.d. user SNRs γ_k, respectively. In (3.4) and (3.5), the slot index is dropped due to the time independence of the statistics. We define the average sum rate of the channel with the $n_T n_R$-fold antenna diversity and K-fold multiuser diversity as

$$C_{n_T n_R}^K \triangleq \int_0^\infty \log_2(1+\gamma) \cdot y(\gamma) d\gamma \tag{3.6}$$

in bits/sec/Hz. This is the expected achievable sum rate, given the effective user SNRs in (3.2). Under the assumptions of the perfect and instantaneous transmitter knowledge of user channel quality information (CQI) in (3.2), so that greedy scheduling and rate adaptation can be performed at the transmitter, the average sum rate in (3.6) is bounded away from zero. Throughput this paper, we use the average sum rate metric to evaluate the spatial diversity of the system. We first derive the asymptotic multiuser diversity in the limit of a large number of users, i.e., as $K \to \infty$, in the remainder of this section.

To calculate the limiting distribution of the γ in (3.3) as $K \to \infty$, we use [33, Lemma 2], which we restate below as *Lemma 3.1*.

Lemma 3.1: Let $z_1, \cdots, z_K$ be i.i.d. random variables with a common cdf $F(\bullet)$ and pdf $f(\bullet)$ satisfying $F(\bullet)$ is less than 1 for all finite z and is twice differentiable for all z, and is such that

$$\lim_{z\to\infty}\left[\frac{1-F(z)}{f(z)}\right]=c>0 \tag{3.7}$$

for some constant c. Then

$$\left(\max_{1\le k\le K} z_k\right)-l_K$$

converges in distribution to a limiting random variable with cdf

$$\exp(-e^{-x/c}).$$

As $K\to\infty$, where l_K is given by $F(l_K)=1-1/K$.

Lemma 3.1 states that the maximum of K such i.i.d. random variables grows like l_K as $K\to\infty$.

3.1.2 Symmetric Rayleigh Fading

With the i.i.d. Rayleigh fading channels, γ_k in (3.2) is a chi-square random variable with $2n_Tn_R$ degrees of freedom. Define $n\triangleq n_Tn_R-1$, the pdf and cdf of γ_k are [34], respectively

$$f(\gamma_k)=\frac{1}{n!\left(\gamma_s/n_t\right)^{n_Tn_R}}\gamma_k^n e^{-\frac{\gamma_k}{\gamma_s/n_T}},\quad \gamma_k>0 \tag{3.8}$$

and

$$F(\gamma_k)=1-e^{-n_T\gamma_k/\gamma_s}\sum_{m=0}^{n}\frac{\left(n_T\gamma_k/\gamma_s\right)^m}{m!},\quad \gamma_k>0. \tag{3.9}$$

We can easily see that

$$\lim_{\gamma_k\to\infty}\frac{1-F(\gamma_k)}{f(\gamma_k)}=\frac{\gamma_s}{n_T}=c_1>0. \tag{3.10}$$

Solving for l_K from *Lemma 3.1*, we have

$$l_K = \begin{cases} \dfrac{\gamma_s}{n_T}\ln K = \gamma_s \ln K, & n_T = n_R = 1 \\ \dfrac{\gamma_s}{n_T}\ln K + O\left(\ln\ln K\right), & \text{otherwise} \end{cases}. \qquad (3.11)$$

Therefore, with i.i.d. Rayleigh fading and for a large number of users, the maximum equivalent SNR γ grows like l_K in (3.11), which is a function of the number of transmit antennas n_T and the number of users K for the fixed average total transmit power, and is independent of the number of receive antennas n_R. The asymptotic sum rate as $K\to\infty$ can therefore be approximated by

$$\lim_{K\to\infty} C^K_{n_T n_R} = \log_2(1+l_K) \approx \log_2(1+\frac{\gamma_s}{n_T}\ln K). \qquad (3.12)$$

Figure 3.1 shows the numerical evaluation of the average sum rate $C^K_{n_T n_R}$ for the cases of n_T, n_R=1, 2 and 4. We see the surprising result that with multiple users (K>1), the average sum rate with transmit diversity is lower than that without transmit diversity for all cases of n_R=1, 2 and 4, although the opposite is true for the single user cases. This result is contrary to what is commonly observed over a single wireless link, where transmit diversity always improves link performance in fading. An intuitive explanation for the difference is that for fixed average transmit power, transmit diversity reduces the variation in the received signal power which is exploited by the greedy scheduler for multiuser diversity. The same is true for receive diversity, but the array gain from coherent receive combining more than compensates for the loss of multiuser diversity. Therefore, care should be exercised in employing open-loop antenna diversity, which may improve single-link performance, but can also degrade system-level performance in the presence of greedy scheduling. We note from Figure 3.1 that in order for the n_T=n_R=1 case to outperform the cases of n_T=n_R=2 and n_T=n_R=4, the minimum numbers of users required are about 2^8=256 and 2^{12}=4096, respectively. Consistent with (3.12), for large K, there is an approximate 1 bits/sec/Hz capacity increase

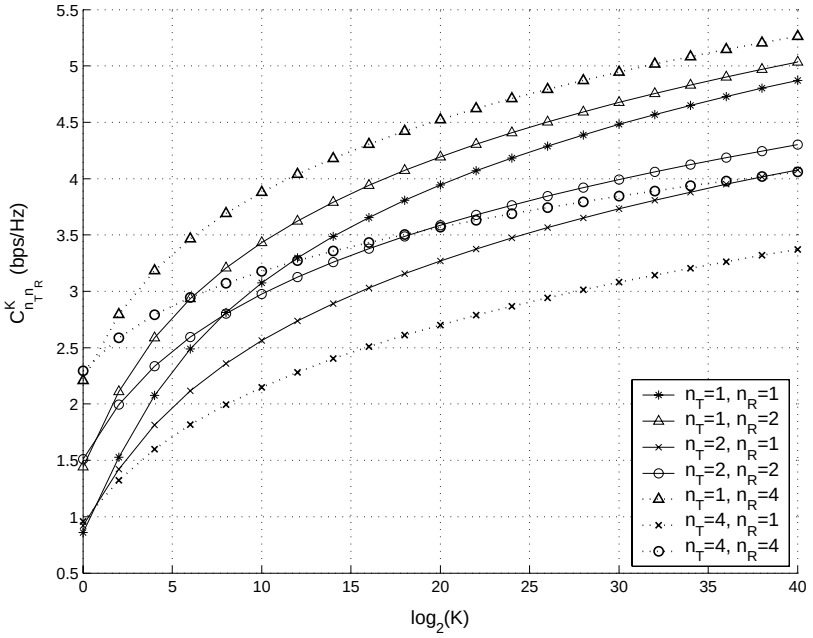


Figure 3.1. Average sum rate under i.i.d. flat Rayleigh fading with $\gamma_s = 0$ dB.

for every 3-dB increase in γ_s or a half of n_T. Finally, it is worth noting that although the asymptotic sum rate in (3.12) is independent of the number of receive antennas, this asymptote is approached at a very slow rate in the number of users, even more so for large n_T.

3.1.3 Symmetric Rician Fading

For the i.i.d. Rician fading channels, γ_k in (3.2) is a noncentral chi-square random variable with $2n_Tn_R$ degrees of freedom. Defining $n \triangleq n_T n_R - 1$, $\sigma^2 \triangleq 1/\left(2(1+K_r)\right)$, and $s^2 \triangleq K_r n_T n_R /(1+K_r)$, where K_r is the Rice factor for all users, the pdf and cdf of γ_k are [34], respectively

$$f(\gamma_k) = \frac{n_T}{2\sigma^2\gamma_s}\left(\frac{n_T\gamma_k/\gamma_s}{s^2}\right)^{\frac{n}{2}} e^{-\frac{s^2+n_T\gamma_k/\gamma_s}{2\sigma^2}} I_n\left(\frac{s\sqrt{n_T\gamma_k/\gamma_s}}{\sigma^2}\right), \quad \gamma_k > 0 \qquad (3.13)$$

and

$$F(\gamma_k) = 1 - e^{-\frac{s^2 + n_T\gamma_k/\gamma_s}{2\sigma^2}} \sum_{m=0}^{\infty} \left(\frac{s}{\sqrt{n_T\gamma_k/\gamma_s}} \right)^m I_m\left(\frac{s\sqrt{n_T\gamma_k/\gamma_s}}{\sigma^2} \right) -$$

$$e^{-\frac{s^2 + n_T\gamma_k/\gamma_s}{2\sigma^2}} \sum_{m=1}^{n} \left(\frac{\sqrt{n_T\gamma_k/\gamma_s}}{s} \right)^m I_m\left(\frac{s\sqrt{n_T\gamma_k/\gamma_s}}{\sigma^2} \right), \quad \gamma_k > 0 \qquad (3.14)$$

where $I_n(\bullet)$ is the nth-order modified Bessel function of the first kind. The tails of the cdf and pdf of γ_k can be approximated respectively by

$$f(\gamma_k) \sim \frac{n_T}{\gamma_s} \frac{s^{-n-\frac{1}{2}}}{2\sigma\sqrt{2\pi}} \left(\frac{n_T\gamma_k}{\gamma_s} \right)^{\frac{n}{2}-\frac{1}{4}} e^{-\frac{\left(\sqrt{n_T\gamma_k/\gamma_s} - s\right)^2}{2\sigma^2}}$$

and

$$1 - F(\gamma_k) \sim \frac{\sigma}{\sqrt{2\pi}} s^{-n-\frac{1}{2}} \left(\frac{n_T\gamma_k}{\gamma_s} \right)^{\frac{n}{2}-\frac{1}{4}} e^{-\frac{\left(\sqrt{n_T\gamma_k/\gamma_s} - s\right)^2}{2\sigma^2}}.$$

These approximations are in the sense that the ratio of the left- and right-hand sides approaches 1 as $\gamma_k \to \infty$. Hence

$$\lim_{\gamma_k \to \infty} \left[\frac{1 - F(\gamma_k)}{f(\gamma_k)} \right] = \frac{2\sigma^2\gamma_s}{n_T} = \frac{\gamma_s}{n_T(1+K_r)} = c_2 > 0. \qquad (3.15)$$

By solving $F(l_K)$=1-1/K, we have

$$l_K = \frac{\gamma_s}{n_T}\left(s + \sqrt{2\sigma^2 \ln K} \right)^2 + O(\ln\ln K)$$

$$= \frac{\gamma_s}{n_T}\left(\sqrt{\frac{1}{1+K_r}\ln K} + \sqrt{\frac{K_r}{1+K_r} n_T n_R} \right)^2 + O(\ln\ln K). \qquad (3.16)$$

According to *Lemma 3.1*, for i.i.d. Rician fading, the maximum equivalent SNR γ grows like l_K in (6.19) as $K \to \infty$. For the fixed average total transmit power, l_K depends on the number of users, the Rice factor, and both the number of transmit and the number of receive antennas. This differs from that of i.i.d. Rayleigh fading in (3.11), where l_K is independent of the number of receive antennas. In fact, (3.16) can be viewed as a generalization of (3.11) for

K_r>0 and n>1. Equations (3.11) and (3.16) reduce to those in [33] for n_T=n_R=1 and γ_s=0 dB. Similarly, we can have the asymptotic sum rate approximated by

$$\lim_{K\to\infty} C^K_{n_T n_R} = \log_2(1+l_K) \approx \log_2\left(1+\frac{\gamma_s}{n_T}\left(\sqrt{\frac{1}{1+K_r}\ln K}+\sqrt{\frac{K_r}{1+K_r}n_T n_R}\right)^2\right). \quad (3.17)$$

This is plotted in Figure 3.2 for K=2^{40} and n_T, n_R=1, 2 and 4. We can see that the approximate asymptotic sum rate goes up as we increase n_R or decrease n_T with the other term remaining fixed. It also decreases as the Rice factor K_r increases at sufficiently large K_r. This is due to the reduced potential channel diversity gain with increased K_r. As the channel approaches an additive white Gaussian noise (AWGN) channel about the mean SNR $\gamma_s n_R$ with $K_r\to\infty$, (3.17) can be further approximated by

$$\lim_{K\to\infty, K_r\to\infty} C^K_{n_T n_R} \approx \log_2(1+\gamma_s n_R) \quad (3.18)$$

which is independent of n_T. This trend is shown in Figure 3.2 with K_r extended to 30 dB. For the intermediate values of K_r, both the values of n_T and n_R affect the average sum rate as shown in the figure.

We conclude this section by pointing out that in an i.i.d. flat Rayleigh fading environment with more than one user, the orthogonal transmit antenna diversity gain is negative, in that the average sum rate is less than that without transmit diversity in the presence of greedy scheduling. In the limit of $K\to\infty$, the maximum effective SNR γ is inversely proportional to the number of transmit antennas and independent of the number of receive antennas. In the corresponding Rician fading environment, both the number of transmit and the number of receive antennas impact the asymptotic sum rate as $K\to\infty$, with negative gain from transmit diversity and positive gain from receive antennas. However, as the Rice factor becomes sufficiently large and the channel approaches an AWGN channel, only the positive power gain from the coherent receive antenna combining remains.

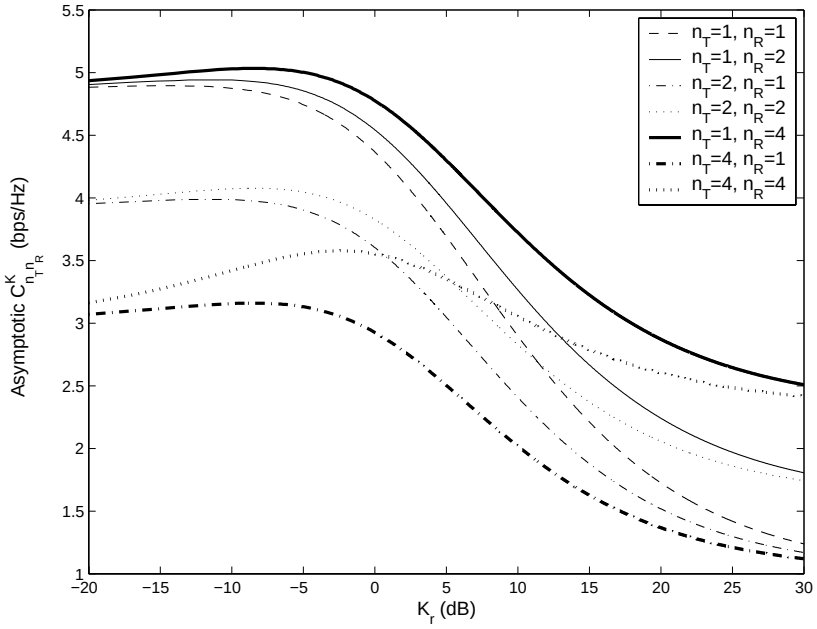


Figure 3.2. Approximate asymptotic sum rate for K=2^{40} users under i.i.d. flat Rician fading with γ_s=0 dB.

3.2 Antenna Diversity and Multiuser Diversity

We numerically evaluate the average sum-rate improvement from the antenna and the multiuser diversity separately for a finite number of users. For simplicity, we restrict ourselves to symmetric flat Rayleigh fading channels. We define the antenna improvement as the percentage increase of the average sum rate relative to the single antenna case of n_T=n_R=1 in a system of K users

$$I_a \triangleq \frac{C^K_{n_T n_R} - C^K_{11}}{C^K_{11}} \tag{3.19}$$

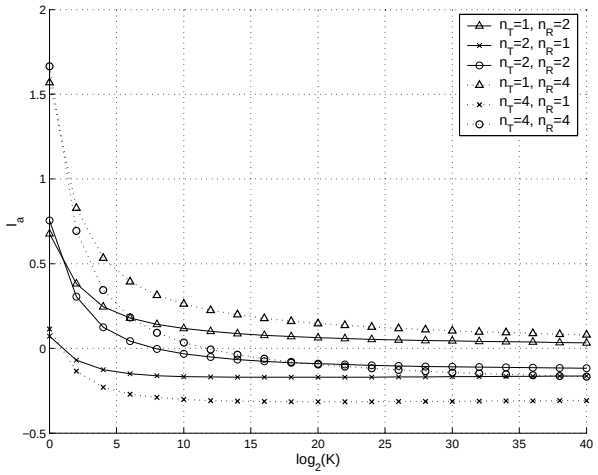
and the multiuser diversity improvement as the percentage increase of the average sum rate relative to the single user case

$$I_u \triangleq \frac{C^K_{n_T n_R} - C^1_{n_T n_R}}{C^1_{n_T n_R}}. \tag{3.20}$$

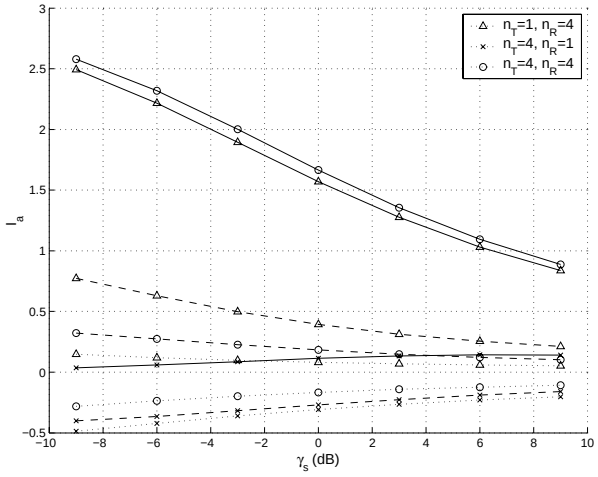
Figures 3.3(a) and (b) illustrate the antenna improvement I_a versus the number of users K at user average SNR of γ_s=0 dB, and I_a versus γ_s with K=1, 64 and 2^{40} users, respectively, both for the cases of n_T, n_R=1, 2 and/or 4. The zero level reference is for n_T=n_R=1. In Figure 3.3(a), the n_T=2 and 4 cases have less antenna improvement than the corresponding n_T=1 case when there are more than one user, and the antenna improvement decreases as K increases, and becomes saturated at sufficiently large K. This trend is more pronounced for the cases with transmit diversity only. In Figure 3.3(b), we note that in all cases, the improvement of the receive antennas of n_T=1 and n_R=4 decreases with γ_s and remains positive, and the improvement of the transmit antennas of n_T=4 and n_R=1 increases with γ_s consistently at a lower percentage level.

Figures 3.4(a) and (b) depict the multiuser diversity improvement I_u versus K at γ_s=0 dB, and I_u versus γ_s with K=64 and 2^{40} users, respectively, also both for the cases of n_T, n_R=1 or 2 and/or 4. The plots show that I_u monotonically increases with K, but monotonically decreases with γ_s, both at percentage levels much higher than those in Figure 3 for a sufficient number of users. In all cases, the improvement of multiuser diversity is positive. Comparing Figure 3.3 with Figure 3.4, we see that in most cases the average sum-rate improvement is dominated by the multiuser diversity in a network for a finite number of users, e.g., K>64 for n_T, n_R=1 or 4.

In evaluating the improvement of antenna and multiuser diversity in terms of the relative increase of the average sum rate, we have assumed the instantaneous and perfect feedback of a continuum of SNR values to the base station. We have also assumed that over each slot, the sum rate can be achieved physically at those continuous SNR values by some powerful coding and modulation techniques. We now evaluate the effect of finite channel



(a) $\gamma_s = 0$ dB .

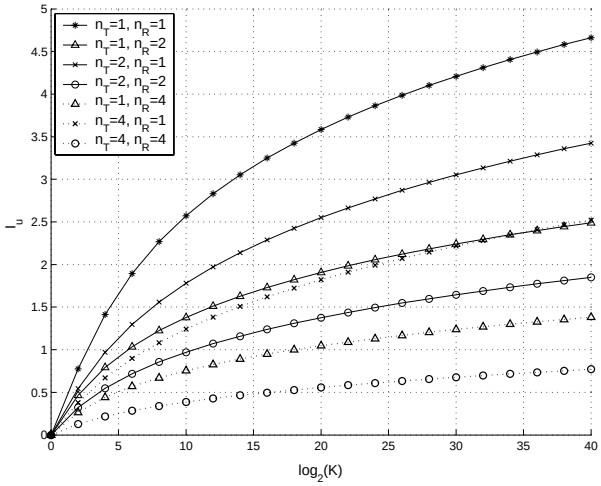


(b) Solid curves: K=1; dashed curves: K=64; dotted curves: K=2^{40}.

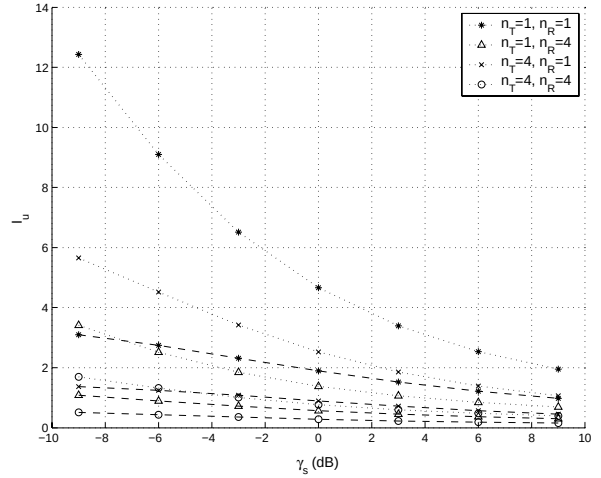
Figure 3.3. Antenna improvement in i.i.d. flat Rayleigh fading.

SNR quantization on the antenna and multiuser diversity improvement. We still retain the other assumptions given above, except that the equivalent channel SNRs γ_k are now quantized into one of Q+1 values, ranked in an ascending order from zero as follows,

$$\gamma_{Q,k} \in \left\{0, \hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_{Q-1}\right\}, \quad k = 1, \ldots, K . \tag{3.21}$$



(a) $\gamma_s = 0$ dB.



(b) Dashed curves: K=64; dotted curves: K=2^{40}.

Figure 3.4. Multiuser diversity improvement in i.i.d. flat Rayleigh fading.

The quantization is to the lower value over each successive interval of SNR in (3.21), with 0 and $\hat{\gamma}_{Q-1}$ being the lower and upper limit, respectively. The probability mass function (pmf) of the discrete random variable $\gamma_{Q,k}$ can be calculated as

$$p(\gamma_{Q,k}) = \left(\int_0^{\hat{\gamma}_0} f(\gamma_k) d\gamma_k \right) \cdot \delta[\gamma_{Q,k}] + \sum_{i=0}^{Q-2} \left(\int_{\hat{\gamma}_i}^{\hat{\gamma}_{i+1}} f(\gamma_k) d\gamma_k \right) \cdot \delta[\gamma_{Q,k} - \hat{\gamma}_i]$$
$$+ \left(\int_{\hat{\gamma}_{Q-1}}^{\infty} f(\gamma_k) d\gamma_k \right) \cdot \delta[\gamma_{Q,k} - \hat{\gamma}_{Q-1}]$$
$$= F(\hat{\gamma}_0) \cdot \delta[\gamma_{Q,k}] + \sum_{i=0}^{Q-2} \left(F(\hat{\gamma}_{i+1}) - F(\hat{\gamma}_i) \right) \cdot \delta[\gamma_{Q,k} - \hat{\gamma}_i]$$
$$+ \left(1 - F(\hat{\gamma}_{Q-1}) \right) \cdot \delta[\hat{\gamma}_{Q,k} - \hat{\gamma}_{Q-1}] \tag{3.22}$$

where $\delta[n-m]$ is the Kronecker delta function defined by

$$\delta[n-m] \triangleq \begin{cases} 1, & n = m \\ 0, & n \neq m \end{cases}.$$

The cdf of $\gamma_{Q,k}$ is, therefore, the stair function

$$P(\gamma_{Q,k}) = F(\hat{\gamma}_0) \cdot u(\gamma_{Q,k}) + \sum_{i=0}^{Q-2} \left(F(\hat{\gamma}_{i+1}) - F(\hat{\gamma}_i) \right) \cdot u(\gamma_{Q,k} - \hat{\gamma}_i) + \left(1 - F(\hat{\gamma}_{Q-1}) \right) \cdot u(\gamma_{Q,k} - \hat{\gamma}_{Q-1})$$
$$= \begin{cases} F(\hat{\gamma}_0), & 0 \le \gamma_{Q,k} < \hat{\gamma}_0 \\ F(\hat{\gamma}_{i+1}), & \hat{\gamma}_i \le \gamma_{Q,k} < \hat{\gamma}_{i+1},\ i = 0, \ldots, Q-2 \\ 1, & \gamma_{Q,k} \ge \hat{\gamma}_{Q-1} \end{cases} \tag{3.23}$$

where $u(x - x_0)$ is the unit step function defined by

$$u(x - x_0) \triangleq \begin{cases} 0, & x < x_0 \\ 1, & x \ge x_0 \end{cases}.$$

According to the order statistics for i.i.d. discrete random variables [32], the cdf and pmf of the maximum quantized SNR γ_Q,

$$\gamma_Q \triangleq \max_{k=1,\cdots,K} \gamma_{Q,k}$$

are, respectively,

$$Y_Q(\gamma_Q) = \left(P(\gamma_Q)\right)^K \tag{3.24}$$

and

$$\begin{aligned} y_Q(\gamma_Q) &= Y_Q(0)\cdot\delta[\gamma_Q] + \left(Y_Q(\hat{\gamma}_0) - Y_Q(0)\right)\cdot\delta[\gamma_Q - \hat{\gamma}_0] \\ &\quad + \sum_{i=1}^{Q-1}\left(Y_Q(\hat{\gamma}_i) - Y_Q(\hat{\gamma}_{i-1})\right)\cdot\delta[\gamma_Q - \hat{\gamma}_i] \\ &= F(\hat{\gamma}_0)^K\cdot\delta[\gamma_Q] + \sum_{i=0}^{Q-2}\left(F(\hat{\gamma}_{i+1})^K - F(\hat{\gamma}_i)^K\right)\cdot\delta[\gamma_Q - \hat{\gamma}_i] \\ &\quad + \left(1 - F(\hat{\gamma}_{Q-1})^K\right)\cdot\delta[\gamma_Q - \hat{\gamma}_{Q-1}]. \end{aligned} \tag{3.25}$$

The average sum rate with Q nonzero-SNR quantization levels can be calculated from

$$\begin{aligned} D_{n_T n_R}^K &\triangleq \sum_{i=0}^{Q-1}\log_2(1+\hat{\gamma}_i)\cdot y_Q(\hat{\gamma}_i) \\ &= \sum_{i=0}^{Q-2}\log_2(1+\hat{\gamma}_i)\cdot\left(F(\hat{\gamma}_{i+1})^K - F(\hat{\gamma}_i)^K\right) + \log_2(1+\hat{\gamma}_{Q-1})\cdot\left(1 - F(\hat{\gamma}_{Q-1})^K\right). \end{aligned} \tag{3.26}$$

Figure 3.5 illustrates $D_{n_T n_R}^K$ with $Q = 8$ uniform SNR quantization levels (in decibels), using $\hat{\gamma}_0 = -11.5$ dB and $\hat{\gamma}_{Q-1} = 9.5$ dB for the i.i.d. flat Rayleigh fading channels with $\gamma_s = 0$ dB for the cases of $n_T, n_R = 1,\ 2$ and 4. For $Q = 8$, $D_{n_T n_R}^K$ in (3.26) is above 78% of the value of $C_{n_T n_R}^K$ in (6.6). For the cases of $Q = 32$ and 16 (not shown in Figure 3.5), $D_{n_T n_R}^K$ is above 94% and 89%, respectively, of $C_{n_T n_R}^K$ for $n_T, n_R = 1$ or 2, and above 85% and 85% respectively of $C_{n_T n_R}^K$ for $n_T, n_R = 1$ or 4. Comparing Figure 3.5 with Figure 3.1, we note that at fixed quantization limits, the average sum rate with SNR quantization is reduced more by receive diversity. The reason is given below.

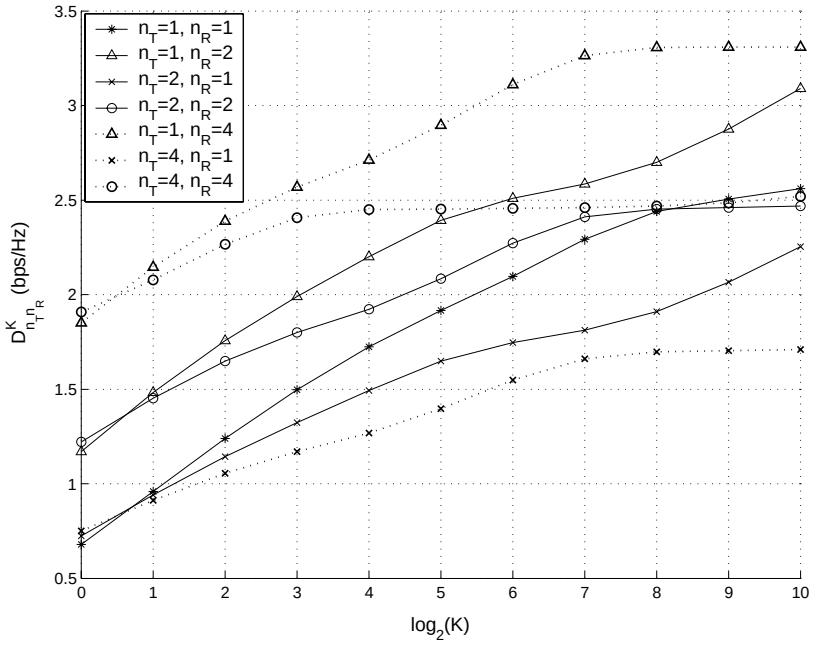


Figure 3.5. Average sum rate in i.i.d. flat Rayleigh fading at $\gamma_s = 0$ dB. $\hat{\gamma}_i$ is uniformly spaced over $[-11.5\ 9.5]$ dB with $Q = 8$.

In the process of user SNR quantization, γ_k greater than $\hat{\gamma}_{Q-1}$ will be quantized to $\hat{\gamma}_{Q-1}$ according to (3.21), so the maximum SNR increase above γ_{Q-1} contributes no sum-rate increase. We define the relative probability of γ greater than $\hat{\gamma}_{Q-1}$ as the SNR clipping rate

$$R_{clip} \triangleq \frac{\Pr\{\gamma > \hat{\gamma}_{Q-1}\}}{\Pr\{\gamma > \hat{\gamma}_0\}} = \frac{1 - Y(\hat{\gamma}_{Q-1})}{1 - Y(\hat{\gamma}_0)} = \frac{1 - F(\hat{\gamma}_{Q-1})^K}{1 - F(\hat{\gamma}_0)^K} \tag{3.27}$$

with $0 < R_{clip} < 1$, and $\lim_{K \to \infty} R_{clip} = 1$. In a large network employing the diversity techniques, for power-efficiency purposes, we want the SNR clipping rate to be as small as possible, given the SNR quantization. The clipping rate R_{clip} is plotted in the Figure 3.6 with the same limiting SNRs above, at $\gamma_s = 0$ dB for the cases of $n_T, n_R = 1$, 2 and 4.

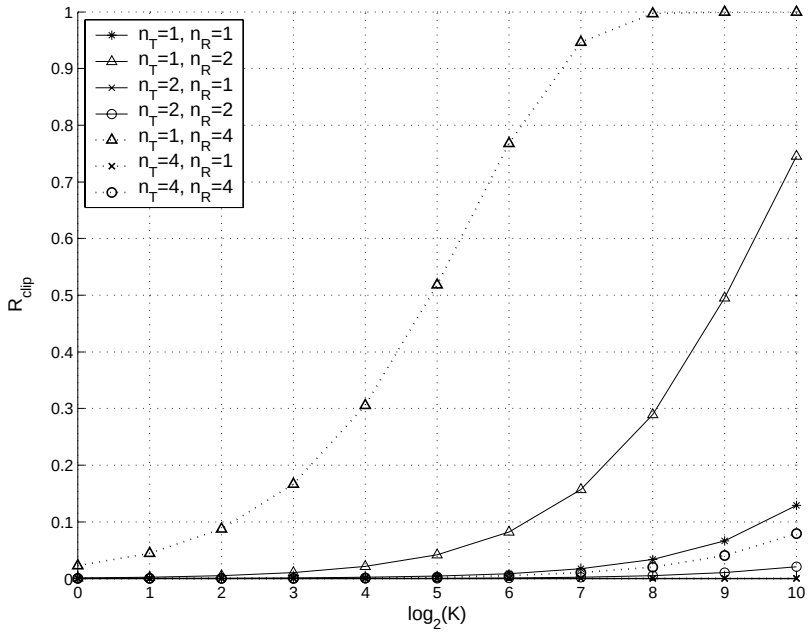


Figure 3.6. Clipping rate in i.i.d. flat Rayleigh fading for $\hat{\gamma}_0 = -11.5$ dB and $\hat{\gamma}_{Q-1} = 9.5$ dB at $\gamma_s = 0$ dB.

As shown in Figure 3.6, the SNR clipping rate can be nontrivial, even at $\gamma_s = 0$ dB, for a modest number of users. For the case of $n_T = 1$ and $n_R = 2$, and the number of users $K = 512$, there is about a 50% probability that the maximum received SNR will exceed the 9.5-dB upper limit at $\gamma_s = 0$ dB. For $n_T = 1$ and $n_R = 4$, the same clipping rate occurs at K less than 32. As the maximum user effective SNR increases, the SNR clipping rate increases. This suggests an adaptive SNR quantization scheme which accounts for the expectation of the spatial diversity improvement. It should be noted that the clipping rate of $n_T = 1$ and $n_R = 4$ at $\gamma_s = 0$ dB is the same as that of $n_T = 2$ and $n_R = 2$ at $\gamma_s = 3$ dB. This is due to the same cdf $F(\gamma_k)$ in (3.9) for two cases, which leads to the same clipping rate R_{clip} in (3.27).

3.3 Receive Antenna Diversity under Proportional Fair Scheduling

We now quantitatively evaluate the improvement from receive antenna diversity under the more practical proportional fair scheduling. Our focus is on the throughput improvement of two-fold receive diversity over that without antenna diversity in a multiuser scenario.

3.3.1 Proportional Fair Scheduling

The concept of proportional fairness was introduced by Kelly for the issues of charging, rate control and routing in a communication network carrying elastic traffic [9]. It is based on a tractable mathematical model from which the max-min fairness [35] emerges as a limiting special case, and the proportional fairness characterizes a system optimum in the user rate allocations. According to [9], a vector of rates $\mathbf{x}$ is *proportionally fair* if it is feasible, and if for any other feasible vector $\mathbf{x}^*$, the aggregate of proportional changes is zero or negative:

$$\sum_{r\in\mathcal{R}} \frac{x_r^* - x_r}{x_r} \le 0 . \tag{3.28}$$

Kelly has proved that a system optimum is achieved when users' choices of charges and network choice of allocated rates are in equilibrium [9]. Based on this, a proportional fair (PF) scheduling algorithm was proposed in [36] for the downlink scheduling in IS-856 (also known as 1xEV-DO or HDR), which is a packet-switched cellular data system. The performance limitations and improvement of PF scheduling algorithm were shown in [33].

The PF scheduler maintains resource fairness by providing a fair sharing of transmission time proportional to the past throughputs of users over a fixed window length. On the time-slotted forward link transmission, at each time slot *t*, user *k* feeds back to the base station a requested data rate $R_k(t)$, which is supportable by its current channel quality.

At the assumption of an instantaneous and error-free rate feedback, the scheduler assigns the time slot *t* to the user k^*, which has the largest ratio

$$\frac{R_k(t)}{T_k(t)}$$

among all active users for which the base station has data to send. The rate $T_k(t)$ is the *k*th user's average throughput in a past window of length W_c, and is updated slot-wise according to

$$T_k(t+1) = \begin{cases} (1-\frac{1}{W_c})T_k(t) + \frac{1}{W_c}R_k(t), & k = k^* \\ (1-\frac{1}{W_c})T_k(t), & k \neq k^* \end{cases}, \tag{3.29}$$

where W_c is the time scale of interest and is constrained by the maximum delay tolerance. In IS-856, W_c is about 1.67 seconds, which is usually much greater than small-scale fading coherence time of most users. The pseudocode for the algorithm is given in Table 3.1.

This PF scheduler does provide the proportional fairness as defined in (3.28) for the user average throughputs T_k as $W_c \to \infty$, and it was shown in [33] that the PF scheduler maximizes

$$\sum_{k=1}^{K} \ln T_k = \ln\left(\prod_{k=1}^{K} T_k\right) \tag{3.30}$$

almost surely among the class of all schedulers when $W_c \to \infty$ for joint ergodic processes $\{R_k(t),\ k = 1,...,K\}$, with *K* being the total number of active users. Equation (3.30) indicates that the PF scheduler 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

Table 3.1. Proportional fair scheduling algorithm pseudocode.

*// **Definitions***   *$DRC_i(t)$: current requested rate from user i at slot t, i=1, …, K.*   *$R_i(t)$: moving-average data rate of user i at slot t , i=1, …, K.*   *R_c: current transmission rate of user i, i=1,…, K.*   *W_c: time constant of user moving-average data rate*   *// **Scheduling at each new packet transmission***   *1. Initialize $R_i(0)=1$, i=1,…, K.*   *2. decide the highest $DRC_m(t)/R_m(t)=max\{DRC_i(t)/R_i(t)$, i=1,…, K.$\}$*   *3. Send data to user m*   *4. randomly break the ties if any*   *// **Update average user data rate at each slot***   *$R_i(t+1) = (1-1/W_c)R_i(t) + R_c/W_c$ i=1,…, K,*   *where $R_c = DRC_i(t)$ if user i is receiving data, $R_c = 0$ otherwise*

share, a PF scheduler brings the maximum revenue to the network operator according to [9]. Equation (3.30) is upper bounded by

$$\ln\left(\prod_{k=1}^{K} T_k\right) \le \ln\left(\frac{1}{K}\sum_{k=1}^{K} T_k\right)^K = K\ln\left(\sum_{k=1}^{K} T_k\right) - K\ln K\,, \tag{3.31}$$

where equality holds for symmetric user channels with $T_k = T$, $k = 1,...,K$. In this case, (3.30) reduces to

$$\sum_{k=1}^{K} \ln T_k = K\ln T\,. \tag{3.32}$$

Therefore, T serves as an upper bound on the limit of the long-term average user data rate.

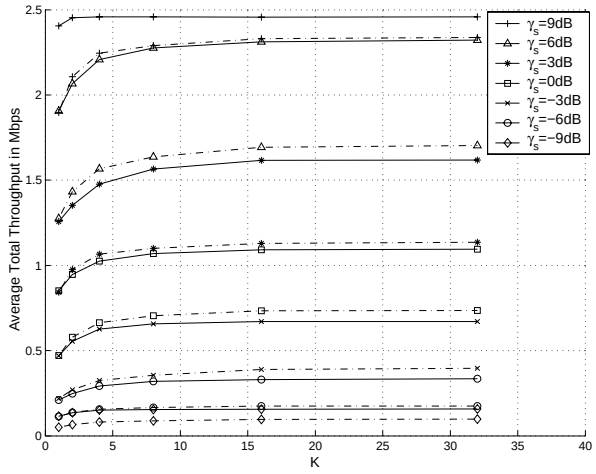
The PF scheduler in (3.29) exploits the multiuser diversity by assigning the radio resource to user with relatively better channel conditions, i.e., when the SINR at or near its peak. In this sense, the PF scheduling can be thought as an approximation of greedy

scheduling with the 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. In real systems, neither the average SINR per user nor their fading statistics can be expected to be the same. This is particularly important to systems without forward link power control as in 1xEV-DO, in which the full transmit power is assigned to each scheduled user for the best possible total throughput. Therefore, users with higher SINR and greater fading variations get higher throughput than those with the opposite condition. However, as indicated in [37], regardless of the user average SINR, the PF 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. Therefore, it is in this radio resource sharing sense that fairness is maintained, rather than in performance.

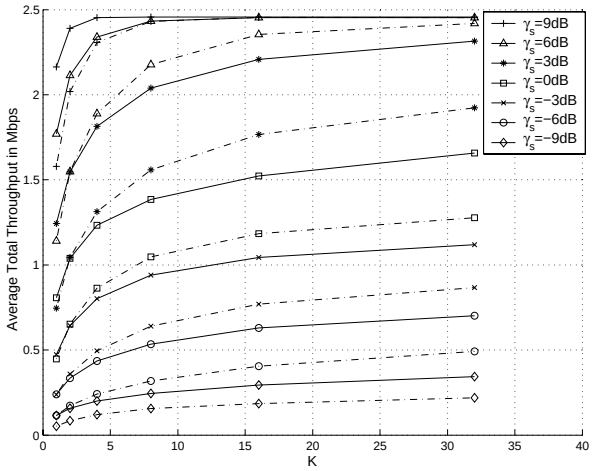
3.3.2 Impact of PF Scheduling on Receive Antenna Diversity

This section presents the numerical throughput improvement with two receive antennas over a single receive antenna in a multiuser scenario, and we focus on the IS-856 system downlink applications. As before, we assume that the channel fading is symmetric. However, instead of the i.i.d. block-fading model, we assume that the fading processes are ergodic with continuous fading variations across slots, and have low-to-medium Doppler rates corresponding to pedestrian-to-low-vehicular speed.

Figures 3.7(a) and (b) plot the average total throughput for the 1.5 Hz-Doppler Rician fading with a Rice factor $K_r = 10$ and the 55.6 Hz-Doppler Rayleigh fading channels, respectively, by using the mapping table in [38] for the SNR-to-data-rate mapping without hybrid automatic repeat request (H-ARQ) on the physical layer. The adaptive data rate is achieved on a single 1.25-MHz physical channel. The mapping is shown in Table 3.2. For



(a) In i.i.d. 1.5Hz-Doppler flat Rician (K_r=10) fading
over 8000 slots with 16.7 ms per slot and W_c=100 slots.



(b) In i.i.d. 55.6Hz-Doppler flat Rayleigh fading
over 2000 slots with 8.3 ms per slot and W_c=200 slots.

Figure 3.7. Average total throughput using PF scheduling with γ_s from 9 dB to –9 dB in a step size of 3 dB from top to bottom. Solid lines: $n_T = 1$ and $n_R = 2$; dash-dotted lines: $n_T = 1$ and $n_R = 1$.

simplicity purposes, we replace variable-length packet scheduling with fixed-length packet scheduling over each slot. As we can see from these figures, the higher fading rate and the larger fading dynamics are beneficial from the total throughput point of view, but they are

Table 3.2. SNR-to-Data Rate Mapping without H-ARQ.

Data Rate Index (4 bits)	Data Rate (Kbps)	$\gamma_k(t)$ Threshold (dB)
0	0	C/I < -11.5
1	38.4	-11.5 ≤ C/I < -9.0
2	76.8	-9.0 ≤ C/I < -6.5
3	153.6	-6.5 ≤ C/I < -3.0
4	307.2	-3.0 ≤ C/I < -0.5
5	614.4	-0.5 ≤ C/I < 2.0
6	921.6	2.0 ≤ C/I < 4.0
7	1228.8	4.0 ≤ C/I < 7.0
8	1843.2	7.0 ≤ C/I < 9.5
9	2457.6	≥ 9.5

subject to higher receive SNR clipping, which is especially pronounced at high average SNR regions. Compared with Figure 3.1, the spectral efficiency is more modest in Figure 3.7. In addition to the SNR quantization, the lower spectral efficiency also comes as a price for fairness over the window length W_c of about 1.67 s. In Figure 3.7(a) for the slow fading case with a strong line-of-sight component, the spectral efficiency suffers more due to limited multiuser diversity over W_c. Plots in Figures 3.7(a) and (b) also show that a doubling of throughput with 100% user average SNR increase cannot always be expected since both the user fading statistics and the number of users vary in a multiuser network. Generally, the higher improvement can be expected under the conditions of the lower user average SNR γ_s, the smaller fading variation, or the smaller number of active users. The first condition is due to the low SNR approximation of the achievable rate by

$$C = \log_2(1+\rho) = \left(\log_2 e\right)\cdot \ln(1+\rho) \approx \rho \log_2 e$$

in bits/sec/Hz for SNR $\rho \ll 1$. Therefore, in the vicinity of very low SNR, a 3-dB increase in

SNR does double the total throughput. The last two conditions are related to limited multiuser diversity via PF scheduling with finite window length. In those cases, the receive antenna diversity improvement can become more pronounced given that the potential channel diversity is not fully exploited via multiuser diversity.

3.4 Summary

In a large data network of multiple users, multiuser diversity can greatly increase system throughput. We have shown that under greedy packet scheduling, open-loop transmit antenna diversity over each point-to-point wireless link can have an adverse effect on the total system throughput. The asymptotic sum-rate analysis also shows that receive antenna diversity can be a waste of resources in the limit of a large number of users for the i.i.d. Rayleigh fading channels. However, in the application environments where the potential multiuser diversity gain is constrained by fairness concerns, benign fading channels with insufficient scattering, strong light-of-sight signal components, or a slow fading rate, the implementation of receive antenna diversity may be justified. With CQI at the transmitter, closed-loop antenna diversity can also be exploited. An example is selection transmit diversity which selects the best among n_T independent transmit antennas for each user using CQI feedback. In our channel model, this is equivalent to greedy scheduling with a single transmit antenna and n_T times the number of users. Therefore, unlike open-loop transmit diversity, selection transmit diversity can improve system throughput with greedy scheduling, although the gains may still be small.

In addition to diversity, multiple transmit antennas can provide spatial multiplexing and improve system throughput with scheduling. One goal of this dissertation is evaluating

downlink throughput improvement by using multiple transmit antennas effectively and efficiently. The following chapters are devoted to the study of multi-antenna downlink transmission using spatial multiplexing under multiuser scheduling, specifically, we provide solutions built on suboptimal dirty paper coding techniques.

Chapter 4
Suboptimal Dirty Paper Coding and Scheduling

In the previous chapter, we demonstrated that open-loop antenna diversity techniques degraded the impact of channel-aware scheduling. In this chapter, we investigate the interaction of spatial multiplexing and channel-aware scheduling. Specifically we examine a transmitter-centric technique termed dirty paper coding (DPC), which permits low complexity receivers while attaining spatial multiplexing through precoding. We will show that channel-aware scheduling leads to substantial capacity gains when paired with DPC. Additionally we show that intelligent reordering allows for significant performance improvement, with minimal rate reduction.

4.1 Review of Zero-Forcing Dirty Paper Coding (ZF-DPC)

Dirty paper coding provides a new coding technique for the broadcast channel, which is superior to the single-user design that ignores the multiuser interference. For a vector Gaussian broadcast channel with single-antenna users, Caire and Shamai [10] proposed a suboptimal DPC scheme which forces to zero the interference from users $j > k$. This suboptimal scheme was named zero-forcing DPC (ZF-DPC). They also provided a closed-form solution for the sum capacity of two users, which is hard to generalize to more than two users or more than one antenna per user. However, for more than two users, the ZF-DPC was proved to be asymptotically sum-rate optimal at both low and high SNR under certain mild conditions. It is worth noting that this same technique was independently

applied to a multiuser DSL channel by Ginis and Cioffi [39]. In this section, we describe the ZF-DPC scheme.

4.1.1 Maximum Sum Rate

For the vector broadcast channel model in (2.11) with $m_k = 1, k = 1,...,K$, a more intuitive precoding matrix $\mathbf{B}$ can be chosen based on a QR-type decomposition [40]. Let $\mathbf{H}(\mathcal{S}) \in \mathbb{C}^{n_s \times n_t}$ be the channel matrix between n_t transmit antennas and a user subset $\mathcal{S}$ of cardinality $n_s = \text{rank}(\mathbf{H})$, and let $\mathbf{H}(\mathcal{S}) = \mathbf{G}\mathbf{B}^H$ be the unique LQ decomposition of $\mathbf{H}(\mathcal{S})$, such that $\mathbf{G} = [g_{i,j}] \in \mathbb{C}^{n_s \times n_s}$ is lower triangular having positive diagonal elements, and $\mathbf{B} = [\mathbf{b}_1,...\mathbf{b}_{n_s}] \in \mathbb{C}^{n_t \times n_s}$ has orthonormal columns with $\mathbf{B}^H\mathbf{B} = \mathbf{I}_{n_s}$. Following the notation used in the section 2.3.2, over each coherence interval of *T* symbols the transmitter encodes the user data streams $\mathbf{X}_k \in \mathbb{C}^{n_t \times T}$ with $\mathbf{X}_k = \mathbf{b}_k\mathbf{u}_k$ using Gaussian inputs $\mathbf{u}_k \in \mathbb{C}^{1\times T}$. Hence we have the received precoded signal as

$$\mathbf{y}_k = g_{k,k}\mathbf{u}_k + \sum_{j<k} g_{k,j}\mathbf{u}_j + \mathbf{w}_k, \quad k = 1,...,n_s, \tag{4.1}$$

where no information is sent to users $k > n_s$. The interference from users $j > k$ to user k is forced to zero since $g_{k,j} = 0$ for $j > k$, and the interference from users $j < k$ is non-causally known to the transmitter and can be pre-subtracted prior to transmission. From Costa's result assuming unit noise variance, the SINR of user *k* is thus

$$SINR_k^{\text{zfdpc}} = {g_{k,k}}^2, \tag{4.2}$$

and the achievable sum rate is then

$$R_{\text{sum}}^{\text{zfdpc}} = \sum_{k=1}^{n_s} [\log_2(\xi \cdot SINR_k^{\text{zfdpc}})]_+, \tag{4.3}$$

where ξ is the waterfilling solution to

$$\mathrm{E}\left[\sum_{k=1}^{n_s}(\xi - 1/SINR_k^{\mathrm{zfdpc}})_+\right] = P, \tag{4.4}$$

with the expectation over the ergodic channel variations. The maximum sum rate with ZF-DPC is therefore

$$R_{\mathrm{sum}}^{\mathrm{zfdpc\text{-}max}} = \max_{\mathcal{S}} R_{\mathrm{sum}}^{\mathrm{zfdpc}}, \tag{4.5}$$

with the maximization over all the *ordered* user subsets $\mathcal{S}$ of cardinality n_s. The total number of such ordered subsets is upper bounded by $\mathrm{P}_K^{n_s}$, the number of n_s permutations out of K.

For comparison, we present the achievable sum rate of transmit zero-forcing linear beamforming (ZF-LBF) for the channel model above. The transmit ZF-LBF consists of inverting the channel matrix at the transmitter in order to create orthogonal channels between the transmitter and the receivers without cooperation between receivers. Let $\mathcal{S}$ be a user subset of cardinality not greater than n_s, i.e., $|\mathcal{S}| \le n_s$, so that the corresponding channel matrix $\mathbf{H}(\mathcal{S})$ has full row rank, i.e., $\mathrm{rank}\left(\mathbf{H}(\mathcal{S})\right) = |\mathcal{S}|$. Let the pseudoinverse of $\mathbf{H}(\mathcal{S})$ be denoted by $\mathbf{H}(\mathcal{S})^+ \triangleq \mathbf{H}(\mathcal{S})^H \left(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H\right)^{-1}$ [41]. With ZF-LBF, the transmitted signal is pre-multiplied by $\mathbf{X} = \mathbf{H}(\mathcal{S})^+\mathbf{U}$ so that the channel in (2.11) is converted into a set of parallel channels

$$\mathbf{y}_k = \mathbf{u}_k + \mathbf{w}_k, \quad k = 1, \ldots |\mathcal{S}|, \tag{4.6}$$

where no information is sent to users $k \notin \mathcal{S}$. The received SINR at user k is then

$$SINR_k^{\mathrm{zflbf}} = \frac{1}{\left[\left(\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H\right)^{-1}\right]_{k,k}}, \tag{4.7}$$

The achievable sum rate for a given user subset $\mathcal{S}$ is thus

$$R_{\text{sum}}^{\text{zflbf}} = \sum_{k=1}^{|\mathcal{S}|} [\log_2(\xi \cdot SINR_k^{\text{zflbf}})]_+ , \tag{4.8}$$

where ξ is the waterfilling solution to

$$\mathrm{E}\left[\sum_{k=1}^{|\mathcal{S}|} (\xi - 1/SINR_k^{\text{zflbf}})_+ \right] = P . \tag{4.9}$$

The maximum sum rate of ZF-LBF is therefore

$$R_{\text{sum}}^{\text{zflbf-max}} = \max_{\mathcal{S}} R_{\text{sum}}^{\text{zflbf}} , \tag{4.10}$$

where the maximization is over all the *unordered* user subsets $\mathcal{S}$ of cardinality $|\mathcal{S}| \le n_s$. The total number of such unordered subsets is upper bounded by $\sum_{i=1}^{n_s} \mathrm{C}_K^i$. It has been shown in [10] that for any channel matrix $\mathbf{H}$, $R_{\text{sum}}^{\text{zfdpc-max}} \ge R_{\text{sum}}^{\text{zflbf-max}}$.

For user subsets of unit cardinality only, the ZF-LBF reduces to the maximal ratio combining beamforming (MRC-BF). The maximum rate of MRC-BF among all the *K* users is simply

$$R_{\max}^{\text{mrcbf-max}} = \log_2(1 + P \left\| \mathbf{h}_{\max} \right\|^2), \tag{4.11}$$

where $\mathbf{h}_{\max}$ is the channel to the user with the largest Euclidean norm. This is achieved by only transmitting to the user in the best channel condition in the sense of conventional transmit beamforming, and is exactly the sum-rate optimal transmission scheme for the degraded Gaussian broadcast channel with $n_t = 1$. However, as we will see later, this scheme is generally not optimal for $n_t > 1$.

If we allow the receivers to cooperate, the channel in (2.11) is then equivalent to a single-link MIMO channel. In this case, the maximum achievable rate is

$$R^{\text{coop}} = \sum_{k=1}^{n_s} [\log_2(\xi \cdot SINR_k^{\text{coop}})]_+ , \tag{4.12}$$

with $SINR_k^{\text{coop}}$ being the nonzero squared singular values of $\mathbf{H}$ and ξ being the waterfilling solution to

$$\mathrm{E}\left[\sum_{k=1}^{n_s}(\xi - 1/SINR_k^{\text{coop}})_+\right] = P. \tag{4.13}$$

The sum rate R^{coop} is obviously an upper bound to both $R_{\text{sum}}^{\text{zfdpc-max}}$ and $R_{\text{sum}}^{\text{zflbf-max}}$ under the same condition for any channel **H**, according to Sato upper bound [26]. Since for any unitary matrix **Q**, **QH** has the same singular values of **H**, the cooperative achievable rate R^{coop} is independent of user ordering.

We have seen that the sum rate of ZF-DPC depends on the ordered user subset of cardinality n_s, whereas the sum rate with ZF-LBF is related to the unordered user subset of cardinality not greater than n_s. It has been shown in [10] that for ZF-DPC, the maximum sum rate $R_{\text{sum}}^{\text{zfdpc-max}}$ is achieved by an ordered set of n_s users for every $P \geq 0$. This implies spatial multiplexing is always achieved for sufficiently large P. Moreover, ZF-DPC becomes independent of the user ordering for asymptotically large SNR if $\mathbf{H}$ has full row rank, i.e., if n_s=min(n_t,K). For ZF-LBF with a fixed power constraint P, the maximum sum rate $R_{\text{sum}}^{\text{zflbf-max}}$ may be achieved by a user subset of cardinality strictly less than n_s. However, there still exists a finite value P_0 depending on $\mathbf{H}$, such that for all $P > P_0$, $R_{\text{sum}}^{\text{zflbf-max}}$ is achieved by a subset of cardinality n_s. This result implies that, when channel conditions permit, the spatial multiplexing in a multi-antenna broadcast channel can be provided with both ZF-DPC and ZF-LBF, whereas for the same multiplexing gain, the power requirement for ZF-LBF will be higher than for ZF-DPC in general.

4.1.2 Asymptotic Sum-Rate Optimality

The asymptotic optimality of $R_{\text{sum}}^{\text{zfdpc-max}}$ was established in [10] at both low and high

SNR. The conclusions are summarized in *Theorem 4.1* and *4.2* below.

Theorem 4.1 (Theorem 3 in [10]): For any channel matrix **H** with full row rank,

$$\lim_{P\to\infty}(R^{\text{coop}} - R_{\text{sum}}^{\text{zfdpc-max}}) = 0, \tag{4.14}$$

and for any channel matrix **H**,

$$\lim_{P\to 0}\frac{R_{\text{sum}}^{\text{zfdpc-max}}}{R_{\text{sum}}^{\text{zflbf-max}}} = 1. \tag{4.15}$$

Theorem 4.1 states that for the condition of a full channel row rank, the maximum sum rate of ZF-DPC asymptotically approaches the sum capacity of the cooperative receivers, or the capacity of a single-link MIMO channel, at high SNR. At low SNR, ZF-DPC asymptotically achieves the maximum sum rate of ZF-LBF for any channel.

Theorem 4.2 (Theorem 4 in [10]): If $\mathbf{H}$ has full row-rank, then

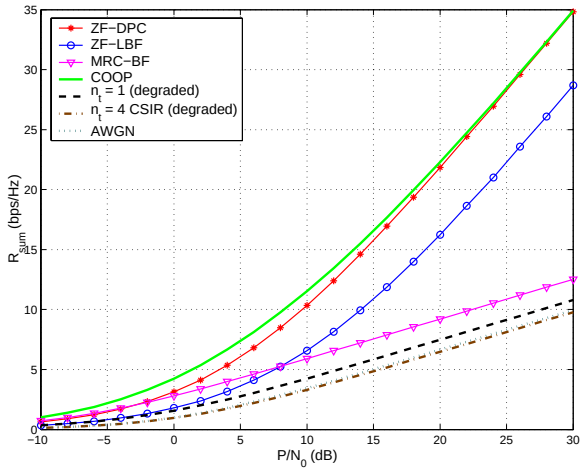
$$\lim_{P\to\infty}(C_{\text{sum}} - R_{\text{sum}}^{\text{zfdpc-max}}) = 0, \tag{4.16}$$

and

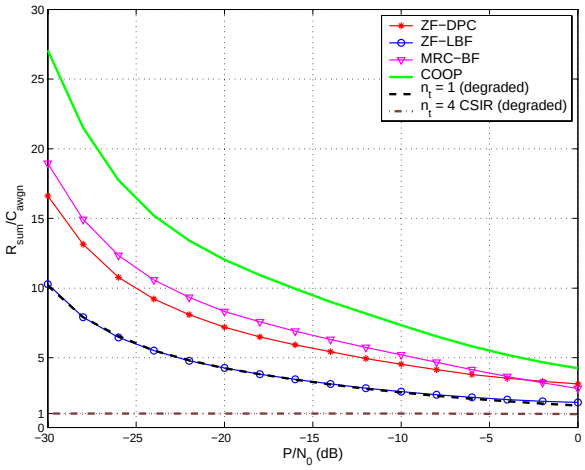
$$\lim_{P\to 0}\frac{C_{\text{sum}}}{R_{\text{sum}}^{\text{zfdpc-max}}} = 1. \tag{4.17}$$

Theorem 4.2 establishes the asymptotic sum-rate optimality of ZF-DPC at both low and high SNR, under the condition of a full channel row rank.

Figures 4.1(a) and (b) depict the average sum rates $R_{\text{sum}}^{\text{zfdpc}}$, $R_{\text{sum}}^{\text{zflbf}}$, $R_{\text{max}}^{\text{marcbf-max}}$ and R^{coop} for n_t=K=4 in i.i.d. flat Raleigh fading channels. The sum rates are averaged over 30,000 independent channel realizations (coherence intervals). The user subset selection and ordering are not taken into account in the plot. For comparison purposes, the capacity of a point-to-point AWGN channel with the same total average power constraint is plotted as a baseline. Also shown in Figure 4.1 is the average sum capacity of the standard degraded Gaussian broadcast channel with n_t=1, in which the sum capacity is achieved by transmitting



(a) At high SNR.



(b) At low SNR. Curves are plotted relative to AWGN channel capacity.

Figure 4.1. Average sum rate in i.i.d. flat Rayleigh fading for $n_t = K = 4$.

to the best user only in each coherence interval [42]. We have assumed that CSI is known to both the transmitter and the respective receivers in all the cases above. Under the assumption that CSI is only known to the receivers (CSIR), and in the condition of a symmetric ergodic channel, where the channel matrix **H** has i.i.d. rows, the marginal user channel statistics are the same, and the vector Gaussian broadcast channel of single-antenna users for any n_t>1 is

stochastically degraded [19]. Hence the capacity region can be obtained by time-sharing. The optimal sum rate in this case is the same as that obtained in [3] with transmit diversity,

$$C_{\text{sum}} = \log_2\left(1+\frac{P}{N_0 n_t}\left\|\mathbf{h}_k\right\|^2\right), \quad k=1,...,K\,. \tag{4.18}$$

This is also plotted in Figure 4.1. While Figure 4.1(a) shows the performance at high SNR, Figure 4.1(b) focuses on low SNR region. The following observations can be made from these plots at $K=n_t$:

- Even without considering user ordering, at high SNR ZF-DPC asymptotically achieves the point-to-point link capacity with cooperative receivers (COOP); while at low SNR, ZF-DPC approaches the average sum rate of MRC-BF. The value of receiver cooperation becomes apparent at vanishingly small SNR.
- When transmitting to the same $K=n_t$ users in a fixed order at low SNR, MRC-BF obtains greater average sum rate than ZF-DPC and ZF-LBF. This means at low SNR, the power gain from only transmitting to the best user brings the most benefit to the average sum rate.
- When transmitting to the same $K=n_t$ users in a fixed order at high SNR, the spatial multiplexing gain of ZF-DPC and ZF-LBF is asymptotically equal to that of COOP, that is, n_t bps/Hz increase per 3dB SNR increase, whereas the MRC-BF achieves only 1 bps/Hz increase per 3dB SNR increase, the same as for the point-to-point AWGN channel and the two degraded broadcast channels.
- With CSI at the transmitter and $K>1$, the average sum rate is above that of the point-to-point AWGN channel, whereas with CSIR only, the average sum rate of the degraded Gaussian broadcast channel with n_t is below that of AWGN due to fading. The advantage of CSI at transmitter is dramatic at both high SNR and low SNR.

We have been using a long-term power constraint for the capacity evaluation, that is we require that the total average transmit power is constrained to P. For an ergodic channel fading process, this is to require

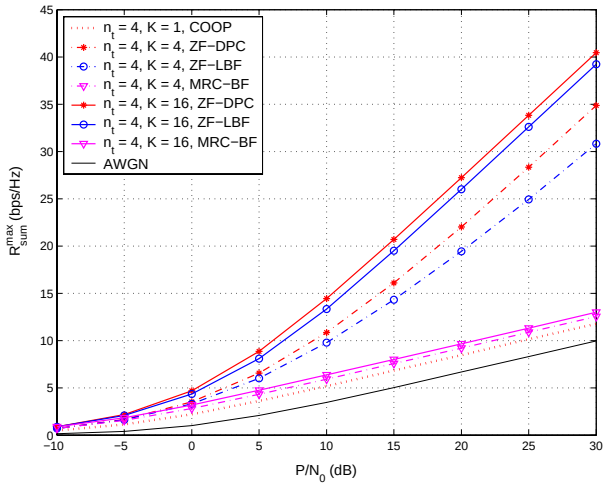
$$\mathrm{E}\left[\sum_{k=1}^{n_s}(\xi - 1/SNR_k)_+\right] = P\,. \tag{4.19}$$

We can also use a strict short-term power constraint over each coherence interval,

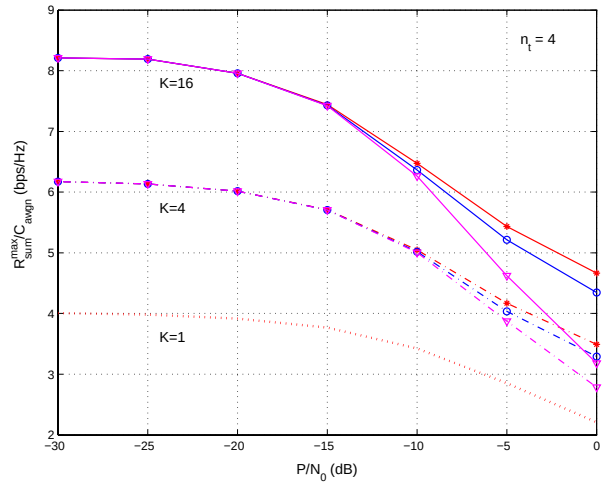
$$\sum_{k=1}^{n_s}(\xi - 1/SNR_k)_+ = P\,, \tag{4.20}$$

which also meets the long-term power constraint. Our interest is to show the impact of multiuser scheduling and ordering on the achievable sum rate of ZF-DPC. We evaluate the maximum sum rates $R_{\text{sum}}^{\text{zfdpc-max}}$, $R_{\text{sum}}^{\text{zflbf-max}}$, $R_{\text{max}}^{\text{mrcbf-max}}$, and R^{coop} with this short-term power constraint below in Figures 4.2(a) and (b), using $n_t = 4$ for different number of users in i.i.d. flat Rayleigh fading. As before, the capacity of a point-to-point AWGN channel is shown as a baseline.

Figure 4.2 indicates that under a power constraint over each block, the ZF-DPC achieves the largest maximum average sum rate over the entire SNR region among ZF-DPC, ZF-LBF and MRC-BF. Comparing the K=4 case with that in Figure 4.1 without user ordering and under a long-term power constraint, we see that the optimal user ordering brings negligible sum rate gain to ZF-DPC at moderate to high SNR. This result confirms the statement made earlier for the asymptotic user-ordering independence of ZF-DPC performance at high SNR. However, the maximum achievable sum rate of ZF-LBF is increased through the optimal user subset at moderate to low SNR, which is no less than that of MRC-BF over the entire SNR region. We see in Figure 4.2 that when K>n_t, a significant power gain from multiuser scheduling leads to a parallel up shift of the capacity curves at high SNR, and a parallel up shift of relative capacity curves at vanishingly small SNR, which



(a) At high SNR.



(b) At low SNR. Curves are plotted relative to AWGN channel capacity.

Figure 4.2. Average maximum sum rate in i.i.d. flat Rayleigh fading for $n_t = 4$ and $K = 1, 4, 16$.

converges to the transmit beamforming to the best user. As K continues to increase, the maximum sum rate of ZF-LBF approaches that of ZF-DPC. At K=1, the sum rates of ZF-DPC, ZF-LBF and MRC-BF are the same.

Therefore, similar to the sum capacity over a scalar multiacess [43] or broadcast channel [42], in a large data network with K>n_t, we can expect a multiuser diversity gain in

average sum rate from dynamically scheduling resources among the users as a function of their channel states, but unlike a scalar multiuser channel, the multiple antennas at the base station can provide parallel independent data transmissions at the expense of more channel state knowledge. We will show later that that this sum-rate greedy scheduler, which always grants the radio resource to the best user subset, can be closely approximated with an efficient suboptimal scheduler, which allows for both an equal transmit power allocation and an adaptive number users to be scheduled. Further, on top of this suboptimal scheduling, a minimum-error reordering can be applied, similar to the optimal ordering in the V-BLAST algorithm [11]. In the following, we first show the dual relationship between ZF-DPC and V-BLAST, and its performance advantages over V-BLAST.

4.1.3 ZF-DPC versus V-BLAST

Theoretically, both ZF-DPC and V-BLAST achieve a lower bound on the MIMO link capacity through successive interference cancellation (SIC) and nulling. For ZF-DPC, this refers to a composite MIMO link with n_t co-located transmit and n_s distributed receive antennas, with the joint processing performed at transmitter assuming CSI is available. For the MIMO broadcast channel model in (2.11), the spatial multiplexing supportable by ZF-DPC is of order $n_s \le \min(n_t, m)$. In the case of V-BLAST, the joint processing is performed at each receiver, and the sum rate is upper bounded by individual MIMO links. For comparison, we focus on the zero-forcing V-BLAST scheme [11]. For the channel model in (2.11), the achievable rate using V-BLAST with $m_k \ge n_t, k = 1,...,K$, and joint processing at receiver *k* is [2]

$$R_k^{\text{vblast}} = \sum_{i=1}^{n_t} \log_2\left(1 + P_T q_{i,i}^2(k)/n_t\right) \tag{4.21}$$

assuming i.i.d. Gaussian noise of unit variance, where $q_{i,i}(k)$ are the real diagonal entries of

the upper triangular matrix from the QR decomposition of the *k*th user channel matrix. Due to the lack of CSI at the transmitter, V-BLAST uses equal transmit power allocation across transmit antennas. Therefore, the maximum achievable sum rate is

$$R_{\text{sum}}^{\text{vblast-max}} = \max_{1 \le k \le K} R_k^{\text{vblast}}, \tag{4.22}$$

where the maximization is over K users each with $m_k \ge n_t, k = 1,...,K$, receive antennas, and consequently, the maximum order of spatial multiplexing supportable is n_t with transmission to a single user each time. If we also use equal power allocation in (4.3), a direct statistical comparison with (4.21) indicates that ZF-DPC and V-BLAST actually achieve the same average sum rate in the absence of scheduling, and can be thought as dual techniques with transmitter- and receiver-centric processing, respectively. However in reality, V-BLAST suffers from the problem of error propagation and ZF-DPC requires full CSI at the transmitter. The transmitter-centric processing of ZF-DPC greatly reduces receiver complexity and power consumption. Also, the maximization of the sum rate for ZF-DPC is over $\mathrm{P}_m^{n_t}$ (n_t permutations out of m) receive antenna subsets for $m \ge n_t$, rather than only K users with V-BLAST. This last point suggests a potentially much larger scheduling gain for ZF-DPC than for V-BLAST since $\mathrm{P}_m^{n_t} \gg K$ for $m \gg n_t$. This is true even for a relatively small number of m and n_t. However, to obtain this scheduling gain, an efficient implementation of scheduling is needed for ZF-DPC.

4.2 Suboptimal Scheduling for ZF-DPC

For the maximization in (4.5), a suboptimal greedy scheduling algorithm was

proposed in [44], which was shown to obtain a nearly optimal average maximum sum rate[1] using optimal power allocation. Optimal waterfilling power allocation puts a high demand on the linear range of transmit power amplifiers, which is extremely costly from a practical point of view if not impossible, especially due to the need for multiple transmit RF chains. From a theoretical point of view, the waterfilling operation also makes it difficult to analyze the scheduling performance. Therefore, we present a suboptimal greedy scheduler which allocates equal power to all selected users. Asymptotically, we show that when K is sufficiently large, the maximum sum rate with equal power allocation approaches that with optimal power allocation. Our algorithm further streamlines the greedy scheduler in [44] by making use of the descending ordering of channel gains for selected users, and more importantly, it allows for the derivation of a tight upper bound on the average maximum sum rate with scheduling.

4.2.1 Suboptimal Greedy Scheduling with Equal Power Allocation

For the MIMO broadcast channel model (2.11) with a total average transmit power constraint P_T and $m_k = 1, k = 1,..., K$, define the maximum sum rate of ZF-DPC with equal user power as

$$\tilde{R}_{\text{sum}}^{\text{zfdpc-max}} = \max_{\tilde{\mathcal{S}}} \sum_{k \in \tilde{\mathcal{S}}} \log_2 (1 + P_T \tilde{g}_{k,k}^2 / |\tilde{\mathcal{S}}|), \tag{4.23}$$

where $\tilde{\mathcal{S}} \subset \{1,..., K\}$ with $1 \le |\tilde{\mathcal{S}}| \le \min(n_t, K)$. The maximization in (4.23) is performed over all *ordered* user subsets $\tilde{\mathcal{S}}$ of cardinality not greater than $\min(n_t, K)$. Let $\mathcal{P}(\mathbf{X})$ be the orthogonal projection onto the orthogonal complement of the subspace spanned by the rows

[1] The average maximum sum rate is averaged over the channel variations with i.i.d. block-wise fading. For ergodic channel fading processes, the term ergodic maximum sum rate can be used interchangeably here.

of matrix $\mathbf{X}$, let $\|\mathbf{x}\|$ be the Euclidean norm of vector $\mathbf{x}$, and let $\{\mathcal{A}\}\setminus\{\mathcal{B}\}$ be the set $\mathcal{A}$ excluding set $\mathcal{B}$. An efficient suboptimal greedy scheduling algorithm for the maximization in (4.23) is given in Table 4.1, where both the users and their number are scheduled adaptively with the objective being the maximization of sum rate. We define $\mathcal{P}\left(\mathbf{H}(\varnothing)\right) \triangleq 1$.

Table 4.1. Suboptimal greedy scheduler
with equal user power allocation.

Step 1: Initialization

$\mathcal{S}^* = \varnothing$, $T^* = 1$

Step 2:

For $n = 1$ to $\min(n_t, K)$

$\gamma_k = \left\|\mathbf{h}_k \mathcal{P}(\mathbf{H}(\mathcal{S}^*))\right\|^2$, $k \in \{1,...,K\}\setminus\mathcal{S}^*$

$l = \arg\max_k \gamma_k$

$T = (1 + P_T\gamma_l / n)\prod_{m=1}^{n-1}(1 + P_T\gamma_m^* / n)$

if $T > T^*$

$T^* = T$, $\gamma_n^* = \gamma_l$, $\mathcal{S}^* = \mathcal{S}^* \bigcup \{l\}$

else

go to Step 3

End

Step 3: Deliver $\widehat{R}_{\text{sum}}^{\text{zfdpc-max}} = \log_2 T^*$.

We have the following result for the suboptimal greedy scheduling algorithm in Table 4.1:

Lemma 4.1: The incremental user selection in Table 4.1 leads to a final subset $\mathcal{S}^*$ of a cardinality $1 \le |\mathcal{S}^*| \le \min(n_t, K)$ and the channel gains of the selected users are chosen in a non-increasing order with $\gamma_1^* \ge \gamma_2^* \ge ... \ge \gamma_{|\mathcal{S}^*|}^*$. For a proof of this lemma see Appendix A.

Remark 4.1: At Step 2 in Table 4.1, we can use $T = (1+P_l^*\gamma_l)\prod_{m=1}^{n-1}(1+P_m^*\gamma_m^*)$ for the optimal power allocation via waterfilling, which obtains the same result as that in [44]. However, the scheduler in Table 4.1 terminates the for-loop whenever adding the best remaining user does not increase the sum rate by using an if-else switch. This if-else switch is based on the result of non-increasing ordering of $\gamma_k^*, k = 1,...,|\mathcal{S}^*|$, in *Lemma 4.1*. We note that this suboptimal scheduler adaptively adjusts the number of admitted users and can be thought of as an approximation to the scheduler with optimal power allocation using adaptive equal power allocation. The implementation in Table 4.1 is more computationally efficient than that in [44], not only because we avoid the waterfilling computation at each step, but also because of the possible early termination of the selection loop. This reduces the necessary computations, especially at low SNR when the maximum sum rate is often achievable with a user subset of cardinality strictly less than $\min(n_t, K)$.

Remark 4.2: We can also maximize spatial multiplexing by setting $|\mathcal{S}^*| = \min(n_t, K)$ and using the suboptimal scheduling in Table 4.2. In this case, the for-loop is carried to the loop end, and the maximum sum rate is no greater than that in Table 4.1. Again, the channel gains of selected users are in a non-increasing order $\gamma_1^* \geq \gamma_2^* \geq ... \geq \gamma_{|\mathcal{S}^*|}^*$ (see Appendix A). It is worth noting that this spatial-multiplexing maximization, even though it decreases the maximum achievable sum rate, increases the fairness in slot allocation since a maximum number of users are admitted over each block. This will be further justified by the numerical results in the following. We note that the decrease in the maximum sum rate is negligible at high SNR or *K* sufficiently large. In the next section, we will show that this simplification also allows us to derive a tight upper bound on the average maximum sum rate with scheduling.

Table 4.2. Suboptimal greedy scheduler with equal power allocation and maximum spatial multiplexing.

Step 1: Initialization

$\mathcal{S}^* = \varnothing$, $T^* = 1$

Step 2:

For $n = 1$ to $\min(n_t, K)$

$\gamma_k = \left\| \mathbf{h}_k \mathcal{P}(\mathbf{H}(\mathcal{S}^*)) \right\|^2$, $k \in \{1, ..., K\} \setminus \mathcal{S}^*$

$l = \arg\max_k \gamma_k$

$\gamma_n^* = \gamma_l$

$T^* = T^* \cdot \left(1 + P_T \gamma_n^* / \min(n_t, K)\right)$

$\mathcal{S}^* = \mathcal{S}^* \bigcup \{l\}$

End

Step 3: Deliver $\breve{R}_{\text{sum}}^{\text{zfdpc-max}} = \log_2 T^*$.

Consider the i.i.d. flat Rayleigh fading channel matrix $\mathbf{H} \in \mathbb{C}^{K \times n_t}$ with i.i.d. circular symmetric complex Gaussian elements of zero mean and unit variance. We have the following result.

Theorem 4.3: For a finite n_t and a positive total transmit power constraint $P_T > 0$, $\lim_{K \to \infty} \left(R_{\text{sum}}^{\text{zfdpc-max}} - \tilde{R}_{\text{sum}}^{\text{zfdpc-max}} \right) = 0$ with optimal scheduling over the i.i.d. flat Rayleigh fading channels defined above. For a proof of this theorem see Appendix B.

Theorem 4.3 also holds when the spatial multiplexing is maximized in (4.23), i.e., $|\tilde{\mathcal{S}}| = \min(n_t, K)$, since under optimal power allocation with $P_T > 0$, the maximum sum rate is achievable with $|\mathcal{S}| = \min(n_t, K)$ as stated before. Further, it is known that

$\lim_{P_T \to \infty} \left(R_{\text{sum}}^{\text{zfdpc-max}} - \tilde{R}_{\text{sum}}^{\text{zfdpc-max}} \right) = 0$ with the total transmit power being equally divided among the $\min(n_t, K)$ best users, and $\lim_{P_T \to 0} \left(R_{\text{sum}}^{\text{zfdpc-max}} - \tilde{R}_{\text{sum}}^{\text{zfdpc-max}} \right) = 0$ with the total transmit power devoted to a single best user for $P_T > 0$ [10].

With perfect CSI at both the transmitter and receiver and under a long-term average transmit power constraint, the average sum rate of $R_{\text{sum}}^{\text{zfdpc}}$ in (4.3) over the i.i.d. block-fading channels is the same as that achievable over the ergodic fading channels, over which the random fading processes vary independently during each transmission block so that the channels are information stable [45]. As a result, a corollary of *Theorem 4.3* is given below.

Corollary 4.1: For a finite n_t and a positive total transmit power constraint $P_T > 0$, $\lim_{K \to \infty} \left(\mathrm{E}_{\mathbf{H}} \left(R_{\text{sum}}^{\text{zfdpc-max}} \right) - \mathrm{E}_{\mathbf{H}} \left(\tilde{R}_{\text{sum}}^{\text{zfdpc-max}} \right) \right) = 0$ under optimal scheduling over ergodic and independent flat Rayleigh fading channels which are information stable, where the operator $\mathrm{E}_{\mathbf{H}}(\cdot)$ denotes the expectation over the channel matrix $\mathbf{H}$.

Figure 4.3 shows the average maximum sum rate under suboptimal scheduling using equal user power allocation for both adaptive user selection and maximum spatial multiplexing. The user channel matrix $\mathbf{H}$ was assumed to be i.i.d. circular symmetric complex Gaussian elements of unit variance. For comparison, the average maximum sum rate under suboptimal scheduling with the optimal transmit power allocation is also plotted. As we can see, the average maximum sum rate with equal power allocation is very close to that of optimal power allocation. In fact, the two curves are almost indistinguishable. For the average maximum sum rate with maximum spatial multiplexing and equal power allocation, the deficiency is very small at both high SNR and low SNR, and decreases with increasing *K*.

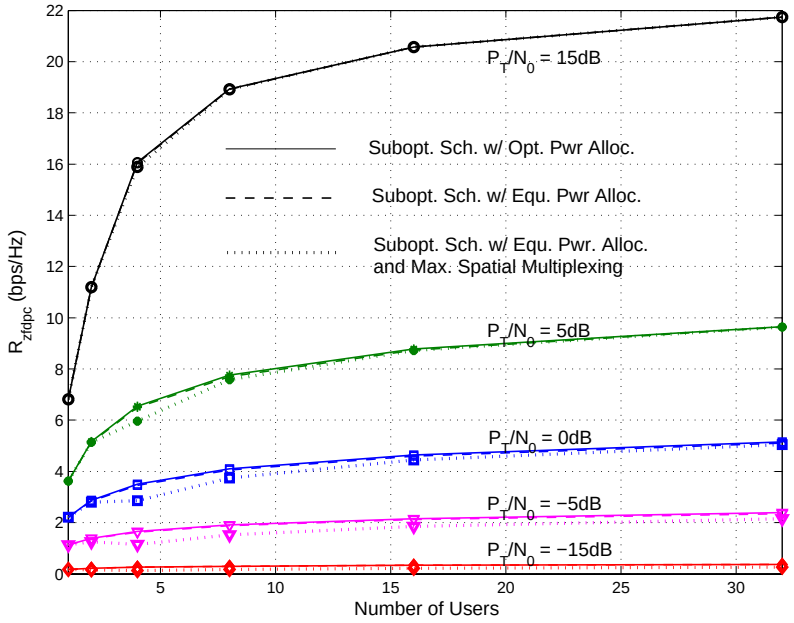


Figure 4.3. Average maximum sum rate under the suboptimal scheduling with optimal and equal user power allocation in i.i.d. flat Rayleigh block-wise fading.

4.2.2 An Upper Bound on the Average Maximum Sum Rate

We now derive an upper bound on the average maximum sum rate under the suboptimal greedy scheduling. To make the analysis tractable, we assume that the channel matrix $\mathbf{H}$ has i.i.d. circular symmetric complex Gaussian elements of unit variance, and we use the suboptimal scheduling in Table 4.2 with the maximum spatial multiplexing of $\min(n_t, K)$ independent user streams. From *Remark 4.2* in Section 4.2.1, we know that when we fix maximum spatial multiplexing $|\mathcal{S}^*| = \min(n_t, K)$, the channel gains of the selected users follow a descending order $\gamma_1^* > \gamma_2^* > \ldots > \gamma_{|\mathcal{S}^*|}^*$ with probability 1 for the channel matrix $\mathbf{H}$ assumed. The following result is based on this observation.

Lemma 4.2: For the user channel matrix $\mathbf{H}$ with i.i.d. circular symmetric complex Gaussian elements of zero mean and unit variance, under the suboptimal scheduling in Table

4.2 with maximum spatial multiplexing of $|\mathcal{S}^*| = \min(n_t, K)$, the selected user channel gains γ_k^*, $k = 2,..., \min(n_t, K)$, are upper bounded by $U_k \triangleq \min(U, \max_{1 \le i \le K-k+1} \chi^2_{2(n_t-k+1)})$, where $U \triangleq \min(\max_{1 \le i \le \max(K/n_t,1)} \chi^2_{2n_t}, \max_{1 \le i \le \max(K/n_t,1)} \chi^2_{2(n_t-1)}, ..., \max_{1 \le i \le \max(K/n_t,1)} \chi^2_{2(n_t-k+2)})$, $k \ge 2$, and $\max_{1 \le i \le N} \chi^2_n$ denotes the maximum of N i.i.d. chi-square random variables distributed as χ^2_n. The first selected user channel gain γ_1^* is distributed as $U_1 \triangleq \max_{1 \le i \le K} \chi^2_{2n_t}$ exactly. For a proof of this lemma see Appendix C.

We note that the upper bound in *Lemma 4.2* is independent of the transmit power constraint P_T. In the next section, we will show that the bound is quite tight over the entire region of SNR, even for a relatively small number of users. Since only the upper bounding of γ_k^*, $k = 2,..., \min(n_t, K)$, is considered in *Lemma 4.2*, which introduces correlation among γ_k^*, $k = 1,..., \min(n_t, K)$, this upper bound is appropriately tight for calculating the average maximum sum rate, rather than the outage sum rate, which is more dependent on lower bounding, especially at low outage rate. Therefore, we have the following result.

Theorem 4.4: For ergodic and independent flat Rayleigh fading channels $\mathbf{H} \in \mathbb{C}^{K \times n_t}$ each with unit average power gain, under the suboptimal scheduling in Table 4.1 with maximum spatial multiplexing of $|\mathcal{S}^*| = \min(n_t, K)$, the average maximum sum rate is upper bounded by

$$\sum_{k=1}^{\min(n_t,K)} \int_0^\infty \log_2(1 + P_T u_k / \min(n_t, K)) f_k(u_k) du_k , \qquad (4.26)$$

where $f_k(u_k)$ are the pdfs of U_k as defined in *Lemma 4.2*. The proof follows directly from *Corollary 4.1* and *Lemma 4.2*.

The pdfs of $U_k, k = 1,..., \min(n_t, K)$, can be calculated using the results from order statistics. The pdf and cdf for the maximum of N i.i.d. random variables $X_i, i = 1,..., N$, are [32], respectively,

$$f_{(N)} = Nf(x)F(x)^{N-1}, \tag{4.27a}$$

and

$$F_{(N)}(x) = F(x)^N \tag{4.27b}$$

where $f(x)$ and $F(x)$ are the pdf and cdf of X_i, respectively. The pdf for the minimum of N independent and non-identically distributed (i.n.d.) random variables X_i, $i = 1,..., N$, is [32]

$$f_{(1)}(x) = \sum_{i=1}^{N} f_i(x) \prod_{\substack{j=1 \\ j \neq i}}^{N} (1 - F_j(x)) \tag{4.28}$$

where $f_i(x)$ and $F_i(x)$ are the pdfs and cdfs of $X_i, i = 1,..., N$, respectively.

Figure 4.4 shows two upper bounds on the maximum average sum rate under the suboptimal scheduling with the maximum spatial multiplexing. The upper bound 1 only uses the upper bounds $\max_{1 \le i \le K-k+1} \chi^2_{2(n_t-k+1)}$ on $\gamma_k^*, k = 1,..., \min(n_t, K)$, and the upper bound 2 is from (4.26). The difference between the two theoretical upper bounds indicates the dependence of the average maximum sum rate on the correlation among $\gamma_k^*, k = 1,..., \min(n_t, K)$. We note that the theoretical bound given in *Theorem 4.4* is very close to the simulated result over the entire region of SNR, even for a relatively small number of users.

4.2.3 Suboptimal PF Scheduling with Equal Power Allocation

The greedy scheduler ignores the fairness of time-slot allocation to users, thus a proportional fair (PF) scheduler, which was introduced in Section 3.3, can be used to achieve

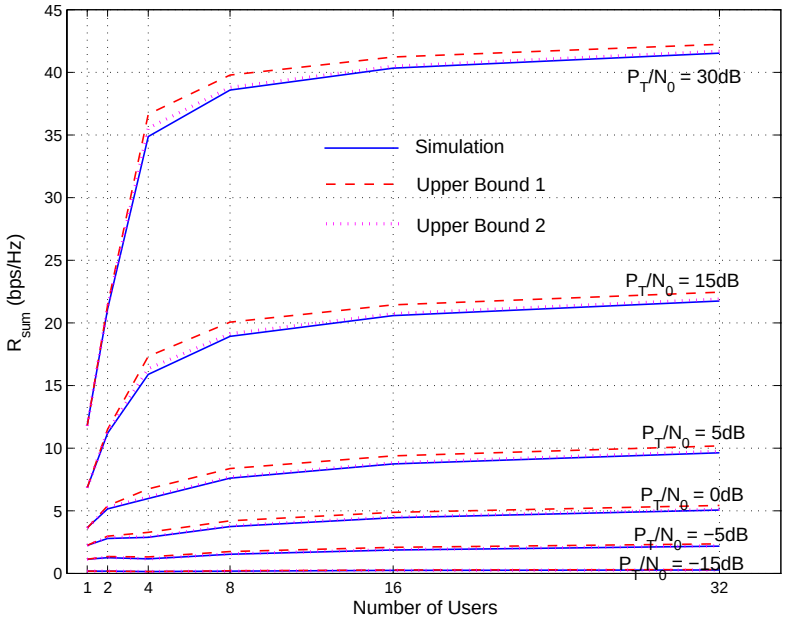


Figure 4.4. Upper bounds on the average maximum sum rate under the suboptimal scheduler in Table 4.2 in i.i.d. flat Rayleigh block-wise fading.

a balanced tradeoff between multiuser diversity and fair allocation of time slots [33]. For V-BLAST transmission, the same PF scheduler in single-antenna system can be used for single-user transmission over each slot. In the general case of ZF-DPC with $m = \sum_{k=1}^{K} m_k$, the PF scheduler must be extended for multiuser transmissions, which assigns the slot t to the antenna subset $\mathcal{S}_{\text{PF}}$ satisfying

$$\mathcal{S}_{\text{PF}} = \arg \max_{\mathcal{S}} \sum_{i \in \mathcal{S}} \frac{R_k^i(t)}{T_k(t)} \tag{4.29}$$

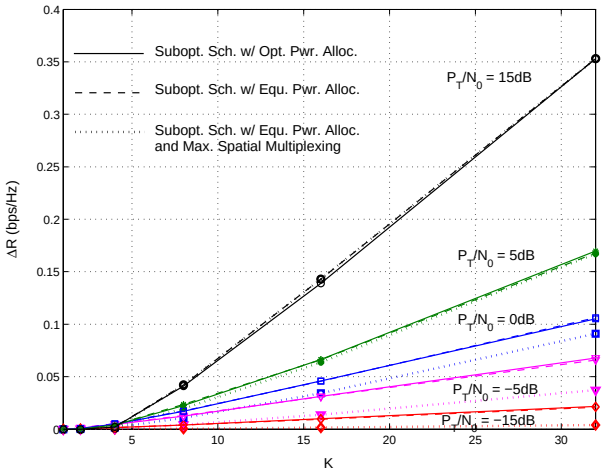
among all receive antenna subsets of $|\mathcal{S}| \le n_t$, where $R_k^i(t)$ is the achievable rate of antenna i of user k at slot t, and $T_k(t)$ is the throughput of user k averaged over a past window of length W_c in slots and is updated slot-wise as

$$T_k(t+1) = \begin{cases} (1-1/W_c)T_k(t) + (1/W_c)\sum_{i \in \text{user } k} R_k^i(t), & i \in \mathcal{S}_{\text{PF}} \text{ and } i \in \text{user } k \\ (1-1/W_c)T_k(t), & \text{otherwise} \end{cases}.$$

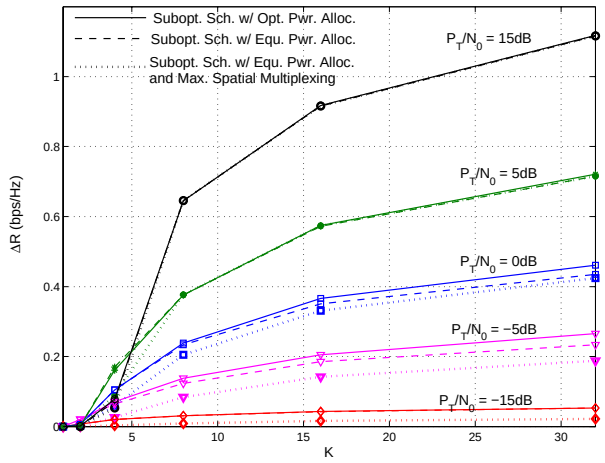
Suboptimal PF schedulers can be used, similar to the suboptimal greedy schedulers in Tables 4.1 and 4.2. It is worth noting that unlike greedy scheduling, PF scheduling for ZF-DPC does depend on the number of receive antennas at each user, since the measure of resource fairness is associated with users, rather than with receive antennas.

Figures 4.5(a) and (b) depict the average maximum sum-rate difference between the suboptimal greedy and PF scheduling, using $n_t = 4$ and $m_k = 1, k = 1, ..., K$. In Figure 4.5, the time scale of interest W_c is set to 100 slots, and the slot width in 1.5-Hz flat Rician fading is set to 16.7ms. Comparing Figures 4.5(a) and (b), we can see that the average maximum sum-rate decrease is more pronounced in a slow fading environment. However, unlike with a single transmit antenna, with maximum spatial multiplexing via multiple transmit antennas, the degradation is constrained to below 1 bps/Hz for a wide range of average SNR and a moderate number of users. Therefore, the spatial multiplexing with ZF-DPC partially provides the benefits of dumb antennas [33], which create relatively faster artificial fading to boost the scheduling gain in an environment with either slow fading or limited fading variations. In i.i.d. flat Rayleigh fading, the degradation grows linearly with the number of users, while in a slow flat Rician fading with strong line-of-sight component, the degradation rate is much slower with the number of users. We also note that for suboptimal scheduling with maximum spatial multiplexing, the degradation in the average maximum sum rate is the smallest for a wide range of average SNR, especially in the low SNR region. This indicates that the scheduler with maximum spatial multiplexing provides more fairness in time-slot allocation than the other two schedulers, and is less sensitive to the fairness constraint.

For a fair comparison between ZF-DPC and V-BLAST, Figure 4.6(a) shows the 1%-



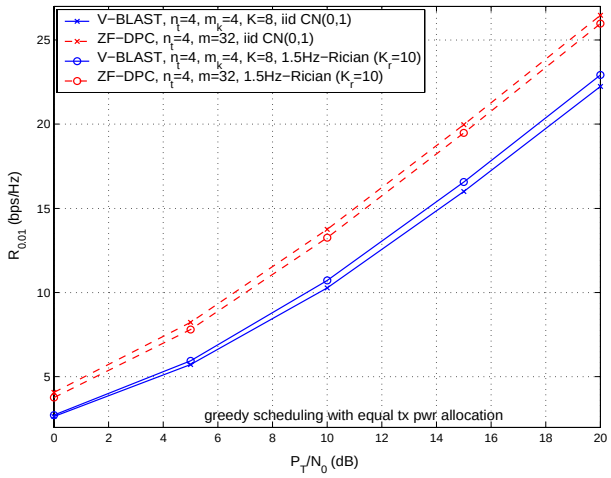
(a) i.i.d. Rayleigh flat fading .



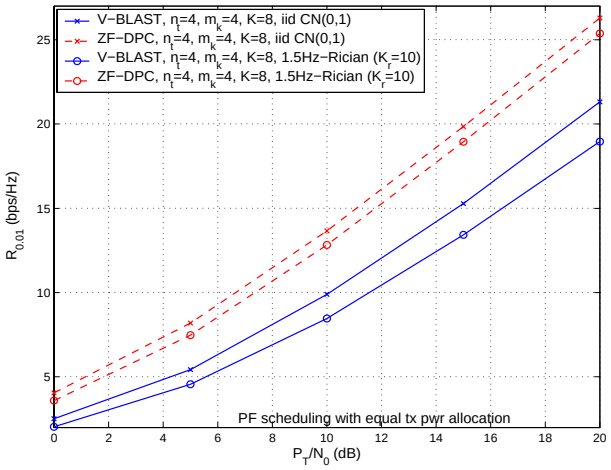
(b) 1.5Hz-Doppler flat Rician fading with a Rice factor of 10.

Figure 4.5. Average maximum sum rate difference between suboptimal greedy and PF scheduling.

outage maximum sum rate with the scheduler in Table 4.2 in both i.i.d. flat Rayleigh and 1.5Hz-Doppler ergodic flat Rician fading (with a Rice factor of 10) using V-BLAST and ZF-DPC. For V-BLAST, we use $n_t = m_k = 4$ for $k = 1,...,K$, and $K = 8$. For ZF-DPC, we use $m = 32$. We see that with greedy scheduling, about 2dB additional scheduling gain is



(a) Suboptimal greedy scheduling



(b) Suboptimal PF scheduling .

Figure 4.6. 1%-outage maximum sum rate in i.i.d. flat Rayleigh fading and flat Rician fading.

possible for ZF-DPC in the scenario simulated. Also for ZF-DPC, the 1%-outage maximum sum rate is higher in Rayleigh fading, while for V-BLAST with $K = 8$, it is slightly higher in Rician fading. However, as K increases and/or SNR decreases, the low-outage maximum sum rate gets higher in Rayleigh fading (not shown in the figure). This is due to the impact of

the line-of-sight signal in Rician fading for a small number of users. Figure 4.6 (b) illustrates the 1%-outage maximum sum rate under the suboptimal PF scheduling for $n_t = m_k = 4$, $k = 1,...,K$, and $K = 8$ for both ZF-DPC and V-BLAST. Again, we assume $W_c = 100$ slots, and for the ergodic Rician fading, we use a slot width of 16.7 ms. Comparing Figures 4.6 (a) and (b), we see that under PF scheduling, there is about 3 to 4 dB scheduling gain improvement with ZF-DPC over V-BLAST, which is greater than that under greedy scheduling. Moreover, the degradation in Rician fading gets more pronounced, especially for V-BLAST, as compared to greedy scheduling.

4.3 Regularized ZF-DPC and Dual Suboptimal Scheduling

In this section, we define a suboptimal dirty paper coding strategy based on the mean square error (MSE) criterion, and name it the regularized zero-forcing dirty paper coding (RZF-DPC). Again, we consider an equal transmit power allocation to scheduled users, for both its simplicity and a relaxed linearity demand on transmit power amplifiers. In the following, we first show that provided a feed-forward of effective channel gains $g_{k,k}$, $k = 1,...,n_s$, in (4,2) to the scheduled users, or equivalently a knowledge of the average transmit SNR at the scheduled users, RZF-DPC provides a more robust precoding solution than ZF-DPC when the user channel matrix is ill-conditioned [46].

Consider the channel model in (2.11) with $m_k = 1$, $k = 1,...,K$, a general form of ZF-DPC using equal power allocation can be expressed as

$$\mathbf{Y} = \mathbf{D} \cdot (\mathbf{H}(\mathcal{S}) \cdot \mathbf{F} \cdot (\mathbf{L} + \mathbf{I}_{n_s})^{-1} \mathbf{X} + \mathbf{W}), \tag{4.30}$$

where $\mathbf{D} \in \mathbb{R}^{n_s \times n_s}$ is a real diagonal matrix whose diagonal entries are sent to the respective

scheduled users. The transmission matrix $\mathbf{F} \in \mathbb{C}^{n_t \times n_s}$ is used to null out the interference from users $j > k$ to user k, and the strictly lower triangular matrix $\mathbf{L} \in \mathbb{C}^{n_s \times n_s}$ cancels the non-causally known interference from users $j < k$ to user k. Given the channel realization, the MSE between the received and the transmitter signals is, therefore,

$$e^2(\mathbf{D},\mathbf{F},\mathbf{L}) = \mathrm{E}[\|\mathbf{Y}\text{-}\mathbf{X}\|_F{}^2] = \mathrm{E}[\|\mathbf{DH}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}\mathbf{X}+\mathbf{DW}\text{-}\mathbf{X}\|_F{}^2], \qquad (4.31)$$

where $\|\mathbf{A}\|_F{}^2 \triangleq \mathrm{tr}(\mathbf{A}\mathbf{A}^H) = \mathrm{tr}(\mathbf{A}^H\mathbf{A})$ is the square of Frobenius norm of matrix $\mathbf{A}$. Rewriting (4.31) for unit-variance noise samples:

$$\begin{aligned} e^2(\mathbf{D},\mathbf{F},\mathbf{L}) &= \mathrm{E}[\|(\mathbf{DH}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1} - \mathbf{I}_{n_s})\mathbf{X} + \mathbf{DW}\|_F{}^2] \\ &= TE_s(\|\mathbf{DH}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1} - \mathbf{I}_{n_s}\|_F{}^2 + \|\mathbf{D}\|_F{}^2 / E_s). \end{aligned} \qquad (4.32)$$

Using equal transmit power allocation, the average transmit symbol energy $E_s = P_T / n_s$. To minimize MSE, we can first minimize $e^2(\mathbf{D},\mathbf{F},\mathbf{L})$ in (4.32) over the set of real diagonal matrices $\mathbf{D} \in \mathbb{R}^{n_s \times n_s}$, and this leads to

$$\begin{aligned} \mathbf{D} &= [\mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}]^H \left[\mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}[\mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}]^H + \frac{n_s}{P_T}\mathbf{I}_{n_s}\right]^{-1} \\ &= \left[[\mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}]^H \mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1} + \frac{n_s}{P_T}\mathbf{I}_{n_s}\right]^{-1} [\mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}]^H, \end{aligned} \qquad (4.33)$$

which approximates a regularized pseudoinverse of $\mathbf{H}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1}$, subject to the real diagonal matrix constraint. Next, we minimize $e^2(\mathbf{D},\mathbf{F},\mathbf{L})$ with respect to the matrices $\mathbf{F}$ and $\mathbf{L}$. We note from (4.32) that given $\mathbf{D}$, the MSE is minimized with

$$\mathbf{DH}(\mathcal{S})\mathbf{F}(\mathbf{L}+\mathbf{I}_{n_s})^{-1} = \mathbf{I}_{n_s}. \qquad (4.34)$$

For the general form of ZF-DPC in (4.30), the minimum MSE (MMSE) can be obtained through a sophisticated tradeoff between the requirements in (4.33) and (4.34).

However, for simplicity, we consider the Cholesky factorization $\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H \triangleq \mathbf{R}_{\text{zf}}{}^H\mathbf{R}_{\text{zf}}$, where $\mathbf{R}_{\text{zf}} \in \mathbb{C}^{n_s \times n_s}$ is an upper triangular matrix with positive real diagonal entries, and let $\mathbf{F}_{\text{zf}} = \mathbf{H}^H\mathbf{R}_{\text{zf}}{}^{-1}$, $\mathbf{D}_{\text{zf}} = \text{diag}^{-1}(\mathbf{R}_{\text{zf}})$, and $\mathbf{L}_{\text{zf}} = \mathbf{D}_{\text{zf}}\mathbf{R}_{\text{zf}}{}^H - \mathbf{I}_{n_s}$, with $\text{diag}^{-1}(\mathbf{A})$ being the inverse of the diagonal matrix formed by the diagonal entries of matrix **A**, then from (4.34), we have the ZF-DPC solution with an MSE of

$$e_{\text{zf}}{}^2(\mathbf{D}_{\text{zf}}, \mathbf{F}_{\text{zf}}, \mathbf{L}_{\text{zf}}) = T \,\|\, \mathbf{D}_{\text{zf}} \,\|_F{}^2 = T\sum_{k=1}^{n_s}(1/g_{k,k})^2, \tag{4.35}$$

where $1/g_{k,k}$ are the diagonal entries of $\mathbf{D}_{\text{zf}}$ and $g_{k,k}$ are as given in Section 4.1.1.

Now consider the Cholesky factorization $\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H + n_s\mathbf{I}_{n_s}/P_T \triangleq \mathbf{R}_{\text{rzf}}{}^H\mathbf{R}_{\text{rzf}}$, which regularizes the ZF-DPC solution by a diagonal matrix $n_s\mathbf{I}_{n_s}/P_T$. This regularization gives a much more reliable result than ZF-DPC when the matrix $\mathbf{H}(\mathcal{S})\mathbf{H}(\mathcal{S})^H$ is ill-conditioned [46] and the estimation of the channel is noisy, analogous to the MMSE-VBLAST implementation [48]. Similar to ZF-DPC, the RZF-DPC solution is obtained by setting $\mathbf{F}_{\text{rzf}} = \mathbf{H}^H\mathbf{R}_{\text{rzf}}{}^{-1}$, $\mathbf{D}_{\text{rzf}} = \text{diag}^{-1}(\mathbf{R}_{\text{rzf}})$, and $\mathbf{L}_{\text{rzf}} = \mathbf{D}_{\text{rzf}}\mathbf{R}_{\text{rzf}}{}^H - \mathbf{I}_{n_s}$. The MSE of RZF-DPC can be calculated from (4.32) directly. Comparing the MSEs for ZF-DPC and RZF-DPC, we note that the MSE of ZF-DPC is independent of the average transmit symbol energy E_s in the absence of scheduling and ordering of users, while the MSE of RZF-DPC is related to E_s. Figure 4.7 illustrates an example of the weighted MSE (by $1/T$) for ZF-DPC and RZF-DPC at P_T/N_0=15 dB and n_t=4 for K=4 users, without scheduling and ordering. The "regulation" effect of RZF-DPC can be easily seen from the plot in terms of reduced peak MSEs over the 200 i.i.d. flat Rayleigh fading blocks.

For the RZF-DPC, suboptimal scheduling can be performed in the dual multiaccess channel, which is described in the following. From [20-22], we know that subject to the same

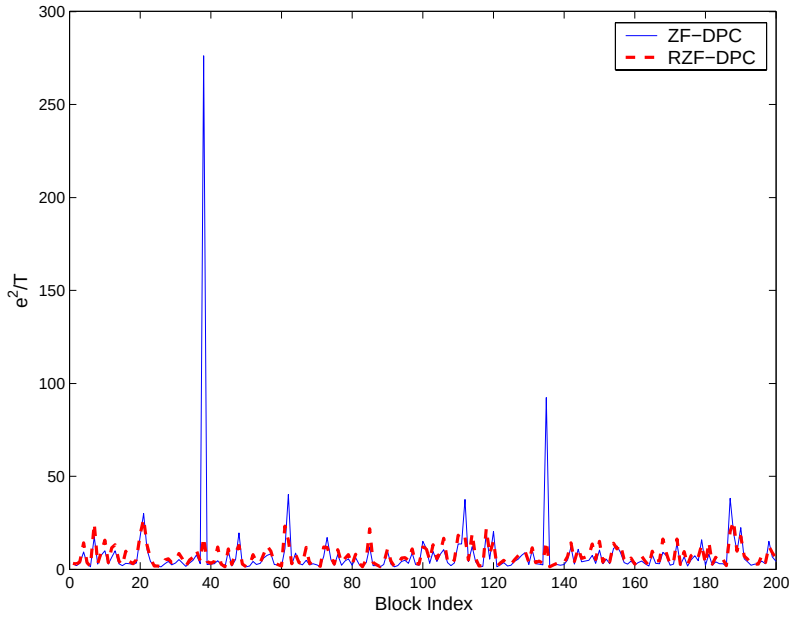


Figure 4.7. Weighted mean square errors of ZF-DPC and RZF-DPC without scheduling and ordering.

total transmit power, the same maximum sum rate is achievable in a broadcast channel and its dual multiaccess channel. Further, in a multiaccess channel with multiple receive antennas, it has been shown in [49] that MMSE successive interference cancellation (SIC) achieves the sum capacity. Therefore, we can build our suboptimal downlink scheduler by scheduling in the dual multiaccess channel. For the broadcast channel model in (2.11) with $m_k = 1$, $k = 1,...,K$, the dual multiaccess channel is defined as [20]

$$\tilde{\mathbf{Y}} = \sum_{k=1}^{K} \mathbf{h}_k{}^H \tilde{\mathbf{x}}_k + \tilde{\mathbf{W}} = [\mathbf{h}_1{}^H \mathbf{h}_2{}^H \cdots \mathbf{h}_K{}^H][\tilde{\mathbf{x}}_1{}^H \tilde{\mathbf{x}}_2{}^H \cdots \tilde{\mathbf{x}}_K{}^H]^H + \tilde{\mathbf{W}} \triangleq \mathbf{H}^H \tilde{\mathbf{X}} + \tilde{\mathbf{W}}, \quad (4.36)$$

with the composite channel matrix $\mathbf{H}^H$ the Hermitian of the dual broadcast channel $\mathbf{H}$ in (2.11), and is subject to the same total average transmit power constraint P_T. In (4.36), matrices $\tilde{\mathbf{x}}_k \in \mathbb{C}^{1\times T}$, $\tilde{\mathbf{Y}} \in \mathbb{C}^{n_t \times T}$ and $\tilde{\mathbf{W}} \in \mathbb{C}^{n_t \times T}$ are the *k*-th user's transmit signal, received signal and complex white Gaussian noise of zero mean and unit variance, respectively, over

each fading block of T symbols, and $\mathbf{h}_k^H \in \mathbb{C}^{n_t \times 1}$ is the channel vector of user *k*, which has zero-mean entries of unit variance. For simplicity, we assume each user is subject to the same transmit power constraint, i.e., the power P_T is equally divided among the scheduled users. In this case, the signal to noise and interference ratio (SINR) at the *k*th user is [49]

$$\gamma_k = (P_T / n_s)\mathbf{h}_k \left(\mathbf{I}_{n_t} + (P_T / n_s) \sum_{j=k+1}^{n_s} \mathbf{h}_j^H \mathbf{h}_j \right)^{-1} \mathbf{h}_k^H$$

$$= \mathbf{h}_k \left((n_s / P_T) \cdot \mathbf{I}_{n_t} + \sum_{j=k+1}^{n_s} \mathbf{h}_j^H \mathbf{h}_j \right)^{-1} \mathbf{h}_k^H \qquad (4.37)$$

where n_s is the number of scheduled users, which satisfies $n_s \le n_t$. The SINR in (4.37) assumes a decoding order of $1, 2, ..., n_s$. To this end, we can schedule the users in the order $n_s, ..., 2, 1$ using (4.37), such that the maximum sum rate in the dual multiaccess channel is maximized with the equal user power constraint and the fixed total power. At step 1, we select the n_s-th user which has the maximum

$$P_T \mathbf{h}_k \mathbf{h}_k^H,$$

which is the single-user SINR with a power constraint P_T, and at step *n*, $n = 2, ..., n_s$, the $(n_s - n + 1)$-th user is chosen for a maximum

$$\prod_{k=n_s-n+1}^{n_s} \left(1 + \mathbf{h}_k \left((n / P_T) \cdot \mathbf{I}_{n_t} + \sum_{j=k+1}^{n_s} \mathbf{h}_j^H \mathbf{h}_j \right)^{-1} \mathbf{h}_k^H \right),$$

which is equivalent to maximizing the sum rate with total power constraint P_T equally shared by the scheduled users. It should be noted at this point that the equal user power constraint in a multiaccess channel does not necessarily mean equal user power allocation in the dual broadcast channel. However, we use both in the dual channels to simplify the scheduling and

transmission of RZF-DPC, and it turns out that the performance degradation in term of sum rate is small, especially with reasonably high average transmit SNR. Table 4.3 gives the suboptimal greedy scheduling algorithm for RZF-DPC using the dual multiaccess channel with equal user power constraint.

Table 4.3. Suboptimal greedy scheduler for RZF-DPC with equal user power constraints in dual multiaccess channel.

Step 1: Initialization

$\mathcal{S}^* = \varnothing$, $T^* = 1$

Step 2:

For $n = \min(n_t, K)$ to 1

$n_s = \min(n_t, K) - n + 1$

For $m = \min(n_t, K)$ to $(n+1)$

$$\beta_{m^*} = \mathbf{h}_{m^*} \left((n_s / P_T)\mathbf{I}_{n_t} + \sum_{j=(m+1)}^{\min(n_t,K)} \mathbf{h}_{j^*}{}^H \mathbf{h}_{j^*} \right)^{-1} \mathbf{h}_{m^*}{}^H,\ m^* \in \mathcal{S}^*$$

End

$$\beta_k = \mathbf{h}_k \left((n_s / P_T)\mathbf{I}_{n_t} + \sum_{j=n+1}^{\min(n_t,K)} \mathbf{h}_{j^*}{}^H \mathbf{h}_{j^*} \right)^{-1} \mathbf{h}_k{}^H,\ k \in \{1,...,K\} \setminus \mathcal{S}^*$$

$$n^* = \arg \max_{k \in \{1,...,K\} \setminus \mathcal{S}^*} (1+\beta_k) \prod_{i=n+1}^{\min(n_t,K)} (1+\beta_{i^*})$$

$$T = \prod_{i=n}^{\min(n_t,K)} (1+\beta_{i^*})$$

if $T > T^*$

$T^* = T$, $\mathcal{S}^* = \mathcal{S}^* \bigcup \{l\}$

else

$n_s = n_s - 1$

go to Step 3

End

Step 3: Deliver n_s and $\widehat{R}_{\text{sum}}^{\text{mmsesic}} = \log_2 T^*$.

The matrix inverse in (4.37) can be calculated using the Sherman-Morrison-

Woodbury formula [46], such that at each step

$$\left((n_s / P_T) \cdot \mathbf{I}_{n_t} + \sum_{j=n+1}^{\min(n_t,K)} \mathbf{h}_{j^*}{}^H \mathbf{h}_{j^*} \right)^{-1} = \left((n_s / P_T) \cdot \mathbf{I}_{n_t} + \sum_{j=n+2}^{\min(n_t,K)} \mathbf{h}_{j^*}{}^H \mathbf{h}_{j^*} + \mathbf{h}_{(n+1)^*}{}^H \mathbf{h}_{(n+1)^*} \right)^{-1}$$

is updated according to

$$\left(\mathbf{A} + \mathbf{v}\mathbf{v}^H \right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^H\mathbf{A}^{-1}}{1 + \mathbf{v}^H\mathbf{A}^{-1}\mathbf{v}}, \quad \text{for} \quad \mathbf{v}^H\mathbf{A}^{-1}\mathbf{v} \neq -1,$$

where $\mathbf{A} = (n_s / P_T) \cdot \mathbf{I}_{n_t} + \sum_{j=n+2}^{\min(n_t,K)} \mathbf{h}_{j^*}{}^H \mathbf{h}_{j^*}$ and $\mathbf{v} = \mathbf{h}_{(n+1)^*}{}^H$.

4.4 Scheduling and Reordering

The duality between ZF-DPC and V-BLAST allows us to apply the ordering criterion of V-BLAST to minimize the total error rate given the selected users from ZF-DPC scheduling. We indicate that this reordering is generally *not* optimal in terms of sum rate. However, the loss in sum rate is small as we will show.

For the channel model in (2.11) with $m_k = 1$, $k = 1, ..., K$, let $\mathrm{ch}(\mathbf{X})$ be the lower triangular matrix from the Cholesky factorization of matrix $\mathbf{X}$, and all the other notations are as defined before. Table 4.4 gives an efficient suboptimal greedy scheduling algorithm using equal power allocation and a minimum-error reordering for ZF-DPC, where the scheduling step is the same as that in Table 4.1. After scheduling, a separate reordering is used to minimize the total average error rate. For the RZF-DPC, the scheduling in Table 4.3 is used, then for the reordering, the pseudoinverse matrix $\mathbf{G}$ in Step 2 is replaced with

$$\tilde{\mathbf{G}} = \mathbf{H}_{\mathrm{o}}{}^H \left(\mathbf{H}_{\mathrm{o}}\mathbf{H}_{\mathrm{o}}{}^H + \frac{|\mathcal{S}^*|}{P_T}\mathbf{I}_{|\mathcal{S}^*|} \right)^{-1}, \tag{4.38}$$

Table 4.4. Suboptimal greedy scheduler for ZF-DPC with minimum-error reordering.

Step 1: Scheduling

Initialize $\mathcal{S}^* = \varnothing$ and $T^* = 1$;

For $n = 1$ to $\min(n_t, K)$

$l = \arg \max_{k \in \{1,\ldots,K\} \setminus \mathcal{S}^*} \gamma_k$, where

$\gamma_k = \left[\mathrm{ch}\left(\mathbf{H}(\mathcal{S}^* \cup \{k\}) \mathbf{H}^H (\mathcal{S}^* \cup \{k\}) \right) \right]_{nn}^2$;

$T = (1 + P_T \gamma_l / n) \prod_{m=1}^{n-1} (1 + P_T \gamma_m^* / n)$;

if $T > T^*$

$T^* = T$; $\gamma_n^* = \gamma_l$; $\mathcal{S}^* = \mathcal{S}^* \cup \{l\}$;

else

go to Step 2;

End

Step 2: Reordering

Initialize $\mathbf{H}_{\mathrm{o}} = \mathbf{H}(\mathcal{S}^*)$, $T_{\mathrm{o}}^* = 1$ and $\mathcal{J} = \varnothing$;

For $n = 1$ to $|\mathcal{S}^*|$

$\mathbf{G} = \mathbf{H}_{\mathrm{o}}^+$;

$l = \arg \min_{j \in \{1,\ldots,|\mathcal{S}^*|\} \setminus \mathcal{J}} \left\| [\mathbf{G}]_j \right\|_2^2$;

$T_{\mathrm{o}}^* = T_{\mathrm{o}}^* \cdot \left(1 + P_T / (|\mathcal{S}^*| \cdot \left\| [\mathbf{G}]_l \right\|_2^2) \right)$;

$\mathcal{J} = \mathcal{J} \cup \{l\}$;

Set the lth row of $\mathbf{H}_{\mathrm{o}}$ to zero;

End

Step 3: Reverse the element order of $\mathcal{J}$ and reorder the elements of $\mathcal{S}^*$ according to $\mathcal{J}$;

Step 4: Deliver reordered $\mathcal{S}^*$ and $\hat{R}_{\mathrm{sum}}^{\max} = \log_2 T_{\mathrm{o}}^*$.

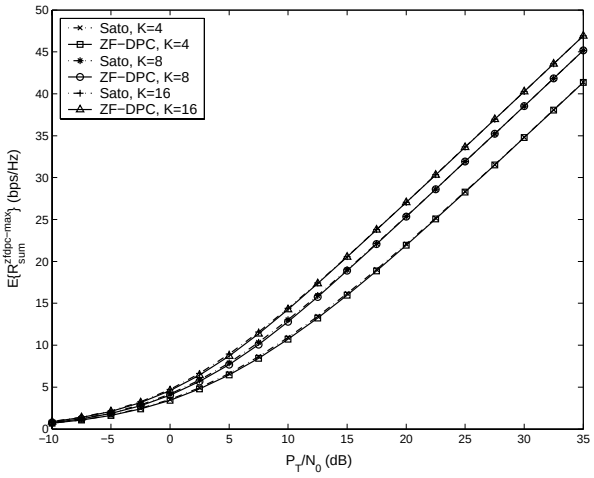
where $\mathbf{I}_{|\mathcal{S}^*|}$ is the identity matrix of size $|\mathcal{S}^*|$.

We note that the suboptimal greedy schedulers in Tables 4.1 and 4.3 both adaptively adjust the number of admitted users while using equal power allocation. However, for the scheduling with reordering, after the best user subset is chosen, users are reordered according to an ordering criterion similar to that for V-BLAST to minimize the error rate. Similarly, the reordering criterion can be based on either zero-forcing or MMSE interference nulling. Nevertheless, unlike V-BLAST where the best stream (the one with the highest post-detection SNR) is decoded first, the precoder encodes the best stream last. Therefore, the minimum-error ordering for one is opposite the other given that their channel realizations are dual to each other [20], [22]. Furthermore, the suboptimal PF scheduling can be used in Table 4.4 for resource fairness considerations. In either case with $K \geq n_t$, the total number of subset evaluations in scheduling and reordering is upper bounded by

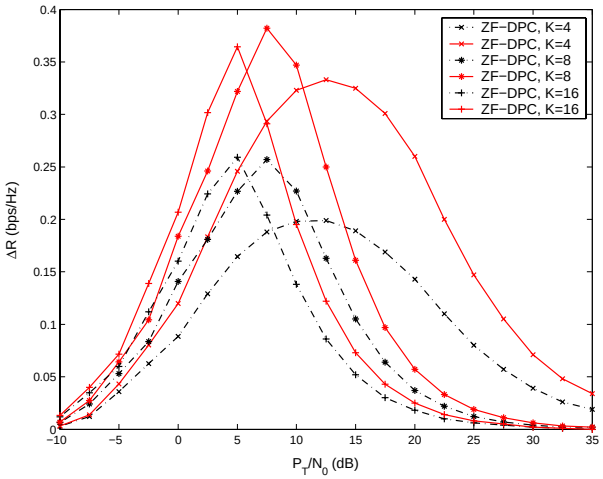
$$\underbrace{[K + (K-1) + ... + (K - n_t + 1)]}_{\text{scheduling}} + \underbrace{[n_t + (n_t - 1) + ... + 1]}_{\text{reordering}} = (K+1) \cdot n_t, \qquad (4.31)$$

which is still linear in the number of users for a finite number of transmit antennas. It is worth noting that the minimum-error reordering usually results in a sum-rate loss as compared to the ordering dictated by the suboptimal scheduler. However, as corroborated by the simulation in the following, the sum-rate loss is insignificant.

We define the sum-rate loss here as the sum-rate loss from that of Sato upper bound [26], which is achieved with perfect dirty paper coding in the MIMO broadcast channel. Figure 4.8(a) depicts the average sum rate for both Sato upper bound and ZF-DPC using equal transmit power and scheduling in Table 4.1 over 10,000 channel realizations for $n_t = 4$. The Sato upper bound is computed using the iteration waterfilling algorithm proposed in [50]. The average sum-rate loss (ΔR) vs. the total transmit SNR ($P_T / N_0 = P_T$) is plotted in Figure 4.8(b) for both scheduling only and scheduling with reordering. The total transmit SNR,



(a) Average Sum Rate.



(b) Average Sum-Rate Loss. Dashed curves: scheduling only;
Solid curves: scheduling and reordering.

Figure 4.8. Average sum rate and sum-rate Loss.

rather than the ratio of bit energy to noise power spectral density (E_b / N_0), is used here due to the infeasibility for an explicit expression of E_b / N_0 when using an adaptive number of users in this case [51]. We note that for adaptive user scheduling with equal power allocation

and no reordering, the average sum-rate loss is very small. When the reordering is imposed, the loss is larger but still less than 0.4bps/Hz over a wide range of SNR. Similar results have been reported previously that the suboptimal scheduling using optimal power allocation suffers insignificant sum-rate loss from the Sato upper bound [44]. Therefore, Figures 4.8(a) and (b) expand previous results by indicating that the sum-rate loss from the Sato bound by using equal power allocation, an adaptive number of users, and minimum-error-rate reordering is also insignificant.

4.5 Summary

Following Costa's work in dirty paper coding, this chapter showed performance of some suboptimal DPC techniques for MIMO systems using various suboptimal scheduling algorithms. We first reviewed a suboptimal DPC strategy, named ZF-DPC, which has been shown to achieve asymptotic sum-rate optimality at both high and low SNR through semi-analytic simulation. The significance of pre-interference cancellation for the spatial multiplexing gain at high SNR and the array gain (via coherent MRC combining) at low SNR, and that of channel-aware scheduling for the multiuser diversity were highlighted from a network point of view. The performance of a suboptimal greedy scheduling algorithm for ZF-DPC was provided, and its performance advantage over the V-BLAST scheme was shown in a symmetric fading environment. A regularized ZF-DPC was proposed using the duality with the MMSE SIC in the dual uplink channel, which also admits a simplified suboptimal scheduling in the dual channel space. Finally, a reordering on top of the suboptimal scheduling was shown to keep minimal the average error rate, which only incurs an insignificant sum-rate loss from the Sato upper bound.

Chapter 5
Tomlinson-Harashima Precoding and Scheduling

This chapter introduces Tomlinson-Harashima precoding (THP) in the context of the multi-antenna Gaussian broadcast channel. We start with a spatial THP implementation proposed previously, and present original performance analysis of THP in a broadcast channel in flat Rayleigh fading. Further, the performance results of THP with scheduling are presented. The aim is to validate the suboptimal scheduling and reordering algorithms in Chapter 4 with a practical precoding solution.

5.1 Spatial Tomlinson-Harashima (T-H) Precoder

The problem with ZF-DPC described in last chapter is that a possibly very high transmit power is needed to null out the un-cancelled user interference, particularly when the channel matrix is ill-conditioned. For this reason, a universal coding scheme [30] was proposed to use a modulo-lattice precoding at the transmitter. This precoding technique built on inflated lattices brings us closer to a real system implementation. In fact, a suboptimal one-dimensional implementation of modulo-lattice precoding named Tomlinson-Harashima precoding (THP) was proposed in [54-56] for intersymbol interference (ISI) cancellation about thirty years ago, and it was adopted for DSL systems around twenty years later. Recently, THP was applied to multiuser interference cancellation in the context of DSL [39]. The application is directly connected to ZF-DPC by using a similar QR-type decomposition, and is based on the realization that multiuser interference cancellation is similar to ISI

cancellation. In flat fading channels, the application of THP to a MISO (multiple input and single output) broadcast channel was shown in [53] as a direct extension of the V-BLAST algorithm. The implementation of THP shifts the joint processing to the transmitter and adds the following benefits to the system. First, it can provide high spectrally efficient transmission free of error propagation, which is inherent with a decision feedback equalizer (DFE) structure. Secondly, unlike DFE, THP can be combined nicely with coded modulation and the shaping techniques used in DSL applications over a low-dimensional lattice. Finally, the transmitter-centric processing greatly reduces the receiver complexity and power consumption, which is in particular attractive for downlink transmission.

5.1.1 Spatial Implementation

Figure 5.1 shows a spatial zero-forcing T-H precoding (ZF-THP) system equivalent to the model in [53]. The model actually forms a distributed MIMO system with an n_t-element antenna array at the transmitter, and K receive antennas distributed across K users with each user having a single-element antenna. At each time slot, the base station sends independent and synchronized data packets to n_s users simultaneously, with $n_s \le n_t$ constrained by the n_t degrees of freedom at transmitter. Let $\mathbf{H}(\mathcal{S})$ be the $n_s \times n_t$ channel matrix between the transmit antenna array and the selected n_s receive antennas defined by set $\mathcal{S}$, with its entries being i.i.d. fading coefficients of unit variance. The Hermitian matrix $\mathbf{A} = \mathbf{H}(\mathcal{S}) \cdot \mathbf{H}(\mathcal{S})^H$ is almost surely positive definite, and has a unique Cholesky factorization as $\mathbf{A} = \mathbf{L} \cdot \mathbf{L}^H$, where the $n_s \times n_s$ lower triangular matrix $\mathbf{L}$ has real positive diagonal entries. The superscript ^{H} denotes conjugate transposition. The $n_t \times n_s$ transmit matrix $\mathbf{F} = (\mathbf{L}^{-1} \cdot \mathbf{H}(\mathcal{S}))^H$ has orthonormal columns. The matrix $\mathbf{L}$ is further decomposed into

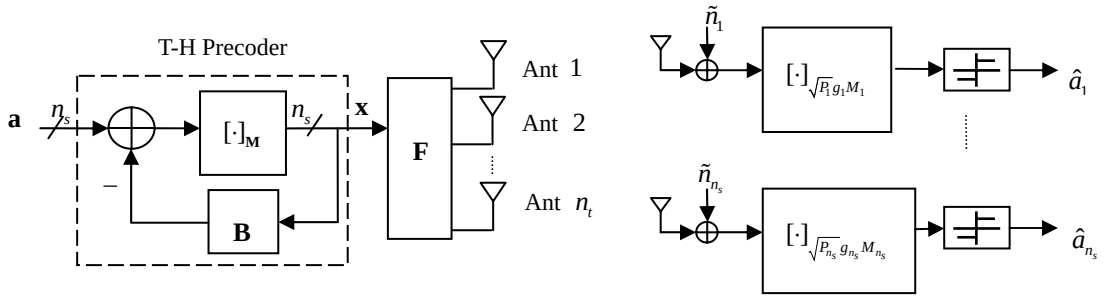


Figure 5.1. The block diagram of a downlink system with spatial zero-forcing T-H precoding (ZF-THP).

$\mathbf{L} = \mathbf{G} \cdot (\mathbf{B} + \mathbf{I}_{n_s})$, where $\mathbf{G}$ is a diagonal matrix with the main diagonal entries g_i of $\mathbf{L}$, $i = 1, ..., n_s$, the matrix $\mathbf{B}$ is lower triangular with zero main diagonal entries, and $\mathbf{I}_{n_s}$ is the identity matrix of size n_s. At the selected receiver inputs, the spatially and temporally white complex Gaussian noises of zero mean and unit variance, $\tilde{n}_i$, $i = 1, ..., n_s$, are added to the signals as shown in Figure 5.1.

The zero-forcing T-H precoder in Figure 5.1 uses a symmetric vector modulo operator $[\cdot]_{\mathbf{M}}$ for data streams, with the modulus vector $\mathbf{M} = [M_1, ..., M_{n_s}]^T$ being set equal to the positive Voronoi boundary values [54] of input signal constellations of unit average symbol energy, possibly adaptive over each slot. For simplicity, we ignore the modulo loss [54] and assume that over each slot, the total transmit power P_T is equally split over the scheduled users with $P_i = P_T / n_s$, $i = 1, ..., n_s$. At the scheduled receivers, the moduli M_i, $i = 1, ..., n_s$, of the symmetric modulo operators are weighted by $\sqrt{P_T / n_s} \cdot g_i$, $i = 1, ..., n_s$, which are assumed perfectly known from the transmitter. Each receiver simply performs a maximum-likelihood detection at its modulo output. In the following analysis, we first set $n_s = n_t$ for maximum spatial multiplexing gain with ZF-THP. For channel matrix $\mathbf{H}$ having i.i.d. circular symmetric complex Gaussian entries of zero mean and unit variance, Bartlett

decomposition [55] indicates that the elements g_i^2 are central chi-square random variables with $2(n_t - i + 1),\ i = 1,...,n_t$, degrees of freedom.

5.1.2 Achievable Rate

It can be shown that through the ZF-THP above, the downlink channels are decoupled into n_t independently faded channels with SNRs $\rho_i = P_T g_i^2 / n_t,\ i = 1,...,n_t$, respectively, and the decoupled signals at receive modulo operator outputs are

$$\tilde{r}_i = \left\lfloor \sqrt{P_T / n_t}\, g_i \cdot a_i + \tilde{n}_i \right\rceil_{\sqrt{P_T / n_t}\cdot g_i M_i} = [a_i + \tilde{n}_i / (\sqrt{P_T / n_t}\, g_i)]_{M_i},\ i = 1,...,n_t, \qquad (5.1)$$

Following the analysis in [56], it is not hard to derive the achievable rates for the user channels in (5.1) as

$$R_i^{\text{zfthp}}(P_T / n_t) = 2\log_2(2M_i) - h\left([\tilde{n}_i / (\sqrt{P_T / n_t}\, g_i)]_{M_i}\right),\ i = 1,...,n_t, \qquad (5.2)$$

in bps/Hz, where $h(\cdot)$ is the differential entropy function [19]. For uniformly distributed signals over a square Voronoi region $(-M_i, M_i] \times (-M_i, M_i]$ of a unit second-order moment, we have $M_i = \sqrt{3/2},\ i = 1,...,n_t$. Therefore when ignoring modulo loss and using equal transmit power allocation, the achievable rates of ZF-THP become

$$R_i^{\text{zfthp}}(P_T / n_t) = \log_2 6 - h\left([\tilde{n}_i / (\sqrt{P_T / n_t}\, g_i)]_{\sqrt{3/2}}\right),\ i = 1,...,n_t, \qquad (5.3)$$

where $\tilde{n}_i / (\sqrt{P_T / n_t}\, g_i),\ i = 1,...,n_t$, are i.i.d. zero-mean and circular symmetric complex Gaussian noise with variance $n_t / \left(P_T g_i^2\right)$, the reciprocal of the SNRs at the receive modulo inputs. It can be easily shown that at high SNR, (5.3) incurs a rate loss of 0.509 bits per complex symbol from that of an AWGN channel, whereas at low SNR, the loss is larger. Previous results on DPC indicate that with the knowledge of input SNR, THP can do better at low SNR with a noise cooling factor $\alpha = (P_T / n_t)/(1 + P_T / n_t)$ [4], [5]. In Figure 5.1, this

corresponds to replacing the matrix $\mathbf{B}$ with matrix $\tilde{\mathbf{B}} = \left((P_T / n_s)/(1+P_T / n_s)\right)\cdot \mathbf{B}$, and multiplying each receive modulo input with $(P_T/n_s)/(1+P_T/n_s)$. Correspondingly, the achievable data rates with noise cooling are

$$\tilde{R}_i^{\text{zfthp}}(P_T / n_t) = \log_2 6 - 2h\left([\tilde{\alpha}_i \cdot n/(\sqrt{P_T / n_t}\, g_i) + (1-\tilde{\alpha}_i)\cdot x]_{\sqrt{3/2}}\right),\ i = 1,...,n_t, \quad (5.4)$$

in bps/Hz, where $\tilde{\alpha}_i = \rho_i/(1+\rho_i),\ i = 1,...,n_t$, the variable n is a zero-mean real Gaussian random variable of variance 0.5, and x is a real random variable uniformly distributed over the interval $(-\sqrt{3/2}, \sqrt{3/2}]$. Numerical integration can be used to evaluate the differential entropies of modulo random variables in (5.3) and (5.4). Figure 5.2 plots the average user achievable rates in (5.3) and (5.4). As we can see in the figure, at high SNR, THP has an SNR loss of about 1.53 dB, and noise cooling increases the rate at low SNR.

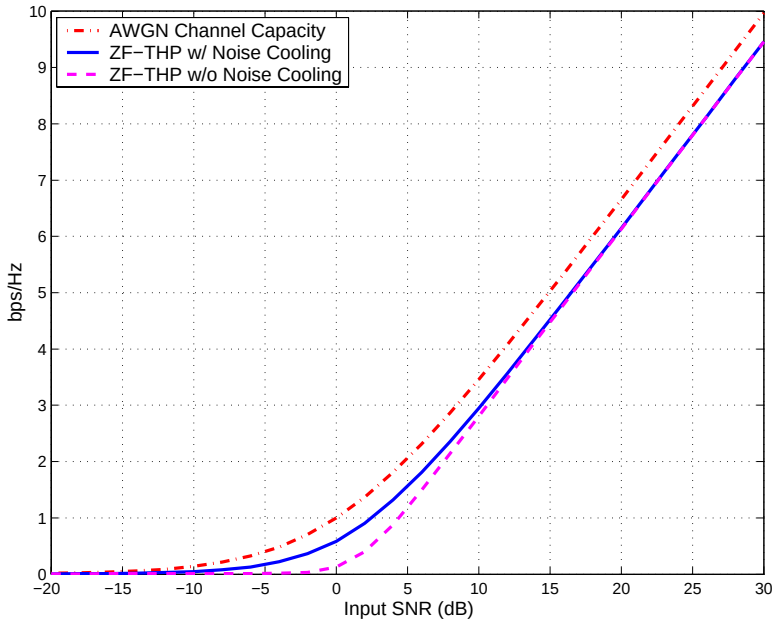


Figure 5.2. User achievable rate of THP.

With CSI at the transmitter and noise cooling, the achievable sum rate of THP is

$$\tilde{R}_{\text{sum}}^{\text{zfthp}} = \sum_{i=1}^{n_t} \tilde{R}_i^{\text{zfthp}}(P_T / n_t) = \sum_{i=1}^{n_t} \left[\log_2 6 - 2h\left([\tilde{\alpha}_i \cdot n / (\sqrt{P_T / n_t}\, g_i) + (1-\tilde{\alpha}_i) \cdot x]_{\sqrt{3/2}} \right) \right], \quad (5.5)$$

for equal power allocation across users. For a network with a large number of users so that $K>n_t$, the maximum sum rate can be achieved through multiuser selection as

$$\tilde{R}_{\text{sum}}^{\text{zfthp-max}} = \max_{\mathcal{S}} \tilde{R}_{\text{sum}}^{\text{zfthp}} \quad (5.6)$$

over all *ordered* user subsets $\mathcal{S}$ with cardinality $|\mathcal{S}| \le n_t$ to obtain multiuser diversity gain. For comparison, we also give the maximum achievable sum rates of zero-forcing DPC (ZF-DPC) and zero-forcing transmit linear beamforming (ZF-LBF) as used in Chapter 4. For ZF-DPC with optimal transmit power allocation,

$$R_{\text{sum}}^{\text{zfdpc-max}} = \max_{\mathcal{S}} \sum_{i=1}^{n_t} \left[\log_2 (\xi \cdot \rho_i^{\text{zfdpc}}) \right]_+ , \quad (5.7)$$

where $\rho_i^{\text{zfdpc}} = g_i^2$, $i = 1, ..., n_t$, are the channel gains of scheduled users, and ξ is the waterfilling solution to

$$\sum_{i=1}^{n_t} \left[\xi - 1/\rho_i^{\text{zfdpc}} \right]_+ = P_T ,$$

over each slot, and the operator $[x]_+$ is defined as $[x]_+ = \max(x, 0)$. The maximization in (5.7) is over all *ordered* user subsets $\mathcal{S}$ with cardinality $|\mathcal{S}| = n_t$. For ZF-LBF with equal power allocation,

$$R_{\text{sum}}^{\text{zflbf-max}} = \max_{\mathcal{S}} \sum_{i=1}^{|\mathcal{S}|} \log_2 (1 + \rho_i^{\text{zflbf}}) , \quad (5.8)$$

where $\rho_i^{\text{zflbf}} = (P_T / n_s) / \left[\left(\mathbf{H}(\mathcal{S}) \mathbf{H}(\mathcal{S})^H \right)^{-1} \right]_{i,i}$, $i = 1, ..., n_s$, are the received SNRs of scheduled users, and the maximization is over all *unordered* user subsets $\mathcal{S}$ with cardinality $|\mathcal{S}| \le n_t$. For $K \ge n_t$, the total number of unordered subsets is $\sum_{i=1}^{n_s} \mathrm{C}_K^i$, with C_m^n being the number of n

combinations out of *m*. When the channel matrix $\mathbf{H}$ has i.i.d. circular symmetric complex Gaussian entries of zero mean and unit variance, the SNRs ρ_i^{zflbf} are weighted central chi-square random variables with $2(n_t - |\mathcal{S}| + 1)$ degrees of freedom [57].

5.2 Multiuser Diversity with Spatial THP

We study the impact of channel-aware scheduling on the performance of THP without reordering. We first consider the i.i.d. symmetric user channels, and then evaluate the system performance in correlated channels.

5.2.1 Independent User Chanenls

We assume that user subset selection is performed by a packet scheduler at the transmitter. We call the scheduler which maximizes the sum rate a greedy scheduler whose objective is defined as in (5.6), (5.7) or (5.8) for different transmission schemes. The greedy scheduler ignores fairness of time-slot allocation to users, thus a proportional fair (PF) scheduler was considered for a balanced tradeoff between multiuser diversity and fair allocation of time slots [15]. The PF scheduler can be extended for multiuser transmission, which assigns the slot t to the user subset $\mathcal{S}^*$ satisfying

$$\mathcal{S}^* = \arg \max_{\mathcal{S}} \sum_{i=1}^{n_s} \frac{R_i(t)}{T_i(t)} \tag{5.9}$$

among all active users (users requesting for services simultaneously), where $R_i(t)$ is the achievable rate of user i over slot t, and $T_i(t)$ is its average throughput in a past window of length T_c and is updated slot-wise as

$$T_i(t+1) = \begin{cases} (1-1/T_c)T_i(t) + R_i(t)/T_c, & i \in \mathcal{S}^* \\ (1-1/T_c)T_i(t), & \text{otherwise} \end{cases}.$$

As stated in Chapter 4, for ZF-THP and ZF-DPC with $K \geq n_t$, a direct maximization of the sum rate requires a search over $\mathrm{P}_K^{n_t}$ total number of subsets for scheduling, i.e., the number of n_t permutations out of K. This is a large number even for moderate values of n_t and K. Fortunately, we know from [46] that the computation of g_i only depends on users j, $j \leq i$, henceforth we can use the suboptimal scheduler as proposed in Table 4.2 with maximum spatial multiplexing. At step k, k=1,…,n_t, we select the user k^* subject to

$$k^* = \arg \max_{k \in \{1^*,\ldots,(k-1)^*\}^c} \left[\sum_{i=1^*}^{(k-1)^*} \mu_i R_i(t) + \mu_k R_k(t) \right], \tag{5.10}$$

where $\{\mathcal{A}\}^c$ denotes the complement of set $\mathcal{A}$, and μ_i are 1 and $1/T_i(t)$ for the greedy and the PF scheduler, respectively. This reduces the total number of subsets over which we must search to

$$N_1 = \sum_{i=0}^{n_t-1} \mathrm{C}_{K-i}^1 = n_t(2K - n_t + 1)/2. \tag{5.11}$$

For K>>n_t, N_1 is $O(K)$. This is analogous to the Viterbi algorithm, since the suboptimal scheduler approximates the global optimum through a local optimization. This principle of optimality was also noted in the V-BLAST algorithm [5], which was proved equivalent to a generalized decision feedback equalizer [58]. Figure 5.3 plots the average maximum sum rate in bps/Hz for ZF-DPC with optimal power allocation, using both optimal (5.7) and suboptimal (5.10) schedulers, for n_t=4 and K=16 in i.i.d. flat Rayleigh fading of 5.6 Hz Doppler over 2000 time slots of 16.7 ms each. For the PF scheduler, we set T_c=1.67 s. For comparison, Figure 5.3 also shows the average total throughput (same as the average sum

rate here) with a round-robin (RR) scheduler which assigns time slots to every subset of n_t users (for a total of $C_K^{n_t}$ subsets) alternately. It can be seen that the suboptimal scheduler for ZF-THP approximates the optimum very closely over a fairly large input SNR range.

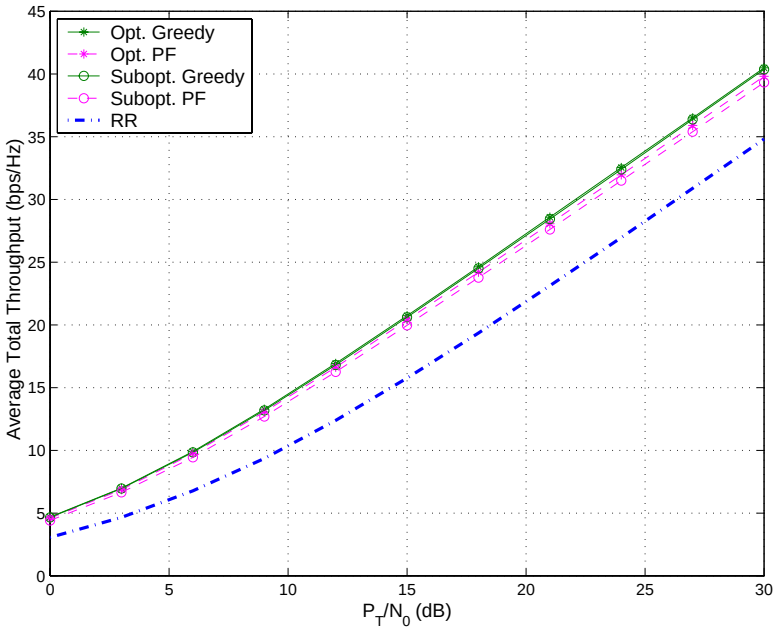
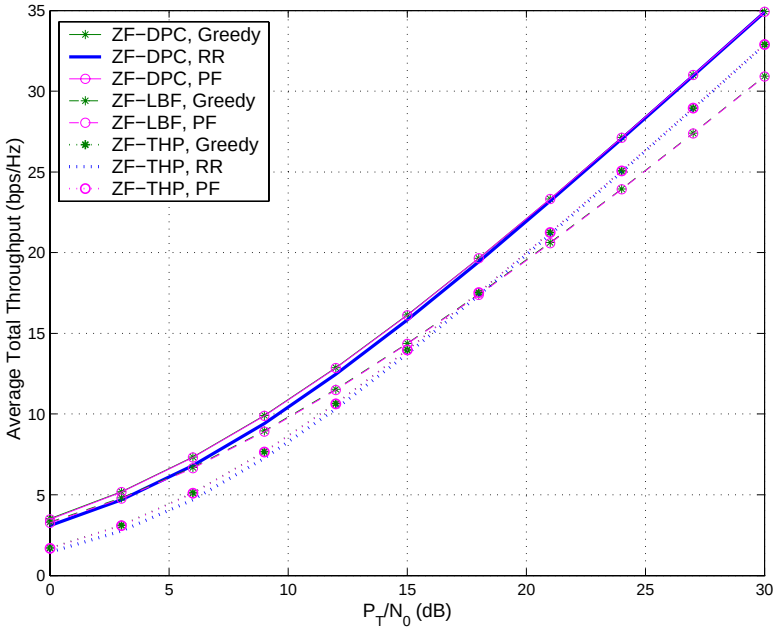
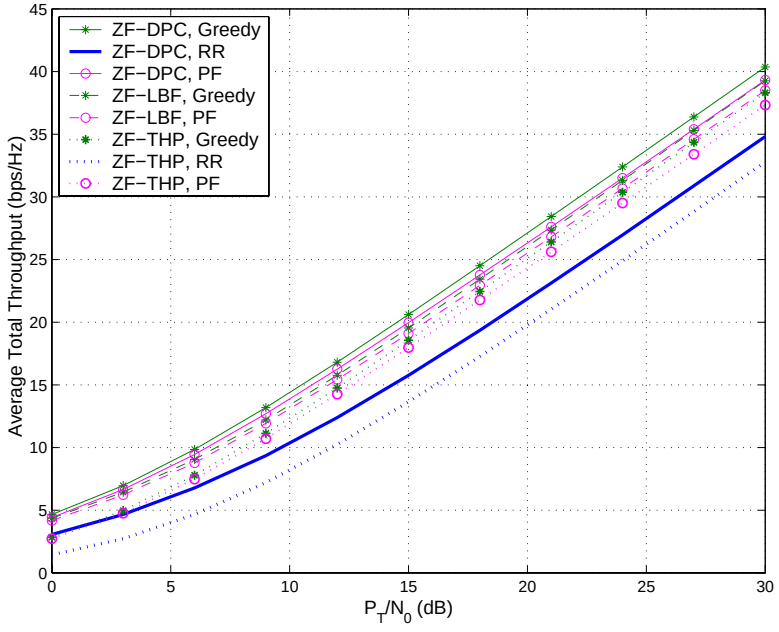


Figure 5.3. Average maximum sum rate of ZF-DPC for n_t=4 and K=16 under different packet scheduling.

Figures 5.4 and 5.5 depict the average maximum total throughputs of ZF-DPC, ZF-THP, and ZF-LBF vs. P_T / N_0 for $n_t = 4$ and $K = 4$ and 16, where N_0 is the complex noise sample variance normalized to 1. Symmetric i.i.d. flat fading channels are assumed for the users under scheduling. Two flat fading scenarios are considered: one is Rayleigh with 5.6 Hz Doppler as used in Figure 5.3; the other is Rician with 1.5 Hz Doppler and a 10 dB Rice factor. The time slot width for both cases is set to 16.7 ms. The ZF-DPC with optimal power allocation is used as an upper bound to both ZF-THP and ZF-LBF with equal power allocations. Throughput with RR scheduling provide baselines for comparisons in these



(a) $K = 4$.



(b) $K = 16$.

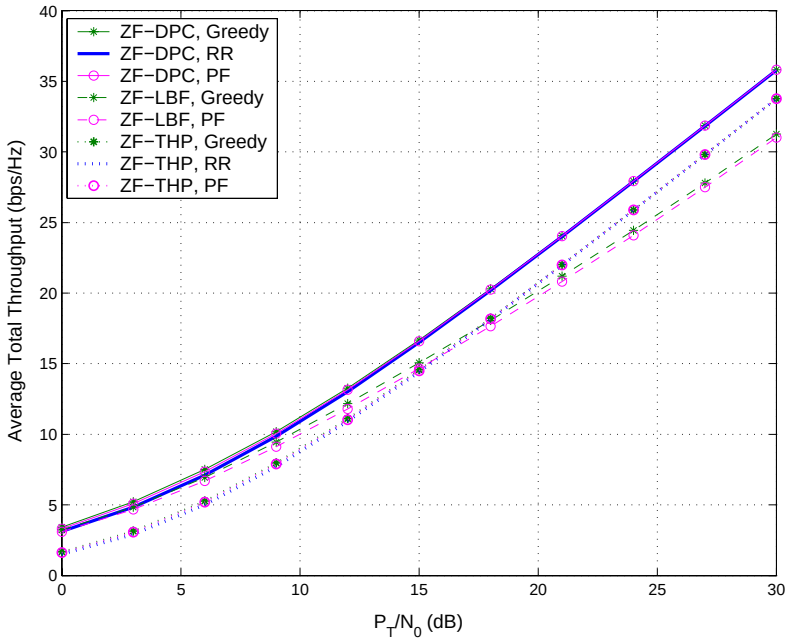
Figure 5.4. Average maximum sum rate for $n_t = 4$ in flat Rayleigh fading.

figures. Both ZF-DPC and ZF-THP use the suboptimal scheduler in (5.10), whereas ZF-LBF uses the extensive scheduler in (5.8). The optimal noise cooling is employed for ZF-THP. The simulation runs for 2000 slots for Rayleigh channels and 8000 slots for Rician channels. For the PF scheduler, we set T_c=100 slots (about 1.67s).

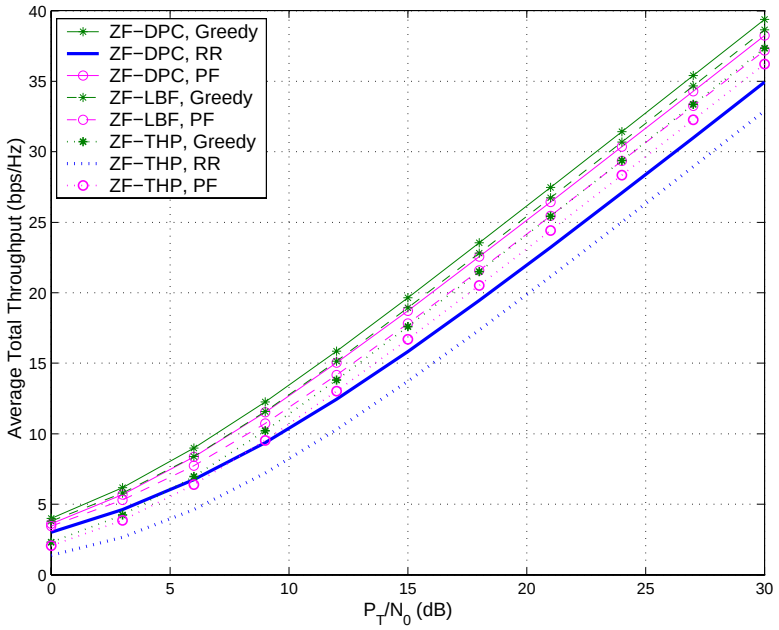
As can be seen from these plots, on the average, more than 16 bps/Hz spectral efficiency is achievable with both ZF-LBF and ZF-THP under PF scheduling at $P_T / N_0 = 15\,\text{dB}$ and $K = 16$. Further, for $K = 16$, the average total throughputs benefit from the multiuser diversity through the channel-aware scheduling (greedy and PF schedulers) as compared to the non-channel- aware scheduling (RR scheduler). We note that the multiuser diversity gain is more pronounced in channels with large fading variations in comparing parts (b) in Figures 5.4 and 5.5, and at high SNR, the ZF-LBF enjoys about 1 bps/Hz higher average spectral efficiency at $K = 16$. In all cases, the rate deficiency of ZF-THP, even with optimal noise cooling, is apparent at low SNR due to the shaping loss. However, at moderate to high SNR, ZF-THP is a competitive alternative to ZF-LBF, especially since the latter normally requires a large transmit peak-to-rms ratio and an extensive user subset search. The advantage of ZF-THP over ZF-LBF becomes more apparent with K=n_t at high SNR. Finally, it should be mentioned that in addition to providing spatial multiplexing, multiple transmit antennas also help to boost the multiuser diversity gain even in a slow fading environment with limited channel variations as shown in Figure 5.5 for ZF-THP and Figure 4.5(b) for ZF-DPC. This is in contrast to the dumb antennas proposed in [33] without full CSI at the transmitter.

5.2.2 Correlated User Channels

While an i.i.d. channel assumption is commonly used in analysis for insight, in reality measurements indicate the presence of transmit and receive correlation when multi-



(a) $K = 4$.



(b) $K = 16$.

Figure 5.5. Average maximum sum rate for $n_t = 4$ in flat Rician fading.

element antennas are used. This is particularly true in a cellular system, where the base station is usually elevated and unobstructed by local scatterers. We assume that each mobile user also uses a multielement antenna in a flat Rayleigh fading environment, and use the channel model

$$\mathbf{H} = \mathbf{R}_r \tilde{\mathbf{H}} \mathbf{R}_t^{H}, \tag{5.12}$$

where $\tilde{\mathbf{H}}$ is an $n_s \times n_t$ circular symmetric complex Gaussian matrix with $\mathrm{vec}(\tilde{\mathbf{H}}) \sim \mathcal{CN}(0, \mathbf{I}_{n_t} \otimes \mathbf{I}_{n_s})$. The covariance matrix at the transmit antenna array is defined as $\boldsymbol{\mathcal{R}}_t \triangleq (\mathbf{R}_t \mathbf{R}_t^{H})^T$, and the covariance matrix at the selected n_s antennas as $\boldsymbol{\mathcal{R}}_r \triangleq \mathbf{R}_r \mathbf{R}_r^{H}$, both with unit diagonal components. It should be noted here that the i.i.d. channel analysis is applicable to a multi-element receiver if the scheduling is over independent antenna elements rather than users, and the scheduling algorithm also works over correlated antennas. The channel model in (5.12) has been used extensively in [59] and [60] for antenna correlation under the assumption that the components of channel matrix **H** are jointly Gaussian.

- *Achievable Rate at High SNR*

Applying a QR-type decomposition on the correlated channel **H** (similar to the analysis of the i.i.d. channel), we have the determinant of $\mathbf{A} = \mathbf{H}\mathbf{H}^{H}$

$$\det(\mathbf{A}) = \prod_{k=1}^{n_s} g_k^2, \tag{5.13}$$

where unlike in i.i.d. channels, the channel gains $g_k^2,\ k = 1, ..., n_s$, are no longer independent chi-square random variables. Instead, we have from (5.12) that

$$\prod_{k=1}^{n_s} g_k^2 = \det(\boldsymbol{\mathcal{R}}_r) \cdot \det(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^{H}) \cdot \det(\boldsymbol{\mathcal{R}}_t) = \det(\boldsymbol{\mathcal{R}}_r) \cdot \prod_{k=1}^{n_s} \tilde{g}_k^2 \cdot \det(\boldsymbol{\mathcal{R}}_t), \tag{5.14}$$

where $\tilde{g}_k^2$ are independent chi-square random variables with $2(n_t - k + 1)$ degrees of

freedom, $k = 1,...,n_s$, and with the cdf and pdf specified in (5.15) and (5.16), respectively,

$$F_k(x) = 1 - e^{-x} \sum_{i=0}^{n_t - k} \frac{x^i}{i!},\ k = 1,...,n_s, \tag{5.15}$$

and

$$f_k(x) = \frac{x^{n_t - k}}{(n_t - k)!} e^{-x},\ k = 1,...,n_s. \tag{5.16}$$

Therefore, the decoupled fading gains are scaled by the determinants of the channel covariance matrices. The eigenvalues of these covariance matrices determine the total spatial multiplexing gain and the diversity gain over each antenna branch. Let r_t and r_r denote the rank of $\boldsymbol{\mathcal{R}}_t$ and $\boldsymbol{\mathcal{R}}_r$, respectively. From (5.2), at high SNR, the total achievable rate of ZF-THP can be approximated by

$$R_{\text{sum}}^{\text{zfthp}} = \sum_{k=1}^{r_s} R_k^{\text{zfthp}} \approx \sum_{k=1}^{r_s} \log_2 \left(\frac{4\bar{M}^2 P_k g_k^2}{\pi e} \right), \tag{5.17}$$

where $r_s \triangleq \min(r_t, r_r)$. If we assume equal power allocation with $P_k = P / r_s,\ k = 1,...,r_s$, and M^2-QAM signaling to all transmit antenna chains, we have

$$R_{\text{sum}}^{\text{zfthp}} = r_s \log_2 \left(\frac{6P}{\pi e r_s} \right) + \sum_{k=1}^{r_s} \log_2 (\tilde{g}_k^2) + \sum_{i=1}^{r_t} \log_2 \lambda_i^+ (\boldsymbol{\mathcal{R}}_t) + \sum_{j=1}^{r_r} \log_2 \lambda_j^+ (\boldsymbol{\mathcal{R}}_r), \tag{5.18}$$

where $\lambda_i^+(\mathbf{R})$ is the i-th positive eigenvalue of $\mathbf{R}$. For any nonnegative definite matrix, such as $\boldsymbol{\mathcal{R}}_t$ and $\boldsymbol{\mathcal{R}}_r$, its determinant is no greater than the product of its main diagonal elements. Therefore, we have $\sum_{i=1}^{r_t} \log_2 \lambda_i^+ (\boldsymbol{\mathcal{R}}_t) \le 0$ and $\sum_{j=1}^{r_r} \log_2 \lambda_j^+ (\boldsymbol{\mathcal{R}}_r) \le 0$.

Compared with the i.i.d. channel case, (5.18) indicates that at high SNR, the sum rate suffers a loss from the product decrease of positive eigenvalues of the channel covariance

matrices. For a system with the number of active users $K>n_t$, the rank of the transmit covariance matrix $\mathcal{R}_t$ determines the total maximum spatial multiplexing gain, whereas the rank of each receive covariance matrix sets its maximum spatial multiplexing gain. The achievable rates of users are affected by the distribution of these eigenvalues, which depends on factors such as local scattering and antenna array geometry. It is worth noting here that while the average achievable sum rate is reduced, the individual achievable user data rates do not necessarily decrease over each coherence interval. In fact, since the maximum eigenvalue of any Hermitian matrix is no less than the maximum diagonal element, when there exists only transmit or receive antenna correlation (practically speaking, this is probably always true since the correlation is most likely at transmitter only), the channel correlation at least increases one user's achievable rate over each coherence interval in the high SNR region, since $\prod_{k=1}^{n_s}\lambda_k = \prod_{k=1}^{n_s} g_k{}^2$ for a nonsingular covariance matrix.

- *Average Symbol Error Rate*

In Appendix D, it is proved that when there is only transmit correlation, the marginal distributions of channel gains g_k^2 in (5.13) have $n_t-(k-1), k=1,...,n_s$, degrees of freedom. Therefore, in terms of total average error performance, channel correlation introduces no diversity loss, but only an average power loss, or less coding gain, without multiuser scheduling. In the following, we focus on the transmit correlation for single-antenna receivers, since this is more typical in a cellular scenario with distributed users. We temporarily ignore the impact of scheduling and reordering. As an example, we consider a transmit covariance matrix given by,

$$\mathcal{R}_t = \begin{bmatrix} 1 & 0.11+0.64j & -0.17+0.06j & -0.01-0.02j \\ 0.11-0.64j & 1 & 0.11+0.64j & -0.17+0.06j \\ -0.17-0.06j & 0.11-0.64j & 1 & 0.11+0.64j \\ -0.01+0.02j & -0.17-0.06j & 0.11-0.64j & 1 \end{bmatrix}, \quad (5.19)$$

which is generated using the GWSSUS model [61], with Gaussian angle-of-departure (AOD) distribution over a linear 4-element transmit antenna array. We assume a carrier frequency of 2GHz, and an array element spacing of 5 times the wavelength, which is about 0.75m. The mean AOD is assumed 0.5654 in radians (about 32 degrees), and the rms angle spread is 0.0349 radians, which is about 2 degrees. The covariance matrix (5.19) has all real positive eigenvalues $\lambda_t = [1.98, 1.44, 0.50, 0.09]^T$. Figure 5.6 plots the distributions of the equivalent orthogonal channel power gains g_k^2, $k = 1, ..., n_s$, with the transmit covariance specified in (5.19). The statistics were collected over more than 10,000 channel realizations at a total average transmit SNR of 6 dB.

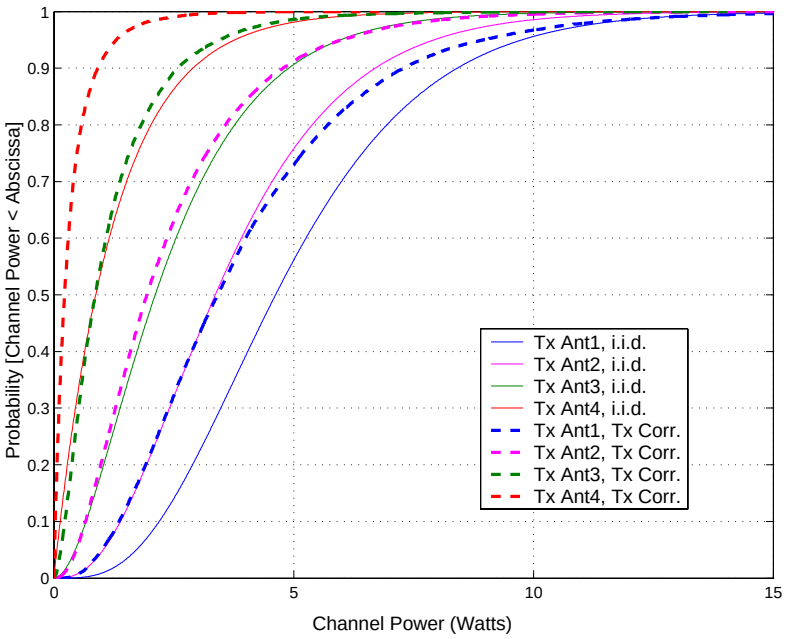


Figure 5.6. CDF of channel power gains at $P_T / N_0 = 6$ dB. From left to right are for g_k^2, $k = n_s, ..., 1$.

The channel gain distributions in i.i.d. flat Rayleigh fading are also plotted for comparison. Figure 5.6 reveals that there is an average power loss over each branch.

Figure 5.7 shows the simulated SER with transmit correlation over 10,000 channel

realizations. Again for comparison, Figure 5.7 also plots the average SER in i.i.d. channels and that for an AWGN channel. The scheduling only is used in these simulations. Each transmit antenna is allocated the same average transmit power P_T/n_t as before. As can be seen from the SER curves, the total average SER suffers no diversity loss, but only an average power loss. For the transmit correlation in (5.19), this loss decreases from about 5 dB with 4 users to about 3 dB with 32 users.

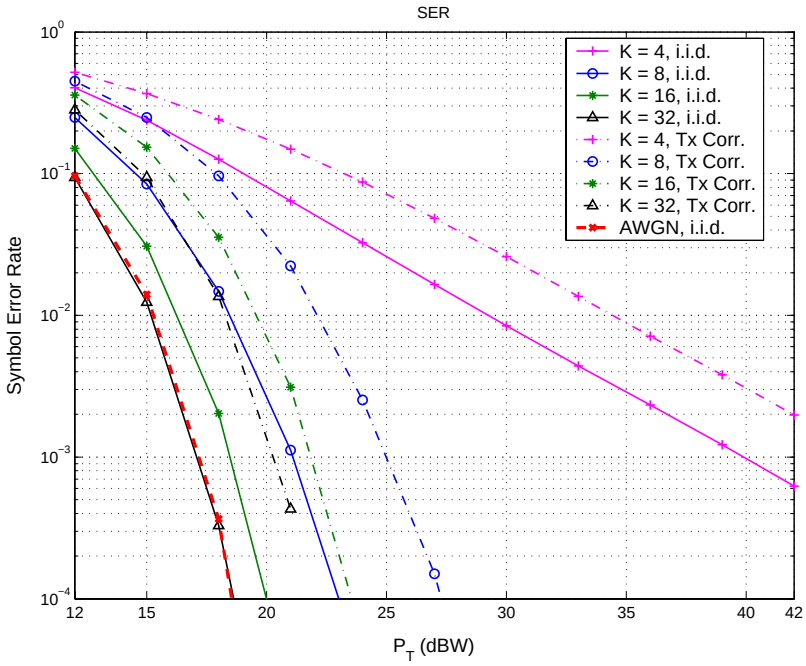


Figure 5.7. SER of 16-QAM transmission for n_t=n_s=4 with transmit correlation in (5.19).

Figure 5.8 plots the total average SER vs. the number of users in both i.i.d. flat Rayleigh fading and correlated fading of (5.19). The total transmit power is fixed at P_T=16 dBW (or P_T/N_0=16 dB for unit noise variance) for 16-QAM signals for each user. When n_s=1, the theoretical SER with correlated channels can be calculated from order statistics [32] as

$$P_{sym} = \int_0^{\infty} P_s(\gamma_s \gamma) \cdot K f_c(\gamma) F_c^{K-1}(\gamma) \cdot d\gamma \qquad (5.20)$$

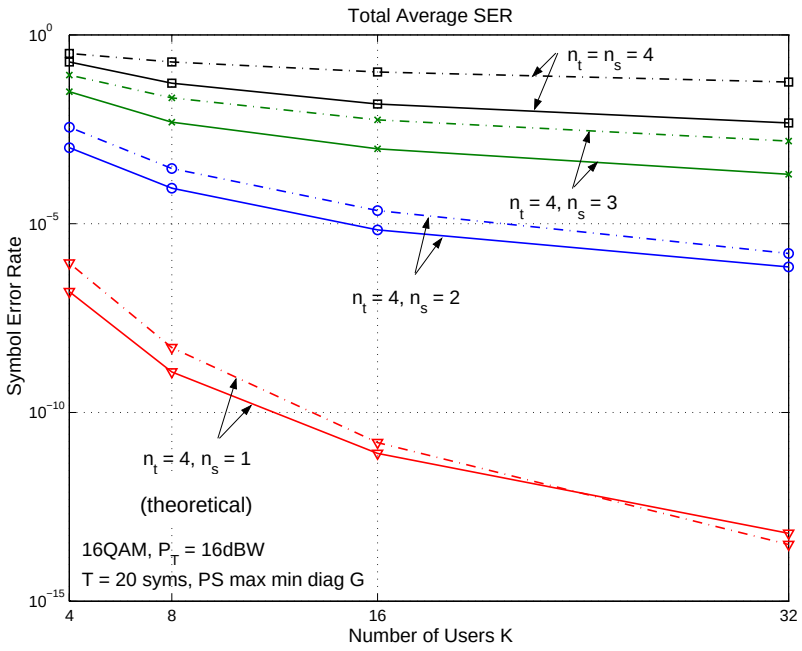


Figure 5.8. SER for 16-QAM at n_t=4 and n_s=1,2,3,4. Solid curves: i.i.d. flat Rayleigh fading; Dash-dotted curves: flat Rayleigh fading with transmit correlation in (5.19).

where $P_s(\cdot)$ is the uncoded SER of each user in the AWGN channel with the same modulation, and γ_s is the average transmit symbol energy to noise power ratio E_s / N_0. The function $f_c(\gamma)$ is the pdf of the channel gain and is defined as (see Appendix D)

$$f_c(\gamma) \triangleq \sum_{i=1}^{n_t} C_i / \lambda_i \exp(-\gamma / \lambda_i), \ \gamma > 0, \tag{5.21}$$

with $C_i \triangleq \prod_{j=1, j\neq i}^{n_t} \lambda_i /(\lambda_i - \lambda_j)$ for the distinct eigenvalues λ_i of a non-singular transmit covariance matrix, and $F_c(\gamma)$ is the corresponding cdf. It can be seen from Figure 5.8 that as n_s decreases, the negative impact of transmit correlation on the total average SER decreases. When n_s=1, we note that there is actually SER improvement with transmit correlation at a sufficiently large number of users. This can be explained by the fact that as the multiuser diversity produces a diminishing return at an increasing number of users, the positive effect

of antenna correlation on transmit beamforming becomes apparent. However, this positive effect of transmit correlation on the transmit beamforming diminishes at high SNR as shown in Figure 5.9.

In the above, we have assumed a fixed channel covariance variance. It should be noted that while the CSI is estimated over each coherence interval, the covariance matrix can be updated over a relatively longer interval. This is because channel covariance statistics generally depend on large-scale scatterers and the distance between the transmitter and receiver, and vary at a larger time-scale than small-scale channel fading. We also note that the channel model (5.12) conveniently decouples the transmit and receive antenna

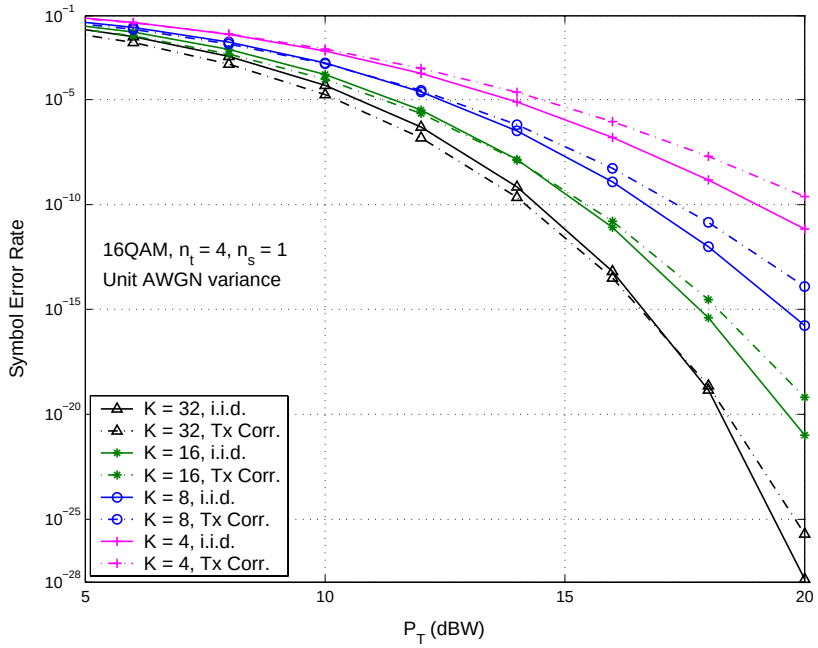


Figure 5.9. Theoretical SER for 16-QAM at $n_t = 4$ and $n_s = 1$. Solid curves: i.i.d. flat Rayleigh fading; Dash-dotted curves: flat Rayleigh fading with transmit correlation in (5.19).

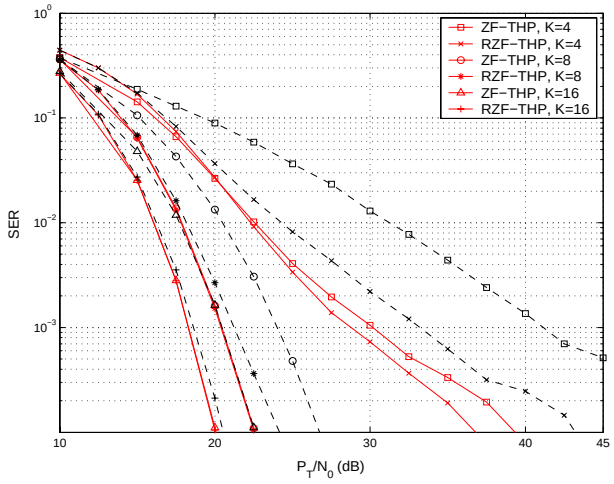
correlation, which can greatly reduce the CSI feedback overhead in an FDD-type network and speed up the packet scheduling. In an FDD MIMO downlink channel, the covariance matrix $\mathcal{R}_r$ can be estimated separately at each user first. Depending on the eigenvalue

distribution of $\mathcal{R}_r$, the user can feed back the channel estimates corresponding to the best channels estimated at its receive antennas. In doing this, the receiver uses an eigenvalue threshold based on the number of active users. If the receive covariance matrices has one dominant eigenvalue, the receiver only needs to send the best channel estimate to the transmitter.

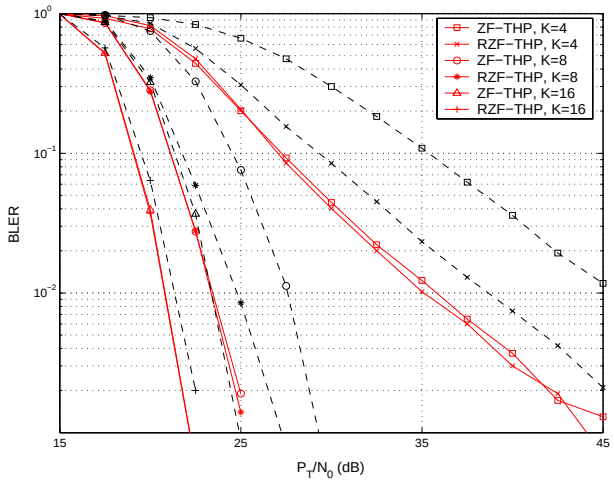
5.3 Scheduling and Reordering for THP

We now evaluate the performance of THP with the suboptimal adaptive scheduling and reordering proposed in Section 4.4. In the following, we shift our focus from the average sum-rate loss to the average error-rate improvement. Figures 5.10(a) and (b) present the simulation results on the average error rate in i.i.d. flat Rayleigh fading over 10,000 channel realizations. The number of transmit antennas is set to n_t=4 and a single receive antenna is used at each user. Both ZF-THP and RZF-THP are simulated using 16QAM signals for each user. For the average block error rate (BLER), each space-time block has 4×100 symbols.

We note from these plots that the multiuser diversity manifests itself in increasing error-rate slopes as K increases. In Figure 5.10(a), with scheduling only, the regularized interference nulling improves SER by about 8dB, 3dB, and 2dB, respectively, for K=4, 8, and 16, at an average uncoded SER of 10^{-3}, as compared with the zero-forcing interference nulling. For the scheduled users, reordering algorithms can improve the error rate over scheduling only by about 10dB, 4dB, and 2dB for K=4, 8, and 16, respectively, at the same SER. However, with reordering imposed, the improvement from regularization (RZF vs. ZF) is barely noticeable for the case of K=4 and increases with SNR. As K is fixed, reordering only improves the average SER by about 4dB, 1dB, and 0dB in each case above, with almost no improvement in the case of large multiuser diversity gain over the SNR range simulated.



(a) SER.



(b) BLER.

Figure 5.10. Average error rate of THP. Solid curves: scheduling and reordering; dashed curves: scheduling only.

Therefore, to a large extent, a large multiuser diversity gain can compensate for the poor error-rate performance of zero-forcing interference nulling, even at relatively low SNR as shown in Figure 5.10. Similar observations can be made from the BLER plot. Simulations show that similar results also hold for larger n_t.

5.4 Summary

This chapter presents a practical precoding scheme built on the spatial Tomlinson-Harashima precoding. This is a one-dimensional suboptimal implementation of modulo-lattice precoding, provided that perfect user SNR is known at each scheduled receiver. Unlike the ZF-LBF solution, ZF-THP has a restricted average transmit power. A shaping loss of 1.53 dB has been identified as the only SNR loss of THP from an AWGN channel at high SNR. In the low SNR region, however, the receiver modulo operation incurs a much larger shaping loss, which is tied to the causal interference cancellation of THP based on a one-dimensional lattice. To recover the shaping loss, non-causal side information must be used to perform the modulo operation on a high-dimensional lattice, which may lead to high latency and complexity. Nonetheless, it turns out that in a large multiuser data network, a gain from multiuser diversity can be obtained by leveraging the user subset selection. With a greedy scheduler to maximize the instantaneous total throughput, the error performance analysis at high SNR substantiates a dramatic multiuser diversity gain, even for a modest number of users in an i.i.d. flat fading environment. Performance in correlated fading indicates that there is a power loss associated with the reduction in the product of the positive eigenvalues of the channel covariance matrices at both the transmitter and the receivers. Simulations of suboptimal scheduling with reordering reveal significant average error-rate improvement over a wide range of SNR.

Chapter 6
Precoding and Scheduling for OFDM Systems

The T-H precoding and scheduling in the previous chapter can be easily extended to an OFDM system in frequency-selective fading. For the resulting THP-OFDM system, a more significant problem is the high peak-to-average power ratio (PAR), due to the multiplexing in both spatial and frequency domains. It turns out that this high PAR can be mitigated by a generalization of T-H precoding on a per-tone basis, which retains the same rate and error performance of THP. In this chapter, we propose such a nonlinear precoder with both low transmit PAR and low average transmit power for a distributed MIMO-OFDM system. System performance under suboptimal scheduling is also presented.

6.1 MIMO Precoding for OFDM Systems

In a rich-scattering environment, multiple transmit and receive antennas can be used to form MIMO channels to increase the link capacity by a factor equal to the minimum number of transmit and receive antennas [2], [3]. In Chapter 2, we have seen the same conclusion holds for the sum capacity in a multi-antenna broadcast channel with distributed receive antennas. However, for wideband transmission in time delay spread environments, space-time processing must be used to mitigate inter-symbol interference (ISI). The complexity of the space-time processing increases with the bandwidth, which limits the supportable data rate of single-carrier solutions. In this case, the use of orthogonal frequency division multiplexing (OFDM) [62] with sufficiently long symbol periods solves this

problem. In addition, OFDM inherently provides frequency diversity over sub-channels (or tones), which offers an opportunity for both interference averaging and interference avoidance in the frequency domain.

In this chapter, we consider an OFDM system using MIMO precoding and scheduling. With transmitter knowledge of the channel states, bit and power loading can also be performed across OFDM tones, in addition to channel-aware scheduling over each tone, to optimize system throughput or performance. However, our focus here is to combine THP with OFDM in a broadcast channel and to use adaptive scheduling over individual tones, rather than power and bit loading across tones. More importantly, we solve a pronounced engineering problem with THP-OFDM (i.e., PAR problem) by using generalized T-H precoding. In the following, we first give a brief review of MIMO-OFDM systems and the associated channel statistics.

6.1.1 Review of MIMO-OFDM systems

A baseband and discrete-time MIMO-OFDM system model is shown in Figure 6.1. The figure depicts a single point-to-point MIMO link using n_t transmit antennas and n_r receive antennas using n_c tones.

In Figure 6.1, the user signals pass through IFFT, parallel-to-serial (P/S) conversion, and cyclic prefix insertion (+CP) at the transmitter side, and the corresponding inverse processing at the receiver. The operations of CP insertion and removal make the effective channel responses into circulant matrices as indicated in Figure 6.1, which can be diagonized by normalized (unitary) IFFT and FFT matrices [63]. Therefore, at tone *p*, an equivalent signal model can be established as [60], [63]

$$\hat{\mathbf{x}}_p = \mathbf{H}_p \mathbf{x}_p + \mathbf{w}_p = \left(\sum_{l=0}^{L} \mathbf{H}_l(t) \exp\{-j2\pi lp / n_c\} \right) \mathbf{x}_p + \mathbf{w}_p, \quad p = 1, ..., n_c \qquad (6.1)$$

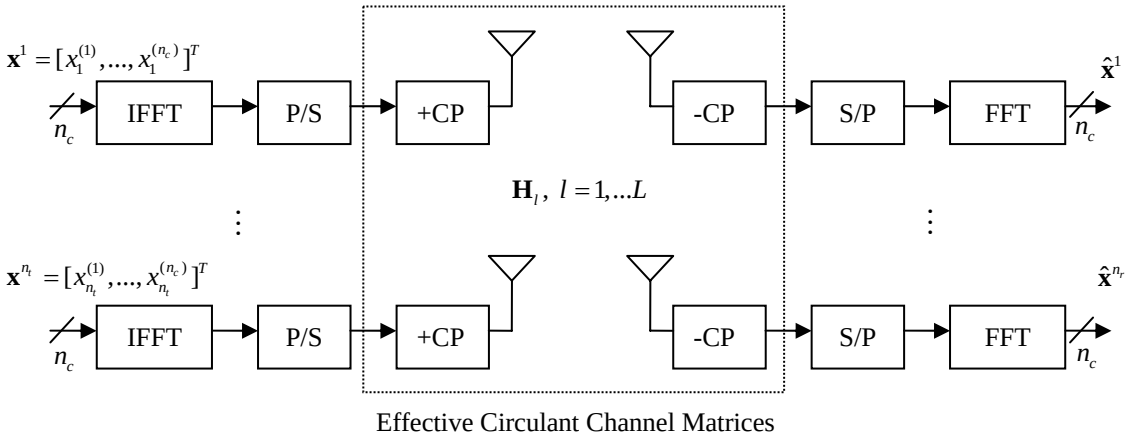


Figure 6.1. A MIMO-OFDM system block diagram.

where $\mathbf{x}_p = [x_1^{(p)}, ..., x_{n_t}^{(p)}]^T \in \mathbb{C}^{n_t \times 1}$, $\hat{\mathbf{x}}_p = [\hat{x}_1^{(p)}, ..., \hat{x}_{n_r}^{(p)}]^T \in \mathbb{C}^{n_r \times 1}$, and $\mathbf{w}_p \in \mathbb{C}^{n_r \times 1}$ are transmitted signal, received signal and AWGN samples over tone *p*. For simplicity, we assume the channel frequency response at tone *p*, $\mathbf{H}_p \in \mathbb{C}^{n_r \times n_t}$, is defined as in (6.1), and is dependent on the *L*-path channel impulse responses $\mathbf{H}_l(t) \in \mathbb{C}^{n_r \times n_t}$, *l*=1,…, *L* at time slot *t*. The elements of the individual $\mathbf{H}_l(t)$are circularly symmetric complex Gaussian random variables, with their variances determined by the corresponding channel power delay profiles (PDP). We also assume that each slot contains one OFDM space-time block with ($n_t n_c$) symbols and no time variation. The AWGN samples $\mathbf{w}_p$, *p*=1,…,n_c, are assumed circularly symmetric complex Gaussian with zero mean and unit variance, and are statistically independent with respect to space, time, and frequency. The model (6.1) presupposes no inter-carrier interference.

The signal model in (6.1) confirms that MIMO-OFDM can provide multiplexing in both frequency and spatial domains, by generating parallel space-frequency pipes $\mathbf{H}_p$, *p*=1,…,n_c. Assuming ideal carrier synchronization, timing, and perfect symbol-rate sampling, the mutually orthogonal tones provide perfectly separated channels in the frequency domain, and what remain to be separated are the spatially multiplexed streams in the spatial domain over each tone. There has been abundant literature on MIMO-OFDM systems focusing on

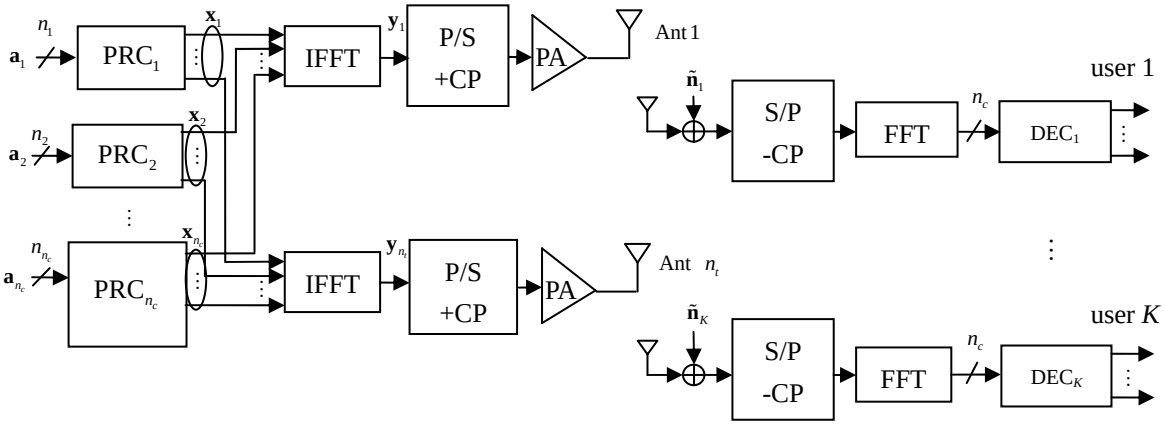
various roles of antennas. For example, [68-74] studied point-to-point MIMO-OFDM systems, and [75-80] investigated multiuser MIMO-OFDM networks. Two good tutorials on MIMO-OFDM systems are [77]-[78].

The focus of this chapter is to extend the preceding T-H precoding results from Chapter 5 to an OFDM system on a per-tone basis. A practical advantage of OFDM is its ability to rapidly measure interference or path loss parameters in parallel on all candidate channels [79], which makes channel-aware scheduling and channel-dependent precoding more affordable. As we will see below, in a MISO broadcast channel, a generalized T-H precoder with scheduling remains an attractive suboptimal solution in terms of the achievable sum rate over all tones.

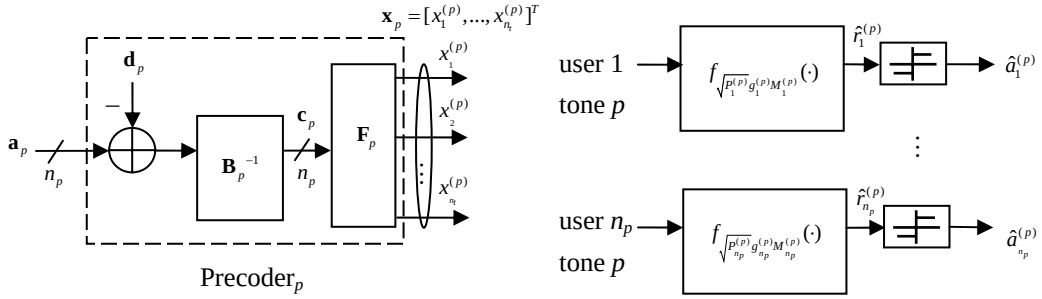
6.1.2 Generalized T-H Precoding

We now extend the THP discussed in Chapter 5 to a MISO-OFDM broadcast channel in frequency-selective fading. Figure 6.2(a) depicts an OFDM downlink system using MIMO precoding. The model in Figure 6.2(a) is similar to the multiuser precoding system in [39] except that it is for a downlink OFDM wireless system using generalized nonlinear precoding. The transmitter has an n_t-element antenna array and the K receivers have a single-element antenna each. At each OFDM space-time block, the base station sends independent and synchronized data packets to n_p users simultaneously over tone p, p=1,…,n_c, with n_c being the total number of tones, and $n_p \le n_t$ being constrained by the n_t degrees of freedom at transmitter on each tone. At the receiver inputs, the spatially and temporally white complex Gaussian noise samples $\tilde{\mathbf{n}}_k$, $k = 1,...,K$, of zero mean and unit variance are added to the user signals.

The corresponding precoder and decoders over tone p are illustrated in Figure 6.2(b). The precoding sequence $\mathbf{d}_p \triangleq [d_1^{(p)},...,d_{n_p}^{(p)}]^T$ is a modulo vector sequence with



(a) System block diagram.



(b) Nonlinear precoder and decoder over tone *p*.

Figure 6.2 OFDM downlink transmission using multiuser precoding.

$d_i^{(p)} = 2m_i^{(p)} \cdot M_i^{(p)}$, where $m_i^{(p)}$ is an integer and $M_i^{(p)}$ is the positive symmetric Voronoi region [54] boundary of the input signal $a_i^{(p)}$ in $\mathbf{a}_p \triangleq [a_1^{(p)}, ..., a_{n_p}^{(p)}]^T$ for the *i*-th user's data over tone p, $i = 1, ..., n_p$, $p = 1, ..., n_c$. For square M^2-QAM signals $a_i^{(p)}$, $M_i^{(p)} = M$. Let $\mathbf{H}_p$ be the normalized $n_p \times n_t$ channel frequency response matrix between the transmit antenna array and the n_p users on tone *p*, with its entries having the same variance. With a probability of one, the Hermitian matrix has a unique Cholesky factorization $\mathbf{H}_p \cdot \mathbf{H}_p^H = \mathbf{L}_p \cdot \mathbf{L}_p^H$, where the $n_p \times n_p$ lower triangular matrix $\mathbf{L}_p$ has real positive diagonal

entries. Similar to a single-carrier system, the matrix $\mathbf{L}_p$ is decomposed into $\mathbf{L}_p = \mathbf{G}_p \cdot \mathbf{B}_p$, where $\mathbf{G}_p$ is a diagonal matrix with the main diagonal entries $g_i^{(p)}$ of $\mathbf{L}_p$, $i = 1,...,n_p$, and the matrix $\mathbf{B}_p$ is lower triangular with unit main diagonal entries. The $n_t \times n_p$ transmission matrix $\mathbf{F}_p = (\mathbf{L}_p{}^{-1} \cdot \mathbf{H}_p)^H$ has orthonormal columns. Hence over each tone, the precoder in Figure 6.2 (b) is a zero-forcing nonlinear precoder since the zero-forcing criterion is used to null out the interference to user i from users j, for $j > i$ and $i, j = 1,...,n_p$. Similarly, the regularized zero-forcing precoder as proposed in Chapter 5 can be used. It is worth noting that although the precoder has a linear form, it is nonlinear due to the nonlinear determination of the precoding sequence. At the decoder, a symmetric modulo operator $f_{\sqrt{P_i^{(p)}} g_i^{(p)} M_i^{(p)}}(\cdot)$ is used for the data stream to user i over tone p, $i = 1,...,n_p$ and $p = 1,...,n_c$, where the modulo function is defined as [39]

$$f_y(x) \triangleq x - 2y\left\lfloor (x+y)/(2y) \right\rfloor \tag{6.2}$$

for each real component of each signal. The modulus is set to the positive symmetric Voronoi boundary value $M_i^{(p)}$ of input signal $a_i^{(p)}$ and is weighed by the transmitter and channel gains, which are assumed known from the transmitter perfectly over each fading block. The user data $a_i^{(p)}$ are assumed to have unit average symbol energy and possibly modulated adaptively over each OFDM block. For simplicity, we also assume that over each OFDM space-time block, the total transmit power P_T is equally split across tones and users such that $P_i^{(p)} = P_T/(n_p n_c)$, $i = 1,...,n_p$ and $p = 1,...,n_c$. Each receiver simply performs a maximum-likelihood detection at its modulo output.

Assuming that the cyclic prefix (CP) of the OFDM system is sufficiently long so that it is no less than the maximum possible delay spread of the user channel matrices, and that the IFFT blocks of all users are perfectly synchronized, then similar to (5.1), we have the received signal at the modulo output for user *i* at tone *p* as

$$\begin{aligned}\hat{r}_i^{(p)} &= f_{\sqrt{P_T/(n_p n_c)}g_i^{(p)}M_i^{(p)}}\left(\sqrt{P_T/(n_p n_c)}g_i^{(p)}(a_i^{(p)}-d_i^{(p)})+w_i^{(p)}\right)\\ &= f_{\sqrt{P_T/(n_p n_c)}g_i^{(p)}M_i^{(p)}}\left(\sqrt{P_T/(n_p n_c)}g_i^{(p)}a_i^{(p)}+w_i^{(p)}\right)\\ &= f_{M_i^{(p)}}\left(a_i^{(p)}+w_i^{(p)}/(\sqrt{P_T/(n_p n_c)}g_i^{(p)})\right),\ i=1,...,n_p,\ p=1,...,n_c,\end{aligned} \tag{6.3}$$

where $\mathbf{w}_p=[w_1^{(p)},...,w_{n_p}^{(p)}]$ are the equivalent complex Gaussian white noise samples of zero mean and unit variance for tone *p* at each user. Therefore, the receive modulo operation makes the precoding sequence $\mathbf{d}_p$ *transparent* to the receivers. The precoder reduces to THP when the precoding sequence $\mathbf{d}_p$ is chosen to force $\mathbf{c}_p$ into the Voronoi region of the input signal constellation, hence it is a generalization of THP and retains the same achievable sum rate and error rate of THP. We first point out the high PAR problem of THP in the following.

A prominent feature of THP is that when the input signals $\mathbf{a}_p$ are independent symbols with equal probability, the components of $\mathbf{c}_p$ tend to be uniformly distributed over the Voronoi regions of the signal space and are independent of each other. For a complex matrix $\mathbf{B}$ as described above, this can be interpreted as the components of $\mathbf{c}_p$ being mutually independent and each statistically approximates a continuous uniform random variable over the Voronoi region. Therefore, the transmission filter outputs at tone *p*,

$$\mathbf{x}_p \triangleq [x_1^{(p)},...,x_{n_t}^{(p)}]=\mathbf{F}_p\mathbf{B}_p^{-1}\cdot(\mathbf{a}_p-\mathbf{d}_p)=\mathbf{F}_p\cdot\mathbf{c}_p=\sum_{i=1}^{n_p}c_i^{(p)}\mathbf{f}_i^{(p)}, \tag{6.4}$$

have component distributions that approximate the convolutions of uniform random variables $c_i^{(p)}[\mathbf{F}_p]_{ji}$, $i=1,...,n_p$, $j=1,...,n_t$, where $\mathbf{c}_p\triangleq[c_1^{(p)},...,c_{n_p}^{(p)}]^T$, and $\mathbf{F}_p\triangleq[\mathbf{f}_1^{(p)},...,\mathbf{f}_{n_p}^{(p)}]$. Over each OFDM block, the PAR of THP at tone *p* is then

$$\begin{aligned}\mathrm{PAR}_{\mathrm{THP}}(p) &= \max_{1\le j\le n_t}\left\{\max_{\mathbf{c}_p}\left|\sum_{i=1}^{n_p}c_i^{(p)}[\mathbf{F}_p]_{ji}\right|^2/\mathrm{E}_{\mathbf{c}_p}\left[\sum_{i=1}^{n_p}|\,c_i^{(p)}[\mathbf{F}_p]_{ji}\,|^2\right]\right\},\\ &= \max_{1\le j\le n_t}\left\{\max_{\mathbf{c}_p}\left|\sum_{i=1}^{n_p}c_i^{(p)}[\mathbf{F}_p]_{ji}\right|^2/\left[\sum_{i=1}^{n_p}\left(M_i^{(p)}\,|[\mathbf{F}_p]_{ji}\,|\right)^2/3\right]\right\}\end{aligned} \tag{6.5}$$

where $\mathrm{E}_{\mathbf{c}_p}[\cdot]$ denotes the expectation over $\mathbf{c}_p$. Using the Cauchy-Schwarz inequality, we have

$$\mathrm{PAR}_{\mathrm{THP}}(p) \le \max_{1\le j\le n_t}\left\{\left[\sum_{i=1}^{n_p} M_i^{(p)} |[\mathbf{F}_p]_{ji}|\right]^2 / \left[\sum_{i=1}^{n_p}\left(M_i^{(p)} |[\mathbf{F}_p]_{ji}|\right)^2 / 3\right]\right\} \qquad (6.6)$$

for the peak-power maximizing $c_i^{(p)} = M_i^{(p)}[\mathbf{F}_p]_{ji}^{\ H} / |[\mathbf{F}_p]_{ji}|$. Assuming all the streams over tone p use the same modulation with square constellation, we have $M_i^{(p)} = M^{(p)},\ i = 1,...,n_p$, and

$$\begin{aligned}\mathrm{PAR}_{\mathrm{THP}}(p) &= \max_{1\le j\le n_t} 3\cdot\left[\sum_{i=1}^{n_p}|[\mathbf{F}_p]_{ji}|\right]^2 / \sum_{i=1}^{n_p}|[\mathbf{F}_p]_{ji}|^2 \\ &= \max_{1\le j\le n_t} 3\cdot\left[1+\sum_{m=1}^{n_p-1}\sum_{n=m+1}^{n_p}|[\mathbf{F}_p]_{jm}|\cdot|[\mathbf{F}_p]_{jn}|\right],\ \ p = 1,...,n_c . \end{aligned} \qquad (6.7)$$

Equation (6.7) indicates that over each tone, the PAR of THP increases with the number of streams n_p and can have a quite high PAR even for moderate n_p. When $n_p = 1$, the PAR reduces to that of $\mathbf{c}_p$, which is 3 for continuous and uniformly distributed components $c_i^{(p)}$ over the square Voronoi region $[-M_i^{(p)}, M_i^{(p)}) \times [-M_i^{(p)}, M_i^{(p)})$. For an OFDM system, the transmit PAR is further increased with IFFT processing in a way similar to the transmission matrix processing above and increases with n_c. In the long run, the system PAR is contingent on the statistics of transmission matrices at all tones. We note in passing that for square M^2-QAM signals with equal symbol probabilities, the PAR is

$$\mathrm{PAR}_{M^2\text{-QAM}} = 3(M-1)/(M+1),$$

which is strictly less than 3 for $M > 1$ and approaches 3 as M increases.

6.2 Complex Sphere Precoding for OFDM Systems

The main technical advantages of THP in [39] and [53] are its simplicity and the low average transmit power increase (modulo loss [54]) due to precoding. The high PAR problem

with THP, however, motivates us to search for better precoding techniques with lower PAR, but which still retain the low modulo loss. In other words, we wish to decrease the PAR without increase the average power. Observing that these criteria can be met by choosing the precoding sequence $\mathbf{d}_p$ in Figure 6.2(b) appropriately, we perform a comprehensive and efficient lattice search for $\mathbf{d}_p$ over each tone using a modified complex version of the Fincke-Pohst algorithm [80], [81]. We name the resulting precoder a complex sphere precoder (CSP) since the search is within a lattice sphere subject to a total transmit power constraint. We start this section with a description of Fincke-Pohst algorithm over a *single* tone, followed by its application to PAR reduction for an OFDM system using MIMO precoding.

6.2.1 Complex Fincke-Pohst Algorithm

The Fincke-Pohst algorithm was used as a sphere decoder to compute maximum-likelihood symbol estimates of space-time signals in a multi-antenna system [82], where it was shown that the sphere decoder has a comparable complexity to V-BLAST processing at high SNR. In [81], a fast complex version of Fincke-Pohst algorithm was proposed to decode M-PSK and M-QAM signals, and this was used for lattice encoding to reduce the transmit power in a multiuser MIMO system using channel inversion [83]. We extend the results in [81] and [83] to the nonlinear precoding in Figure 6.2(b) and evaluate both the PAR and the average transmit power of the system.

Suppose the transmit power constraint over the p-th tone is $P^{(p)}$, we have

$$\mathbf{x}_p{}^H \mathbf{x}_p = (\mathbf{a}_p - \mathbf{d}_p)^H \mathbf{B}_p{}^{-H} \mathbf{B}_p{}^{-1} (\mathbf{a}_p - \mathbf{d}_p) \le P^{(p)} \tag{6.8}$$

with the total transmit power subject to $\sum_{p=1}^{n_c} P^{(p)} \le P_T$. The objective is to search for the optimum $\tilde{\mathbf{d}}_p$ using some criterion over an n_p-dimensional complex lattice subject to the power constraint (6.8) at tone p. Figure 6.3 depicts a modulo lattice using 16-QAM signals for each component of $\tilde{\mathbf{d}}_p$, which differs from a conventional lattice decoder in that it contains a potentially much larger number of lattice points (check points only shown in 3 squares)

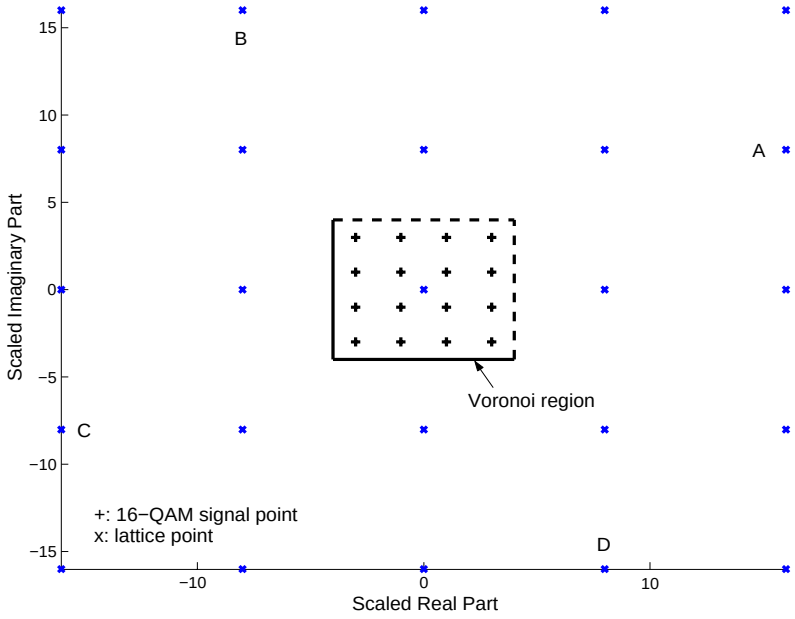


Figure 6.3. Lattice for 16-QAM over tone p.

under the constraint in (6.8), rather than confined to a fixed constellation. A direct exhaustive search over this lattice has to examine a number of n_p-dimensional complex candidate points, which increases exponentially with $P^{(p)}$. Within the complex lattice boundary of fixed $P^{(p)}$, an exhaustive search still has to evaluate a number of points exponential in n_p, which can be very computationally intensive even for a small value of n_p. The Fincke-Pohst algorithm, however, avoids the direct search by rewriting (6.8) as

$$\sum_{i=1}^{n_p} \left| a_i^{(p)} - d_i^{(p)} + \sum_{j=1}^{i-1} [\mathbf{B}_p^{-1}]_{ij} (a_j^{(p)} - d_j^{(p)}) \right|^2 \le P^{(p)}, \tag{6.9}$$

where we have used lower triangular matrix $\mathbf{B}_p^{-1}$ with unit diagonal elements $[\mathbf{B}_p^{-1}]_{ii} = [\mathbf{B}_p]_{ii} = 1$, $i = 1, ..., n_p$. Equation (6.9) makes possible a successive calculation of $d_i^{(p)}$ from $i = 1$ to n_p for each candidate point $\mathbf{d}_p$, and obtains a complexity polynomial in n_p provided that the lengths of the columns of $\mathbf{B}_p$ stay bounded [80].

For the square-QAM based lattice in Figure 6.3, we note that the complex lattice

points can be identified with concentric circles, with four points on each circle indexed with a unique pair of radius and phase offset. The four points on each circle are uniformly spaced with an angular spacing of $\pi/2$. Therefore, we can perform the Fincke-Pohst lattice search on these concentric circles each with a fixed phase offset. For example, the four points labeled A, B, C, and D in Figure 6.3 are identified with a (scaled) radius of $8\sqrt{5}$ and a phase offset of $\tan^{-1}(1/2)$. The phases for the four points are therefore $\tan^{-1}(1/2)+m\cdot\pi/2$, m=0, 1, 2, 3. Using the complex version Fincke-Pohst algorithm for i=1, we have from (6.9)

$$|a_1^{(p)}-d_1^{(p)}|^2\le P^{(p)}.$$

Let $\Gamma_1^{(p)}=P^{(p)}$, $r_a^{(p)}=|a_1^{(p)}|$, $\theta_a^{(p)}=\measuredangle(a_1^{(p)})$, $r_d^{(p)}=|d_1^{(p)}|$, $\theta_d^{(p)}=\measuredangle(d_1^{(p)})$, with $0\le\measuredangle(\cdot)<2\pi$, and define $\eta\triangleq(r_a^{(p)2}+r_d^{(p)2}-\Gamma_i^{(p)})/(2r_a^{(p)}r_d^{(p)})$, we can solve for the magnitude values of $d_1^{(p)}$ as

$$0\le r_d^{(p)}\le\sqrt{\Gamma_i}+r_a^{(p)}, \tag{6.10}$$

and the phases of $d_1^{(p)}$ by following [81]

$$\triangle\theta(r_d^{(p)})+\frac{\pi}{2}\left\lceil\frac{\theta_a^{(p)}-\cos^{-1}\eta-\triangle\theta(r_d^{(p)})}{\pi/2}\right\rceil\le\theta_d^{(p)}\le\triangle\theta(r_d^{(p)})+\frac{\pi}{2}\left\lfloor\frac{\theta_a^{(p)}+\cos^{-1}\eta-\triangle\theta(r_d^{(p)})}{\pi/2}\right\rfloor \tag{6.11}$$

for $-1\le\eta\le1$ at radius $r_d^{(p)}$, where $\triangle\theta(r_d^{(p)})$ is the phase offset at radius $r_d^{(p)}$ and $0\le\cos^{-1}(\cdot)\le\pi$. In (6.10), we used $P^{(p)}\ge a_1^{(p)}$ and $r_d^{(p)}\ge0$. For the phase bounds in (6.11), when $\eta<-1$, all four points on $r_d^{(p)}$ are solutions for $d_1^{(p)}$; when $\eta>1$, there is no solution for $d_1^{(p)}$ at $r_d^{(p)}$. Proceed to i=2,…,n_p, and at each step update

$$r_d^{(p)}=|d_i^{(p)}|,$$

$$\theta_d^{(p)}=\measuredangle(d_i^{(p)}),$$

$$r_a^{(p)}=\left|a_i^{(p)}+\sum_{j=1}^{i-1}[\mathbf{B}_p^{-1}]_{ij}(a_j^{(p)}-d_j^{(p)})\right|,$$

$$\theta_a^{(p)} = \angle\left(a_i^{(p)} + \sum_{j=1}^{i-1} [\mathbf{B}_p{}^{-1}]_{ij} (a_j^{(p)} - d_j^{(p)}) \right),$$

and

$$\Gamma_i^{(p)} = \Gamma_{i-1}^{(p)} - \left| a_{i-1}^{(p)} - d_{i-1}^{(p)} + \sum_{j=1}^{i-2} [\mathbf{B}_p{}^{-1}]_{i-1\,j} (a_i^{(p)} - d_j^{(p)}) \right|^2, \quad i = 2, ..., n_p,$$

we can solve for the magnitude and phase values of $d_i^{(p)}$ for $2 \le i \le n_p$ using (6.10) and (6.11). A comprehensive search over the concentric circles satisfying (6.10) at each step provides all $\mathbf{d}_p$ in (6.8).

6.2.2 Bounded Lattice Search with THP Initialization

In the complex Fincke-Pohst algorithm above, we can apply an upper bound on the lattice boundary to ensure the search over a confined lattice at each tone. This may become necessary when the lattice boundary of (6.10) is too large. Recall that $d_i^{(p)}$ is an even integer multiple of $M_i^{(p)}$, we simply use an upper bound on the boundary in (6.10), which is rounded to the nearest even integer not greater than the value in (6.10). Therefore, the maximum normalized search radius can be upper bounded by

$$L_{\max}^{(p)}(i) \triangleq \min\left\{ 2\left\lfloor \left(\sqrt{\Gamma_i^{(p)}} + r_a^{(p)} \right) / 2 \right\rfloor, L_{\text{ub}}^{(p)} \right\}, \quad i = 1, ..., n_p, \tag{6.12}$$

where $L_{\text{ub}}^{(p)}$ is an input upper bound on $L_{\max}^{(p)}(i)$ at each step i. The specification of $L_{\text{ub}}^{(p)}$ can also be used as a threshold by a multi-tone scheduler to avoid transmission over some tones having extremely large $L_{\max}^{(p)}(i)$. The multi-tone scheduling is beyond the scope of this section. In the next section, it will be shown that an $L_{\text{ub}}^{(p)}$ as large as 100 is sufficient in most cases.

The additional constraint on the lattice boundary with an input $L_{\text{ub}}^{(p)}$ may exclude the THP solution $\mathbf{d}_p^{\text{thp}}$ for $\mathbf{d}_p$. In this case, we can explicitly include $\mathbf{d}_p^{\text{thp}}$ by initializing $\mathbf{d}_p$ with $\mathbf{d}_p^{\text{thp}}$. For THP, we have the following equation

$$f_{\mathbf{M}_p}\left(\mathbf{B}_p{}^{-1} \cdot (\mathbf{a}_p - \mathbf{d}_p^{\text{thp}}) \right) = \mathbf{B}_p{}^{-1} \cdot (\mathbf{a}_p - \mathbf{d}_p^{\text{thp}}) \tag{6.13}$$

where $\mathbf{M}_p \triangleq [M_1^{(p)},...,M_{n_p}^{(p)}]^T$ are the corresponding modulus values from the components of $\mathbf{a}_p$. Using $f_{\mathbf{M}_p}(\mathbf{a}_p) = \mathbf{a}_p$ and $f_{\mathbf{M}_p}(\mathbf{x}+\mathbf{d}_p^{\text{thp}}) = f_{\mathbf{M}_p}(\mathbf{x})$, we can have $d_{p,1}^{\text{thp}} = 0$, and a successive solution for the remaining components of $\mathbf{d}_p^{\text{thp}}$ through a forward substitution as

$$d_{p,i}^{\text{thp}} = a_i^{(p)} + \sum_{j=1}^{i-1}[\mathbf{B}_p{}^{-1}]_{ij}(a_j^{(p)} - d_{p,j}^{\text{thp}}) - f_{M_i^{(p)}}\left[a_i^{(p)} + \sum_{j=1}^{i-1}[\mathbf{B}_p{}^{-1}]_{ij}(a_j^{(p)} - d_{p,j}^{\text{thp}})\right],\ i = 2,...,n_p. \qquad (6.14)$$

The complex Fincke-Pohst algorithm with input lattice bounding and THP initialization is summarized in Appendix E at tone p.

6.2.3 Complex Sphere Precoder (CSP) for OFDM System

We now apply the modified complex Fincke-Pohst algorithm above to the OFDM system with precoding. In particular, we are interested in finding the $\mathbf{d}_p$, p=1,…,n_c, for both a low PAR and a low total average transmit power. For an OFDM system using a complex sphere precoder (CSP) over each tone, an optimum solution would require a search over a $\sum_{p=1}^{n_c} n_p$-dimensional complex lattice subject to a total transmit power constraint P_T, which is obviously not practical for a reasonable n_c value of tens or hundreds of tones. Therefore, we perform a suboptimal search over each tone and find the best $\tilde{\mathbf{d}}_p$ among all the searches. We consider a search successively over each tone using one of the following four criteria

$$\tilde{\mathbf{d}}_p = \arg\min_{\mathbf{d}_p}\left\{\text{tr}(\mathbf{Y}^H\mathbf{Y})\,\middle|\,\hat{\mathbf{d}}_q, q = 1,...,n_c; q \neq p\right\} \qquad (6.15a)$$

$$\tilde{\mathbf{d}}_p = \arg\min_{\mathbf{d}_p}\left\{\max_{\substack{1\le i\le n_t\\ 1\le j\le n_c}} |[\mathbf{Y}]_{ij}|^2\,\middle|\,\hat{\mathbf{d}}_q, q = 1,...,n_c; q \neq p\right\} \qquad (6.15b)$$

$$\tilde{\mathbf{d}}_p = \arg\min_{\mathbf{d}_p}\left\{\text{tr}(\mathbf{Y}^H\mathbf{Y})\max_{\substack{1\le i\le n_t\\ 1\le j\le n_c}} |[\mathbf{Y}]_{ij}|^2\,\middle|\,\hat{\mathbf{d}}_q, q = 1,...,n_c; q \neq p\right\} \qquad (6.15c)$$

$$\tilde{\mathbf{d}}_p = \arg\min_{\mathbf{d}_p} \left\{ \max_{\substack{1\le i\le n_t \\ 1\le j\le n_c}} |[\mathbf{Y}]_{ij}|^2 / \mathrm{tr}(\mathbf{Y}^H\mathbf{Y}) \middle| \hat{\mathbf{d}}_q, q=1,...,n_c; q\neq p \right\} \tag{6.15d}$$

for $p=1,...,n_c$, where $\mathbf{Y} \triangleq [\mathbf{y}_1,...,\mathbf{y}_{n_t}]^T = \left(\mathbf{F}_{\mathrm{u}}{}^H \cdot [\mathbf{x}_1,...,\mathbf{x}_{n_c}]^T\right)^T$ is an $n_t \times n_c$ matrix consisting IFFT outputs over n_t transmit antennas, with the n_c symbols at the m-th antenna given as

$$\mathbf{y}_m = \sum_{p=1}^{n_c} \mathbf{f}_u^{(p)} \sum_{i=1}^{n_p} (a_i^{(p)} - d_i^{(p)}) [\mathbf{F}_p \mathbf{B}_p{}^{-1}]_{mi}, \quad m=1,...,n_t, \tag{6.16}$$

The $n_c \times n_c$ matrix $\mathbf{F}_{\mathrm{u}}$ is the normalized (unitary) Fourier matrix with its elements defined as

$$[\mathbf{F}_{\mathrm{u}}]_{kn} \triangleq \frac{1}{\sqrt{n_c}} \exp\{-j2\pi(k-1)(n-1)/n_c\}, \quad k,n=1,...,n_c, \tag{6.17}$$

and $\mathbf{F}_{\mathrm{u}}^H \triangleq [\mathbf{f}_{\mathrm{u}}^{(1)},...,\mathbf{f}_{\mathrm{u}}^{(n_c)}]$ in a column partition. The precoding sequences $\hat{\mathbf{d}}_q$ are set to

$$\hat{\mathbf{d}}_q = \begin{cases} \tilde{\mathbf{d}}_q, & q=1,...,p-1 \\ \mathbf{d}_q^{\mathrm{thp}}, & q=p+1,...,n_c \end{cases}. \tag{6.18}$$

The minimized metrics in (6.15a) to (6.15d) correspond to (a) the total transmit power (min AP), (b) the maximum peak power (min-max PP), (c) the product of total transmit power and maximum peak power (min-max AP*PP), and (d) the maximum PAR over each OFDM space-time block (min-max PAR), respectively. The minimization of these metrics is performed over each tone sequentially using the algorithm in Appendix E, given the precoding sequences over all the other tones. The precoding sequences are initialized with the THP solution for each tone. It should be pointed out that each of the sequential searches in (6.15a) to (6.15d) uses a natural order of tones form tone 1 to tone n_c, and no effort was made to search in a specific order. In fact, for a multiuser OFDM system with unequal search boundaries at different tones, we can reorder the tones according to

$$d_{\mathrm{ub}}^{(p)} = \sum_{i=1}^{n_p} \left\{ \frac{1 + \sum_{j=1}^{i-1} \left[|\mathrm{Re}([\mathbf{B}_p{}^{-1}]_{ij})| + |\mathrm{Im}([\mathbf{B}_p{}^{-1}]_{ij})| \right]}{\min_{1\le j\le i-1} \left(1, |\mathrm{Re}([\mathbf{B}_p{}^{-1}]_{ij})|, |\mathrm{Im}([\mathbf{B}_p{}^{-1}]_{ij})|\right)} \right\}, \quad p=1,...,n_c, \tag{6.19}$$

and perform the sequential search over tones in an ascending order of (6.19). The $d_{\text{ub}}^{(p)}$ in (6.19) is an upper bound on $\sum_{i=1}^{n_p}\left|d_i^{(p)}\right|$ with $d_i^{(p)}$ being normalized by $M_i^{(p)}$ such that $c_i^{(p)}$ are confined to Voronoi regions at tone p. A sequential search in such an order gives the system an overall high potential for PAR reduction over tones with large search boundaries. The performance effect of this search ordering will be studied in the next section through simulation. Finally, we note that the above metrics also hold for single-carrier transmission, for which the output matrix $\mathbf{Y}$ can be replaced with $\mathbf{x}$, the precoder output over a single carrier. In [83], the total transmit power was minimized based on (6.15a) for a single-carrier transmission using a multiuser channel inversion, which can be further simplified as shown in [80], [84].

We reiterate that for the fixed total transmit power, the efficiency of the CSP is closely tied to the lattice search boundary value $L_{\max}^{(p)}(i)$ over each tone. For a fixed search boundary, it has been reported in [81] that the complex Fincke-Pohst algorithm has a complexity roughly cubic in n_p for an n_p value less than 10, a dramatic improvement from an exhaustive search which is exponential in n_p over tone p. Therefore, with a moderate number of transmit antennas in an OFDM system, the suboptimal sequential search of CSP over all tones in (6.15) has a complexity roughly on the order of $O(n_c n_t^3)$.

6.2.4 Average Transmit Power and PAR Comparison

This section demonstrates the performance of CSP for MIMO precoding in both a single-carrier and an OFDM system. Specifically, we show the PAR and the total average transmit power of each system using CSP. The performance of a system using the criteria (6.15a)-(6.15d) are contrasted with each other and with the THP system proposed in [39] and [53]. Due to the symmetry of systems, we consider the PAR metrics averaged over the

multiple transmit antennas and the total transmit power in the following.

Figures 6.4 and 6.5 depict the CSP performance for a single carrier in flat fading. The channel coefficients over fading blocks are assumed i.i.d. complex circular symmetric Gaussian with unit variance.

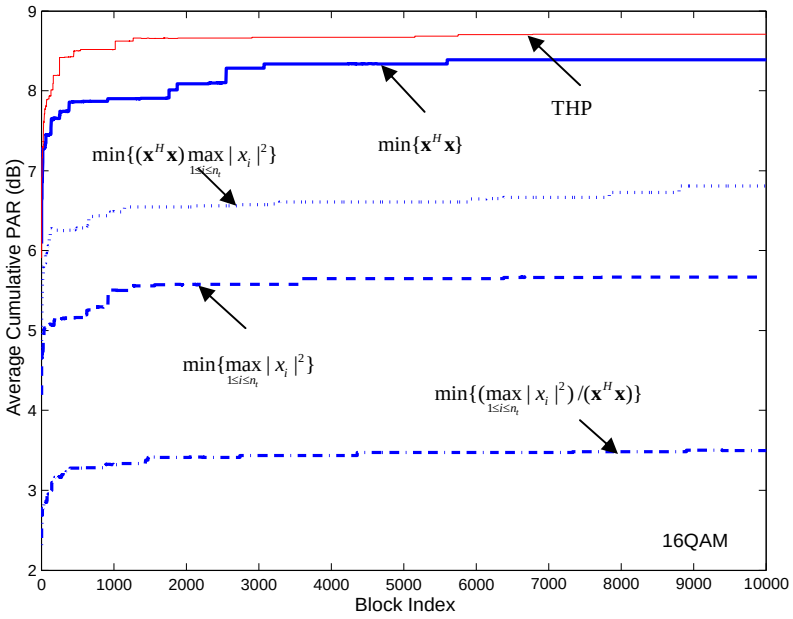
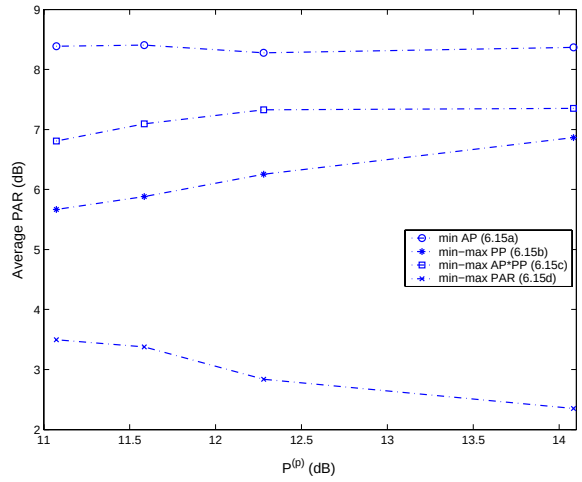
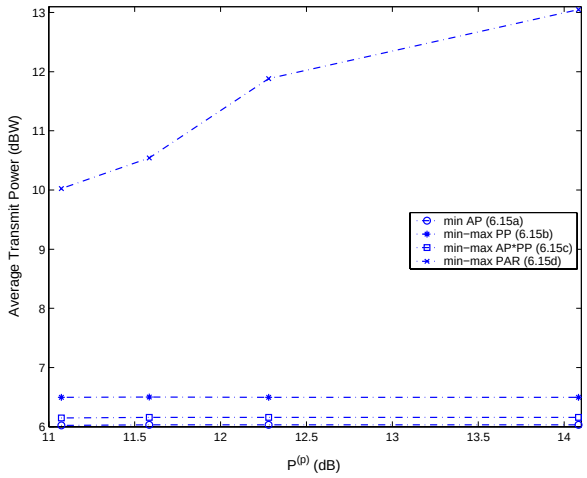


Figure 6.4. Cumulative PAR averaged over 4 transmit antennas for 16-QAM signals.

Figure 6.4 shows the average cumulative PAR of CSP and THP for 16-QAM signals from 4 transmit antennas to 4 fixed users over 10,000 blocks, with 100 symbols over each block. We note that even over a single carrier in flat fading, THP has an average PAR close to 9dB. By minimizing transmit power (criterion (6.15a)), the average PAR is still over 8dB. When accounting for the peak power constraint in the lattice search, average PAR reduction of about 2dB, 1dB, and 4dB are obtained for the criteria (6.15b)-(6.15d), respectively. The average transmit power of THP and CSPs using (6.15a)-(6.15d) are, respectively, 6.2dBW, 6.0dBW, 6.5dBW, 6.2dBW, and 10.0dBW for the input $\mathbf{a}$ of 0dBW. A normalized input upper



(a) Average PAR vs. $P^{(p)}$.



(b) Average transmit power vs. $P^{(p)}$.

Figure 6.5. Average PAR and transmit power of CSP for 16-QAM signals.

bound $L_{\text{ub}}^{(p)} = 100$ on the lattice provides a good compromise between performance and search time in this case. We note that a minimized maximum-peak-power based on criterion (6.15b) achieves an average PAR of less than 6dB, while it still has an average transmit power less than 0.5dB above the minimum-power case (6.15a). Similar performance has been observed for 4-QAM and 64-QAM signals. In Figure 6.4, the total transmit power constraint P_T in (6.8)

is specified as the squared radius of an n_t-dimensional complex sphere circumscribing about the Voronoi region (n_t-dimensional complex cube for square QAM signals) of inputs. This is the minimum power constraint in the form of (6.8) to preserve the input signals. For 4-dimensional complex 16-QAM inputs of unit average symbol energy (i.e., 1-Watt average power), the minimum P_T is 12.8 Watts (11.07dBW).

Figures 6.5(a) and (b) show the impact of increasing the power constraint P_T on PAR and average transmit power, respectively, for 4-dimensional complex 16-QAM inputs of unit symbol energy. It can be seen that as P_T increases, the PAR of (6.15b) increases, but the PAR of (6.15d) decreases, both change by 1 dB. However, the 1-dB PAR reduction using (6.15d) is accompanied by an approximate 8dB boost in the average transmit power, while the other cases have no noticeable change in the average transmit power. The performance using the criterion (6.15a) is insensitive to the change in P_T, while that using (6.15c) is in-between (6.15a) and (6.15b). Based on these observations, in the following we use the minimum transmit power constraint P_T over each tone for the overall best performance.

Figure 6.6 depicts the average cumulative PARs for a MIMO-OFDM system using THP and CSP over each tone as illustrated in Figure 6.2. For simplicity, we assume an equal transmit power constraint and a maximum number of users over each tone, i.e., $P^{(p)} = P_T / n_c$ and $n_p = n_t$ for $p = 1,...,n_c$, where $n_t = 4$, $n_c = 128$, and $P_T = 11.07$ dBW. The power delay profile (PDP) of each user is assumed to have 8 resolvable and uniformly symbol-spaced paths of exponentially decreasing power, with each tap being 3.1dB lower than the previous tap for a unit total power. Compared with Figure 6.4, a much larger PAR reduction has been obtained by using the suboptimal sequential single-tone search in Section 6.2.3. In Figure 6.6, a reordered sequential tone-search is used in an ascending order of $d_{\text{ub}}^{(p)}$. Both THP and CSP using (6.15a) lead to a PAR of close to 12 dB, and CSPs using (6.15b)-(6.14d) lead to a PAR

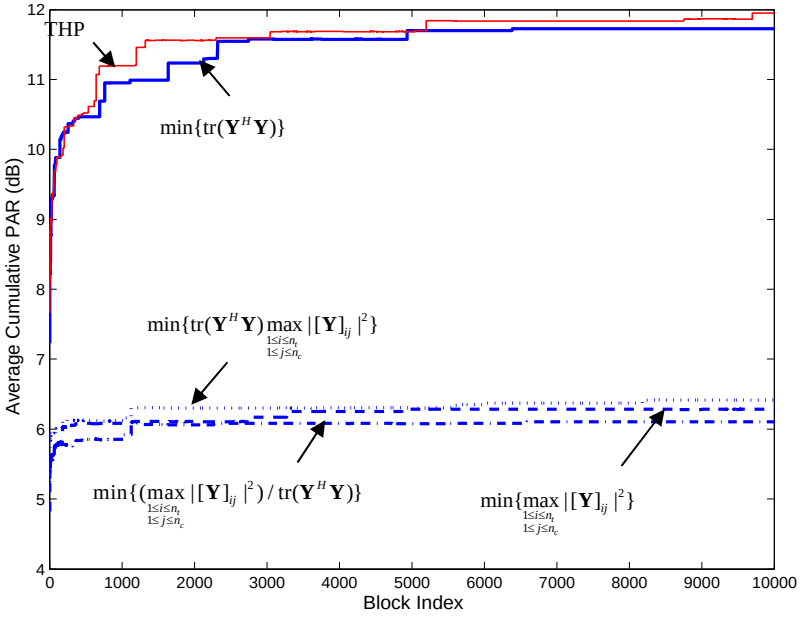


Figure 6.6. Average cumulative PAR of CSP in an OFDM system.

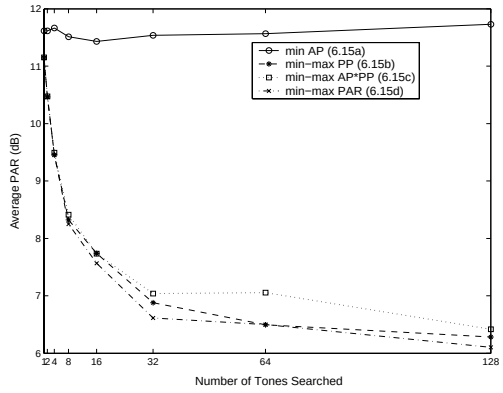
less than 6.5 dB over 10,000 OFDM space-time blocks, more than a 5dB PAR reduction. We note that with the suboptimal search, the PARs for (6.15b)-(6.15d) approach each other due to a search over a subspace only. The total average transmit power for THP and CSPs using (6.15a)-(6.15d) are, respectively, 6.2dBW, 6.0dBW, 6.9dBW, 6.7dBW, and 7.2dBW. Therefore, compared with a comprehensive search in a single-carrier system, the average transmit powers of (6.15b) and (6.15c) are increased slightly, exceeding the minimum power by about 0.9dB and 0.7dB, respectively, and the power boost of (6.15d) is mitigated to about 1.2dB above that of minimum-power case (6.15a). The total average transmit power of THP is the same in both systems, about 0.2dB above the minimum power using CSP. Over each tone, a lattice bound $L_{\mathrm{ub}}^{(p)}$ of 100 is used, and we see that this is sufficient for significant performance improvement while still keeping the simulation time reasonable.

The CSP with reduced-tone search has also been simulated using the criteria in

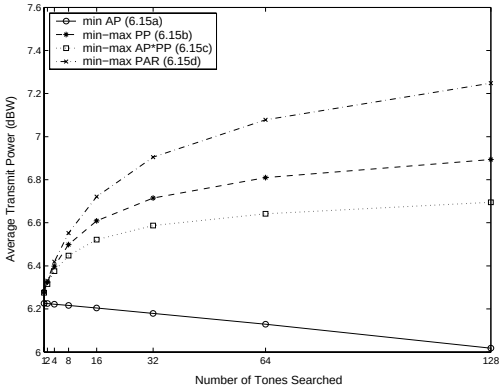
(6.15a)-(6.15d), where only a fraction of tones with the largest $d_{\text{ub}}^{(p)}$ are searched in an ascending order of $d_{\text{ub}}^{(p)}$. The corresponding PAR and total average power are shown in Figures 6.7(a) and (b) as a function of the number of tones searched for the same scenario as Figure 6.6. We note that as the number of searched tones increases, there is no apparent change in PAR with minimized transmit power criterion (6.15a), but the PARs of other criteria decrease dramatically. Meanwhile, the total average transmit power of (6.15a) decreases slightly (about 0.2dB), while those of (6.15b)-(6.15d) increase respectively by approximately 0.4 dB, 0.6 dB, and 1.0 dB, as the number of searched tones increases from 1 to 128. From Figure 6.7(a), we see that for the system under consideration, a sequential CSP using (6.15b) over 32 tones with the largest $d_{\text{ub}}^{(p)}$ keeps the average PAR below 7dB and an average transmit power only about 0.7dB above the minimum power using full-tone sequential search. The same experiment has been performed using a natural ordering of tones (without reordering), and the result is shown in Figure 6.7(c). We note that compared with reordered tones, over 1 dB PAR improvement can be achieved for the cases (6.15b) to (6.15d) with reduced tone search, and there is no improvement in the case of (6.15a). No noticeable change was observed in the total average transmit power in all cases. Similar performance improvement has been observed for user channels of 8 equal-gain multipaths.

6.3 Suboptimal Scheduling Performance

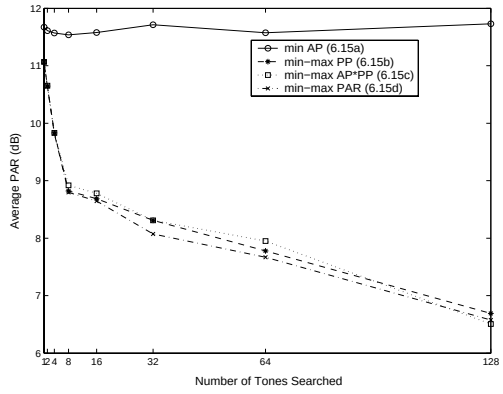
We extend the suboptimal scheduling algorithms in Chapter 4 to a precoded OFDM system on a per-tone basis. For simplicity, we consider the system performance in terms of achievable sum rate and average error rate in a single cell setting, where suboptimal precoding and scheduling are implemented at the base station. Following the precoding



(a) PAR with tone reordering.



(b) Average transmit power.



(c) PAR without tone reordering.

Figure 6.7. PAR and average transmit power vs. the number of tones searched.

discussion, the performance is the same for both the simple THP and the more sophisticated CSP in an OFDM system.

6.3.1 Achievable Sum Rate

We first present the average achievable sum rate of the system. The OFDM system under study has the same system parameters as those in the previous section, where an equal transmit power constraint is applied to each data tone. According to the 802.11/Hiperlan2 standards, we presuppose that each OFDM block is composed of 64 tones, among which 47 are for the user data. For simplicity, symmetric user fading channels are assumed, where each user channel has 8-path exponential PDP, same as that in the previous section. What differs from the previous section lies in the OFDM frame structure. In this section, we assume each OFDM frame consists of 47 OFDM blocks[1], where each block has 47 tones devoted to user data. The central 17 tones are reserved for control and channel estimation over each OFDM block. The channel frequency responses over each OFDM frame are assumed static, therefore the previous block-wise static fading is extended here to the frame-wise channel fading in the frequency domain. As before, the total number of transmit antennas is set to 4 and each user has a single receive antenna. Figure 6.8 depicts a time-frequency frame of 5 data tones and 5 blocks used at each transmit antenna.

Figure 6.9 illustrates the total average achievable sum rate verses the average per-tone transmit SNR, P_T/N_0 , using the suboptimal scheduler algorithm in Table 4.1 and the suboptimal scheduling and reordering algorithm in Table 4.3 over each data tone.

The simulation is over 1000 OFDM frames, or 47,000 OFDM blocks with N_{dat}=47 data symbols over each block. For comparison, the average sum rate of the suboptimal

[1] The frame structure assumed here is consistent with Chapter 7 for the reason given in Chapter 7.

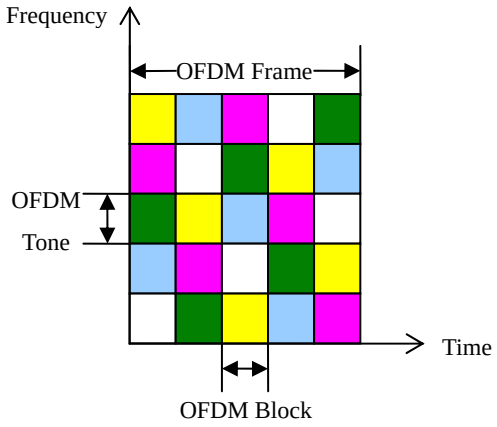


Figure 6.8. OFDM frame structure in the time-frequency domain.

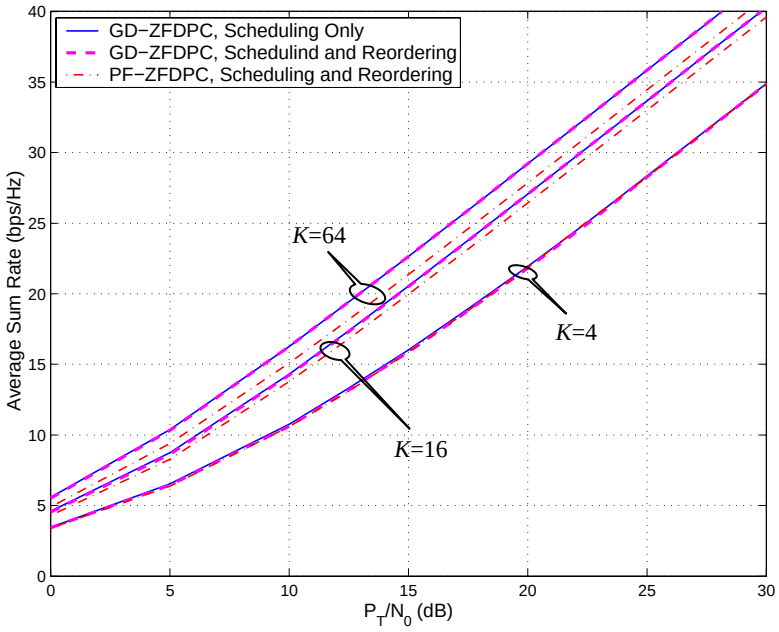


Figure 6.9. Average sum rate of ZF-DPC in the OFDM system.

scheduling and reordering under the proportional fair criterion is also plotted. For PF scheduler, the window-average throughput of user k, $T_i(t)$, is updated block-wise as

$$T_k(t+1) = \begin{cases} (1-1/W_c)T_k(t) + (1/W_c)\sum_{\substack{\text{tone } p \in \text{user } k, \\ p=1,\ldots,n_c}} R_k^p(t), & k \in \bigcup_{p=1}^{n_c} \mathcal{S}_{\text{PF}}^p \\ (1-1/W_c)T_k(t), & \text{otherwise} \end{cases},$$

where $\mathcal{S}_{\text{PF}}^p$ is the user subset satisfying

$$\mathcal{S}_{\text{PF}}^p = \arg\max_{\mathcal{S}} \sum_{k \in \mathcal{S}} \frac{R_k^p(t)}{T_k(t)}$$

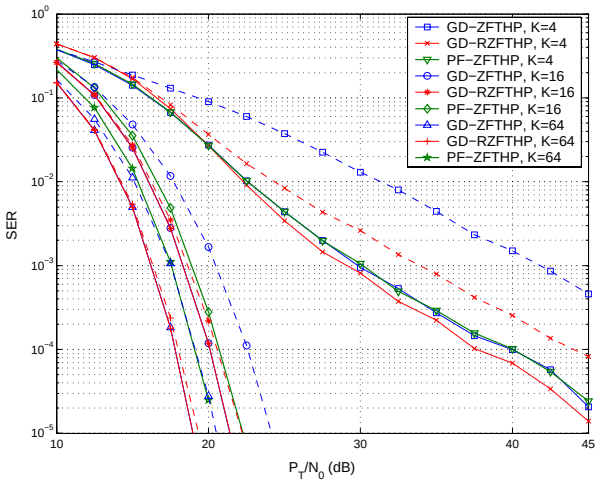
among all the user subsets of $|\mathcal{S}| \le n_t$ over tone p and at OFDM block t, and $R_k^p(t)$ is the achievable rate of user k on tone p at OFDM block t. The window length W_c is set to 100 OFDM blocks in the simulation. We note from Figure 6.9 that in the OFDM system using ZF-DPC, the average sum rate suffers negligible loss from both the minimum-error reordering (provided in Chapter 4) and PF scheduling when the number of active users is less than the number of data tones. This result holds even though the sum-rate loss due to PF scheduling increases slightly with the number of users. Nevertheless, as the number of active users exceeds the number of data tones, e.g., K=64 for N_{dat}=47, a noticeable sum-rate loss (about 1dB in average transmit power at high SNR) can be observed under PF scheduling. Comparing Figure 6.9 with Figures 5.4(a) and (b) in a flat fading channel, we conclude that under the proportional fairness criterion, ZF-DPC using suboptimal scheduling and reordering still achieves a significant amount of sum rate improvement and benefits greatly from the multiuser diversity. As revealed in Figure 6.9, with n_t=4, a scheduling gain of approximately 3 to 4dB can be obtained for 16 users over 4 users at an average sum rate of 15bps/Hz in all cases simulated, and it reaches about 5dB for 64 users. This scheduling gain grows with the average transmit power constraint, and saturates at very high SNR values, which confirms our conclusion in Chapter 4 that, at high SNR, the throughput is constrained by the maximum degrees of freedom, the number of transmit antennas in this case.

6.3.2 Average Error Rate

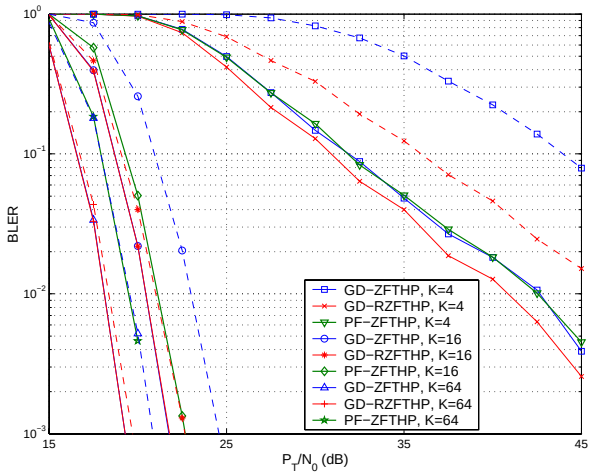
The average SER and BLER of the THP-OFDM system are shown in Figures 6.10(a) and (b). The plots are very close to a single-carrier system in flat fading as shown in Figure 5.10 except that a different block length is used here. Also plotted in the figures are the error rates of the ZF-THP using the suboptimal PF scheduler with the minimum-error reordering. As can be seen, in all cases, the suboptimal PF scheduler with reordering obtains about the same error-rate slopes as the corresponding suboptimal greedy scheduler with reordering. For ZF-THP, a direct comparison of PF and greedy schedulers with reordering indicates about a 0.7dB SNR loss at K=16, and less than 1.5dB SNR loss for K=64, which can be interpreted as coding-gain losses. No apparent error-rate difference is observed at K=4. Just as with the average sum rate, the impact of multiuser diversity on error rates rapidly saturates as the number of users increases. But unlike the sum rate, the minimum-error reordering is more beneficial when the number of users is small. The same error rates can be obtained by using the complex sphere precoder to reduce the PAR of an OFDM system.

6.4 Summary

This chapter presented a complex sphere precoding algorithm as a generalization of T-H precoding. The CSP achieves the same throughput and error performance as THP while obtaining low values for average transmit power and transmit PAR. Performance simulations of CSP indicate that more than 3dB PAR reduction is possible by using a reordered and sequential search over half of the data tones. A direct extension of the previous suboptimal scheduling and reordering algorithm is readily available to the OFDM system on a per-tonebasis, and similar performance can be obtained as in flat fading channels. Moreover, simulation results showed that with the proportional resource fairness, the degradation of



(a) Average SER.



(b) Average BLER.

Figure 6.10. Total average error rate of THP-OFDM. Solid curves: scheduling and reordering; dashed curves: scheduling only.

system performance in terms of total throughout and average error rate is insignificant, a result ascribed to the multiplexing in both spatial and frequency domains.

Chapter 7
Comparison of Two Design Philosophies for Multicell OFDM Systems

In a multicell environment where the downlink system performance is limited by interference, the design of MAC protocols is essential for an efficient mulituser transmission. This chapter investigates the MAC-layer impacts on the system-level performance of an OFDM system in a multicell setting, which employs the nonlinear precoding for spatial multiplexing as described in the preceding chapters. We compare and contrast the impact of interference avoidance by channel-aware scheduling and that of interference averaging from orthogonal frequency hopping through simulation.

7.1 Frequency-Hopped OFDM System (FH-OFDM)

We first describe a frequency-hopped OFDM (FH-OFDM) cellular system based on time division multiple access (TDMA). A TDMA-based FH-OFDM also provide the multiple access basis for both the IEEE802.11a [85] local area network (LAN) standard and the IEEE 802.16a [86] metropolitan area network (MAN) standard. For the cellular architecture under consideration, a set of orthogonal frequency hopping patterns are adopted at different base stations to provide an effect of interference averaging. These mutually orthogonal tone hopping patterns at interfering transmitters in fact provide a topology-transparent TDMA scheduling design, as proposed originally for multihop packet radio networks with unicast traffic in [87] and later were extended for broadcast traffic in [88].

The same frame structure as in Figure 6.8 is assumed for the OFDM system in this chapter. In a multicell system, the basic idea is that we hop the tones at neighboring transmitters such that each transmitter sees co-channel interference from each neighboring transmitter exactly once over an OFDM frame and the interference is widely spread over tones, an effect called *interference diversity* [7]. To ensure the minimum time-and-tone overlap among neighboring transmitters, we use the mutually orthogonal latin squares (MOLS) [89] as the periodic hopping patterns for the transmitters, which repeat every OFDM frame of N_{dat} OFDM blocks. The construction of MOLSs based on Galois Field theory can be found in [90]. However, it turns out that when N_{dat} is prime, there is a simple construction of a family of N_{dat}-1 mutually orthogonal latin squares, with the entries (virtual channel indices) of the m-th square, $m = 1, ..., N_{\text{dat}} - 1$, are specified as [91]

$$[R^m]_{i,j} = (m \cdot i + j) \bmod N_{\text{dat}}, \; i, j = 0, ..., N_{\text{dat}} - 1 \qquad (7.1)$$

where $[R^m]_{i,j}$ is the (i, j)-th entry of the matrix R^m, and $(x) \bmod y$ denotes x modulo y. Figure 7.1 illustrates a set of 4 mutually orthogonal latin squares using the modulo-based construction (7.1), each with 5 tones and 5 OFDM blocks for an OFDM frame.

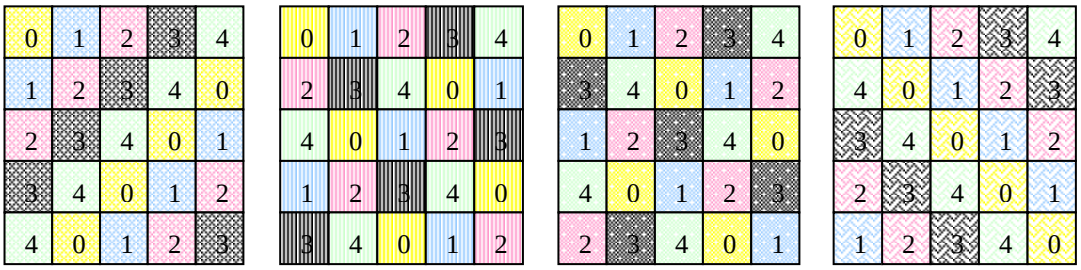


Figure 7.1. Mutually orthogonal latin squares of 5 tones and 5 OFDM blocks.

For each latin square, the rows corresponds to physical tones and the columns represent OFDM blocks (or OFDM symbols). The entries represent "virtual channels" that use the corresponding tones at a particular OFDM symbol time. Assuming perfect tone

separation, there is no in-cell co-channel interference. We note that within a latin-square hopping pattern, each virtual channel hops over different physical tones at different OFDM symbols and spans all the tones over each OFDM block, and consequently exploits frequency diversity of wideband system. These latin squares are mutually orthogonal in the sense that, for each pair of latin squares, all their ordered pair entries (*i*, *j*) are different for a total of N^2 such entries, with *N* being the size of each latin square [89]. The interference diversity through such orthogonal hopping patterns at interfering transmitters therefore ensures that no single strong interference from a virtual channel dominates the performance degradation and the full interference diversity is harnessed (similar to the interference averaging used in CDMA).

For a MIMO-OFDM system with the nonlinear precoding as discussed in Chapter 6, the maximum number of simultaneous transmissions at each base station is $n_t \cdot N_{\text{dat}}$, with n_t being the number of transmit antennas. Suppose we associate a distinct set of users, say *m* users, to each virtual channel, then each transmitter can support up to $\min(m, n_t) \cdot N_{\text{dat}}$ users simultaneously. For maximum interference averaging, the total number of active users, *K*, must meet $K \geq m \cdot N_{\text{dat}}$, so that each virtual channel is allocated a distinct set of users. In the following simulations, we assume that the number of pre-allocated users per virtual channel, *m*, satisfies $m = n_t$, and the total number of users $K = m \cdot N_{\text{dat}}$. Specifically, we assign the *i*-th virtual channel, $i = 0, ..., N_{\text{dat}} - 1$, to the n_t users indexed by

$$\left((i \cdot n_t) \bmod K + k\right) \bmod K = \left(i \cdot n_t + k\right) \bmod K, \quad k = 0, ..., n_t - 1,$$

which amounts to a round-robin multiuser allocation with a number of n_t users to each virtual channel. With the orthogonal frequency hopping based on MOLS, an additional advantage of the precoded OFDM system is that, over each tone, only the channels of the pre-allocated n_t

users need to be estimated, rather than all the user channels as required in scheduling among all the available users. For a single-antenna system without user channel knowledge at the transmitter, this orthogonal frequency-hopped OFDM system overcomes the devastating interference problems with both the purely TDMA-based narrowband system (e.g., GSM) and the wideband CDMA system (e.g., cdma2000) of full frequency reuse. The former has its spectral efficiency limited by the out-of-cell co-channel interference, while the latter suffers from a generally lower SINR after averaging both in-cell and out-of-cell co-channel interference.

7.2 Opportunistic Interference-Nulling in OFDM System

The concept of opportunistic interference-nulling was first described in [33] in the context of opportunistic beamforming. For a MISO broadcast channel with out-of-cell co-channel interference, the authors pointed out that beamforming to the best user alone at each time slot not only exploits the significant multiuser diversity gain in a single cell, but also opportunistically nulls out the out-of-cell interference in a multicell deployment. In this section, we apply the opportunistic interference-nulling to a precoded OFDM system in a MISO broadcast channel. However, more appropriately in our context, we view the opportunistic interference nulling as a special form of interference avoidance as provided by packet scheduling. Correspondingly, we name the OFDM system with opportunistic interference-nulling as the OIN-OFDM system. Previous results [92] in a capacity study of single-carrier wireless systems indicate that interference averaging techniques can perform better than fixed channel assignment technique, whereas interference avoidance techniques can outperform interference averaging techniques by a factor of 2-3 in spectrum efficiency. Similar performance gain has also been observed in an OFDM system with dynamic packet

assignment [79]. Furthermore, it has been shown that even without power control, interference avoidance can outperform interference averaging with power control in a CDMA system [92]. This result is particular advantageous for packet transmission when effective power control is problematic due to the rapid arrival and departure of interfering packets.

Now consider the wideband OFDM system of Chapter 6 in an interference-limited environment, the multiuser diversity of a large network is exploited through the channel-aware scheduling at the transmitter over each data tone. As for the FH-OFDM system in the previous section, the scheduling and reordering algorithms for the OIN-OFDM system use the received SINR of users over each data tone. The difference lies in that, for the OIN-OFDM system, the scheduling is performed over all the active users on each data tone, rather than only among the pre-assigned users as in an FH-OFDM system. Therefore, the OIN-OFDM system assumes that all the user channel responses at all N_{dat} tones are available to the transmitter over each OFDM frame, which are often more costly than only knowing channels of the pre-assigned users at each virtual channel in an FH-OFDM system, particular for a number of active users. With a fully channel-aware scheduler, this channel knowledge requirement also may lead to a very low per-tone SNR when N_{dat} is large for a fixed total transmitter power, hence poor channel estimation performance over each data tone. To ease the channel estimation, we can restrict the number of tones over each OFDM block for each user receiver, therefore ensuring a minimum per-tone transmit power allocated. In a multiuser network, the unused tones for one receiver can be dynamically used by others, therefore need not go wasted. Together with multiuser scheduling, this dynamic tone allocation at the transmitter can also improve the robustness of system in adverse communication environment.

7.3 Multicell Simulation Results

We simulate the total average sum rate of a spatially precoded OFDM system with interfering base stations surrounding the interested base station over 1000 OFDM frames. We limit computations by using 6 interfering base stations, gathering sum-rate and error-rate statistics only for the center cell. The use of a small number of interfering transmitters in the downlink channel can be justified by the observation that, for the downlink transmission, performance is typically dominated by a few close high-power base stations in a wideband system with full frequency reuse [93]. This "law of small numbers" actually makes the downlink a performance bottleneck in a CDMA system [7], rather than the uplink as was initially widely believed.

In addition, we assume a high-speed data system with a far denser deployment of base stations in our simulations. Under this assumption, it is unreasonable to stipulate that the base stations be located high above the ground with minimal local scattering. In an urban environment, chances are that there is substantial local scattering around a base station and the gain of sectorization is minimal. Thus, users in a sector see interference from the same base station intended for another sector due to local scattering. For this reason, we presuppose omni-directional antennas at each base station. As a matter of fact, the gain of sectorization can be automatically realized using adaptive scheduling in this case as an extension of opportunistic beamforming [33]. Another reason for not using sectorization is that, in a system with full frequency reuse across both cells and sectors, sectorization tends to reduce the dynamic range of user SINR. Consequently, sectorization creates a negative interaction with channel-aware scheduling, similar to the interaction presented in Chapter 3 between diversity and scheduling.

Both the suboptimal scheduling and the minimum-error reordering are performed

over each tone using the received SINR for each user, instead of the SNR as in a single cell. However, for FH-OFDM, the scheduling over each virtual channel is confined to the pre-assigned users to that virtual channel, while all the active users are considered by the scheduler for OIN-OFDM on each data tone. To simulate an interference-limited system, the user average power losses from all the interfering base stations are assumed uniformly distributed on [20, 40] dB, so that the average out-of-cell co-channel interference from each interfering transmit antenna is 20-40 dB below the average transmit signals of the base station of interest. Since we are only interested in the co-channel interference impact on the system performance at relatively high SNR in an interference-limited environment, we consider the composite large-scale channel fading (path loss and shadowing) by only specifying a single long-term average power attenuation, while including frame-wise small-scale fading in the frequency domain. The co-channel interference is generated at each interfering base station using scheduling and reordering.

A worst interference scenario is simulated by assuming full-power and continuous transmissions of all the interfering base stations. As in Chapter 6, we use 47 data tones (N_{dat}=47) among the total 64 tones available. It is assumed that each base station transmitter has 4 transmit antennas (n_t=4) while each user receiver has a single receive antenna. Figure 7.2 depicts the average sum rate vs. the average transmit SNR per tone (denoted as P_T/N_0 here), for both the OFDM systems with out-of-cell co-channel interference. These are the achievable sum rates assuming Gaussian signals and interference in AWGN. For comparison, the average sum rates with a single transmit antenna (n_t=1) are also plotted. We use m=n_t and $K = m \cdot N_{\text{dat}}$ for maximum interference averaging with FH-OFDM, and the same number of users for multiuser diversity with OIN-OFDM. Both the suboptimal greedy scheduler and the PF scheduler are simulated, with the minimum-error reordering performed for n_t>1. As shown in Figure 7.2, as the average transmit SNR at each base station increases, the total

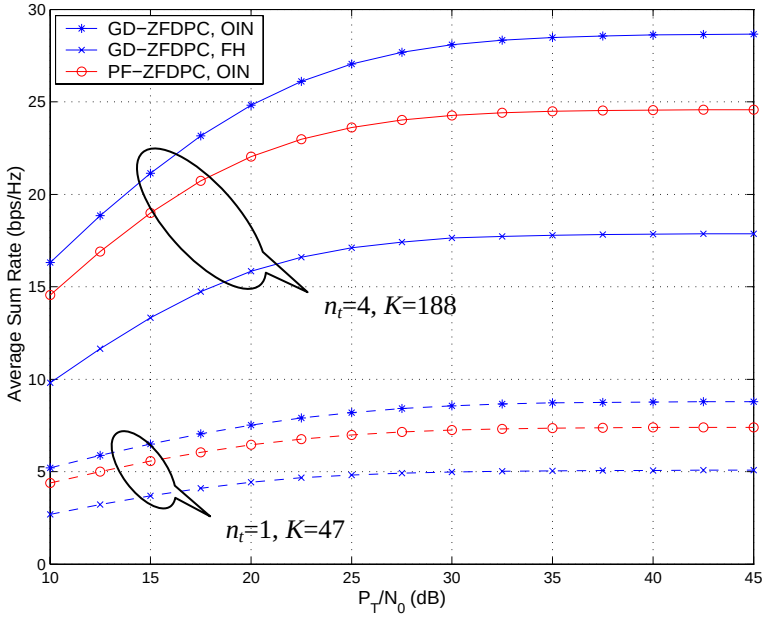
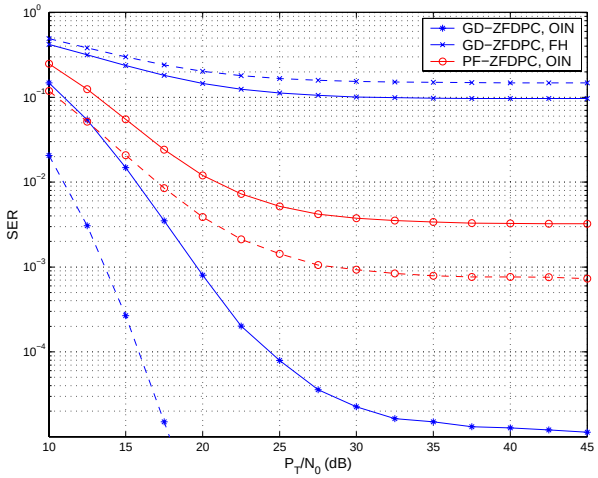


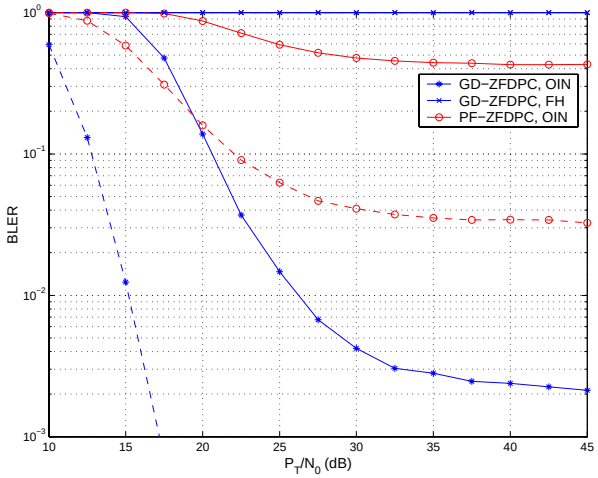
Figure 7.2. Average sum rates of OIN-OFDM and FH-OFDM using ZF-DPC.

average sum rates of both systems are constrained by the unknown out-of-cell interference. However, the total achievable throughput is dominated by the multiuser scheduling gain, rather than by interference averaging. The throughput advantage of OIN-OFDM remains even with a proportional fairness constraint. At high SNR when the performance becomes seriously interference-limited, average sum rate gains of 2.3 bps/Hz and 6.7 bps/Hz can be obtained by OIN-OFDM with PF scheduler over the FH-OFDM for n_t=1 and n_t=4, respectively. Furthermore, the larger the number of transmit antennas, the greater the throughput benefit from interference avoidance.

The corresponding SER and BLER using ZF-THP are plotted in Figures 7.3(a) and (b). The error floors at high SNR indicate the interference-limited performance of the systems. Similar to the sum rates in Figure 7.2, the error rates benefit more from multiuser scheduling than from multitone hopping. However, unlike the sum rates in Figure 7.2, the OIN-OFDM system with n_t=1 and K=47 has the best error-rate performance, and the OIN-



(a) Average SER.



(b) Average BLER.

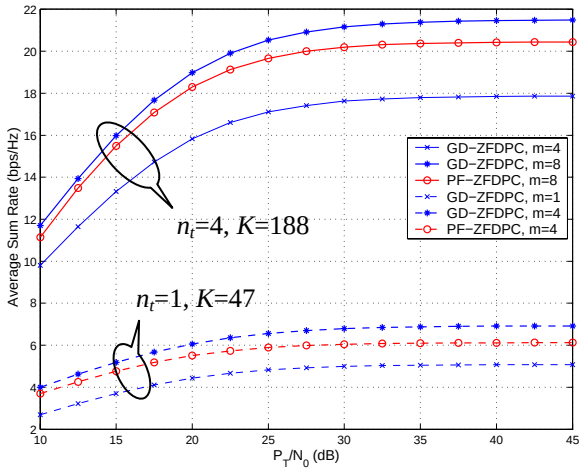
Figure 7.3. Average error rates of OIN-OFDM and FH-OFDM using ZF-THP. Solid curves: n_t=4, K=188; dashed curves: n_t=1, K=47.

OFDM with n_t=4 and K=188 next. This is due to the absence of spatial multiplexing, therefore the signal is free of in-cell co-channel interference over each data tone in the former case. This error-rate difference is maitained with a suboptimal PF scheduler. We note that the error-rate performance is very sensitive to the out-of-cell interference, even though this

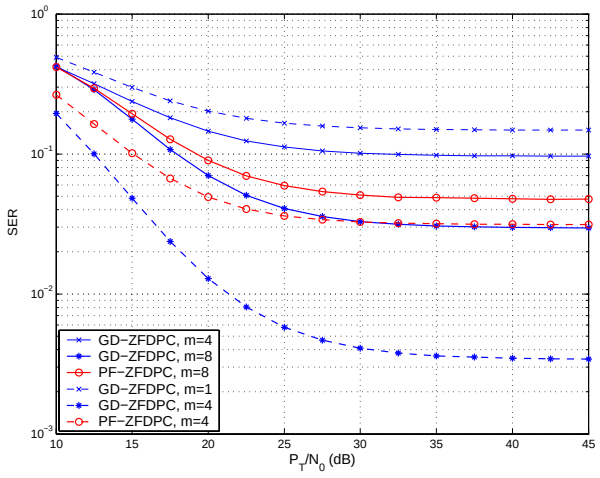
problem can be partly relieved with adaptive scheduling. For the FH-OFDM system, slightly lower average error rate is obtained with n_t=4 and K=188 than with n_t=1 and K=47, as the former provides better interference averaging with a larger number of users over virtual channel.

We have only considered the cases for $m = n_t$ and $K = m \cdot N_{\text{dat}}$. When $K < m \cdot N_{\text{dat}}$, the round-robin multiuser allocation still works for the FH-OFDM system, but with reduced interference averaging since multiple virtual channels may share the same user subsets. Therefore, the performance degradation tends to be dominated by a few virtual channels of some close interfering base stations, similar to the problem with a narrowband TDMA system. With a reduced number of users over each data tone, the OIN-OFDM suffers from a loss in multiuser diversity. However, owing to the rapid saturation of muliuser scheduling gain, a moderate number of users per tone often is sufficient to reap the scheduling benefit.

The above comparison also suggests that, in a multi-antenna broadcast channel, we can combine the interference averaging and interference avoidance. This can be achieved by pre-allocating a number $m>n_t$ users to each virtual channel in an FH-OFDM system and performing adaptive scheduling and reordering among the pre-allocated users over all virtual channels. As stated above, the system in this case has a reduced interference averaging effect. However, considering the rapid convergence of the multiuser scheduling gain, we can appropriately choose the m value and allow the system to enjoy both interference averaging and interference avoidance. Moreover, without full user scheduling over each tone, the system only needs to estimate the channel frequency responses of pre-assigned users on each virtual channel. Figure 7.4(a) shows the average sum rates for the FH-OFDM systems using adaptive multiuser scheduler and reordering over each virtual channel for n_t=4, m=8, K=188 and n_t=1, m=4, and K=47. The performance results without the multiuser scheduler are also shown for comparison. It can be seen from these plots that significant multiuser scheduling



(a) Average sum rate of FH-OFDM using ZF-DPC.



(b) Average SER of FH-OFDM using ZF-THP.

Figure 7.4. Rate and error performance of FH-OFDM systems with and without adaptive scheduler. Solid curves: n_t=4, K=188; dashed curves: n_t=1, K=47.

gain can be obtained by using $m>n_t$ users over each virtual channel. At n_t=4 and K=188, using m=8 users over each virtual channel can bring more than 5bps/Hz average sum rate improvement over pure FH-OFDM with m=4. The sum-rate improvements are apparent under both greedy and PF scheduling, and become more pronounced at n_t=4. The sum rate

can be further increased with larger *m*, but the price is that more channel estimation is needed, and poor interference diversity can lead to performance degradation dominated by a few strong interferers, especially under resource fairness constraints. The associated symbol error rates in Figure 7.4(b) also indicate a significant multiuser diversity gain from the adaptive scheduling, but the error-rate improvement over pure FH-OFDM is greater with n_t=1 and K=47 than with n_t=4 and K=188.

7.4 Summary

This chapter compared two design philosophies for wideband OFDM systems using spatial precoding. Simulations showed that the system performance is dominated by interference avoidance through adaptive multiuser scheduling, rather than by interference averaging from orthogonal multitone hopping. While the technique of interference averaging has proved to be very successful for CDMA-based 3G systems, its limitations are revealed by its incapability to flexibly avoid strong interference in a multiuser packet data network, which tends to dominate the system performance with parallel transmissions. Significantly greater capacity improvement was demonstrated even under suboptimal adaptive scheduling and with resource fairness constraints. To ease channel estimation and to leverage interference diversity, a tradeoff design was delineated which combined both frequency hopping and adaptive schedulin. Simulations indicated that this tradeoff effectively improved performance by granting the system some freedom to avoid multisuer interference. It should be mentioned that, while we only compared system performance using simple simulation models, such comparisons serve chiefly to motivate more detailed engineering investigation of what is actually required to achieve higher capacity.

Chapter 8
Conclusion

The spectrally efficient data transmission on lossy wireless links is beginning to take off as wireless communication evolves from a circuit-switched to a packet-based infrastructure. In this context, MIMO channels serve as short and fat pipes over the space dimension, and a joint design of space-time transmission and scheduling schemes is of great importance for future high-speed packet data services.

This dissertation illustrates the role of multiple antennas in a broadcast channel with multiuser scheduling. First, it is shown that, from a diversity perspective, the use of transmit antenna diversity based on orthogonal design actually reduces the total throughput in the presence of channel-aware scheduling. This negative interaction between open-loop antenna diversity and closed-loop mulituser diversity motivates us to study other roles of multiple antennas. Second, from a sum-capacity point of view, it is shown that the parallel transmission through a suboptimal dirty paper coding approach can realize a significant portion of the achievable throughput. An efficient channel-aware scheduling algorithm is designed to further increase the throughput through multiuser diversity, while keeping the average error rate to a minimum. Furthermore, the DPC-based pre-interference cancellation is shown to allow for a large scheduling gain and to avoid error propagation, with the latter being a serious problem for BLAST transmission. At medium to high SNR, a practical solution using Tomlinson-Harashima precoding and PF scheduling substantiates very high spectral efficiency and significant multiuser diversity gain.

Both the precoding and scheduling algorithms can be easily extended to a wideband

OFDM system in frequency-selective fading. Moreover, the concept of T-H precoding is generalized by using a modified complex version of the Fincke-Pohst algorithm. This generalization brings about 5dB reduction in transmit peak-to-average power ratio and generally a low average transmit power. Finally, in an interference-limited multi-cell environment, system performance benefits significantly more from interference avoidance than from pure interference averaging. For a MIMO-OFDM system with a large number of sub-carriers, the problem of channel estimation can be eased with a combination of both interference averaging and interference avoidance.

Overall, the key issue in a multiuser MIMO system is to provide an optimal radio resource sharing of various degrees of freedom, in time, frequency, or space domain, among multiple users. In this dissertation, for both the spatial multiplexing gain (through interference cancellation) at high SNR and the array gain (via coherent maximal-ratio combining) at low SNR, and that the scheduling gain from multiuser diversity, we stress the importance of channel knowledge at the transmitter for the achievement of these gains. In comparison with the traditional transmission schemes only with receiver-side channel knowledge, the performance gain presented in this dissertation warrants the additional implementation complexity for channel knowledge at the transmitter. Considering that communication systems typically evolve along a path from analog designs with minimal cooperation between transmitter and receiver, to designs involving sophisticated digital techniques and feedback controls, we anticipate that a system-wide information sharing among different layers and the efficient feed-forward and feedback between transmitter and receiver would be the future direction to take.

Appendix A: Proof of Lemma 4.1

Consider the scheduling algorithm in Table 4.1, we have the channel power gain of the first selected user as the maximum squared Euclidean norm of user channel vectors, i.e.,

$$\gamma_1^* = \max_{1 \le k \le K} \|\mathbf{h}_k\|_2^2 \tag{A.1}$$

According to the Gram-Schmidt algorithm [46], the best remaining user at loop n, $2 \le n \le \min(n_t, K)$, has a channel power gain of

$$\gamma_n = \max_{k \in \{1,\ldots,K\} \setminus \{1^*,\ldots,(n-1)^*\}} \left\| \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-1} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H \right) \mathbf{h}_k^H \right\|_2^2, \tag{A.2}$$

where $\mathbf{I}_{n_t}$ is the identity matrix of size n_t, and $\mathbf{q}_{i^*} \in \mathbb{C}^{n_t \times 1}$, $i^* \in \mathcal{S}^*$, are the orthonormal vectors in the subspace spanned by the selected user channel vectors, and are obtained from Gram-Schmidt orthogonalization. Since we have

$$\begin{aligned} \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-1} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H \right)^H &= \mathbf{I}_{n_t} - \sum_{i=1}^{n-1} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H \\ &= \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-1} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H \right) \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-1} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H \right), \quad 2 \le n \le \min(n_t, K) \end{aligned}$$

the $n_t \times n_t$ matrix $\mathbf{I}_{n_t} - \sum_{i=1}^{n-1} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H$ is Hermitian and idempotent, hence is an orthogonal projection matrix. From the properties of projection matrices, if $\mathbf{P}$ and $\mathbf{Q}$ are projection matrices and $\mathbf{P} - \mathbf{Q}$ is also a projection matrix, then $\|\mathbf{Px}\| \ge \|\mathbf{Qx}\|$ for all $\mathbf{x}$ [47]. Therefore, we have

$$\gamma_n = \max_{k \in \{1,\ldots,K\} \setminus \{1^*,\ldots,(n-1)^*\}} \left\| \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-2} \mathbf{q}_{i^*} \mathbf{q}_{i^*}^H - \mathbf{q}_{(n-1)^*} \mathbf{q}_{(n-1)^*}^H \right) \mathbf{h}_k^H \right\|_2^2$$

$$\leq \max_{k\in\{1,\ldots,K\}\setminus\{1^*,\ldots,(n-1)^*\}} \left\| \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-2} \mathbf{q}_{i^*}\mathbf{q}_{i^*}^H \right) \mathbf{h}_k^H \right\|_2^2$$

$$\leq \max_{k\in\{1,\ldots,K\}\setminus\{1^*,\ldots,(n-2)^*\}} \left\| \left(\mathbf{I}_{n_t} - \sum_{i=1}^{n-2} \mathbf{q}_{i^*}\mathbf{q}_{i^*}^H \right) \mathbf{h}_k^H \right\|_2^2$$

$$= \gamma_{n-1}, \quad 2 \leq n \leq \min(n_t, K). \tag{A.3}$$

For the user channel matrix $\mathbf{H}(\mathcal{S}^*)$ of full row rank, the strict inequality holds with probability 1. According to the scheduling algorithm given in Table 4.1, over each loop, the admission of the best remaining user depends on whether it contributes to an increase of sum rate. If not, the loop is terminated and the user is not selected. Therefore, we have the number of the final selected users as $1 \leq \left|\mathcal{S}^*\right| \leq \min(n_t, K)$, and $\gamma_1^* \geq \gamma_2^* \geq \ldots \geq \gamma_{|\mathcal{S}^*|}^*$.

Appendix B: Proof of Theorem 4.3

Given a channel matrix $\mathbf{H}$ with i.i.d. circular symmetric complex Gaussian elements of zero mean and unit variance, we have the user channel gains $g_{k,k}^2$ in (4.2), $k = 1,...,\min(n_t, K)$, being statistically independent and distributed as $g_{k,k}^2 \sim \chi_{2(n_t-k+1)}^2$ [*Lemma2*, 10], where χ_n^2 denotes the central chi-square distribution with *n* degrees of freedom. Using the asymptotic result on the selected user SNR from (3.11), as $K \to \infty$, the channel gains of the selected user *k* under optimal scheduling can be approximated by

$$\max_{1 \le i \le K-k+1} \chi_{2(n_t-k+1)}^2 = \begin{cases} \ln K, & k = n_t \\ \ln K + O(\ln \ln K), & \text{otherwise} \end{cases} \tag{B.1}$$

for $k = 1,...,\min(n_t, K)$, where $\max_{1 \le i \le N} \chi_n^2$ denotes the maximum of *N* i.i.d. chi-square random variables distributed as χ_n^2 . The asymptotic independence of selected user channel gains as $K \to \infty$ can be established by dividing the users into $\lceil K / n_t \rceil$ independent disjoint sets and using the fact that $\lim_{K\to\infty} \lceil K / n_t \rceil = \lim_{K\to\infty} K = \infty$ for a finite n_t , where the notation $\lceil \cdot \rceil$ denotes the ceiling function. Therefore, we have

$$\begin{aligned} \lim_{K\to\infty} \left(R_{\text{sum}}^{\max} - \tilde{R}_{\text{sum}}^{\max} \right) &= \max_{\mathcal{S}} \sum_{k\in\mathcal{S}} \log_2(1 + P_k^* \ln K) - \max_{\tilde{\mathcal{S}}} \sum_{k\in\tilde{\mathcal{S}}} \log_2(1 + P_T \ln K / |\tilde{\mathcal{S}}|) \\ &\le \max_{\mathcal{S}} |\mathcal{S}| \log_2(1 + \sum_{k\in\mathcal{S}} P_k^* \ln K / |\mathcal{S}|) - \max_{\tilde{\mathcal{S}}} \sum_{k\in\tilde{\mathcal{S}}} \log_2(1 + P_T \ln K / |\tilde{\mathcal{S}}|) \qquad \text{(B.2)} \\ &\le |\mathcal{S}_0| \log_2(1 + \sum_{k\in\mathcal{S}_0} P_k^* \ln K / |\mathcal{S}_0|) - \sum_{k\in\mathcal{S}_0} \log_2(1 + P_T \ln K / |\mathcal{S}_0|) \qquad \text{(B.3)} \\ &= |\mathcal{S}_0| \log_2(1 + P_T \ln K / |\mathcal{S}_0|) - |\mathcal{S}_0| \log_2(1 + P_T \ln K / |\mathcal{S}_0|) = 0 \end{aligned}$$

where (B.2) is due to the concavity of log function of SNR, and (B.3) follows from (4.23). The subset $\mathcal{S}_0$ maximizes the sum rate in (4.5). Since $R_{\text{sum}}^{\max} \ge \tilde{R}_{\text{sum}}^{\max}$, we have

$$\lim_{K\to\infty}\left(R_{\text{sum}}^{\max}-\tilde{R}_{\text{sum}}^{\max}\right)=0\,.$$

Appendix C: Proof of Lemma 4.2

A first upper bound of $\max_{1\le i\le K-k+1}\chi^2_{2(n_t-k+1)}$ on γ_k^* can be easily established by considering $K+(K-1)+...+(K-n_t+1)=n_t(2K-n_t+1)/2$, rather than K, independent users under scheduling. Using *Lemma 2* in [10] (also in the proof of *Theorem 4.3*) and ignoring the correlation between γ_k^*, the suboptimal scheduler in Table 4.2 upper bounds each γ_k^* with $\max_{1\le i\le K-k+1}\chi^2_{2(n_t-k+1)}, k=1,...,\min(n_t,K)$. According to the classical Gram-Schmidt algorithm [46], the selected user channel gains γ_k^* only depend on γ_j^*, $j\le k$, i.e., the previous selected user channels. Therefore, the suboptimal scheduler has a maximum of $\max(K/n_t,1)$ sets of independent $\mathcal{S}^*$ with $|\mathcal{S}^*|=\min(n_t,K)$, where for the non-integer K/n_t, the remaining users are in one set. Considering the descending ordering of $\gamma_1^*>\gamma_2^*>...>\gamma_{|\mathcal{S}^*|}^*$ with probability 1, we have a second upper bound on γ_k^*, $k\ge 2$, as U defined in *Lemma 4.2*. Therefore, a tighter upper bound on $\gamma_k^*, k\ge 2$, using the scheduling in Table 4.2 is $\gamma_k<U_k$, where U_k is defined as $U_k\triangleq\min(U,\max_{1\le i\le K-k+1}\chi^2_{2(n_t-k+1)})$. It can be easily seen from the scheduling algorithm that the first user selected has a channel gain γ_1^* as the maximum squared Euclidean norm of K independent user channels, i.e., it is distributed as $\max_{1\le i\le K}\chi^2_{2n_t}$ exactly.

Appendix D: A Quadratic Form of Complex Normal Variates

Consider the $n \times n$ complex matrix $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{R}_t^H$, where the entries of $n \times n$ matrix $\mathbf{H}$ are independent complex circular symmetric Gaussian random variables with zero mean and unit variance, i.e., $\text{vec}(\mathbf{H}) \sim \mathcal{CN}(0, \mathbf{I}_n \otimes \mathbf{I}_n)$, and $\boldsymbol{\mathcal{R}}_t \triangleq (\mathbf{R}_t \mathbf{R}_t^H)^T$ is an $n \times n$ positive definite Hermitian matrix with unit diagonal entries. Let the Hermitian matrix $\mathbf{S}$ is defined as

$$\mathbf{S} \triangleq \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H = \mathbf{H}\mathbf{R}_t^H \mathbf{R}_t \mathbf{H}^H = \mathbf{H}\Psi_t \mathbf{H}^H, \tag{D.1}$$

with the $n \times n$ Hermitian matrix $\Psi_t \triangleq \mathbf{R}_t^H \mathbf{R}_t$. Therefore, Ψ_t is also positive definite, and the vector $\text{vec}(\mathbf{S})$ has a covariance matrix of $\boldsymbol{\mathcal{R}}_t \otimes \mathbf{I}_n$. We note that for a general non-idempotent covariance matrix Ψ_t, $\mathbf{S}$ does not have a Wishart distribution. Let the eigenvalue decomposition of Ψ_t be $\Psi_t = \mathbf{Q}_t \Lambda_t \mathbf{Q}_t^H$, where $\mathbf{Q}_t$ is an $n \times n$ unitary matrix and $\Lambda_t = diag\{\lambda_i\}_{i=1}^n$ is an $n \times n$ diagonal matrix with λ_i, $i = 1, ..., n$ being the real and positive eigenvalues of Ψ_t, which are the same as those of $\boldsymbol{\mathcal{R}}_t$. Therefore, (D.1) can be written as

$$\mathbf{S} = \mathbf{H}\mathbf{Q}_t \Lambda_t \mathbf{Q}_t^H \mathbf{H}^H \sim \mathbf{H}\Lambda_t \mathbf{H}^H, \tag{D.2}$$

where the notation $\sim$ indicates that the joint distribution of $\mathbf{S}$ is identical to the joint distribution of $\mathbf{H}\Lambda_t \mathbf{H}^H$. Let the unique QR decomposition of $\tilde{\mathbf{H}}^H$ be $\tilde{\mathbf{H}}^H = \mathbf{Q}\mathbf{R}$, where the unitary matrix has a column partitioning $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n]$, and the upper triangular matrix $\mathbf{R} = (r_{ij})$, $1 \le \text{i} \le \text{j} \le \text{n}$, with $r_{ii} > 0$, $1 \le \text{i} \le \text{n}$. Let the column partitioning of $\tilde{\mathbf{H}}^H$ be $\tilde{\mathbf{H}}^H = \left[\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, ..., \tilde{\mathbf{h}}_n\right]$, and from (D.1) and (D.2) we have $\tilde{\mathbf{h}}_k = [\sqrt{\lambda_1} h_{k1}, \sqrt{\lambda_2} h_{k2}, ..., \sqrt{\lambda_n} h_{kn}]^H$, $k = 1, ..., n$. According to the classical Gram-Schmidt computation of $\mathbf{R}$ [46], we have

$$\mathbf{z}_k = \tilde{\mathbf{h}}_k - \sum_{i=1}^{k-1} r_{ik}\mathbf{q}_i \,, \tag{D.3}$$

$$r_{ik} = \mathbf{q}_i^H \tilde{\mathbf{h}}_k, \; i = 1,...,k-1, \tag{D.4}$$

and

$$r_{kk} = \left\| \mathbf{z}_k \right\|_2, \; i = 1,...,n \,. \tag{D.5}$$

Taking (D.4) into (D.3), we have

$$\mathbf{z}_k = \tilde{\mathbf{h}}_k - \sum_{i=1}^{k-1} (\mathbf{q}_i^H \tilde{\mathbf{h}}_k) \cdot \mathbf{q}_k = \tilde{\mathbf{h}}_k - \sum_{i=1}^{k-1} (\mathbf{q}_i \mathbf{q}_i^H) \cdot \tilde{\mathbf{h}}_k = \left[\mathbf{I}_n - \sum_{i=1}^{k-1} \mathbf{q}_i \mathbf{q}_i^H \right] \cdot \tilde{\mathbf{h}}_k = \left[\sum_{i=k}^{n} \mathbf{q}_i \mathbf{q}_i^H \right] \cdot \tilde{\mathbf{h}}_k \,, \tag{D.6}$$

for orthonormal vectors $\mathbf{q}_i$. The final equation in (D.6) indicates that $\mathbf{z}_k$ is the sum of projections of $\tilde{\mathbf{h}}_k$ over the complex orthonormal vectors $\{\mathbf{q}_k,...,\mathbf{q}_n\}$, $k = 1,...,n$. Therefore, r_{kk}^2 have the degrees of freedom $2(n-k+1), \; k = 1,...,n$. For $k = 1$, we have

$$r_{11}^2 = \left\| \mathbf{z}_1 \right\|_2^2 = \sum_{i=1}^{n} \lambda_i \left| h_{1i} \right|^2 \,, \tag{D.7}$$

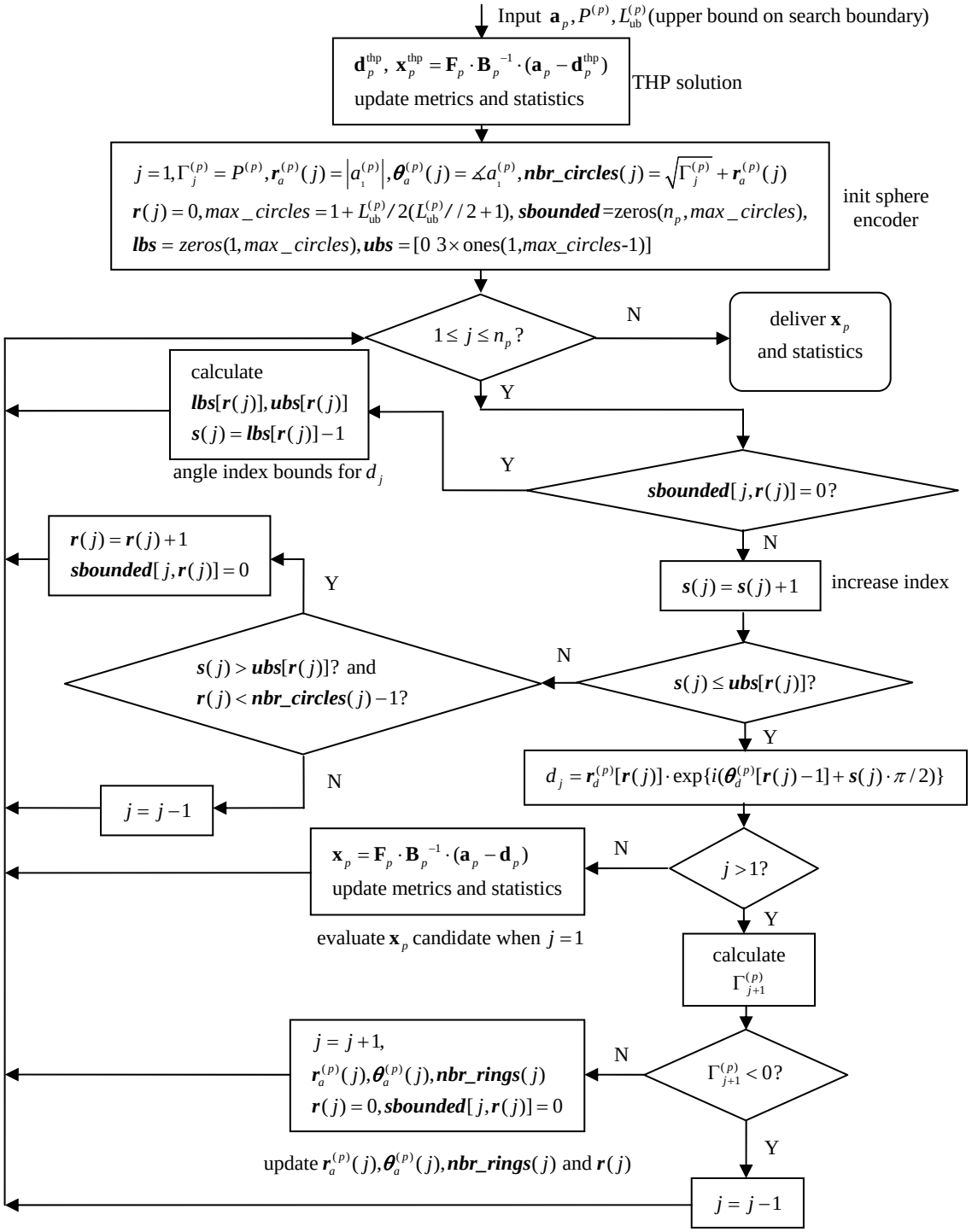
which is a weighted sum of independent standard chi-square random variables, and for distinct λ_i, the pdf of r_{11}^2 is [95],

$$f(x) = \sum_{k=1}^{n} \frac{C_k}{\lambda_k} \exp(-\frac{x}{\lambda_k}), \; x > 0, \tag{D.8}$$

where

$$C_k \triangleq \prod_{\substack{i=1 \\ i \neq k}}^{n} \frac{\lambda_k}{\lambda_k - \lambda_i} \,. \tag{D.9}$$

Appendix E: Complex Sphere Precoding



Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379-423, 623-656, July, Oct. 1948.

[2] G. Foschini and J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311-335, Mar. 1998.

[3] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *AT&T-Bell Labs, Internal Tech. Memo.*, pp. 287-290, Jun. 1995. Available for download at http://mars.bell-labs.com/. See also *European Trans. on Telecomm. (ETT)*, vol. 10, pp. 585-595, Nov. 1999.

[4] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. on Inform. Theory*, vol. 29, pp. 439-441, May 1983.

[5] G. J. Foschini and A. H. Diaz, "Dirty paper coding: Perturbing off the infinite dimensional lattice limit," presented at *Workshop on Sig. Proc. for Wireless Transmission (DIMACS 2002)*, Rutgers University, New Jersey, Oct. 2002. In *Multi-Antenna Channels: Capacity, Coding and Signal Processing*, DIMACS Series in Discrete Mathematics and Computer Science, G. J. Foschini and S. Verdu, Eds. Providence, RI: AMS, 2003.

[6] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Upper Saddle River, New Jersey: Prentice-Hall PTR, 1996.

[7] A. J. Viterbi, *CDMA: Principles of Sprad Spectrum Communication*. Reading, MA: Addison-Wesley, 1995.

[8] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1456-1467, July 1999.

[9] F. Kelly, "Charging and rate control for elastic traffic," revised version, *European Transactions on Telecommunications*, vol. 8, pp. 33-37, 1997.

[10] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1691-1706, July 2003.

[11] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electron. Lett.*, vol. 35, pp. 14-16, Jan. 1999.

[12] G. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Select. Areas Commun.*, vol. 17, pp.1841-1852, Nov. 1999.

[13] M. Tomlinson, "New automatic equalizer employing modulo arithmetic," *Electron. Lett.*, vol. 7, pp. 138-139, Mar. 1971.

[14] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Trans. Commun.*, vol. COM-20, pp. 774-780, Aug. 1972.

[15] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. 18, pp. 2-14, Jan. 1972.

[16] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. 19, pp. 197-207, Mar. 1973.

[17] R. G. Gallager, "Capacity and coding for degraded broadcast channels," *Problemy Peredachi Informatsii*, vol. 10, pp. 3-14, July-Sept. 1974.

[18] P. P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 20, pp. 279-280, Mar. 1974.

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

[20] 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, pp. 2658-2668, Oct. 2003.

[21] W. Yu and J. Cioffi, "Sum capacity of Gaussian vector broadcast channels," to appear in *IEEE Trans. Inform. Theory*, submitted Nov. 2001 and revised Mar. 2002.

[22] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1912-1921, Aug. 2003.

[23] D. Tse and P. Viswanath, "On the capacity of the multiple antenna broadcast channel" in *Multiantenna Channels: Capacity, Coding and Signal Processing*, DIMACS Series in Discrete Mathematics and Computer Science, G. J. Foschini and S. Verdu, Eds. Providence, RI: AMS, 2003, pp. 87-105.

[24] S. Vishwanath, G. Kramer, S. Shamai (Shitz), S. Jafar, and A. Goldsmith, "Capacity bounds for Gaussian vector broadcast channels," presented at the DIMACS Workshop on *Signal Processing for Wireless Transmission*, Rutgers University, Piscataway, NJ, Oct. 2002. In *Multi-Antenna Channels: Capacity, Coding and Signal Processing*, DIMACS Series in Discrete Mathematics and Computer Science, G. J. Foschini and S. Verdu, Eds. Providence, RI: AMS, 2003.

[25] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inform. Theory*, vol. 25, pp. 306-311, May 1979.

[26] H. Sato, "An outer bound to the capacity region of broadcast channels," *IEEE Trans. Inform. Theory*, vol. 24, pp. 374-377, May 1978.

[27] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19-31, 1980.

[28] W. Yu and J. Cioffi, "Trellis precoding for the broadcast channel," in *Proc. of the IEEE Global Telecommun. Conf. (Globecom)*, San Antonio, Texas, Nov. 2001, pp. 1344-1348.

[29] A. Cohen and A. Lapidoth, "The Gaussian watermarking," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1639-1667, Jun. 2002.

[30] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for canceling known interference," submitted to *IEEE Trans. Inform. Theory*. See also *ISITA*

2000, Honolulu, Hawaii (USA) Nov. 2000.

[31] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. on Inform. Theory*, vol. 48, pp. 1250-1276, Jun. 2002.

[32] H. A. David, *Order Statistics*, 2nd ed. New York: Wiley, 1981.

[33] P. Viswanath, D. N. C. Tse and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp.1277-1294, June 2002.

[34] J. G. Proakis, *Digital Communication*, 3rd ed. Boston, MA: MacGraw-Hill, 1995.

[35] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1991.

[36] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR: A high efficiency-high data rate personal communication wireless system," *IEEE 51st Vehicular Technology Conference*, Tokyo, Japan, Vol. 3, pp. 1854-1858, May 2000.

[37] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," *12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Vol. 2, pp. 33-37, Sept./Oct. 2001.

[38] Y. Jou, "Development in third generation (3G) CDMA technology," in *Proc. IEEE 6th Int. Symp. On Spread Spectrum Techniques and Applications*, pp. 460-464, vol. 2, 2000.

[39] G. Ginis and J. M. Cioffi, "A multiuser precoding scheme achieving crosstalk cancellation with application to DSL systems" in *34th Asilomar Conf. on Signals, Systems and Computers*, vol. 2, pp. 1627-1631, Pacific-Grove, CA, USA, Nov. 2000.

[40] R. A. Horn and C. R. Johnson, *Matrix Analysis*, New York: Cambridge Univ. Press, 1990.

[41] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, New Jersey:

Prentice Hall, 2002.

[42] D. N. C. Tse, "Optimal power allocation over parallel Gaussian broadcast channels", in *Proc. Int. Symp. Information Theory*, ISIT, Ulm. Germany, June 1997, pp. 27.

[43] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, Seattle, WA, pp. 331-335, June 1995.

[44] Z. Tu and R. S. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Commun. Lett.*, vol. 7, pp. 370-372, Aug. 2003.

[45] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2619-2692, Oct. 1998.

[46] G. H. Golub, C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: John Hopkins Univ. Press, 1996.

[47] M. Brookes, *Matrix Reference Manual*, available online at: http://www.ee.ic.ac.uk/hp/staff/dmb/matrix.

[48] J. Benesty, Y. Huang, and J. Chen, "A fast recursive algorithm for optimum sequential signal detection in a BLAST system," *IEEE Trans. Signal Processing*, vol. 51, pp. 1722-1730, July 2003.

[49] M. K. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Asilomar, CA, 1997, pp. 1405-1409.

[50] S. Vishwanath, W. Rhee, N. Jindal, S. Jafar, and A. Goldsmith, "Sum power iterative waterfilling for Gaussian vector broadcast channels," in *Proc. ISIT 2003*, Yokohama, Japan, June/July, 2003, pp. 467.

[51] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1319-1343, June 2002.

[52] M. Miyakawa and H. Harashima, “A method of code conversion for a digital communication channel with intersymbol interference,” *Trans. Institute of Electronics Communications Engineering of Japan*, vol. 53-A, pp. 272-273, Jun. 1969.

[53] R. F. H. Fischer, C. Windpassinger, A. Lampe, and J. B. Huber, “MIMO precoding for decentralized receivers,” in *Proc. of IEEE Intern. Symp. Inform. Theory*, ISIT’02, Lausanne, Switzerland, Jun./Jul. 2002, p. 496.

[54] G. D. Forney and M. V. Eyuboglu, “Combined equalization and coding using precoding,” *IEEE Commun. Mag.*, pp. 25-34, Dec. 1991.

[55] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, New York: Wiley, 1982.

[56] R. D. Wesel and J. M. Cioffi, “Achievable rates for Tomlinson-Harashima precoding,” *IEEE Trans. on Inform. Theory*, vol. 44, no. 2, pp. 824-831, Mar. 1998.

[57] D. A. Gore, R. W. Heath, and A. J. Paulraj, “Transmit selection in spatial multiplexing systems,” *IEEE Commun. Lett.*, vol. 6, no. 11, pp. 491-493, Nov. 2002.

[58] G. Ginis and J. M. Cioffi, “On the relation between V-BLAST and the GDFE,” *IEEE Commun. Lett.*, vol. 5, pp. 364-366, Sept. 2001.

[59] D. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, “Fading correlation and its effect on the capacity of multi-element antenna systems,” *IEEE Trans. Commun.*, vol. 48, pp. 502-513, Mar. 2000.

[60] H. Bolcskei, D. Gesbert, and A. J. Paulraj, “On the capacity of OFDM-based spatial multiplexing systems,” *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 225-234, Feb. 2002.

[61] 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. 41, pp. 10-22, Feb. 1998.

[62] O. Edfors, M. Sandell, J. van de Beek, D. Leandstrom, and F. Sjoberg, "An introduction to orthogonal frequency-division multiplexing," *Tech. Report*, Lulea University of Technology, Sept. 1996.

[63] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications – where Fourier meets Shannon," *IEEE Signal Processing Magazine*, pp. 29-48, May 2000.

[64] Y. Li, J. H. Winter, and N. R. Sollenberger, "MIMO-OFDM for wireless communications: Signal detection with enhanced channel estimation," *IEEE Trans. Commun.*, vol. 50, pp. 1471-1477, Sept. 2002.

[65] Z. Liu, Y. Xin, and G. B. Giannakis, "Space-time-frequency coded OFDM over frequency-selective fading channels," *IEEE Trans. Signal Processing*, vol. 50, pp. 2465-2476, Oct. 2002.

[66] R. J. Piechocki, P. N. Fletcher, A. R. Nix, C. N. Canagarajah, and J. P. McGeehan, "Performance evaluation of BLAST-OFDM enhanced Hiperlan/2 using simulated and measured channel data," *Electr. Letts*, vol. 37, pp. 1137-1139, Aug. 2001.

[67] Y. Xin and G. B. Giannakis, "High-rate space-time layered OFDM," *IEEE Commun. Letts.*, vol. 6, pp. 187-189, May 2002.

[68] A. F. Molisch, M. Z. Win, and J. H. Winters, "Space-time-frequency (STF) coding for MIMO-OFDM systems," *IEEE Commun. Letts.*, vol. 6, pp. 370-372, Sept. 2002.

[69] H. Bolcskei and A. J. Paulraj, "Space-frequency coded broadband OFDM systems," in *Proc. IEEE WCNC-2000*, Chicago, IL, Sept. 2000, pp. 1-6.

[70] H. Bolcskei and A. J. Paulraj, "Space-frequency coded MIMO-OFDM with variable multiplexing-diversity tradeoff," in *Proc. IEEE Int. Conf. on Communications (ICC'03)*, vol. 4, May 2003, pp.2837-2841.

[71] N. Geng, I. Viering, and M. Kiessling, "Multi-user MIMO-OFDM cell throughput under real-world propagation conditions," in *Proc. IEEE 56th Vehicular Tech. Conf. (VTC2002-Fall)*, vol. 2, Sept. 2002, pp. 879-883.

[72] W. Wang, T. Ottosson, M. Sternad, A. Ahlen, and A. Svensson, "Impact of multiuser diversity and channel variability on adaptive OFDM," in *Proc. IEEE 58th Vehicular Tech. Conf. (VTC2003-Fall)*, vol. 1, Oct. 2003, pp. 547-551.

[73] M. Sternad, T. Ottosson, A. Ahlen, and A. Svensson, "Attaining both coverage and high spectral efficiency with adaptive OFDM downlinks," in *Proc. IEEE 58th Vehicular Tech. Conf. (VTC2003-Fall)*, vol. 4, Oct. 2003, pp. 2486-2490.

[74] S. Ye and R. S. Blum, "Optimum signaling for MIMO-OFDM systems with interference," in *Proc. European Signal Processing Conf.*, Toulouse, France, Sept. 2002.

[75] L. Giangaspero, L. Agarossi, G. Paltenghi, S. Okamura, M. Okada, and S. Komaki, "Co-channel interference cancellation based on MIMO OFDM systems," *IEEE Wireless Commun. Mag.*, pp. 8-17, Dec. 2002.

[76] S. Visuri and H. Bolcskei, "MIMO-OFDM multiple access with variable amount of collision," to appear in *Proc. IEEE International Conf. on Communications (ICC'04), Paris, France*, June 2004.

[77] G. L. Stuber, J. R. Barry, S. W. McLaughlin, Y. Li, M. A. Ingram, and T. Pratt, "Broadband MIMO-OFDM wireless communications," *Proc. of the IEEE*, vol. 92, pp. 271-294, Feb 2004.

[78] H. Bolcskei, "Principles of MIMO-OFDM wireless systems," chapter in CRC handbook on *Signal Processing for Communications*, M. Ibnkahla, Ed., 2004.

[79] J. Chuang and N. Sollenberger, "Beyond 3G: Wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Commun. Magazine*, pp. 78-87, July 2000.

[80] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, pp. 463-471, Apr. 1985.

[81] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.*, vol. 51, pp. 389-399, Mar. 2003.

[82] O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice code decoder for space-time codes," *IEEE Commun. Lett.*, vol. 4, pp. 161-163, May 2000.

[83] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multi-antenna multi-user communication," submitted to *IEEE Trans. Commun.*, June 2003. Available for download at http://mars.bell-labs.com/.

[84] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in Lattices," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2201-2214, Aug. 2002.

[85] *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band*, IEEE Standard 802.11a-1999.

[86] *Local and Metropolitan Area Networks—Part 16, Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Standard IEEE 802.16a.

[87] J. H. Ju and V. O. K. Li, "TDMA scheduling design of multihop packet radio networks based on latin squares," *IEEE J. on Select. Areas in Commun.*, vol. 17, pp. 1345-1352, Aug 1999.

[88] Z. Cai, M. Lu, and C. N. Georghiades, "Topology-transparent time division multiple access broadcast scheduling in multihop packet radio networks," *IEEE Trans. Vehicular Technology*, vol. 52, pp. 970-984, July 2003.

[89] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2001

[90] J. Denes and A. D. Keedwell, *Latin Squares and Their applications*. New York: Academic, 1974.

[91] D. Tse, EE224B class notes—*Fundamentals of Wireless Communication*, UC Berkeley.

[92] G. J. Pottie, "System design choices in personal communications," *IEEE Pers. Commun.*, vol. 2, pp. 50-67, Oct. 1995.

[93] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Magazine*, vol. 38, no. 7, pp. 70-77, July 2000.

[94] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channels," in *Conf. on Inform. Systems and Science (CISS)*, Mar. 2004.

[95] M. K. Simon, *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists*. Boston, MA: Kluwer Academic Publishers, 2002.

Vita

Jing Jiang was born on Nov. 25, 1969, in Nanjing, P. R. China. She received her Bachelor and Master's degrees in radio engineering from the University of Electronic Science and Technology of China (UESTC), P. R. China, in 1995, with honors. From 1995 to 1996, she worked for Ranger Communications, Inc., Shanghai, P. R. China, in the area of vehicular/marine/CB transceiver design. From 1996 to 1997, she was employed by Shanghai Bell, Shanghai, P. R. China, as a system application engineer working on wireless local loop development. She received her MSEE degree in communications from the Mississippi State University in Dec. 1998. From Jan. to Aug. of 1999, she was a staff scientist at R&D group of Adtran Inc., Huntsville, AL, focusing on HDSL2 applications. Since the fall of 1999, Jing has been working toward her Ph.D. degree at Virginia Tech. In the summers of 2000 and 2001, she interned at Lucent Technologies, Whippany, NJ, and Nokia Research Center, Irving, Tx.

Jing received Daniel E. Noble Fellowship jointly from IEEE Vehicular Technology Society and Motorola in 2002 for her contribution to the cross-layer design of wireless networks. Her current research interests include network information theory, space-time signal processing, and multiuser packet scheduling. She authored and co-authored about 12 papers in various journals and conferences.