Spectrum Sharing in the Spatial Domain

H. Erik A. Yngvesson

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Institute for Communication Systems
Faculty of Engineering and Physical Sciences
University of Surrey
Surrey, U.K.

October 2017

© H. Erik A. Yngvesson 2017
Abstract

The problem of spectrum sharing by exploiting the spatial domain is investigated in this thesis. The ultimate purpose of such a scheme is to mitigate the under-utilization of the scarce spectrum resource. By taking into account the availability of multiple antennas at the communicating nodes, an additional level of freedom can be exploited. Multiple-input multiple-output antenna systems have previously been shown to hold great promise: a linear growth in capacity without bandwidth expansion, enhanced transmission reliability using for instance space-time codes, and effective interference handling.

The question of how these properties can be harnessed is explored by considering two perspectives: no cooperation and cooperation between users. For the cooperative scenario, a spatial-domain interweave spectrum sharing scheme is introduced that enables opportunistic transmission at a controlled cost to the license holders. The proposed scheme demonstrates three excellent characteristics: that exploitation of the spatial domain allows opportunistic communication in a “spatial hole,” that spectrum sharing is effectively enabled by inter-tier cooperation, and finally that in this scenario spatial-domain interweave is feasible with a “small” (as compared to the number of receive antennas at the incumbent) number of transmit antennas. In essence, this opens the possibility of the incumbents’ performance to be traded against opportunistic transmission. In the non-cooperative scenario, a spectrum sharing model between a small and large MU-MIMO system is proposed and analysed. The significant service antenna number asymmetry poses unique challenges and opportunities. In the limit of an infinite number of service antennas at one of the access point, the interference and noise power tends to zero and the transmit power can also be scaled back accordingly. These traits seem ideal for use in a spectrum sharing scenario, but in the present case with the coexistence of a conventional MIMO system and with a finite number of service antennas, how will the system behave? The resulting interference scenario is analysed explicitly both in the uplink and downlink, assuming linear receive and transmit equalizers, respectively. Characterization of the mean SINR operating point and required transmit power are presented, and concise transmit power scaling laws are derived. The scaling laws offer insight into how the system behaves with the number of service antennas and system load.

Keywords: Cooperative spectrum sharing, multiple-input multiple-output (MIMO), non-cooperative spectrum sharing
Acknowledgements

I am most grateful to Dr Yi Ma for the unwavering support shown throughout my graduate studies. His unbounded enthusiasm and energy, continual encouragement, and loyalty to his students has been truly inspirational. I’d also like thank my co-supervisors Prof. Rahim Tafazolli and Dr Na Yi for trying to instil in me what the essence of research is and how to conduct it. In addition, I recognize that I have been privileged in being part of the fifth generation mobile communications project at the University of Surrey spearheaded by Prof. Tafazolli, and wholeheartedly supported by the academic staff at the Institute for Communication Systems. It has resulted in a vibrant work place, bubbling with exciting ideas.

I would also like to thank Dr Tim Brown for his support and advice, and for giving me the opportunity to interact with undergraduate students as a lab demonstrator and tutor. It gave me a first taste of teaching and the chance to review my radio-frequency related engineering skills.

I’m deeply indebted to friends and colleagues at the University of Surrey, including Parisa Cheraghi, Zhengwei Lu, Chuyi Qian, Jiancao Hou, Guangyi Wang, Carlos De Luna Ducoing, and Abderraouf Yamani. It was always interesting to hear of problems related to other areas of digital communications, and they were all brilliant as sounding boards for new ideas or perceived problems. In addition, I’d like to thank them for their patience and understanding when faced with the occasional rant about graduate life and its particular challenges.

Finally I would like to express my warmest gratitude to Eva and Elsa Johansson.
Contents

List of Figures v

Abbreviations vi

List of Symbols viii

1 Introduction 1

1.1 Background .................................................. 1
1.2 Motivation and Objectives ................................. 1
1.3 Contributions ................................................. 4
1.4 Outline of Thesis ............................................ 5
1.5 Publications .................................................. 6

2 Background and Related Literature 7

2.1 Introduction .................................................. 7
2.2 The Multiple-Input Multiple-Output Channel .............. 8
  2.2.1 The Point-to-Point Channel ............................ 8
  2.2.2 The Multiple Access Channel .......................... 11
  2.2.3 The Broadcast Channel ................................. 12
  2.2.4 The Interference Channel ............................... 13
  2.2.5 MIMO Detection ........................................ 16
  2.2.6 MIMO Precoding ........................................ 21
  2.2.7 A Note on Channel Assumptions ....................... 23
2.3 Spectrum Sharing .......................................... 24
5 Conclusion and Future Work

5.1 Conclusion ................................................................. 95
5.2 Future Work ............................................................... 98

A Appendix
A.1 Distribution of MF Interference Term ............................. 100
A.2 ZF SINR Distribution .................................................. 101
A.3 ZF Distribution with Interferer ..................................... 102
A.4 Solution Model to Problem (3.23) .................................. 103

References ................................................................. 104
List of Figures

2.1 The MIMO channel. The signal transmitted from each antenna is received (through different “paths”) at each receive antenna. 7

2.2 Approximate mean SINR performance for an arbitrary user or stream versus load. The figures (a), (b), (c) and (d) depict an SNR (i.e., no interference) set to −10, 0, 10, 20 dB, respectively. 19

2.3 Spectrum access categorization. Note the attempt to include the idea of a polar model of exclusive use and commons at either extreme, and possible combinations in between. 25

3.1 Exact and approximate $p_1$ transmit powers versus $p_2$ for the single UT scenario with MF equalizers (Chapter 3.2.2). The AP has $M_2 = 4$ service antennas. 46

3.2 Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 100, M_2 = 4, p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 56

3.3 Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 100, M_2 = 4, p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 56

3.4 Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 1000, M_2 = 4, p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 57

3.5 Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 1000, M_2 = 4, p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 57

3.6 Empirical cdf of SINR (detected at both AP 1 and AP 2) for an arbitrary UT associated to AP 2. The antenna configuration is $\{M_1, M_2\} = \{100, 4\}$, the load is set to $\beta_1 = 0.1$, $\rho = 1$ and $p_2 = 10$ dB. This implies a transmit power $p_1 = -11.6$ dB for the ZF receiver type and $p_1 = -2.27$ dB for the MF equalizer. 58
3.7 Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 100$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. .............................................. 65

3.8 Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 100$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. .............................................. 66

3.9 Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 1000$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. .............................................. 66

3.10 Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 1000$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. .............................................. 67

4.1 $K$-user interference channel where the transceivers in set $\mathcal{P}_s$ are the secondary user pairs. ......................................................... 71

4.2 Probability that PU $k$ remains idle due to its inability to reach the rate constraint within the power constraint. ......................... 84

4.3 Average achieved SU rate with respect to PU rate constraint. The SU has $M_K = N_K$ antennas and the PUs have $M_k = 2, N_k = 4$, $\forall k \in \mathcal{P}_p$ and both SU and PUs have power constraints of 30 dB. ......................... 85

4.4 Average PU transmit power under same conditions as in Fig. 4.3. Note that the OIA algorithm does not affect PU system and hence the PU transmit power is equivalent to a system with no SU. ......................... 86

4.5 The effect on average SU achievable rate with varying the PUs power constraint. The SU has $M_K = N_K = 2$ antennas and its power constraint is set to 30 dB. ......................... 87

4.6 Multi-user IWF method. SU rate and PU transmit power as a function of PU rate constraint. Plots (a) and (c) are with power constraint 10 dB for all users, and plots (b) and (d) are with 30 dB power constraint. ... 88

4.7 Multi-user sequential admission method. SU rate (a) and PU transmit power (b) as a function of PU rate constraint, for $P_k = 10$ dB, $\forall k \in \mathcal{P}_p \cup \mathcal{P}_s$. 91

4.8 Multi-user sequential admission method. SU rate (a) and PU transmit power (b) as a function of PU rate constraint, for $P_k = 30$ dB, $\forall k \in \mathcal{P}_p \cup \mathcal{P}_s$. 92
Abbreviations

**AWGN** Additive White Gaussian Noise.

**BER** Bit Error Rate.

**ccdf** Complementary Cumulative Distribution Function.

**cdf** Cumulative Distribution Function.

**CDMA** Code Division Multiple Access.

**CSI** Channel State Information.

**dB** Decibel, defined as $y = 10\log_{10}(x)$ where $x$ is dimensionless, e.g., a ratio of powers.

**FDD** Frequency Division Duplex.

**i.i.d.** Identically independently Distributed.

**LOS** Line Of Sight.

**mgf** Moment Generating Function.

**MIMO** Multiple-Input Multiple-Output.

**MISO** Multiple-Input Single-Output.

**MMSE** Minimum Mean Square Error.

**MRC** Maximum Ratio Combiner/Combining.

**MSE** Mean Squared Error.

**MU-MIMO** Multi-User MIMO.

**OFDM** Orthogonal Frequency Division Multiplexing.

**pdf** Probability Density Function.

**SIC** Successive Interference Cancellation.
**Abbreviations**

**SINR**  Signal-to-Interference-and-Noise-Ratio.

**SISO**  Single-Input Single-Output.

**SNR**  Signal-to-Noise-Ratio.

**TDD**  Time Division Duplex.

**w.r.t.**  With Respect To.

**WF**  Water-Filling.

**ZF**  Zero Forcing.
List of Symbols

\( x \triangleq y \) \( x \) is by definition equal to \( y \).

\(|X|\) Determinant of square matrix \( X \).

\( x \sim y \) Random variable \( x \) is distributed as \( y \).

\([X]_{ij}\) Element of matrix \( X \) in row \( i \) and column \( j \).

\( \|x\| \) Euclidean norm defined as \( \sqrt{\text{tr}(XX^H)} \).

\( \mathcal{E}(\cdot) \) Expectation operator.

\( E_1 \) Exponential integral, \( E_1(z) \triangleq \int_z^\infty t^{-1}e^{-t}dt \).

\( k! \) Factorial of the integer \( k \), i.e., \( k! \triangleq (k)(k-1)\ldots(1) \).

\( X^H \) Conjugate transpose or hermitian of the matrix \( X \).

\( I_m \) Identity matrix of size \( m \) by \( m \). If dimension is omitted, then size is inferred from context.

\( \log_n(x) \) Logarithm, base \( n \).

\( \text{diag} X \) Vector containing diagonal elements of \( X \).

\( X \) Matrix \( X \).

\( \max_x f(x) \) Maximum of \( f(x) \).

\( \max(x_1,x_2,\ldots,x_n) \) Maximum of the elements \( x_1,x_2,\ldots,x_n \).

\( \min_x f(x) \) Minimum of \( f(x) \).

\( \min(x_1,x_2,\ldots,x_n) \) Minimum of the elements \( x_1,x_2,\ldots,x_n \).

\( \ln(x) \) Natural logarithm, i.e., \( \ln(x) \triangleq \log_e(x) \).

\( \mathcal{N}(\mu,\sigma^2) \) Normal distribution with mean \( \mu \) and variance \( \sigma^2 \).
$\mathcal{CN}(\mu, \sigma^2)$ Complex normal distribution with mean $\mu$ and variance $\sigma^2$.

$X^\dagger$ Pseudo-inverse of $X$.

$\mathbb{C}$ Set of complex numbers.

$\mathbb{R}$ Set of real numbers.

$\text{tr}(X)$ Trace, i.e., sum of diagonal elements of $X$.

$X^T$ Matrix transpose of $X$.

$x$ Vector $x$. By convention $x$ is a column vector.

$I_{m \times n}$ An $m$ by $n$ matrix with all elements of value 1. If dimensions are omitted, then size is inferred from context.

$0_{m \times n}$ An $m$ by $n$ matrix with all elements of value 0. If dimensions are omitted, then size is inferred from context.
Introduction

1.1 Background

At the time of writing, research institutions, mobile network operators and telecommunications equipment providers, are scrambling to position themselves favourably for the launch of the fifth generation mobile communications standards. Higher data rates, lower latency, and support for a high density of devices, have been highlighted as key (and potentially contradicting) properties for the new system. It is at times difficult to distinguish cause from effect in the supply-and-demand relationship between provider and consumer, but Cooper’s law\textsuperscript{a}, regarding the exponential growth in wireless traffic, has been shown to be accurate for the past century, and unlike Moore’s law\textsuperscript{b}, shows no sign of slowing down. It is this basic observation that sets the scene for this thesis, and indeed much of the research in wireless communication in general. More specifically, in this case it is spectrum usage and spectrum efficiency that are of primal interest.

1.2 Motivation and Objectives

Traditionally, wireless communications technologies have striven to orthogonalize channel usage in order to avoid interference. Examples of this include the common multiple access schemes; time division multiple access, frequency division multiple access, and code division multiple access. In addition, the use of a cellular structure in mobile

\textsuperscript{a}Cooper’s law, coined by Martin Cooper [1], states that the effective data rate doubles every two-and-a-half years [2, 3].

\textsuperscript{b}Moore’s law describes the observation that the number of transistors on an integrated circuit roughly doubles every two years [4]. It seems that, at the time of writing, the pace of integration has slowed.
1.2. Motivation and Objectives

Communications can be said to be a form of space division multiple access, where transmissions over the same frequency occur in (ideally) non-overlapping geographical areas, giving rise to the concept of frequency re-use. These methods have been developed for the simple reason that interference is generally seen as difficult to handle and hence harmful. It suffices to consider the simple example where a receiver (treating interference as noise) suffers interference with power on the same order of magnitude as the AWGN noise power term. In such a situation the SINR is halved compared to the interference-free case. Furthermore, in environments where the desired signal and the individual interference terms from other transmitters are likely to be of the same magnitude, the resulting SINR would be too low for many practical applications, such as voice, video, or streaming. Orthogonalization, in the sense of isolating transmissions, then seems like a perfectly sensible solution. However, it has been shown for various multiple-user channel types (e.g., the (MIMO) multiple access channel and the degraded single antenna broadcast channel (see e.g., the excellent textbook by Cover and Thomas [5])), that the sum capacity (or indeed other points on the capacity region boundary) offer much better use of the spectrum if all transmitters are active. Although the capacity region for the interference channel is unknown in general (see e.g., Kramer’s short review of the available results for the two-user interference channel [6]) it seems that significant spectrum efficiency gains can be obtained in such scenarios as well, compared to the simpler orthogonalization schemes [7].

This simple observation that strict orthogonalization is wasteful has motivated a flurry of research. Two broad such research avenues (with considerable overlap) are spectrum sharing and multiple antenna techniques. Spectrum sharing takes on many guises (and has in many works been facilitated or enabled by multiple antenna technologies), and several attempts have been made to classify or categorize different methods, techniques or paradigms. One of the most common forms of spectrum sharing is known simply as cognitive radio. This is a hierarchical form of spectrum sharing where a primary user or entity has priority and a secondary user only transmits if interference (or in general its impact on the primary system) is kept below a pre-defined limit. For a fascinating account of the available limits of cognitive radio, please refer to the article by Goldsmith et al. [8]. It is also in this article that the three paradigms of cognitive radio are best defined; interweave, underlay, and overlay. In the interweave mode, the secondary user only transmits if it inflicts no interference to the primary user. In underlay, the interference caused at the primary receiver is kept below a limit deemed harmful, and in the overlay paradigm, the secondary user actively aids the primary user transmission. Different schemes can at times be difficult to categorize and Zhang et al. [9] for instance instead suggested the use of only two categories for this type of hierarchical spectrum sharing; opportunistic spectrum sharing, and spectrum sharing. Opportunistic spectrum sharing referred to the case where the secondary user opportunistically
transmitted in spectrum holes, and can hence be equated to the interweave paradigm. The second mode, termed simply spectrum sharing, covered the remaining possibilities. Zhang et al. also stress that this mode, where primary and secondary transmitters are active simultaneously, in the same band, and the same physical area was deemed to offer higher spectral efficiencies. Notable results combining cognitive radio with multiple antennas are the works of Gastpar [10], enforcing the idea of spatial signatures, Zhang et al. [11] for further developing the idea of underlay MIMO cognitive radio, and Perlaza et al. [12] who introduces the idea of space pooling. Finally, cementing the notion that non-orthogonal spectrum sharing (not necessarily in a hierarchical system) can provide improved performance, Jorswieck et al. [13] and Litjens et al. [14] separately showed that spectrum sharing between two mobile networks may be beneficial.

It is worth explicitly stating that multiple antennas are not required to receive multiple information streams or handle interference. Take the simple example of the single-antenna multiple access channel where with the use of a successive interference cancellation receiver it is theoretically possible to operate at any point on the capacity boundary [15]. However, it is also true that multiple antennas may provide a linear increase in the sum capacity with the number of antennas [16], and due to the potentially unique (or at least widely differing) spatial signatures of the different transmitters and receivers, linear processing techniques offer an attractive method to separate the signals of different users, due to their more modest computational complexity requirements.

With the explosion in development of MIMO technologies in the mid-1990s with the seminal works of Telatar [16] and Foschini [17], and similar progress over the same time frame in the areas of spectrum sharing and cognitive radio [18], it is perhaps surprising that at the time of writing no standard is in place covering spatial-domain spectrum sharing. It can be argued that this form of spectrum sharing offering fine-grained control in the spatial domain incurs a cost in terms of for instance hardware (multiple antennas, front-ends) and system complexity (more than one entity operating in the same frequency band), that has caused industry to invest in other technologies with better short term returns. Separately, the research community have brought forward both MIMO technologies and hierarchical access as key elements in the next generation mobile communications standards [19, 20, 21]. However, it has also been emphasised for spectrum sharing in general that in order to enable wide-spread adoption issues such as protocol design for efficient spectrum use, incumbent or priority user incentivisation, definition of harmful interference, and techniques for enabling harmonious coexistence must still be resolved [22, 23]. This state of affairs has resulted in commercial entities hesitating due to a lack of strong evidence in favour of sharing [24, 25]. In an effort to explore the use of multiple antennas in a spectrum sharing setting, two separate policies are proposed in this thesis. The first policy aims to take advantage of the interference
1.3 Contributions

This thesis focuses on the use of the spatial dimension in spectrum sharing applications. The main contributions may be summarized as:

1. A competitive spectrum sharing model between a small and a large MU-MIMO system is proposed and analysed. The significant service antenna number asymmetry poses unique challenges and opportunities. In the limit of an infinite number of service antennas at one of the access points it can be shown that with simple coherent combining (i.e., linear processing techniques) the noise and interference terms vanish and single-user performance is approached. These distinctive traits inherent to massive MIMO have been widely reported (see e.g., the works by Marzetta [26], Rusek et al. [27], Larsson et al. [28], and Björnson et al. [2]). However, in the present case, with the coexistence of a small MIMO system, it is unclear how these systems will interact and what the quantitative impact on performance (transmit power and SINR) will be. In order to investigate the scenario it is proposed to frame the problem as a common mean SINR maximization problem. The solution to the problem formulation leads to the characterization of the transmit power and mean SINR. In addition, concise power scaling laws are presented giving insight into how transmit power and mean SINR behave with service antenna dimensions and number of user terminals. Due to the existence of both a small and a large MIMO system, the study did not resort to use the asymptotic (in the number of service antennas) results available in the literature. Instead as a starting point, the mean SINR of the matched filter and zero-forcing equalizers in the uplink (and maximum ratio transmission and zero-forcing precoding in the downlink) are used. Extensions to these expressions are provided for the cases where a zero-forcing equalizer suffers additional external interference,
and for the case where the interference seen by a receiver originates from a transmitter using zero-forcing precoding. Upper and lower bounds to the mean SINR are presented, and these are used to derive the scaling laws. The characterization of the transmit power and SINR allows an assessment of the behaviour of a small and large MU-MIMO system in a spectrum sharing scenario. The derived scaling laws can be used to gain an intuitive understanding of the effect of changing for instance the system load (i.e., number of user terminals) or how the systems can be dimensioned with respect to the number of service antennas in order to achieve a specific operating point.

2. A MIMO hierarchical access system is investigated that breaks with the traditional non-aware scenario [8]. The principle is that through cooperation, the primary adjusts its linear receive filter to allow the secondary transmitter to employ a subset of its antennas, and place its interference in the null-space. Cooperative feedback, whereby the primary aids in estimating the cross-channel state information has been previously explored [29]. Also, cooperation in the sense that the secondary aids in relaying the primary’s message has also been investigated in the overlay cognitive radio paradigm [30]. However, the use of primary active cooperation to increase spectrum utilization while satisfying the primary users’ rates in the underlay or interweave paradigms has not yet been investigated. In general this enables secondary communication (and hence increasing spectrum usage), while incurring a cost in terms of transmit power at the primary system in order to keep the desired rate requirements. Specifically, the objective is to maximize the rate of the secondary user, subject to a secondary user power constraint and individual primary user rate constraints. The optimal solution involves iterating over all possible secondary antenna combinations and is hence only suitable for small systems. For larger systems, two algorithms are proposed based on a greedy selection and an orthogonality criterion, respectively, that show good performance at high SINRs. It is demonstrated that cooperation enables secondary user communication in situations where zero-forcing beamforming or opportunistic interference alignment remain infeasible. Simulation results highlight the advantage of cooperation over opportunistic interference alignment in situations where the secondary user only just has enough transmit antennas to perform interference alignment.

1.4 Outline of Thesis

The rest of this thesis is organized as follows. Chapter 2 provides an overview and state-of-the-art review of spectrum sharing technologies, focusing on multiple antenna
technologies. Chapter 3 presents the small and large MU-MIMO coexistence study in a competitive environment. The scenario is formulated as a maximum mean SINR problem, and in addition to transmit power and mean SINR characterization, concise power scaling laws are derived that provide insight into how the number of antennas and user terminals affect performance. In Chapter 4 a cooperative interference system is studied, where the spatial dimension is used to enable spectrum sharing. Specifically, the primary system reserves a certain sub-space for interference thereby allowing the secondary system access. The scenario is investigated in terms of secondary user rate and the cost to the primary system in terms of transmit power. Conclusions and possible extensions of this thesis appear in Chapter 5.

1.5 Publications

Journal Publications


Conference Publications

Chapter 2

Background and Related Literature

2.1 Introduction

For a basic understanding of spatial-domain spectrum sharing, this chapter begins by reviewing multiple antenna techniques. The point-to-point MIMO channel is introduced, as well as extensions to multi-user channel models, in order to illustrate the potential benefits of MIMO in terms of spectral efficiency, spectrum utilization and diversity enhancements. The commonly used interference channel is also mentioned since it captures scenarios in which multiple users wish to communicate simultaneously in the presence of mutual interference. The principles of spectrum sharing and the potential advantages of employing the spatial-domain for this purpose are then reviewed.

Figure 2.1: The MIMO channel. The signal transmitted from each antenna is received (through different “paths”) at each receive antenna.
2.2 The Multiple-Input Multiple-Output Channel

2.2.1 The Point-to-Point Channel

Consider a narrowband time-invariant channel, as in Fig. 2.1, with \( N \) transmit antennas and \( M \) receive antennas described by the deterministic channel matrix \( H \in \mathbb{C}^{M \times N} \). Given an input vector \( x \in \mathbb{C}^{N \times 1} \), the output of the channel is described by

\[
y = Hx + v \tag{2.1}
\]

where \( v \in \mathbb{C}^{M \times 1} \) denotes the added noise. The noise term is commonly seen as the sum of different unknown signals such as the internal thermal noise of the receiver or interference from other sources. With the assumption that there are many (ideally infinite) small such noise contributions the term \( v \) is assigned a complex circularly symmetric Gaussian distribution by calling on the central limit theorem [31]. The noise variance is normalized to \( \mathbb{E}(vv^H) = I \) in this chapter to reduce notation clutter.

The relationship in (2.1) is generally called the vector Gaussian channel. It is interesting to note that although the capacity of the scalar version of the Gaussian channel was derived by Shannon in 1948 [32], extensions of this result to the MIMO channel were not given until the mid-1990s by Foschini [17] and Telatar [16]. For the vector Gaussian channel the achievable rate, with coherent detection, is defined as

\[
R = \log |I + HQH^H|, \tag{2.2}
\]

where \( Q = \mathbb{E}(xx^H) \) is the transmit covariance matrix. The notation \(|·|\) and \((·)^H\) is used to denote the determinant and the Hermitian operation, respectively. The derivation of (2.2) will not be repeated here (the interested reader is recommended to review Telatar’s work [16]), but it is still worth pointing out the main assumptions and steps to gain an understanding of the use and limitations of the relation:

1. The noise is assumed complex circularly symmetric Gaussian. Not only does the previous argument using the central limit theorem seem natural, it can also be shown that no other distribution possesses a larger differential entropy for a given variance (hence the assumption is actually of a “worst case” type).

\[\text{It must be noted that work pre-dating Foschini or Telatar on the MIMO channel do exist, but its impact seems to have been limited. For instance, Tsybakov [33] derived the MIMO capacity with transmitter side information for a deterministic channel using a WF approach in 1965. (original text in Russian, paper result explained by Tulino et al. [34]). Brandenburg and Wyner [35] considered the MIMO channel capacity with memory in 1974. From a detection perspective, Shnidman seems to have been one of the earliest researchers to consider the MISO channel in 1967 [36].}\]
2. With an average transmit power constraint, $E(x^H x) \leq P$, the input distribution that maximizes the mutual information is also a complex circularly symmetric Gaussian (since it is an entropy maximizer).

3. Using the expression for the mutual information, $I(x; y) = h(y) - h(v)$, and the differential entropy expression for a Gaussian random variable, $h(z) = \log(\pi e |\Sigma|)$, with $E(zz^H) = \Sigma$, an explicit form for the achievable rate in (2.2) is derived.

4. Limiting the case to where $N = 1$, the expression in (2.2) reduces to the case of a single stream (or user) capacity. Now using the expressions in the previous point gives some insight into why all communications engineers are obsessed with the term signal-to-noise-ratio. Crucially, note that due to the logarithm terms in the derivation, the SNR is not the result of $E(a/b)$, where $a$ is the signal power component and $b$ is the noise power portion.

It has to be emphasised that although the optimal (i.e., maximizing the mutual information) input distribution has been used to derive (2.2), the equation can still not be designated the channel capacity since the transmit covariance matrix $Q$ remains to be chosen to maximise the achievable rate. The method of deciding $Q$ generally falls into two separate categories: a) channel state information (CSI) is available at the transmitter, or b) CSI is not known to the transmitter.

For the case with CSI at the transmitter, a remarkable result can be shown. The capacity can be computed by decomposing the vector channel into a set of parallel, independent scalar sub-channels. The steps are as follows. First, by using the singular value decomposition (SVD) technique, the channel matrix can be decomposed into

$$H = U \Sigma V^H,$$

(2.3)

where $U \in \mathbb{C}^{M \times M}$ and $V \in \mathbb{C}^{N \times N}$ are unitary matrices (i.e., $AA^H = A^H A = I$) and $\Sigma \in \mathbb{C}^{M \times N}$ is a diagonal matrix with non-negative real elements. Now if the covariance matrix is chosen as $Q = VD V^H$, where $D$ is another diagonal matrix where the diagonal elements contain the transmit power allocations, it can be shown that the expression inside the log $\cdot |\cdot|$ in (2.2) reduces to a diagonal matrix. According to Hadamard’s inequality, this diagonal form maximizes the determinant. All that is left to do is to optimize the diagonal entries of $D$ to assign individual transmit power values for the parallel sub-channels. This is carried out with the WF method as

$$|D|_{ii} = \max \left(0, \mu - \frac{1}{|\Sigma|_{ii}^2}\right), \quad i = \{1 \ldots K\}$$

(2.4)

where $\mu$ is designated the water level, $K$ is the rank of the channel matrix $H$, and $[,]_{ij}$ is used to denote the $i, j$th element of the matrix. The water level $\mu$ is chosen to satisfy
2.2. The Multiple-Input Multiple-Output Channel

the power constraint with equality. Explicit calculation of $\mu$ is generally carried out using a search algorithm (see Palomar and Fonollosa’s paper for practical algorithms [37]). It is worth pointing out a few key characteristics of the optimal solution:

- It was shown that by using the SVD decomposition of the channel, the achievable rate with an associated power constraint is maximized. Interestingly, it can also be very easily shown by substitution into (2.1), that if $x$ is left-multiplied by $V$ prior to transmission and, if $y$ is left-multiplied by $U$ after reception, the elements of $x$ are completely separated (orthogonal) and can be individually detected at the receiver. This is what is meant by parallel, independent transmission.

- Note that the power allocation is entirely determined by the gains of the sub-channels (i.e., the associated eigenvalues of $HH^H$). Hence, with WF the “strongest” sub-channel is allocated the most power, while “weaker” sub-channels may not be allocated any power at all. This in turn implies that the number of parallel streams transmitted over a particular MIMO channel is a function of both the available transmit power and the channel realization.

- Lastly, the optimal solution was obtained with the assumption of (ideal) CSI at the transmitter. Gaining knowledge of the channel state is associated with a certain cost that has been ignored in the analysis.

For the case with no CSI at the transmitter, a slightly different view of the nature of the channel has to be taken to arrive at a meaningful result. As an example, consider that a single channel realization is drawn from some distribution, and a predefined rate is used to communicate over the channel. Except for the case where the probability distribution of the channel has been doctored to support the rate, reliable communication is not possible since there will always be a non-zero probability that the desired rate exceeds the capacity of the channel. Strictly speaking, the capacity of such a channel is zero, and the notion of outage probability (i.e., the probability that the channel cannot support the desired rate) can be used instead.

One of the best known results on the MIMO channel is instead derived by considering $H$ to be a random matrix, independent of both $x$ and $v$, and for each use of the channel an independent realization of $H$ is drawn. Under such conditions, and assuming each entry of the matrix $H$ is zero mean, complex circularly symmetric Gaussian, Telatar showed that the capacity of the Gaussian channel with Rayleigh fading is given by

$$C = \mathcal{E}_H \left( \log \left| I + \frac{P}{N} HH^H \right| \right)$$

(2.5)

where $P$ is the total normalized transmit power and $\mathcal{E}(\cdot)_H$ is the expectation with respect to the random matrix $H$. It is from (2.5) that the multiplexing gain, pre-log
factor or number of degrees of freedom relation for the MIMO channel was first shown. In essence it states that in the high SNR regime the capacity of the MIMO Rayleigh fading channel is about $K = \min(M, N)$ times the capacity of the equivalent single-input single-output system. Lozano et al. reflect that it was perhaps this observation that spurred the interest in multi-antenna communication from the mid-1990’s onwards [34].

The presented MIMO channel is a point-to-point communication channel. It is true that by for instance dividing the available resources into separate blocks it is possible to reduce the prevalent multi-user systems into equivalent point-to-point links. Indeed many of the existing cellular systems make use of for instance time division multiplexing (TDM) or frequency division multiplexing (FDM) to share the available resources among users [38]. Ideally, this creates an interference-free environment due to the orthogonality between users, where point-to-point communication techniques can be applied. In addition, the idea of orthogonality has been applied to mitigate interference from adjacent cells with the use of spatial reuse partitioning. A natural question to ask when confronted with the idea of dividing up the available resources in this way is: Are these schemes optimal with respect to resource utilization, fairness or computational complexity (to name a few concerns)?

In particular the question regarding resource utilization in the sense of spectral efficiency is still challenging the information theory society today. For single-hop networks (i.e., the information is passed directly between a source and a destination) three commonly studied multi-user channel models are: a) the multiple access channel (MAC), b) the broadcast channel (BC), c) the interference channel (IC). These models will be reviewed next.

### 2.2.2 The Multiple Access Channel

The MAC models a collection of sources all transmitting to the same destination. For instance this could be a group of mobile terminals communicating to a base station, i.e., in the uplink, or a set of wireless LAN (local area network) enabled devices transmitting to an access point. The discrete-time baseband model for the MIMO MAC is [38]

$$y = \sum_{k=1}^{K} H_k x_k + v$$  \hspace{1cm} (2.6)$$

where $K$ is the number of sources or transmitters and $H_k$ represents the channel between transmitter $k$ and the destination. At first sight, it seems a daunting task to determine the transmitted messages from the superposition of all users’ signals. Indeed, this is one of the reasons why in for instance the GSM system a strict time and
2.2. The Multiple-Input Multiple-Output Channel

frequency scheduling arrangement is kept to separate the signals of all the users and simplify detection. At a second glance at (2.6), the savvy reader may notice that the destination observes a vector, i.e., the destination has multiple views of the set of transmitted symbols $x_k, k = 1\ldots K$, and by viewing the problem as a set of simultaneous equations the transmitted symbols may be estimated as long as the system is not underdetermined (i.e., the number of independent equations $\geq$ the number of variables). This is one way to view spatial division multiple access (SDMA), i.e., the use of multiple antennas (critically in this case at the receiver side) to enable the equalization or separation of more than one user. Simply attempting various solutions such as the two examples above (no matter how good the rationale) does not seem to effectively give an answer to how close the schemes are to the optimal (w.r.t. rate). For such an answer, one must again turn to the information theoretic tools of entropy and mutual information. For the 2 user case, the rate region of the vector Gaussian MAC channel is confined to

$$R_k \leq \log |I + H_k Q_k H_k^H|, \quad k = 1, 2$$  \hspace{1cm} (2.7)

$$R_1 + R_2 = \leq \log |I + \sum_{k=1}^K (H_k Q_k H_k^H)|$$  \hspace{1cm} (2.8)

where $Q_k$ is the covariance matrix of the transmitted signal from source $k$. There are several points to note for this rate region:

- For the Gaussian MAC the optimum input distribution that maximizes the achievable rate is again complex circularly symmetric Gaussian [5].

- For the 2 user case, the inequalities create a rate region in the shape of a pentagon. Surprisingly, this set of equations indicate that it is possible for user 1 to transmit at $R_1$ as defined by (2.7) and user 2 can still achieve a non-negative rate (the difference between the r.h.s of (2.8) and $R_1$). These extreme points in the rate region can be shown to be achieved with an MMSE-SIC (minimum mean square error successive interference cancellation) type receiver [38].

- The MIMO MAC rate region can still be optimized with respect to the transmit covariance matrices. Tse and Hanly [39] exploited the polymatroid structure of the SISO MAC rate region to derive an optimal resource allocation strategy for the fading channel. This method was later extended by Mohseni et al. to the MIMO MAC [40].

2.2.3 The Broadcast Channel

The BC model features a single source (e.g., a base station) transmitting separate information to multiple users (i.e., user terminals). The MIMO downlink channel is
succinctly described by
\[ y_k = H_k x + v, \quad k = \{1, 2 \ldots K\} \tag{2.9} \]

where \( H_k \) is the channel from the source to destination \( k \), and in this instance the input vector \( x \) contains independent messages to be delivered to the \( K \) users. It must be noted that the capacity region of the general MIMO BC is still an open problem. However, in 2006 Weingarten et al. presented a capacity region for the MIMO Gaussian BC [41]. The capacity region itself is described as the union of the dirty paper coding (DPC) rate regions over the set of positive semidefinite covariance matrices \( Q \), where in turn the DPC region is defined over the convex hull of the union taken over the set of all permutations (interference cancellation ordering) and power allocation strategies. For an overview of the MIMO Gaussian BC capacity region, including explicit characterization of the capacity region, the reader is referred to the tutorial by Caire et al. [42].

Although the description of the capacity region for the MIMO Gaussian BC is truly a very abstract concept, the tool used to approach the boundary is perhaps less so. The dirty paper coding scheme, proposed and named by Costa in 1983 [43], is a type of interference pre-cancellation. Costa showed that if the interference is known (non-casually) at the transmitter this signal can be cancelled out and the capacity of this channel is the same as that of a channel devoid of interference. As an example consider a 2 user MIMO Gaussian BC. Encode the message for user 2 using Gaussian coding, then encode user 1 using DPC by treating the signal to user 2 as known interference. At the receiver side, user 2 detects its signal and suffers from an additional Gaussian interference due to the signal intended from user 1, but at the receiver of user 1 the intended message can be decoded without suffering from interference. Of course, the roles of the two users can be exchanged and this gives rise to the permutations issue found in the description of the DPC rate region.

2.2.4 The Interference Channel

The IC models a set of source and destination pairs, where the receivers suffer interference from all undesired sources. For all its close associations with real communications systems, from an information theoretic perspective even the simplest 2 user IC has remained unsolved. The simplest incarnation of the IC, the 2 user SISO (i.e., 2 source and destination pairs) IC has a long history. It appears to have first been mentioned by Shannon in 1961 [44], and early studies of the channel were reported by Ahlswede [45] and Carleial [46, 47]. In addition Carleial mentions in his 1978 paper [47] that some overlap in results with other works by Sato and Bergmans exists.
The main results on the 2 user SISO Gaussian IC are often presented in terms of the class of the interference channel. These categories can best be understood by realizing that the 2 user IC can undergo a scaling transformation to normalize both direct channel gains to unity. In effect, the scaling operation converts the IC into an IC with the same capacity region, only the direct channels are normalized, and new (scaled) cross channels, power constraints and noise variances are assigned (for details see Carleial’s paper [47]). The categories are subsequently defined as [6]:

- Strong interference: both cross channels are equal or greater than unity.
- Moderate or weak interference: either cross channel is less than unity.
- Z-interference: one of the cross channels is zero.

For the strong interference case Carleial proved that the capacity is equal to the case with no interference at all [46]. This result is due to interference cancellation: the received interfering signal is strong enough to be decoded first and subtracted from the received signal, uncovering the desired signal. The capacity regions for the remaining categories, however, are unfortunately still unknown [6].

A general framework for the study of the capacity region is the idea of determining inner and outer bounds. Where the bounds meet, the result is exact. The idea however, critically hinges on finding good bounds, and this seems to be carried out to a large extent by the intuition of the researcher. The greatest intuition therefore, was perhaps shown by Carleial [46] and later by Han and Kobayashi [48], who proposed the use of rate-splitting codes to allow a portion of the interfering signal to be decoded. With this strategy each user employs a part of its transmission power to transmit a common message that is decodable at both receivers, and the remaining power to transmit a private message. In the work by Han and Kobayashi, it is in addition assumed that the transmitters may employ time sharing to reach any point in the convex hull of the rate region. Unfortunately, the resulting optimization problem is not well understood, and hence it is unclear in general how close to optimal such a scheme can be, and under what channel parameters significant improvement would be seen. For the Gaussian interference channel in the high SNR regime, the gap between the outer bound and the inner bound was recently reduced by Etkin et al. [49].

There seems to be even fewer results available for the Gaussian MIMO IC. In fact, to the author’s best knowledge, the capacity region of the MIMO IC is known only for the case termed aligned-strong interference (note similarity with SISO strong interference class) where the direct and cross channel matrices are linked through an equation [50]. Considering the channels as random variables the probability of this relation holding...
2.2. The Multiple-Input Multiple-Output Channel

seem very small (depending on the channel distributions, the probability may well tend to zero).

In general, characterizing the capacity of the IC for even special cases or classes seems a daunting challenge, and although considerable effort has been expended and the gap between the inner and outer approximations to the IC capacity region has been repeatedly improved over the last five decades, the results in general still remain very abstract and difficult to interpret in a way that lead to specific achievable schemes. Perhaps partially as a reaction to the difficulty in obtaining meaningful results for the IC, there has recently been an explosion of research into the IC by looking at the problem from a different perspective. By moving away from attempting to characterize the capacity region as a whole and focusing on the sum rate of the system as the signal-to-noise ratio approaches infinity, i.e., by looking at the degrees of freedom of the system, the concept of interference alignment was first proposed for the MIMO X channel (a channel where each receiver desires messages from both sources) by Jafar and Shamai [51]. A general method to align an arbitrarily large number of interferers was published soon after by Cadambe and Jafar [7], firmly establishing the interference alignment concept and surprisingly showing that wireless networks are not essentially interference limited [52]. A few points to note about interference alignment are:

- Cadambe and Jafar [7] showed that the K user interference channel almost surely (i.e., using a probabilistic argument) has \( \frac{K}{2} \) degrees of freedom.

- The cost of interference alignment however was demonstrated by Grokop et al. [53] who showed that the sum capacity can only be made to scale linearly with the number of users as the number of dimensions is allowed to grow with K. This bandwidth scaling requirement is a common feature of interference alignment schemes [52].

- Determining the feasibility of interference alignment in general has been shown to be NP-hard by Razaviyayn et al. [54]. Simplified results for single streams have however been derived by Yetis et al. [55].

- Numerous algorithms for interference alignment have been proposed for alignment in time, frequency, space or code. For instance, alternating minimization [56, 57], minimizes the interference leakage by iteratively adapting receive and transmit side linear filters. A different approach has been based on minimum mean square error (MMSE) beamforming [58]. A third method applies rank constrained minimization in order to maximize the number of accessible dimensions [59].
It has to be noted that interference alignment algorithms in general, where the objective is related to the alignment of interference, have only been shown to be asymptotically optimal (i.e., in a degrees of freedom sense). Hence IA may not be the optimal solution at finite SNR [60].

2.2.5 MIMO Detection

The derivation of the MIMO point-to-point channel in Chapter 2.2.1 hinted towards a simple precoding, equalization and detection scheme. If the transmitter has CSI knowledge, it is possible to decompose the channel and transmit along the “eigen-channels”. This effectively creates a set of parallel (independent) channels and readily available point-to-point channel coding schemes can be used to approach the theoretical capacity (for an overview of channel coding, refer to the work by Costello and Forney, Jr. [61]). What are the specific challenges of MIMO point-to-point communications, then? Examining the above case, it seems that communication is enabled by three main ingredients: 1) CSI at both transmitter and receiver, 2) the channel decomposition is feasible from a computational complexity perspective, and 3) channel coding and decoding is also possible within the required latency (also a computational complexity issue). Starting from the last point, channel coding and decoding is certainly not a trivial operation, but thanks to modern scalar codes such as Turbo or LDPC codes and appropriate decoding techniques [61], very good performance is achieved with the hardware available at the time of writing. As for the computational complexity of the SVD decomposition of the $H \in \mathbb{C}^{M \times N}$ channel, it is on the order of $MN^2$ operations [62]. This roughly cubic complexity order (assuming $M \approx N$) is acceptable in many applications. The focus therefore is on the availability of CSI at both receiver and transmitter. CSI at the receiver side is a requirement for coherent detection [38], and in practical systems knowledge of the channel is gained through a pilot or training sequence (a known signal is sent by the transmitter for the estimation of the channel). The alternative to coherent detection, i.e., noncoherent detection, is generally seen as a less spectrum efficient approach, even when considering the training overhead. CSI at the transmitter can be gained through some feedback mechanism, and it can be seen that the feedback overhead grows at a rate proportional to $MN$ if the (narrowband) channel coefficients are modelled as i.i.d. random variables. In addition, the finite coherence time of the channel (i.e., the duration the channel is considered static) implies that any feedback mechanism that introduces delay will result in stale or inaccurate

\^Note that the complexity order of an algorithm allows for easy comparison between methods in terms of number of operations, but it is the end application that ultimately dictates the solution. As a counter-example to where a cubic order complexity may be considered too expensive is the use of ZF in CDMA type receivers with very long spreading codes [63].
CSI at the transmitter [64].

These observations led for instance Foschini [17] to propose the Bell Labs Layered Space-Time (BLAST) architecture, where due to the lack of CSI at the transmitter, the separation of the individual streams is handled solely by the receiver. The final message is that if the MIMO channel cannot be “untwisted” into a set of parallel (orthogonal) channels, then this basically forms a multi-user detection problem where optimal joint-decoding at the receiver follows an exponential growth in computational complexity [65]. Multi-user detection approaches therefore often fall back to using three common heuristic equalization methods to first separate the streams and then decode each stream individually. This is certainly a reasonable approach, since for the multi-user scenario, the users may be assumed to transmit independent messages, and without sharing data between the transmitters, joint precoding (in the sense of transmitting a combination of all the users’ symbols) is not feasible. At the receiver therefore, a linear filter, $G$, is applied to equalize the channel as:

$$ z = Gy = GHx + Gv, \quad (2.10) $$

and a decision on the transmitted symbols is made by looking at the individual elements of $z$. Note that the multi-user detection problem has now been reduced to $N$ single user detection problems and all available scalar coding and decoding techniques can be applied to reduce uncertainty about the estimated messages. In general, if the channel does not consist of a set of mutually orthogonal vectors, treating the symbol elements separately leads to a loss (increased probability of error or decreased rate) due to the inherent coupling of the variables. However, it must also be stressed that the two most commonly cited reasons for splitting the received signal in this manner with a linear filter are due to computational complexity constraints and analytical tractability. To gain an overview of linear equalization and an idea of their performance, the three most prevalent linear equalization methods, matched, zero-forcing, and minimum mean squared error filters are covered next [65].

**Matched Filtering**

Consider the output of a linear receive filter for the MIMO channel with a single transmit antenna (strictly speaking this is the filter output of a SIMO channel):

$$ z = gy = ghx + gv, \quad (2.11) $$

where $h \in \mathbb{C}^{M \times 1}$, and $g \in \mathbb{C}^{1 \times M}$ is the linear receive filter. The SNR is simply $\|gh\|^2/\|g\|^2$. This is readily seen to be in the form of a Rayleigh quotient [66, Theorem 4.2.2], and the maximum SNR is the eigenvalue of $hh^H$, i.e., $h^Hh$, and it is obtained
2.2. The Multiple-Input Multiple-Output Channel

when \( g = \alpha h^H \), where \( \alpha > 0 \) is an arbitrary scaling factor. The same result can also be found by examining the Cauchy-Schwarz inequality [66, Theorem 5.1.4].

Hence, having restricted the receive filter to a weighted sum of the observed signals at the receive antennas, the optimal (in a max SNR sense) weights match the channel response. For more than one transmit antenna (and more than one stream), the MF has been generalized to \( G = H^H \) (the arbitrary constant \( \alpha \) set to 1 for all vectors). It can be seen that the filter maximizes the individual signal components, but it does not take the interference from the other streams or users into account at all. Although it may be poorly suited for multi-user separation, it does have one very nice feature: given that the receiver has a channel estimate, the computational complexity of the filter is very modest (i.e., on the order of \( MN \)). However the mean SINR performance of the filter does suffer in the face of interference as Fig. 2.2 shows. The plots in Fig. 2.2 are based on the asymptotic expressions obtained from the work by Tse and Hanly on the use of linear multi-user receivers [39]. Clearly, the MF performs better than the ZF at low SNRs (\( SNR < 0 \text{ dB} \), but at higher SNRs, the ZF equalizer has an advantage over a significant portion of the load due to its nulling of interference.

Zero-Forcing Equalization

While the MF does not take interference into account at all, the aim of the ZF equalizer is to null all interfering terms, leaving only the desired signal. Hence, ignoring the noise term in (2.1), the problem becomes equivalent to finding the solution to a system of linear equations. The filter matrix \( G \) is therefore given by the pseudoinverse of \( H \), or explicitly \( G = H^\dagger = (H^H H)^{-1} H^H \). For the pseudoinverse to exist, the channel \( H \) must be of rank \( N \), and for the case where \( M = N \), this reduces to the channel inverse \( H^{-1} \). The computational complexity of the ZF method does vary somewhat, depending on the approach used. For an exact calculation of the Moore-Penrose pseudoinverse one algorithm claims a complexity of \( MN^2 + N^3 \) [67]. This is on the same order, i.e., cubic, as the SVD decomposition, and in fact many commercial packages make use of the SVD for the calculation of the pseudoinverse. Other approaches look for an approximate solution based on for instance Neumann series or the Cayley-Hamilton theorem (see e.g. Rusek et al. [27]). This can result in substantial savings, but accuracy may suffer. Hence, when an inverse operation is involved it will be assumed in this thesis that the cost is of order three.

Assuming uniform power allocation, \( p \), the SINR of user or stream \( n \in \{1, \ldots, N\} \) can readily be seen to be:

\[
\text{SINR}_n = \frac{p}{(H^H H)^{-1}}_{nn}.
\]  
(2.12)
2.2. The Multiple-Input Multiple-Output Channel

Figure 2.2: Approximate mean SINR performance for an arbitrary user or stream versus load. The figures (a), (b), (c) and (d) depict an SNR (i.e., no interference) set to $-10, 0, 10, 20$ dB, respectively.
2.2. The Multiple-Input Multiple-Output Channel

For the Rayleigh fading channel, potentially with transmit side correlation, Gore et al. [68] showed that the SINR follows a Gamma distribution. For an alternative derivation, please see Appendix A.2. The mean SINR for instance is therefore readily available. Note that the mean ZF sum rate was also considered by Matthaiou et al. [69] for correlated Rayleigh fading channels and uncorrelated Ricean fading channels. These results were presented as a series of lower and upper bounds, due to the intractability of the exact expressions. Unfortunately, the bounds still contain rather unintuitive elements, e.g. Euler’s digamma function.

Using the mean SINR as a metric, the performance of the ZF equalizer is shown in Fig. 2.2. Finally, take note of the performance at either extremes \( N = \{1, M\} \). At \( N = 1 \), i.e., the interference free scenario, the ZF filter reduces to the MF. At the other end when the system is fully loaded, \( M = N \), the per stream or user SINR is abysmal. The performance is closely linked with the idea of the condition number of a matrix [66, Chapter 5.8], and specifically for the case of a matrix with i.i.d. Gaussian elements, the condition number suffers as the matrix approaches the square case [70]. The condition number specifically deals with bounding the error of a matrix inverse operation, and is often stated simply as a ratio of the largest to smallest singular values. The condition number has previously been applied to for instance switch between a ZF approach and another more robust technique in adaptive MIMO detection [71].

**Minimum Mean Squared Error Equalization**

The “interest” of the MF is restricted to simply maximizing the signal power component, while the ZF filter suppresses all interference terms completely. It is natural to ask if there is a linear filter that balances these two extreme operating points. The simple answer is affirmative, and it takes on the form of the linear MMSE filter. Formally the filter is defined as the solution that minimizes the mean squared error [72]:

\[
G = \arg \min_G \mathcal{E} \left( \|x - z\|^2 \right).
\]  

(2.13)

The problem is in essence no different from finding the minimum of a scalar quadratic equation. The issue of finding the derivative of a scalar function of a complex-valued matrix variable is covered by Hjørungnes and Gesbert [73], and so setting the derivative of the mean squared error to zero, the filter matrix can be explicitly stated as [74]:

\[
G = \left( H^H H + Q^{-1} \right)^{-1} H^H,
\]

(2.14)

where \( Q \) is the transmit covariance matrix as previously used in the MIMO channel capacity relation, (2.2). It has to be noted that this filter is also known as the Wiener filter [75]. Considering the case where all streams or users transmit at the same power,
it can be observed that as the power is increased, the first summand in (2.14) begins to dominate and ultimately the MMSE filter tends to the ZF solution. Also, as the power is decreased the second summand in (2.14) begins to dominate and the MMSE tends towards the MF.

A perhaps more intuitive explanation for the performance of the MMSE filter, is given by Tse and Viswanath [15, Chapter 8.3.3] in the multi-user context (the same line of reasoning was also given in another paper by Tse and Hanly [39]). Consider only the recovery of one user’s signal. It was seen above that the MF maximizes the signal portion, but ignores the structure of the noise. Hence, Tse and Viswanath argued that if a pre-whitening filter is applied first to whiten the interference and noise, and then apply the MF, the SINR of the stream is maximized. More formally Gao et al. [76] for instance show that minimizing the mean squared error of a stream, maximizes its SINR.

From a computational complexity perspective, the MMSE suffers roughly the same complexity as the ZF filter due to the inverse operation, i.e., of order three. Finally it is important to note that when the transmit antennas are distributed and do not belong to the same system, estimating CSI may not be trivial. However, it can be noted that whereas the ZF filter requires specific CSI, the MMSE actually only needs an estimate of the interference and noise covariance matrix. This lumped together estimate could in some situations be easier to obtain than separate channel paths.

The distribution of the SINR of the $n$th output ($n \in \{1, \ldots, N\}$) does not seem to be available in a tractable form, however [77]. In the case of Rayleigh fading, the ccdf is given as a polynomial of order $M - 1$ by Gao et al. [76]. Instead of working with the exact distribution, Li et al. approximated it by using a Gamma distribution [78]. However it has to be noted that the fitting procedure also involves a recursive step and hence complicates any further computations. Tse and Hanly [39] derive a concise approximation based on random matrix theory (where the dimensions of the matrix $H$ tend to infinity at a specific ratio). This form was used in Fig. 2.2 to give an appreciation of the mean SINR performance in comparison with the other two linear equalizers. The MMSE, optimally balances between signal gain and interference suppression and hence always dominates over both MF and ZF.

## 2.2.6 MIMO Precoding

MIMO precoding, or the use of a precoding matrix to adjust the transmission from a multiple antenna transmitter, takes on a very similar form as the discussion in Chapter 2.2.5 on MIMO detection. In fact, the three most common linear receiver types have direct equivalents on the transmit side [74]. The main difference is that in the derivation
of the filters, some form of power constraint is required [72]. Letting \( \mathbf{F} \in \mathbb{C}^{N \times N} \) denote the precoding matrix, and the data vector by \( \mathbf{s} \in \mathbb{C}^{N \times N} \), the transmitted combinations over the antennas are:

\[
\mathbf{x} = \mathbf{F}\mathbf{s}.
\]  

(2.15)

Four power constraint alternatives are [79, 72, 74]:

- An instantaneous power constraint: \( \mathbf{x}^H \mathbf{x} = \mathbf{s}^H \mathbf{F}^H \mathbf{F} \mathbf{s} = P \).
- A constraint on the maximum eigenvalue of the precoding matrix: \( \rho(\mathbf{F}^H \mathbf{F}) = P \), where \( \rho(\cdot) \) is used to denote the spectral radius (maximum eigenvalue) operator.
- An average power constraint over the data symbols: \( \mathcal{E}_s (\mathbf{x}^H \mathbf{x}) = \text{tr} (\mathbf{FQ}_s \mathbf{F}) = P \), where \( \mathbf{Q}_s \) is the covariance of the data symbols.
- An average constraint over the data symbols and the channel: \( \mathcal{E}_{s,\mathbf{H}} (\mathbf{x}^H \mathbf{x}) = P \).

The list is roughly ordered in a “stringency” sense. The first constraint type prevents any fluctuation in the transmit power, while the last constraint allows fluctuations both over the channel realization and data symbol amplitudes. It was noted by Scaglione et al. [72] that the second constraint limiting the excursions of the maximum eigenvalue of the precoding matrix can be seen as a form of peak power constraint. From this perspective it may model the limitations of a power amplifier the closest. However, the third power constraint seems to be the most popular in the scientific literature due to its simple explicit form. Other forms of power constraints are of course also possible. For instance Yu et al. [80] attempted to model the more realistic case where each transmit antenna was fed by an individual power amplifier, i.e., where there was a per-antenna power constraint.

The naming convention for the transmit equalization filters is somewhat more complicated. The filter of the form \( \mathbf{F} = \sqrt{1/P} \mathbf{H}^H \), is called the transmit MF by Joham et al. [74]. However in the multi-user scenarios, i.e., where the receive antennas are “distributed”, the term maximum ratio transmission (MRT) [81, 82] is more commonly found. For single stream applications the term “beamforming” has often been appended to contrast it against the case with multiple streams (where the terms precoding or precoding matrix are often used). Also, the same filter structure has been called eigen-beamforming by Hoydis et al. [83] and conjugate beamforming by Yang and Marzetta [84]. The interference nulling transmit equalization filter, \( \mathbf{F} = \sqrt{1/P} \mathbf{H}^H (\mathbf{H}^H \mathbf{H})^{-1} \), is generally called some variant including ZF. However, in an attempt to improve the performance of the ZF filter, Peel et al. [85] added a “regularization” term to stabilize the inverse operation, and called it regularized ZF. The complication arises when one
2.2. The Multiple-Input Multiple-Output Channel

considers the optimal (in terms of maximum SINR) regularization parameters: then the regularized ZF is equal to the MMSE transmit filter. Finally, the transmit Wiener filter (as named by Joham et al. [74]) of the form, \( F = \alpha H^H (HH^H + M/P)I^{-1} \), where \( \alpha \) is a scalar function of the power constraint, and interference and noise components, is also generally known as the linear MMSE precoding matrix.

2.2.7 A Note on Channel Assumptions

The focus of this work is on the investigation of multiple antennas as an enabler for spectrum sharing. In this sense, the spatial signatures of the users (or equivalently the paths between the transmit antennas and receive antennas) is a crucial factor in enabling the separation of different streams, as exemplified by the linear equalizers and precoders listed in Chapters 2.2.5 and 2.2.6.

To take an example in case, the ZF equalizer or ZF precoder requires that the rank of the channel \( H \in \mathbb{C}^{M \times N} \) is \( N \leq M \) or \( M \leq N \), respectively, for the existence of the pseudo-inverse. Can this full rank condition be met in practice? The argument that is often given is that in scattering-rich environments with sufficient separation between the transmit (receive) antenna elements to reduce correlation the full rank assumption is valid. As a rule of thumb, around half a wavelength or more is needed between antenna elements [86] to be able to claim independent fading. The rich scattering environment giving rise to multi-path propagation can be seen to be a possibility in typical urban environments, modern office spaces, etc., but can probably not be fulfilled in rural areas where a lack of obstacles and objects make the two-ray ground-reflection model very accurate [87]. Finally, it can also be seen from the MIMO capacity relation (2.5), that the rank of the channel dictates the pre-log factor and hence has a significant impact on the rate performance of the system.

In the preceding discussion, a flat fading channel was assumed for clarity of exposition. However, the rich scattering environment assumption implies a multitude of paths which in turn implies different path delays and hence a frequency selective channel model. This contradiction does not however constitute a major hindrance. Modern communications systems capitalize on the efficient fast Fourier transform to convert a frequency selective channel into a set of flat fading sub-channels [88]. This transformation, generally known as orthogonal frequency division multiplexing (OFDM), incurs an overhead in the form of for instance a cyclic prefix to ensure that the resulting channel is circulant, but is otherwise capacity preserving (it is a unitary matrix transformation). Hence the methods presented in this thesis can be seen to be on a per sub-channel basis.

Finally, it is assumed throughout that the individual channel elements are distributed
as i.i.d. complex Gaussian random variables. This of course means that the individual channel element gains follow a Rayleigh distribution, and hence the name of the Rayleigh fading channel. As noted by Biglieri et al. [89], the Rayleigh fading channel is the simplest of the three most common channel models. The other two most widely used statistical wireless channel models are the Nakagami-m, and the Ricean fading models. The Nakagami-m model is based on a variation of the Gamma distribution, and with an additional parameter compared to the Rayleigh fading model (basically a shape parameter), it can be easier to fit to actual channel measurements. The Ricean fading model builds on the Rayleigh channel model by allowing for a line-of-sight component, i.e., a non-zero mean component. The motivation to limit the study to the Rayleigh fading channels is three-fold. First, comparison to other works is greatly facilitated since the vast majority of relevant papers also base their simulation work on the Rayleigh fading channel. Secondly, the other two channel models generally restrict the analytical tractability further. Lastly, for non-line-of-sight environments the model is known to give representative results (see e.g. the recent channel measurement work by Gao et al. [90, 91]).

2.3 Spectrum Sharing

2.3.1 Background

The range of usable radio frequencies employed by wireless systems is generally referred to simply as the spectrum in the technical literature. Although the radio spectrum portion of the electromagnetic spectrum is defined as the frequencies spanning from $3 \text{ Hz}$ to $3 \text{ THz}$ (i.e., $3 \times 10^{12} \text{ Hz}$) [92], the usable range, or at least the most highly valued, is between $100 \text{ MHz}$ and $6 \text{ GHz}$ and is affectionately called the “beachfront spectrum” [93]. Glancing the over-crowded U.S. frequency allocation chart [94] and considering that the price for leasing such beachfront spectrum is nearly 10 million pounds per MHz (e.g., Vodafone paid £790 761 000 for a total of 95 MHz in 2013 [95]) confirms the moniker.

The static frequency allocation policies have been known to suffer from poor performance (e.g., high call drop rates, or low per-user rates) due to excessive user density and data requirements [10, 96]. Paradoxically, measurement campaigns (e.g., made by FCCs Spectrum Policy Task Force [97]) have shown that at any given time and location significant portions of the prized spectrum lies idle. Put simply, spectrum allocation is simply not matched with demand in time, frequency and space. This should come as no surprise since contemporary policy only exploits the frequency degree of freedom to ensure different systems, operators, and users are separated: a frequency band is
2.3. Spectrum Sharing

Given or leased to an entity for a specific purpose and the time span is on the order of decades. Spatially, the license covers large geographical areas (up to a national level).

This mismatch in demand and supply has attracted significant attention from the research community and industry. In an attempt to gain an overview of the state of spectrum allocation, Zhao and Sadler [98] divided the problem into a set of broad categories. Fig. 2.3 shows the author’s interpretation of the taxonomy. Note that Zhao and Sadler also prefixed every term with “dynamic” in order to emphasize that future spectrum allocation is envisioned to be a much more active and adaptive process, contrasting it to the static model currently used by regulatory bodies. In addition, the fundamental polar model of exclusive use on one extreme and entirely unlicensed spectrum utilization on the other extreme [99] have also been included in the figure. From the figure it can be seen that spectrum access is divided into three sub-branches:

- **Exclusive use:** this format maintains the idea of a single system in a given band. One approach is spectrum property rights, where access is given exclusively to one party and therefore mirrors the state of current regulations. It is envisioned that some flexibility is included by granting the licensees the right to trade spectrum and giving the licensees more freedom in choosing transmission format. The other approach, dynamic spectrum allocation, is said to offer further flexibility by offering allocations on a much finer time scale (and in geographical space) compared to spectrum property rights. Note that this form has a distinct similarity with Mitola’s idea of spectrum pooling [100].

![Spectrum Access Diagram]

Figure 2.3: Spectrum access categorization. Note the attempt to include the idea of a polar model of exclusive use and commons at either extreme, and possible combinations in between.
2.3. Spectrum Sharing

- **Hierarchical Access**: the use of different tiers allows guarantees on the quality of service (e.g., availability, rate, delay) for primary or priority users, and opens the spectrum to secondary users, under certain conditions. This type of dynamic spectrum access, as at the time of writing, best known simply as CR. It is felt that the term CR best encompasses the capabilities of the transceiver, and these features may be used in scenarios outside hierarchical access. Hence the need to separate the terms hierarchical and cognition. The three basic coexistence scenarios will be further explained in Chapter 2.3.2.

- **Commons**: the spectrum commons [101] models sharing among peers. A typifying example is the use of the unlicensed bands (e.g., industrial, scientific, and medical (ISM), or Unlicensed National Information Infrastructure (U-NII) band) for the operation of wireless services. Systems such as Bluetooth, Wifi, and cordless phones share the 2.4 GHz ISM band and their universal use in modern day life is testament to their success. However, their success may also be their Achilles’ heel: over-crowding increases the interference they cause to each other and may ultimately render the use of such bands useless for communications. Spectrum sharing between equals is further expanded in Chapter 2.3.3.

In addition to the main branches outlined above, a variety of spectrum sharing models have been proposed by industry, regulatory bodies and the research community. These models must meet a host of requirements in order to be adopted. From a technical view for instance, the spectrum sharing method must protect incumbent or priority users, enable sharing under all reasonable circumstances, be compatible with existing systems, and have a realistic computational complexity. From a business stance, the method must justify the additional cost in infrastructure, create a sufficient amount opportunity to generate interest in the model, and preferably allow for simple monitoring to detect malignant behaviour. At the time of writing only the use of television white space (TVWS) has been standardized in for instance the US. In the white space scenario the secondary user must interrogate a database to gain access, as spectrum sensing alone has been deemed insufficient [25]. Other notable advancements are for instance license shared access (LSA), where the incumbent may lease spectrum to a secondary user [102, 22, 103, 104]. A similar approach has also been proposed under the so called authorized shared access scheme. Lastly, collective use of spectrum (CUS) is a spectrum sharing approach proposed by the European Commission based on an unlicensed model, where transmit power restrictions, duty cycle limits, or coordination between users could be introduced to mitigate interference. Also note that TVWS technology at this moment is restricted to a time interweave paradigm, and the first trials of LSA [102] have been limited to utilizing available time slots. The use of the spatial domain has largely remained untapped. It is hypothesised that the added system com-
plexity in using multiple antenna techniques, the availability of suitable hardware and the requirement of at least cross-channel state information has hampered the spread of spatial domain spectrum sharing. The above limiting factors have of course also highlighted the question of how to protect incumbents’ performance effectively.

The breakdown of spectrum access as depicted in Fig. 2.3 is to some extent artificial, and other categorizations exist. For example Berg et al. [99] propose the use a priority-based taxonomy, where each system is classified according to priority level, and some identifiers tied to how the system can be classified as active or inactive. Peha [105] instead applies another grouping based on coexistence or cooperation and as peer sharing or hierarchical sharing, in order to explore how to enable or facilitate spectrum sharing arrangements. No single division of the spectrum sharing or spectrum allocation terminology seems to clearly expose all facets of the problem, and this is of course common in any problem where there are many interactions and the choice of metric (or group of metrics) to assess the system can take many forms.

From the above exposition two observations can already be made:

- Dividing the frequency into separate bands and regulating their use is a simple method to control interference. This is the original reason behind spectrum licensing [99], and enables for instance network operators to give quality of service guarantees. This idea of protecting communications from interference remains with the hierarchical access approaches (the primary or incumbent user has priority).

- The current static spectrum management policies have lead to a spectrum scarcity versus under-utilization dilemma.

Perhaps more importantly, the published literature unanimously agrees that while a static frequency allocation is simple, more flexible and adaptive approaches are needed to combat inefficient use of spectrum. In addition to using time and frequency domains in a more dynamic manner, research has naturally extended to the use of the spatial domain to enable and enhance spectrum sharing [10, 106, 96, 107]. Generally, exploiting the spatial domain through the use of multiple antenna techniques can provide the transceiver the capability to balance its own rate, SINR or outage probability (to name a few performance metrics) and the interference it causes to others. MIMO technologies therefore have the potential to act as an enabler of fine-grained spatial spectrum sharing.

With this initial glance at spectrum sharing, how can the term spectrum sharing be made a little more precise? Certainly, spectrum sharing is a form of resource allocation between systems, users, and even services. This wide definition includes the regulators’ division of spectrum to different systems (with different radio access technologies) or
network operators, all the way down to how the smallest resource block is allocated to a user. In this thesis, however, the interest is restricted to the behaviour of two or more wireless systems operating simultaneously, in the same frequency band, by using the additional degrees of freedom offered by multiple antenna techniques.

2.3. Hierarchical Access

Hierarchical access, i.e., where a primary user has priority over a secondary user, is commonly known simply as CR. However, the term CR as used by Mitola [100] is a radio aware of its surroundings and capable of adapting to its environment. In this sense, the author agrees with Zhao and Sadler [98] that CR enables dynamic spectrum access, but can also be used in a variety of other scenarios that require intelligence.

In the scope of hierarchical access three common approaches are generally advocated. Note that there seems to be no strict consensus on these basic terms, and that in this thesis the definitions suggested by Goldsmith et al. [8] are used.

- **Interweave**: In the classical interweave paradigm the secondary user opportunistically communicates in spectrum holes. These spectrum holes are defined as temporary unused frequency bands where the secondary user can transmit without causing any interference to the primary user. This definition can be straightforwardly extended to the spatial domain, where the secondary user can align its interference into the null-space of the primary user.

- **Underlay**: This paradigm has been commonly used to describe the setting in which the secondary user transmits simultaneously and over the same frequency with the primary user, but is not allowed to significantly degrade the performance of the primary user. Specifically, an interference power constraint, commonly designated the allowed interference temperature [97, 108], has to be strictly obeyed. Note that the question of how the secondary user obtains a sufficiently accurate estimate of the interference at the primary user is not included in the definition. This scheme has been extended to the MIMO case by for instance Scutari et al. [96] who have called it a soft-shaping constraint. In this form a limit on the total interference power summed over the primary user receive antennas is set. In essence this provides a relatively simple constraint, but note that since multiple streams can be deployed over a MIMO system, there is no simple relationship between SINR and achievable rate, as there is for the SISO channel.

- **Overlay**: In the overlay paradigm the secondary user transmits simultaneously along with the primary user, but any potential harm to the primary user is offset
by the secondary user actively assisting in the transmission of the primary user’s message. The first use of this paradigm seems to have been by Devroye et al. [30]. The idea is that the secondary user has knowledge of the message and assigns a proportion of the transmit power to relay the message and the remaining part to transmit its own message. Compared to the other two paradigms this scheme seems to suffer the most from the severe assumption of non-casual knowledge of the primary user’s message at the secondary user’s transmitter. Indeed this scheme may perhaps be too information theoretic, but Goldsmith [8] does give an example where non-casual knowledge at the secondary user transmitter is not required: in scenarios where the primary user applies retransmission (e.g. an ACK / NACK protocol) the secondary user can offer assistance. In the spatial domain this definition remains unchanged.

As previously mentioned, cognition has played a significant role in hierarchical access research. For the interweave paradigm in time and frequency, a large effort has been rightfully spent investigating spectrum sensing techniques suitable for the detection of spectrum holes (e.g. see the work by Axell et al. [109] or by Sharma et al. [110] for an overview). For the overlay paradigm, the level of cognition must be higher still: the secondary user must gain knowledge of the primary user’s message. In the underlay approach, it has been argued that the channel to the primary receiver from the secondary transmitter may be estimated, e.g. by intercepting a pilot signal if operating in TDD mode, or through some blind estimation technique [111].

The realization that users tend to have different spatial signatures (assumption holds in a rich-scattering environment), extends the idea of resource allocation in time and frequency to the space dimension as well. With multiple antennas, additional freedom is provided to the SU to optimize its own rate and control the interference at the PU receiver. By the same token, the spatial-domain may be leveraged by the PU to for instance mitigate received interference or maximize rate. It has been shown for the point-to-point channel that significant spectral efficiencies can be gained with the use of MIMO. It has also been demonstrated that multiple antennas at the receiver of a MAC enables SDMA. Can similar benefits be found with the use of MIMO techniques for the spatial-domain spectrum sharing scenario?

The field of spatial-domain spectrum sharing is in many ways still evolving. On the one hand MIMO techniques have only really taken off with the work of Telatar [16] and Foschini [17] these last two decades, and on the other hand spectrum sharing, although not a fledgling topic, has in many ways also seen most progress over a similar time span with the rise of cognitive radio [18]. Adding to this, in its most elementary form, the two-tier system consists of a single PU and a single SU interfering with each other. This is the very definition of the IC for which there exists very few good answers.
Focus will now be limited to the interweave and underlay paradigms, since the overlay case in general is considered a very optimistic use case. Extending the MIMO channel model to a $K$ transmitter-receiver pair IC, the signals received at the $k$th node is

$$y_k = H_{kk}x_k + \sum_{n=1, n\neq k}^{K} H_{kn}x_n + v_k,$$

(2.16)

where $H_{kk}$ is the channel between the $k$th transmitter and the $k$th receiver, and $H_{kn}$ is the cross-channel from the $n$th transmitter. Note that this model can also be used to describe for instance two or more interfering MACs or BCs. If receiver $k$ is considered the primary user, it has priority and a limit on the interference caused by the other transmitters is imposed (note the implicit assumption that interference is treated as noise).

How this constraint is specifically defined and regulated is still an open regulatory issue. Generally speaking, restricting the interference too much reduces the opportunity of the secondary systems, while setting a loose limit may jeopardize the performance of the primary system. For the MIMO setting Scutari et al. [96] proposed three different interference power constraints, based on the idea of an “interference temperature limit” at the primary receiver [97, 108]. These are: the null-shaping constraint, the soft-shaping constraint, and the peak power constraint. The peak power constraint limits the magnitude of the largest eigenvalue of the interference covariance matrix at the primary receiver. This has a close similarity with the peak transmit power constraint introduced in Chapter 2.2.6. The soft-shaping constraint for secondary transmitter $k$ is defined as

$$\mathcal{E} \left( \text{tr} \left( \sum_{n \in \mathcal{N}} (H_{kn}x_kx_k^H H_{kn}^H) \right) \right) \leq \text{ITC}_k,$$

(2.17)

where $\mathcal{N}$ is the set of primary users, and $\text{ITC}_k$ is the maximum allowed aggregate interference power for transmitter $k$. Setting $\text{ITC}_k = 0$ results in the null-shaping constraint. Note that if the secondary transmitter is equipped with only a single antenna, the null-shaping constraint results in zero transmit power and so no communication. With multiple antennas, the transmitter may be able to “steer” or shape its transmissions to enable information transfer.

The concept of interference temperature limit is somewhat controversial. Firstly, it is not entirely clear how it should be set to safeguard the performance of the primary system and stimulate secondary access. Secondly, the issue of how the secondary systems gains access to this information is still hotly debated. The inherent uncertainty involved with wireless communications for instance prompted the idea of a probabilistic limit instead of the above soft-shaping constraint [98]. In addition, the idea of an aggregate limit such as in (2.17) may also cause issues. A simple limit on the interference power...
may not adequately reflect the effect on user rate. This was observed by Cumanan et al. [112] who instead proposed to use a primary rate constraint directly. While this constraint may prove to be more useful, the question of how to collect the necessary side-information for the secondary system to be able to calculate the limit was left unanswered. A recent resource allocation survey by Tsiropoulos et al. for instance, still exclusively use only the simpler interference temperature limit [113], and many modern heuristic hierarchical access algorithms still adhere to this constraint due to its simplicity [114]. This poses a problem for fine-grained spectrum sharing. For the interference temperature limits to be effective, these need to be set at a very conservative level, and may reduce secondary access opportunities significantly.

After Gastpar’s seminal work considering MIMO capacity in a spectrum sharing scenario with an interference power constraint [10], it seems that Zhang and Liang [11] were among the first to investigate secondary user rate subject to primary interference constraints using linear transmit precoding. Specifically, the scenario consisted of one secondary user, and a group of primary receivers. It was shown that the problem is convex and hence efficient numerical algorithms exist to find the secondary transmitter’s precoding matrix. Using computational complexity as an argument, Zhang and Liang also proposed two algorithms for the case where the secondary user has multiple antennas both at the transmitter and receiver. One algorithm is based on allocating power on the eigen channels of the secondary’s direct channel (cf. with WF on the MIMO channel, (2.3)), while the second algorithm instead limits transmit power to the null-space of the primary receiver (i.e., a null-shaping constraint, and hence a form of “block” ZF). Zhou and Thompson [115] analysed a two user MISO interference channel where one user is subject to an interference power constraint, and the results mirror the work by Zhang and Liang. Zhang et al. [116] then followed up on their initial sum rate maximization problem by investigating the robustness of the solution to partial channel state information.

A central aspect of spectrum sharing (regardless of if it occurs between peers or in a hierarchical fashion) is that the problem is cast as some type of a resource allocation problem. In the above discussion the focus was on sum rate maximization. Two other common objective functions as stated by Pennanen et al. [117], are: SINR balancing and power minimization. For instance, Tajer et al. [118] and Cumanan et al. [119] investigated downlink beamforming using an SINR balancing approach. Zheng et al. [120] designed the beamforming vectors considering channel state uncertainty. From a power minimization perspective with appropriate minimum SINR or rate constraints and an interference temperature limit, hierarchical access has been considered by Phan et al. [121], Gharavol et al. [122], and Huang et al. [123]. The last two works also specifically tackle robust beamforming, in the sense that uncertainty in the
2.3. Spectrum Sharing

estimated link will not lead to intolerable performance degradation for the primary system. It should be explicitly noted that the original work by Zhang and Liang [11] considered rate maximization with a single secondary transmitter-receiver pair. This allowed a convex problem formulation. Similarly, it was shown by Tajer et al. [118], and separately by Zheng et al. [120], that the SINR balancing problem for \( K \) secondary transceivers could be re-cast into a convex form. However, both the power minimization problem and sum rate optimization with more than one secondary transceiver have been shown to be non-convex [124], and hence finding a global optimum is computationally intractable. Therefore, the above works have resorted to approximating the original problem in an effort to find good local solutions. In addition, it must be emphasized that the above works also attempt to deal with the fact that the multiple secondary users are usually considered autonomous and hence distributed algorithms are an essential requirement for any practical implementation.

Tackling the spectrum sharing problem from another direction, Perlaza et al. [125, 12] introduced the concept of interference alignment (as previously reviewed in Chapter 2.2.4) for hierarchical access. In the original work, a strict null-shaping constraint is assumed, and hence the multiple transmit antennas of the secondary transmitter are used to place the interference into unused sub-spaces at the primary receiver. This is of course a ZF method, but the key idea is to “squeeze” a number of secondary transmit streams into an existing null-space. In other words, the principle is to consolidate the interference into a smaller subspace. If this can be achieved the degrees of freedom (or pre-log factor) is not simply divided among the \( K \) users, but could be made to grow at a rate of \( K/2 \) [7]. In the same vein, Nosrat-Makouei et al. [126] proposed an interference alignment scheme to accommodate new arrivals (i.e., the new user could be seen as a secondary user).

The previous body of work considered secondary access, where some form of interference constraint is imposed to safeguard the performance of the primary or incumbent system. However, very little attention was paid as to how the secondary system copes with the “unregulated” interference from the primary transmitters. In the case of the interference alignment schemes, a common technique seems to be to offset the performance degradation with additional receive antennas. As for the sum rate maximization, SINR balancing and minimum transmit power problems mentioned above, the focus was on finding suitable solutions, but the ultimate performance crucially depends on the number of primary transmitters, cross-channel states and primary transmit power. Few works have been found where these issues are treated directly. One example is by Tran et al. [127], who investigates the impact of outage probability and rate of a secondary user with respect to primary transmit power and channel gain. It must also be mentioned that Bixio et al. [128] studied a single primary, single secondary
transceiver system, where the secondary receiver explicitly nulls interference from the primary transmitter.

The vast majority of MIMO techniques require CSI. In a spectrum sharing scenario, knowledge of the cross-channels from the secondary to the primary transceivers is vital, but far from trivial to acquire. Assuming an unaware primary system, the secondary system must be a sufficiently capable cognitive radio and sense or learn the channel properties. This is realised as one of the main concerns in real-world deployments [129]. One method to alleviate the risk of poor estimation, is for instance to allow cooperation between the primary and secondary for the sake of gathering CSI, as outlined by Chen et al. with the concept of cooperative feedback [29]. However, these are not the only challenges facing spatial-domain spectrum sharing. As emphasised by Bhattarai et al. [23], there are several open problems in spectrum sharing in general obstructing wide adoption. Issues such as protocol design for efficient spectrum use, incumbent or priority user incentivisation, definition of harmful interference, and techniques for enabling harmonious coexistence must be clarified in order to push spectrum sharing into a commercial setting.

### 2.3.3 Spectrum Sharing Between Equals

In hierarchical spectrum sharing, the primary user(s) are protected in some manner. This requirement often appears as a constraint in the problem formulation. However when two or more systems share the spectrum as peers, the design of transmission strategies naturally constitutes a multi-objective optimization problem [130, Section 1.4.2], where the performance of each link (SINR, rate, etc.) is a separate single objective function. The idea of global optimality is then extended to the Pareto optimal surface [96]. In terms of rate vectors, this would be interpreted as the rate region boundary. As previously discussed in Chapter 2.2.4 the rate region for even the simplest two user SISO IC is analytically intractable. In addition the solutions to the common system utility functions\(^*\) (i.e., specific points on the Pareto boundary) [124, 130]:

- Weighted arithmetic mean (e.g. sum-rate),
- Proportional fairness,
- Harmonic mean,

\(^*\)A prudent point made by Björnson and Jorswieck [130, Section 1.6] is that all utility function are subjective to some extent. In other words there is room to argue on the merits and weaknesses of any utility function. This in turn suggests that the four listed utility functions, although in widespread use, are not unique and other metrics can of course be used (if a suitable argument can be conjured).
have been shown to involve prohibitively high computational complexity (i.e., they have been classified as NP-hard problems) [124]. Consequently, research has focused on finding “good” less complex algorithms that work in various situations. Two common ways to characterize such situations are centralized versus distributed, and cooperative versus competitive.

For the scenario where a set of MIMO secondary users share the spectrum with a primary system, Scutari et al. [96], explored the resource allocation problem from a non-cooperative (or competitive) game-theoretic perspective. This allowed the use of the Nash equilibrium concept in designing an iterative distributed algorithm. The Nash equilibrium in this sense is used to establish a stable operating point. Arslan et al. [131] described the equilibrium as a steady-state situation where no individual link has an incentive to change its choice. In the situation described by Scutari et al. [96] this lead to the development of an asynchronous iterative WF algorithm, where essentially each secondary user optimizes its own transmit precoding matrix in reaction to its interference and noise covariance matrix. The same method was used (without the primary interference temperature constraint) for the multi-user MIMO system by Scutari et al. [132]. A very similar approach was also proposed by Arslan for the MIMO interference channel [131]. Note that one of the key elements in designing such algorithms is determining if the algorithm actually converges to a stable point. The iterative WF algorithm proposed by Scutari et al. [132] does not possess universal convergence, but conditions on when the algorithm is stable has been extensively documented [133].

The predecessor to the MIMO iterative WF algorithm was proposed by Yu et al. [134] for the frequency selective digital subscriber line. Although iterative WF is certainly an elegant idea, and due to its distributed nature has a small implementation threshold, it was pointed out by Popescu et al. [135] that the resulting equilibrium point may be well inside the rate region. By introducing cooperation, Huang et al. [136] attempt to overcome this limitation of the iterative WF algorithm. The idea is based on exchanging interference prices to take into account the effect one user or system has on the other parties. In fact, in one of the first studies on cooperative game theory for the MISO interference channel Larsson and Jorswieck [107] use the argument that the Nash equilibrium corresponds in many cases to an outcome which is “bad for all players”. This “price of anarchy” (i.e., selfish users attempting to maximize their own profit) [137], lead Larsson and Jorswieck [107] to consider a Nash bargaining solution to the two user MISO interference channel. This requires some mode of communication between the users (in direct similarity with the work by Huang et al. [136]) where the operating point can be negotiated, and always results in an element-wise greater than or equal rate allocation compared to the Nash equilibrium. Interestingly, Etkin et al.
2.3. Spectrum Sharing

[138] came to similar conclusions for the sharing scenario over a frequency band for non-cooperative entities. Crucially, the results hinge on the use of spectrum sharing rules (that are self-enforced) that gives enough incentive for the users to work in the same direction. It must be pointed out that both Etkin et al. [138] and Larsson and Jorswieck [107] emphasize that when the systems are very asymmetric (e.g. channel gains are very different, or transmitters have widely varying capabilities) neither bargaining or spectrum sharing rules provide any form of fairness.

The potential poor result of the Nash equilibrium (i.e., an operating point well inside the rate region boundary), and the generally intractable form of the Bargaining solution [107], has lead a number of researchers to propose heuristic algorithms. For instance Yu and Lui [139] attempt to approximate a non-convex function as a convex equivalent using a time-sharing argument, and investigate SISO spectrum sharing algorithms using the dual of the optimization problem. The work was originally triggered by the paper by Cendrillon et al. [140] who claimed vastly improved performance with a centralized algorithm approach. Centralized, iterative approaches were also proposed for the MIMO interference channel by for instance Ye and Blum [141], and Rong and Hua [142]. Spurred by the promise of a linear growth in the degrees of freedom by using interference alignment, Negro et al. [143] proposed an iterative (between transmitter and receiver) scheme to maximize the sum rate of the MIMO interference channel, basing the design on an MMSE idea. Peters and Heath, Jr. [60] similarly designed cooperative algorithms for MIMO networks, specifically with the objective to either minimize interference, minimize the mean squared error, or maximize a function of the SINR. All the proposed MIMO approaches apply some form of iterative updating of the transmit and receive filters and have been stated to converge to a stable operating point. Due to the iterative nature of the algorithms, and at times the objective function, it is difficult to gauge performance in any other way but to compare numerical results. In addition, the question of how CSI is gained still remains unanswered. In most works, local CSI is assumed implicitly, i.e., the receiver is able to estimate the channels from all other transmitters. In addition if the systems operate in time division duplex mode, channel reciprocity may be exploited [144] to acquire local CSI at the transmitters. With more capable transceivers the other users’ CSI may be intercepted by overhearing a feedback channel, and finally learning mechanisms have been suggested to acquire CSI. All these methods may be put into question in terms of feasibility and accuracy, and the relative overhead cost is heavily dependent on the coherence time of the system.
2.4 Discussion

In this chapter the basics of MIMO were explored from an information theoretic perspective covering the point-to-point channel and the more common multi-user channels. This was used as a foundation for introducing the spatial-domain aspect to spectrum sharing. Noteworthy points are that MIMO techniques can offer substantial increases in capacity compared to SISO (shown for the point-to-point channel) and for instance enable SDMA for the MAC channel. The guiding idea behind spatial-domain spectrum sharing is to harness the additional space dimension and just as in the point-to-point case or the MAC channel demonstrate the benefits of multi-antenna systems [65].

The fundamental principle of utilizing multiple antennas to enable fine-grained spectrum sharing in either hierarchical access or in a general authorization regime has been reviewed. It was shown that although few problem statements have been found that offer optimal solutions at a manageable computational complexity, there are a whole host of heuristic methods that generally find a good compromise between performance and complexity. Such methods have been suggested for beamforming (SIMO or MISO) applications to control the level of interference caused to others or to mitigate interference. The use of MIMO has likewise been effectively used to balance rate and interference, but the investigations have also highlighted the fact that multiple streams generally complicates the already difficult problem structure. Although the problem structure represents a challenge in itself, it is not the only obstacle in the path of spatial domain spectrum sharing. As previously noted, there are several key open problems in spectrum sharing in general [23]: protocol design for efficient spectrum use and utilization, incumbent or priority user incentivisation, definition of harmful interference, and techniques for enabling harmonious coexistence. The above issues amount to a lack of strong evidence in favour of sharing, that has to date impedes business entities such as mobile network operators from proceeding with any significant practical deployments [24, 25].

For the specific case with multiple antennas it is worth emphasising the interference aspect. The vast majority of available spatial domain spectrum sharing studies (with a few notable exceptions such as the work by Cumanan et al. [112]) have assumed a simple sum interference power constraint. The main issue with this approach is that the interference power is not an effective means to judge the performance impact on the incumbents (or other users in the case of general authorization) [145]. This in turn leads to significant difficulty in attempting to assess the level of protection a particular spectrum sharing scheme offers an incumbent, and increased uncertainty with regards to spectrum access. One of the aims of this thesis is to attempt to frame the spatial-domain spectrum sharing problem in a way that focuses on the relationship
between the systems. In a first step this implies using either the SINR or rate metrics explicitly, instead of an interference temperature constraint. Secondly, recognising that spatial-domain spectrum sharing generally introduces additional interference, the gap in performance should be investigated, either as a function of the SINR or rate, or indirectly as an increase in transmit power. The interplay of the systems is evidently critical in a spectrum sharing regime, and the incorporation of multiple antennas add a greater level of flexibility. As an example, the use of multiple receive antennas has previously been explored in order to handle additional interferers, but it has not yet been employed to explicitly reduce transmit power in order to allow coexistence. Such coexistence may be extremely well suited to general authorization schemes where the systems generally have equal priority. It is envisioned that harmonious coexistence may be possible with the correct scaling of the systems in terms of load, number of service antennas and transmit power. Another use of multiple antennas is in an active cooperation scheme, where a primary user adapts its post-processing matrix to accommodate the secondary user. Previously, primary-secondary cooperation has been set in the overlay cognitive radio paradigm, but it has not been extended to the interweave or underlay regimes, and particularly not with the use of multiple antennas as the enabling technology.
Chapter 3

Competitive Spectrum Sharing between a Small and a Large MU-MIMO System

3.1 Introduction

With significant research focused on both spectrum sharing in unlicensed bands and on the use of a large number of service antennas to improve area spectral efficiency, it is perhaps not unlikely that systems equipped with widely varying antenna numbers may have to coexist in the near future. Both hierarchical access [146] and massive MIMO [147, 148, 20] have also been highlighted as key technologies for the next generation mobile communications standard, 5G.

The concept of spectrum sharing was covered in Chapter 2.3 and for the present case may be narrowed down to the sharing of the same time-frequency resource by equal priority users. Multiple antenna deployments allow the transceivers more freedom (i.e., not only power and rate control, but may also harness the varying spatial signatures of the users) to combat interference. In this scenario, the widely touted characteristics of massive MIMO deployments; inherent interference insensitivity and low (approaching zero) transmit power, makes it an ideal candidate.

A key ingredient of massive MIMO, or large scale antenna systems, is a large number of service antennas [26, 27, 2, 149]. However, this is not the complete definition. Perhaps a more apt explanation of massive MIMO is “growing the number of service antennas relative to the number of active users” [150]. With the increase in service antennas to users proportion the distinctive traits become apparent. The users spatial signatures’ become more orthogonal (given a full rank channel assumption), enabling the use of simple linear techniques to separate users without a significant penalty in SINR, and
any additional interference can easily be rejected. The array gain contributes directly to lower transmit power [151], and the large number of service antennas lead to a channel hardening effect [152] where the effective channel gain becomes independent of channel realization. The orthogonal quality of massive MIMO may be seen in Fig. 2.2, where at low loads the SINR remains level. From the plots it is also clear that for instance that transmit power influences the behaviour of the MF and MMSE designs, and hence it does not seem reasonable to define massive MIMO in terms of a specific load or number of service antennas.

Key differences between conventional multi-user MIMO and massive MIMO are that for smaller systems, the addition of a user substantially decreases SINR, and scheduling mechanisms need to be in place to account for channel quality fluctuations. For 5G systems it is primarily the spatial-multiplexing capability of massive MIMO with the computational complexity of a linear receiver that is attractive [150]. Massive MIMO has also been proposed for millimetre-wave bands in 5G [2], but it must be emphasized that for multi-user operation to work effectively a rich-scattering environment is required. There is still much debate whether the much more significant absorption of millimetre-wave radiation may hamper such an approach.

For the present case, it is unclear to what extent these massive MIMO characteristics hold with a large but finite number of service antennas, and how these systems should be scaled to control the amount of interference one system causes the other. The key contribution of this chapter is the performance analysis of an interference system consisting of one small and one large MU-MIMO system. The main challenges included finding an appropriate framework to allow analysis and ways to handle both the small and large number of service antennas at the access points. The main outcome are concise transmit power scaling laws, clearly showing the behaviour of transmit power (and hence mean SINR) with respect to the number of service antennas and system load (number of user terminals).

The remainder of the chapter introduces the system model and challenges, describes the method used to analyse the interference system in a cohesive manner and presents mean SINR results along with approximations for added insight. The chapter concludes with potential implications for spectrum sharing with very asymmetric access points.

3.2 Massive MIMO in Unlicensed Bands: System Modelling

Consider the situation where two wireless systems share spectrum in an unlicensed band in the local area. A massive MIMO AP (AP, System 1) having $M_1$ service antennas
communicates with a set of (say $N_1$) single-antenna UTs. Simultaneously, a small MU-MIMO system operates in the same band, where an AP employing $M_2$ service antennas serves $N_2$ UTs. It is assumed that these two systems do not cooperate, and they compete for radio resources under certain policies (see Chapters 3.3 and 3.4 for the analysis in the uplink and downlink, respectively).

The above scheme could be used to describe the physical-layer model of indoor application scenarios, where the massive MIMO based cellular radio system utilizes unlicensed bands for mobile traffic offloading and has to share spectrum with another system (at the time of writing a concrete example is the MU-MIMO version of Wi-Fi). For the sake of communications efficiency and reliability, there is a great need for efficient spectrum-sharing strategies. The two main questions pursued in this chapter; how to view or model this system, and further how the transmit power and SINR behave, aim to investigate what performance can be achieved. The main challenge arises from the setup of very unequal number of service antennas at the two systems. The term “very unequal” is elaborated by the condition:

$$c1) \quad M_1 \gg M_2 \text{ (by convention, much greater is taken as at least an order of magnitude larger)}$$

It can also be seen that in general, the number of UTs, or load, will also vary greatly between the two systems. The relative load however, Defined as $\beta_i = N_i/M_i$, $i \in \{1, 2\}$, may not necessarily be very different. Intuitively, condition c1) reflects the fact that the massive MIMO system enjoys a much larger antenna gain than the small MU-MIMO system. Impact of this condition for spectrum sharing will be discussed for both uplink and downlink.

### 3.2.1 System Description and Modelling

#### Uplink

Consider both systems operating in the uplink. Despite the fact that competition occurs across all channels (or sub-channels) of the unlicensed band, the focus will be on the competition behaviour in one of the sub-channels, with emphasis on the physical-layer procedure. Moreover, the commonly utilized assumptions that the channels are flat fading, and that the two systems are synchronized in time and frequency, will be adopted [132, 107, 153]. This enables an appreciation of the main issues in a simpler setting, and aids in the construction of the key ideas based on well established MU-MIMO channel models; the same assumptions also apply to the downlink procedure in Chapter 3.4. The flat-fading assumption is reasonable in many situations where
3.2. Massive MIMO in Unlicensed Bands: System Modelling

OFDM is in use (see also the discussion in Chapter 2.2.7), such as LTE [154] or the 802.11a, c, g, and n based Wi-fi standards [155]. As for the assumption of time and frequency synchronization (as well as cross-link channel knowledge), it does requires a more sophisticated massive MIMO system. However, several techniques exist that make this assumption reasonable. Where TDD is in use, channel reciprocity can be exploited to acquire CSI. Feedback channels often form an integral part of a wireless communication protocol [156] and may provide valuable indirect information of channel state. This can also be combined with learning mechanisms [56] for instance to iteratively learn the required CSI (and the response of the other system).

Denote $x$ to be the information-bearing symbols sent by the UTs. Also, designate the subscript $(\cdot)_n$ to be the user index, with $n \in \mathcal{N}_1 = \{1, 2, \ldots N_1\}$ indicating UTs from System 1 and $n \in \mathcal{N}_2 = \{(N_1 + 1), (N_1 + 2), \ldots (N_1 + N_2)\}$ indicating UTs from System 2. Analogously, service antennas are indexed by the subscript $(\cdot)_m$, with $m \in \{1, 2, \ldots M_1\}$ indicating service antennas within System 1, and $m \in \{(M_1 + 1), \ldots (M_1 + M_2)\}$ indicating those within System 2. Every symbol goes through their corresponding UT to service antenna channels, termed $h_{m,n}$. The discrete-time equivalent form of received baseband signals at the APs are

$$y_1 = \sqrt{p_1} H_{11} x_1 + \sqrt{p_2} H_{12} x_2 + v_1,$$

$$y_2 = \sqrt{p_1} H_{21} x_1 + \sqrt{p_2} H_{22} x_2 + v_2,$$

where $y_1$ is an $M_1 \times 1$ signal vector received at AP 1, $y_2$ an $M_2 \times 1$ signal vector received at AP 2; $x_1 \triangleq [x_1, \ldots, x_{N_1}]^T$ and $x_2 \triangleq [x_{N_1+1}, \ldots, x_{N_1+N_2}]^T$ are the transmitted symbol vectors; and $v_1$ and $v_2$ are white complex Gaussian noise vectors at AP 1 and AP 2, respectively, with covariance $\sigma^2 I$. Note that the superscript $(\cdot)^T$ denotes matrix (or vector) transpose, and $I$ is the identity matrix. The channel matrices $H_{11}, H_{12}, H_{21}, H_{22}$ are the corresponding sub-matrices of the compound MIMO channel matrix $H$, whose $(m, n)$th entry is $h_{m,n}$. It is assumed throughout that the channel entries are i.i.d. zero-mean circularly symmetric complex Gaussian random variables, i.e., $h_{m,n} \sim \mathcal{CN}(0, 1)^a$.

By normalizing the channel gains it is assumed that geometric attenuation and shadow fading are approximately the same for all users. It is a simplifying assumption made to enable the statistical analysis, and common practise in the MIMO interference channel literature [158, 159, 160, 161]. Other common approaches to circumvent the above limitation include the use of random matrix theory and approximating the interference and noise terms with other probability distributions. Tackling the issue from yet another

---

\(^a\)Not only is the i.i.d. Rayleigh fading channel one of the most widely used statistical models [89] and hence facilitates comparison. It has also been pointed out that this model is representative of non-line-of-sight propagation [157].
direction is the work by Basnayaka et al. [162, 163] concerned with macro-diversity. The derived results for rate and SINR, based on the Laplace approximation method, however do not seem appropriate for the present work. In any case, the results derived here are valid where the channels suffer similar path loss, and such situations have been found in measurement campaigns [90].

The information-bearing symbols are assumed independent and transmitted at equal power, leading to a covariance matrix \( p_i I = p_i \mathcal{E}(xx^H) \), \( i \in \{1, 2\} \). The assumption of a uniform transmit power for each system is adopted not only for simplifying the analysis, but also for the sake of keeping physical-layer user fairness on a long-term basis. Given the received signals \( y_1 \) and \( y_2 \), the APs reconstruct their desired symbol blocks via

\[
\hat{x}_1 = \mathcal{D}(y_1 | H_{11}, H_{12}),
\]

\[
\hat{x}_2 = \mathcal{D}(y_2 | H_{22}, H_{21}),
\]

where \( \mathcal{D}(\cdot) \) is the signal detection function. It is stressed that the terms \( H_{12} x_2 \) and \( H_{21} x_1 \) in (3.1) and (3.2) are interference terms at corresponding APs, which should be appropriately handled in the signal reconstruction functions.

**Downlink**

Consider AP 1 sending a block of symbols \( x_1 \), destined to their corresponding UTs. Prior to transmission, pre-processing \( f_1(\cdot) \) is performed on \( x_1 \) with the output being an \( M_1 \times 1 \) signal block

\[
s_1 = f_1(x_1).
\]

(3.5)

The same procedure also takes place at AP 2, and the transmitted signal is an \( M_2 \times 1 \) block

\[
s_2 = f_2(x_2),
\]

(3.6)

where \( f_2(\cdot) \) is the processing function employed at AP 2. Define a super symbol block \( s \triangleq [s_1^T, s_2^T]^T \). The received signal at the \( n \)th UT is

\[
y_n = h_n^H s + v_n, \quad n \in \mathcal{N}_1 \cup \mathcal{N}_2,
\]

(3.7)

where \( h_n \triangleq [h_{1,n}, \ldots, h_{M_1+M_2,n}]^T \). Then, the signal \( x_n \) is reconstructed by

\[
\hat{x}_n = \mathcal{D}(y_n).
\]

(3.8)
3.2.2 New Challenges and Problems

The mathematical model presented in Chapter 3.2.1 is generally applicable to MU-MIMO interference channels, which have been employed in the investigations of various wireless use-cases, as illustrated in the introduction. Considering spectrum sharing between a massive MIMO system and a small MU-MIMO system, new challenges arise from condition c1). As a specific example, contemplate the coexistence between a mobile communications network and a local Wi-Fi system in an unlicensed band. It does not seem far-fetched that with the significant investment in LTE in unlicensed bands [164, 165], next generation 5G may also include similar techniques where a massive MIMO base station may attempt to access the unlicensed bands and hence share the spectrum with for instance multiple antenna capable Wi-Fi systems.

To illustrate the impact of the significant antenna number asymmetry (condition c1)) consider the following example in the uplink: a massive MIMO system with a single UT coexists with a small MIMO system with a single UT. Assume both APs apply MF equalization (not only does this filter maximize SNR in an interference-free scenario [166], it is also very modest in terms of computational complexity and requires no knowledge of for instance noise variance). Signal detection at AP $i \in \{1, 2\}$ is performed on

$$z_i = w_i y_i, \quad (3.9)$$

$$= \sqrt{p_1} \mathbf{H}_{i1}^H \mathbf{x}_1 + \sqrt{p_2} \mathbf{H}_{i2}^H \mathbf{x}_2 + \mathbf{H}_{ii}^H \mathbf{v}_i, \quad (3.10)$$

where in this special case, $\mathbf{H}_{ij}, i, j \in \{1, 2\}$ are all column vectors (i.e., $N_1 = 1, N_2 = 1$). The SINR of the $n$th user branch ($n \in \mathcal{N}_1 \cup \mathcal{N}_2$), denoted by $\gamma_n$, is (see e.g. [74] or [167, Ch. 13.3.4])

$$\gamma_n = \frac{(\mathbf{H}_{mn}^H \mathbf{H}_{nn})^2 p_n}{||\mathbf{H}_{mn} \mathbf{H}_{nj}||^2 p_j + ||\mathbf{H}_{nn}||^2 \sigma^2}, \quad (3.11)$$

where $j \in \{1, 2\} \setminus n$, and $\| \cdot \|$ denotes the Euclidean norm. From (3.11), it is clear that the power level of the massive MIMO system impacts the SINR of the small MIMO system, and vice versa. As stated by Marzetta [26], if the number of service antennas at System 1 is allowed to go to infinity, then the interference and noise terms at system 1 vanish. In turn, the transmit power of the UTs in $\mathcal{N}_1$ can be scaled back (inversely proportional to $M_1$) while sustaining the same SINR level [151], and this effectively reduces the interference at AP 2. This is certainly one way to decouple the systems (and approach a noise limited performance level), but the question of how this interference scenario behaves with a large, but finite number of antennas at AP 1 remains unanswered.
In order to gain some insight into the behaviour of the example system consider the long-term, or average SINR, $\gamma_n = E(\gamma_n)$. This metric nicely summarizes the SINR performance over every possible channel realization, and is often used in ergodic capacity expressions where the capacity is lower-bounded using a Jensen’s inequality argument (see e.g. [69, 151, 157]). Reformulating (3.11) as

$$\gamma_n = \frac{H_H H_{nn} p_n}{I_1 + \sigma^2},$$

(3.12)

with

$$I_1 = \frac{(H_H H_{nj}) (H_H H_{nj})^H}{H_H H_{nn}},$$

(3.13)

it can be shown (see Appendix A.1) that $I_1$ follows a Gamma distribution with shape factor $k = 1$ and scale factor $\theta = p_j$. A random variable $z$ is said to be Gamma distributed if its pdf is described by $G(z, k, \theta) \triangleq z^{k-1} \theta^{-k} e^{-z/\theta} / (k-1)!$. Furthermore, the random variable $I_1$ is independent of $H_H H_{nn}$ and so the average SINR can be written as

$$\gamma_n = E\left(\frac{1}{I_1 + \sigma^2}\right),$$

(3.14)

$$= M_n p_n \frac{1}{I_1 + \sigma^2},$$

(3.15)

Expanding the expectation involving the $I_1$ term results in

$$\gamma_n = M_n p_n \int_0^\infty \left( z + \sigma^2 \right)^{-1} G(z, 1, p_j) dz$$

(3.16)

$$= M_n p_n \frac{\sigma^2}{p_j} E_1 \left( \frac{\sigma^2}{p_j} \right),$$

(3.17)

where $E_1(z) \triangleq \int_z^\infty t^{-1} e^{-t} dt$, is commonly known as the exponential integral. Equation (3.17) is used in the numerical results (software packages such as Matlab or Gnu Octave provide the expint function for the efficient calculation of $E_1$), but it offers little insight into how the SINR behaves and how the two interfering systems coexist.

Instead a different approach will be adopted. First, in order to compare the power levels of system 1 and system 2, the operating point of this example interference system will be set to $\gamma_1 = \gamma_2$. The case where the UTs target a common average SINR threshold can for instance occur when the users request the same type of service or content with the same quality. Secondly, the $E_1$ function will be approximated for large transmit power. Therefore, equating the SINRs of the two systems, using (3.17),

$$\frac{M_1 p_1}{p_2} e^{\frac{\sigma^2}{p_1}} E_1 \left( \frac{\sigma^2}{p_1} \right) = \frac{M_2 p_2}{p_1} e^{\frac{\sigma^2}{p_2}} E_1 \left( \frac{\sigma^2}{p_2} \right),$$

(3.18)
and denoting $\alpha = p_1/p_2$ as the ratio of the transmit powers between system 1 and system 2, equation (3.18) can be written as
\[
\frac{M_1}{\alpha} e^{\frac{\alpha^2}{p_1}} E_1 \left( \alpha \sigma^2 / p_1 \right) = M_2 e^{\frac{\sigma^2}{p_2}} E_1 \left( \sigma^2 / p_1 \right). \tag{3.19}
\]

Letting $p_1 \to \infty$, both the exponential and the exponential integral on the left-hand side and right-hand side approach the same limiting value\(^{1}\), and the ratio $\alpha$ reduces to
\[
\alpha = \sqrt{\frac{M_2}{M_1}}. \tag{3.20}
\]

It has to be emphasized that (3.20) is only asymptotically tight (with $p_1$ or $p_2$), but as shown in Fig. 3.1, the approximation is reasonable, and perhaps more importantly, offers a concise way to characterize the system’s behaviour in terms of transmit power. Note that convergence of the approximate and exact result as $p_2 \to \infty$ is quicker for smaller $M_1$ simply because the power levels between the massive MIMO system and the small MIMO system are closer in magnitude. The approximation’s simple form clearly illustrates the resulting difference in transmit powers between system 1 and system 2 as a ratio of the antenna numbers.

In accordance with condition c1), then in order to have an equivalent SINR level at both systems, the UT within System 2 needs to transmit at a much higher power than the transmitter associated with System 1. This is a new phenomenon, which does not occur in the context of spectrum sharing either between two massive MIMO systems or two small MU-MIMO systems.

Using $p_1 = \alpha p_2$, with $\alpha$ as in (3.20) and substituting into (3.17), the common average SINR scales as
\[
\gamma \approx \sqrt{\frac{M_1}{M_2}} e^{\frac{\sigma^2}{p_2}} E_1 \left( \sigma^2 / p_2 \right). \tag{3.21}
\]

In visualizing the common SINR performance it may be helpful to recall the following inequality for the exponential integral [168, eq. 5.1.20]
\[
\frac{1}{2} \ln \left( 1 + \frac{2}{z} \right) < e^z E_1 \left( z \right) < \ln \left( 1 + \frac{1}{z} \right), \tag{3.22}
\]
where $\ln (\cdot)$ is the natural logarithm. Note the condition $z > 0$, and hence it is not straightforward to investigate asymptotic behaviour. However Larsson and Jorswieck [107, Proposition 4] show that the rate reaches an asymptote (in fact the inclusion of the log(\cdot) function makes the integral well-behaved). It is also worth pointing out that the common average SINR includes a $\sqrt{M_1}$ factor, whereas in the absence of system 2,

\(^{1}\)More formally, it can be shown that $\lim_{p_1 \to \infty} \left( e^{\frac{\alpha^2}{p_1}} E_1 \left( \alpha \sigma^2 / p_1 \right) / e^{\frac{\sigma^2}{p_2}} E_1 \left( \sigma^2 / p_1 \right) \right) = 1$, by for instance using a series expansion (Puiseux series) at infinity.
the average SINR would have been proportional to $M_1$ (see (3.15)). This bears some resemblance to the result by Ngo et al. [151] who showed that by incorporating a channel estimation error, the power scaling law also included a $\sqrt{M_1}$ term.

The above example was limited to MF receivers and a single UT for each system. An interesting research issue, therefore, is to explore spectrum sharing uplink approaches between massive MIMO and small MU-MIMO. This question is addressed in Chapter 3.3. A similar problem also occurs in the downlink of spectrum sharing, which will be addressed in Chapter 3.4.

### 3.3 Uplink Spectrum Sharing Approaches

The example in Chapter 3.2.2 highlighted several issues:

1. For a given common average SINR, a significant transmission power difference exists between the small MIMO system and the massive MIMO system.

2. With a finite number of service antennas at AP 1 (a practical range between $10 \leq M_1 \leq 200$ has been suggested by Björnson et al. [157]), external interference cannot be ignored: it has a non-negligible impact on average SINR.
3. The use of a massive number of service antennas at AP 1 can be leveraged to arbitrarily increase $\gamma$ (see (3.21)), but the square root factor results in a “diminishing returns” factor.

Do the above conclusions hold in the multi-user case? Can another type of receiver improve performance? The aim of the following section is to provide an answer to these questions in the uplink scenario.

For the massive MIMO system, the scope will be limited to the two most prominent receiver types [151, 84]: the MF and the ZF receiver. An extension to the MF and ZF receivers including the cancellation of the relatively much stronger interference, referred to as partial SIC, will also be examined. For the small MIMO system, it will be assumed throughout that a ZF receiver is in use. It is a practical linear equalizer type, requiring no knowledge of the noise variance, and in contrast to the MF receiver it is capable of handling strong intra-interference. The term intra-interference is used here to denote interference between UTs within the same system. It will be further assumed that the small MIMO system treats inter-interference as noise (i.e., interference from system 1). Not only does this assumption simplify the receiver design, but (as shown in Figs. 3.2 to 3.5) nulling interference from a massive MIMO UT (transmitting at relatively low power) does not in general compensate for the loss in diversity order. System 1 on the other hand is assumed to be sufficiently sophisticated to gain knowledge of the interfering channels and use this CSI to null the inter-interference from system 2 when the ZF receiver is used.

As illustrated by the example in Chapter 3.2.2, comparing the required transmit powers between system 1 and system 2 for a given operating point proved very useful. It is possible to formalize this concept with the following problem formulation:

\[
\begin{align*}
\text{maximize} & \quad \gamma \\
\text{subject to} & \quad \gamma_1 \geq \rho \gamma \\
& \quad \gamma_2 \geq \gamma \\
& \quad p_n \leq p_{\text{max}} \ \forall n \in \{1, 2\}
\end{align*}
\]

The average common SINR metric is denoted by $\gamma$; and $\gamma_n$, $n \in \{1, 2\}$ is the average per UT SINR of system $n$. Note that the problem formulation allows for an arbitrary operating point by adjusting the coefficient $\rho \geq 0$. A similar concept was also employed by Mohseni et al. [40] and Zhang et al. [169], where instead of weighting the SINR directly, a rate profile was applied. The UTs associated to system $n$ transmit at power $p_n$, and are constrained by a maximum transmit power $p_{\text{max}}$.

One approach to interpret the problem formulation is to recognize that it is in fact a slight reformulation of the $\max_{p_1, p_2} \min_k (\gamma_k)$ problem. The key property employed
throughout to efficiently solve the problem, is the fact that if all transmit powers, $p_n$, are increased by a common factor, all SINRs, $\gamma_k$, also increase. This monotonicity property may also be used to characterize the solution [170]. The global optimum is found when the SINR constraints in (3.23b) and (3.23c) are active. The details are presented in Appendix A.4.

3.3.1 ZF Receiver

Concerning the uplink model where the linear ZF channel equalizer is assumed at both the massive MIMO AP and the small MIMO AP. Signal detection at AP 1 is performed on

$$z_1 = W_{zf}^1 y_1,$$

$$= \begin{bmatrix} \sqrt{p_1} x_1 \\ \sqrt{p_2} x_2 \end{bmatrix} + W_{zf}^1 v_1,$$

where $W_{zf}^1$ is an $(N_1 + N_2) \times M_1$ channel equalization matrix. Note that for the ZF equalizer, AP 1 separates all signals (including UTs associated to AP 2), but then for symbol detection the $x_2$ vector component is simply discarded. Likewise, at AP 2, the signal after channel equalization is

$$z_2 = W_{zf}^2 y_2,$$

$$= \sqrt{p_2} x_2 + \sqrt{p_1} W_{zf}^2 H_{21} x_1 + W_{zf}^2 v_1,$$

with $W_{zf}^2 \in C^{N_2 \times M_2}$. For both systems (i.e., at AP $n \in \{1, 2\}$) the linear channel equalizer is designed to satisfy $W_{zf}^n H = I$, with $H = [H_{11}, H_{12}]$ at AP 1, and $H = H_{22}$ at AP 2. This is accomplished through the use of the common pseudo-inverse $W_{zf}^n = H^\dagger = (H^H H)^{-1} H^H$, and the separation of the transmitted symbols is guaranteed as long as the rank of the matrix $H$ is equal to or greater than the number of independent streams. Assuming a full rank channel (an assumption fulfilled (with probability 1) by the independent Rayleigh fading channel), AP 2 can support a maximum of $M_2$ UTs. Given the much larger number of service antennas at AP 1, and hence potentially large number of UTs, AP 2 cannot (except for the corner case $N_1 + N_2 \leq M_2$) eliminate all inter-interference. Hence the decision to limit the design of the $W_{zf}^2$ filter to the direct channel, $H_{22}$. As a consequence, AP 2 does not need to estimate interfering channels. In addition, as is shown in the numerical results, nulling weak interferers is generally detrimental, compared to treating these as additional noise and retaining all receive dimensions for diversity gain.

Based on (3.25), the SNR of an arbitrary UT associated to AP 1 is (see e.g. [68])

$$\gamma_1 = \frac{P_1}{\sigma^2 W_{zf}^1 (W_{zf}^1)^H}_{11}. $$

(3.28)
While using (3.27), the SINR of an arbitrary UT belonging to AP 2 can be expressed as

\[
\gamma_2 = \frac{p_2 \mathbf{W}_2^f \mathbf{H}_{21} (\mathbf{W}_2^f \mathbf{H}_{21})^H + \sigma^2 \mathbf{W}_2^2 \mathbf{H}_{21}^H \mathbf{W}_2^f \mathbf{H}_{21}^H}{\text{interference power, } I_1}
\]  

(3.29)

where \([\mathbf{A}]_{ij}\) denotes the \((i,j)\)th element of \(\mathbf{A}\). Note that both (3.28) and (3.29) single out the first UT’s SINR. This is an arbitrary choice and made with the observation that the SINRs of all users belonging to an AP follow the same distribution (although users’ SINRs are not independent) and hence share the same mean.

It is known that the SNR after the ZF filter is distributed according to a Gamma distribution [68] (for an alternative derivation see Appendix A.2). Specifically, \(\gamma_1 \sim \mathcal{G}((M_1 - N_1 - N_2 + 1), p_1/\sigma^2)\), and its mean is described by the simple relation

\[
\overline{\gamma}_1 = (M_1 - N_1 - N_2 + 1) \frac{p_1}{\sigma^2}.
\]  

(3.30)

Note that having defined the distribution for the ZF SNR, a simple expression for the variance also exists for this case: \(\text{var}(\gamma_1) = k\theta^2\), where \(k = (M_1 - N_1 - N_2 + 1)\) and \(\theta = p_1/\sigma^2\). Taking \(\sqrt{k\theta^2}\) as a measure of the deviation of the random variable \(\gamma_1\) from the mean, it can be seen that the spread of the SNR as a fraction of the mean decreases with increasing antenna size, \(M_1\). However, this view does not perhaps clearly depict the channel hardening effect stated in the massive MIMO literature. Instead, consider that a specific average SNR is desired. From (3.30), the transmit power can hence be scaled as \(p_1 = \overline{\gamma}_1 \sigma^2 / (M_1 - N_1 - N_2 + 1)\) with increasing \(M_1\). If this is the case, \(\text{var}(\gamma_1)\) actually decreases with increasing \(M_1\). In the limit as \(M_1 \to \infty\), the SNR therefore “hardens” around the mean, \(\overline{\gamma}_1\). Hence for the case with a ZF receiver and independent Rayleigh fading the mean SNR has a direct relation (equality) to the asymptotic result (see e.g. Hoydis et al. [83] or Ngo et al. [151]). This relation is worth pointing out, since the majority of massive MIMO literature using asymptotic results claim a good match to numerical simulations with small antenna numbers, without having made the above observation explicit. It is also worth noting, that for instance, Tse and Hanly [39] derived similar results for CDMA with random spreading sequences, but using much less restrictive assumptions on the shape of the random variable’s distribution. This is one of the strengths of the asymptotic results (in terms of number of antennas): they can be used to gauge the performance of systems with widely varying channel statistics.

The average SNR result in (3.30) does not hold for \(\gamma_2\) in (3.29) due to the interference term. It is however still possible to characterize the SINR. It can be shown, see
Appendix A.3, that the interference term $I_1$ in (3.29) follows a scaled $\mathcal{F}$-distribution

$$I_1 \overset{p_1}{\sim} \left[ W_2^2 F_{H_21} (W_2^2 F_{H_21})^H \right]_{11},$$

(3.31)

$$\sim \frac{N_1}{M_2 - N_1 + 1} \mathcal{F}_{2(N_1),2(M_2-N_1+1)}.$$  

(3.32)

In itself, this is perhaps not a very useful characterization, but realizing that the coloured noise term in (3.29) and $I_1$ share a common denominator (see the appendix), $\gamma_2$ can be described by the following equivalent distribution

$$\gamma_2 \sim \frac{p_2/\sigma^2 X}{p_1/\sigma^2 Y + 1},$$

(3.33)

where $X \sim \mathcal{G} ((M_2 - N_2 + 1), 1)$, $Y \sim \mathcal{G} (N_1, 1)$ and the random variables $X$ and $Y$ are independent. In terms of distribution, the SINR after the ZF equalizer (with an additional interference term) is identical to that of a MF receiver with $M_2 - N_2 + 1$ service antennas and $N_1$ interferers, each transmitting with power $p_1$. As expected, with a single UT, $N_2 = 1$, the expression in (3.33) is equivalent to the MF SINR (cf. (3.15)).

The mean SINR of an arbitrary UT at AP 2 is therefore

$$\gamma_2 = \mathcal{E} \left( \frac{p_2/\sigma^2 X}{p_1/\sigma^2 Y + 1} \right),$$

(3.34)

$$= (M_2 - N_2 + 1) \frac{p_2}{\sigma^2} \int_0^\infty \frac{1}{y + 1} \mathcal{G} \left( y, N_1, \frac{p_1}{\sigma^2} \right) dy,$$

(3.35)

$$= (M_2 - N_2 + 1) \frac{p_2}{\sigma^2} \left( \frac{\sigma^2}{p_1} \right)^{N_1} \frac{\sigma^2}{p_1} \Gamma \left( 1 - N_1, \frac{\sigma^2}{p_1} \right).$$

(3.36)

where $\Gamma (\cdot, \cdot)$ is the upper incomplete Gamma function. Moving from (3.35) to (3.36), the following identity was used [171, eq. 3.383.10]:

$$\int_0^\infty \frac{z^{a-1}e^{-bz}}{z+c} dz = e^{a-1}e^{cb} (a-1)! \Gamma (1-a, cb).$$

(3.37)

Note that many texts (and implementations in software packages) limit the first argument of the upper incomplete Gamma function to where the real part has to be non-negative. In the present case however, negative integer values need to be accommodated. The following relation proved useful for the numerical experiments [171, eq. 8.352.5]:

$$\Gamma (-n+1, z) = \frac{(-1)^{n+1}}{(n-1)!} \left( \mathrm{E}_1 (z) - e^{-z} \sum_{m=0}^{n-2} (-1)^m \frac{m!}{z^{m+1}} \right),$$

(3.38)

where $n \in \{2, 3, \ldots\}$. A further observation is that $\Gamma (0, z) = \mathrm{E}_1 (z)$, and hence setting $N_1 = 1$ and $N_2 = 1$, (3.36) reduces to (3.17) in the example in Chapter 3.2.2.
Armed with expressions for both $\gamma_1$ and $\gamma_2$ it is now possible to determine the required transmit powers to reach the maximum average common SINR, $\gamma^*$, using the problem statement (3.23). It can be shown (see Appendix A.4) that the optimal average SINR, $\gamma^*$, is reached when either $p_1$ or $p_2$ equals the transmit power constraint $p_{\text{max}}$. Assuming $p_2 = p_{\text{max}}$, then $p_1$ may be solved for by equating $\gamma_1 = \rho \gamma_2$ and using the derived expressions in (3.30) and (3.36). To assert the validity of the assumption $p_2 = p_{\text{max}}$, set $p_1 = p_{\text{max}}$ and solve for $p_2$. If $p_2 > p_{\text{max}}$ then the initial assumption was true. Note that $p_2$ can be easily isolated, but that solving for $p_1$ requires the use of a root finding method such as bisection or Newton’s method.

The form of the average SINR, $\gamma_2$, in (3.36) does not provide much insight into the behaviour of the interference system. In order to gain some understanding, two approximations will be explored. Approximating the upper incomplete Gamma function itself (particularly with a negative first argument) in (3.36) does not seem to lead to any fruitful results. Instead, equation (3.33) will be used as a starting point. The complicated form of (3.36) is of course due to the denominator in (3.33), and meaningful approximations to such inverse moment problems have been widely studied in the statistics literature (see e.g. [172]). Applying, perhaps the two most elementary, approximations bounds the denominator in (3.33) as

$$\frac{1}{\frac{p_1}{\sigma_2} \mathcal{E}(Y) + 1} \leq \mathcal{E}\left(\frac{1}{\frac{p_1}{\sigma_2} Y + 1}\right) \leq \mathcal{E}\left(\frac{1}{\frac{p_1}{\sigma_2} Y}\right),$$  

(3.39)

where the lower bound is due to the application of Jensen’s inequality, and the upper bound is obtained by ignoring the noise component (i.e., equivalently considering the case where the interference term $p_1/\sigma_2 Y >> 1$). The average, $\mathcal{E}(Y)$, of the Gamma distributed random variable is simply $N_1$, while the inverse moment of the Gamma random variable can concisely be expressed as

$$\mathcal{E}\left(\frac{1}{\frac{p_1}{\sigma_2} Y}\right) = \frac{\sigma^2}{p_1} \int_{0}^{\infty} \frac{1}{y} G(y, N_1, 1) \, dy,$$  

(3.40)

$$= \frac{\sigma^2}{p_1} \int_{0}^{\infty} \frac{y^{N_1-2} e^{-y}}{(N_1 - 1)!} \, dy,$$  

(3.41)

$$= \frac{\sigma^2}{p_1 \, N_1 - 1}. \quad (3.42)$$

The integral in (3.41) was simplified by using the identity $\int_{0}^{\infty} y^{a-1} e^{-y} \, dy \triangleq (a - 1)!$ [171, eq. 8.310.1], and restricts the solution to $N_1 > 1$.

Allowing the transmit power, $p_1$, to increase, it can be seen that the bounds in (3.39) approach each other, and hence asymptotically, the bounds become tight. Therefore, using the upper bound as an approximation, the relationship between the transmit
powers can be approximated as
\[ \gamma_1 = \rho \gamma_2 = \rho (M_2 - N_2 + 1) \frac{p_1}{\sigma^2} \]  
(3.43)

\[ (M_1 - N_1 - N_2 + 1) \frac{p_1}{\sigma^2} \approx \rho \frac{p_2}{\sigma^2} (N_1 - 1). \]  
(3.44)
Solving for \( p_1 \) gives
\[ \frac{p_1}{\sigma^2} \approx \sqrt{\frac{\rho (M_2 - N_2 + 1)}{(M_1 - N_1 - N_2 + 1)(N_1 - 1)}} \sqrt{\frac{p_2}{\sigma^2}}. \]  
(3.45)
Notice the similarity with (3.20) for the single UT case and MF equalizers. Naturally, (3.45) includes terms for the number of UTs, but perhaps the most significant difference is that \( p_1 \) is proportional to \( \sqrt{p_2} \), i.e., it grows at a much slower rate compared to the single UT MF case. Using the approximation for the transmit power in (3.45) to calculate the average SINR in (3.30), again shows that the average common SINR grows with \( \sqrt{M_1} \), but since AP 1 cancels the interference from UTs associated to AP 2 the SINR grows with \( \sqrt{p_2} \). This is a significant improvement over (3.21).

### 3.3.2 MF Receiver

The potential difficulty in acquiring CSI of the interfering channels, and the computational complexity of the ZF receiver, makes it interesting to consider the simpler MF equalizer at AP 1. The signal after the MF is
\[ z_1 = W_1^{MF} y_1, \]  
(3.46)
\[ = \sqrt{p_1} H_{11}^H x_1 + \sqrt{p_2} H_{12}^H x_2 + H_{11}^H v_1. \]  
(3.47)
Focusing on an arbitrary UT (in terms of SINR pdf they are all equivalent), denote \( h \) as the first column of \( H_{11} \) and \( H \) the remaining columns. Separating the desired signal strength and the interference-and-noise power, the SINR can be written as (cf. (3.11))
\[ \gamma_1 = \frac{(h^H h)^2 p_1}{\|h^H H_{12}\|^2 p_2 + \|h^H H\|^2 p_1 + \|h\|^2 \sigma^2}. \]  
(3.48)
Rearranging (3.48) as
\[ \gamma_1 = \frac{h^H h \frac{p_1}{\sigma^2}}{I_1 + I_2 + 1}, \]  
(3.49)
with
\[ I_1 = \frac{(h^H H_{12}) (h^H H_{12})^H p_2}{h^H h}, \]  
(3.50)
3.3. Uplink Spectrum Sharing Approaches

and

\[ \mathcal{I}_2 = \left( h^H H \right) \frac{h^H p_1}{\sigma^2}, \]  

(3.51)

it can be shown (following similar arguments as in Appendix A.1) that \( \mathcal{I}_1 \sim \mathcal{G} \left( N_2, \frac{p_2}{\sigma^2} \right) \) and \( \mathcal{I}_2 \sim \mathcal{G} \left( (N_1 - 1), \frac{p_1}{\sigma^2} \right) \). Furthermore, \( \mathcal{I}_1 \) and \( \mathcal{I}_2 \) are independent, and both are independent of \( h^H h \sim \mathcal{G} (M_1, 1) \). The average SINR can therefore be described by

\[
\bar{\gamma}_1 = \mathcal{E} \left( h^H h \frac{p_1}{\sigma^2} \right) \mathcal{E} \left( \frac{1}{\mathcal{I}_1 + \mathcal{I}_2 + 1} \right),
\]

(3.52)

\[
= M_1 \frac{p_1}{\sigma^2} \mathcal{E} \left( \frac{1}{\mathcal{I}_1 + \mathcal{I}_2 + 1} \right).
\]

(3.53)

With two independent Gamma distributed random variables in the denominator (with different scale factors) no concise expression for the expectation in (3.53) was found. Results on the pdf of sums of Gamma random variables are available in terms of infinite sums (see e.g. the work by Alouini et al. [173] or Moschopoulos’ original work on the sum of Gamma random variables [174]), but these do not seem suitable for the present case. Instead, focus will be placed on how to evaluate (3.53) efficiently, and present simpler bounds on the average SINR to visualize performance.

Two observations can be made to facilitate the computation of (3.53). First, the mgf of a sum of two (or more) independent random variables can nicely be described by the product of its individual mgf [175]. Secondly, there is no need to explicitly transform the mgf back into a pdf to calculate the expectation. Rather, a result highlighted by Cressie et al. [176] can be used to express the inverse moment as a function of the mgf:

\[
\mathcal{E} \left( \frac{1}{aX + b} \right) = \int_0^\infty e^{-bt} M_x (-at) \, dt,
\]

(3.54)

where \( M_x (t) \triangleq \mathcal{E} \left( e^{-tx} \right) \) is the mgf. The mgf of the sum of the random variables \( \mathcal{I}_1 \) and \( \mathcal{I}_2 \) is specifically

\[
M_x (t) = M_{\mathcal{I}_1} (t) M_{\mathcal{I}_2} (t),
\]

(3.55)

\[
= \left( 1 - \frac{p_2}{\sigma^2 t} \right)^{-N_2} \left( 1 - \frac{p_1}{\sigma^2 t} \right)^{-(N_1 - 1)}.
\]

(3.56)

Substituting (3.56) into (3.54), and applying the resulting expression for the expectation in (3.53) leads to a formulation for the average SINR that has two advantages in terms of computation. The double integral has been reduced to a single integral, and by using the mgfs, explicit use of factorials (see the definition for the Gamma random variable) have been avoided. This generally improved the stability of the numerical experiments for large \( N_1 \).

With the characterization of the average SINR at AP 1 employing a MF equalizer, the required transmit powers to reach a specific operating point of the interference...
system can now be studied. It is assumed that AP 2 still employs the ZF receiver introduced in Chapter 3.3.1, and hence the average SINR expression in (3.36) is valid. Using problem (3.23) to formalize the objective, it can be shown (applying the same arguments as in Appendix A.4) that the maximum average common SINR is reached when either $p_1 = p_{\text{max}}$ or $p_2 = p_{\text{max}}$. Assuming $p_2 = p_{\text{max}}$, $p_1$ may be solved for in $\tau_1 = \rho \tau_2$ with the use of (3.36) and (3.53). To test the initial assumption, the obvious method is to set $p_1 = p_{\text{max}}$, solve for $p_2$, and if $p_2 > p_{\text{max}}$ then the initial assumption was true. The bounds (developed in the previous section and below) may also be used to reject or accept the assumption.

To gain an idea of the performance, the expectation in (3.53) can be bounded. Jensen’s inequality can again be directly applied to bound it from below:

$$E\left(\frac{1}{I_1 + I_2 + 1}\right) \geq \frac{1}{N_2 \frac{p_2}{\sigma^2} + (N_1 - 1) \frac{p_1}{\sigma^2} + 1}. \quad (3.57)$$

Note that substituting (3.57) into (3.53) results in a form equal to the asymptotic (in the number of antennas) result (see e.g. [39]). An upper bound for the expectation term in (3.53) is less evident, however. An upper bound derived by Wooff [177] is given in terms of a ratio of polynomials involving the mean and variance of the non-negative random variable $X = I_1 + I_2$. This is a viable alternative, but it does have the limitation that this bound does not converge with the lower bound with increasing $p_1$ or $p_2$. Instead, a different approach is offered, with (3.54) as a starting point:

$$E\left(\frac{1}{X + b}\right) \leq E\left(\frac{1}{X}\right), \quad (3.58)$$

$$= \int_0^\infty M_X(-t) \, dt, \quad (3.59)$$

$$= \int_0^\infty \frac{1}{(1 + \frac{p_2}{\sigma^2} t)^{N_2} (1 + \frac{p_1}{\sigma^2} t)^{N_1-1}} \, dt, \quad (3.60)$$

$$< \int_0^\infty \frac{1}{(1 + N_2 \frac{p_2}{\sigma^2} t)(1 + (N_1 - 1) \frac{p_1}{\sigma^2} t)} \, dt, \quad (3.61)$$

$$= \sigma^2 \left(\ln (N_2 p_2) - \ln ((N_1 - 1) p_1)\right) \frac{N_2 p_2 - (N_1 - 1) p_1}{N_2 p_2 - (N_1 - 1) p_1}, \quad (3.62)$$

where Bernoulli’s inequality, $(1 + z)^a > 1 + az$, was used for (3.61). This upper bound has the nice property that it does converge with the lower bound in the limit $p_2 \to \infty$ (or $p_1 \to \infty$). Unfortunately, due to the existence of the logarithm function the bound does not lead to a simple relation between the transmit powers. Hence with the argument that the upper and lower bounds converge, the lower bound (3.57) will be used for $\tau_1$ and the upper bound will be used for $\tau_2$ in (3.34). At the operating point $\tau_1 = \rho \tau_2$ this expands to

$$\frac{\frac{M_1 p_1}{\rho \sigma^2}}{N_2 \frac{p_2}{\sigma^2} + (N_1 - 1) \frac{p_1}{\sigma^2} + 1} \approx \frac{(M_2 - N_2 + 1) \frac{p_2}{\sigma^2}}{(N_1 - 1) \frac{p_1}{\sigma^2}}. \quad (3.63)$$
Equation (3.63) of course results in a quadratic equation in $p_1$ (or $p_2$) and only the non-negative root is of interest:

$$p_1 \approx a + \sqrt{a^2 + b},$$  \hspace{1cm} (3.64)

where

$$a = \frac{\rho(M_2 - N_2 + 1)p_2}{2M_1},$$  \hspace{1cm} (3.65)

$$b = \frac{\rho(M_2 - N_2 + 1)(N_2p_2 + 1)p_2}{M_1(N_1 - 1)}.$$  \hspace{1cm} (3.66)

It can be seen that (3.64) scales with $p_2$, and is hence in accordance with the single UT scenario presented in the example in Chapter 3.2.2. Also, note that if it can be assumed that $\sqrt{b} \gg a$, and hence $p_1 \approx \sqrt{b}$ the transmit power of the users associated to the massive MIMO AP scales as

$$p_1 \approx \sqrt{\frac{\rho(M_2 - N_2 + 1)N_2}{M_1(N_1 - 1)}}p_2.$$  \hspace{1cm} (3.67)

In similarity with the single UT example and the ZF result in Chapter 3.3.1, the growth of the common average SINR with $\sqrt{M_1}$ and $\sqrt{M_2}$ is evident. Note that the assumption $\sqrt{b} \gg a$ is made all the more reasonable with the condition c1).

### 3.3.3 Extension to partial SIC

Chapters 3.3.1 and 3.3.2 highlighted the significant transmit power difference between the massive MIMO system and the small MIMO system for a similar operating point ($\rho \approx 1$). This immediately brings to mind the use of an SIC scheme, where the high signal power of the interferer, combined with the array gain of massive MIMO, allows it to be decoded and cancelled, leaving only the desired signals to be detected. It can be argued that SIC receivers impose unrealistic assumptions on for instance synchronization (time and frequency), dynamic range of the analogue front end, and knowledge of the interferer’s coding and modulation format. Nevertheless, it is still interesting to investigate whether a more advanced receiver brings about any substantial benefit in the present scenario.

Consider the case where all interfering signals from the UTs associated to AP 2 can successfully be decoded and removed. Then following the same derivation as in Chapter 3.3.1, the average SINR after the ZF equalizer at AP 1 is simply:

$$\bar{\gamma}_1 = (M_1 - N_1 + 1)p_1/\sigma^2,$$ \hspace{1cm} (3.68)

and the approximate transmit power at the operating point $\bar{\gamma}_1 = \rho\bar{\gamma}_2$ is:

$$\frac{p_1}{\sigma^2} \approx \sqrt{\frac{\rho(M_2 - N_2 + 1)}{(M_1 - N_1 + 1)(N_1 - 1)}}\sqrt{\frac{p_2}{\sigma^2}}.$$ \hspace{1cm} (3.69)
Figure 3.2: Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 100$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$.

Figure 3.3: Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 100$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 
3.3. Uplink Spectrum Sharing Approaches

Figure 3.4: Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 1000$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$.

Figure 3.5: Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the uplink (Chapters 3.3.1 and 3.3.2). $M_1 = 1000$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 
Figure 3.6: Empirical cdf of SINR (detected at both AP 1 and AP 2) for an arbitrary UT associated to AP 2. The antenna configuration is \( \{M_1, M_2\} = \{100, 4\} \), the load is set to \( \beta_1 = 0.1 \), \( \rho = 1 \) and \( p_2 = 10 \) dB. This implies a transmit power \( p_1 = -11.6 \) dB for the ZF receiver type and \( p_1 = -2.27 \) dB for the MF equalizer.

As illustrated in Fig. 3.6, given that condition c1) is fulfilled, the probability that the interference cannot be decoded is negligible (regardless of SIC decoding order, and rate adaptation technique used at AP 2). However, (3.68) shows that the performance improvement is also negligible, considering that \( M_1 \gg N_2 = \max (M_2) \).

Following the same derivation as in Chapter 3.3.2, but assuming the interference from the small MIMO system is successfully cancelled, shows that the only difference for the MF is in the \( b \) coefficient of (3.64):

\[
b = \frac{\rho (M_2 - N_2 + 1) p_2}{M_1 (N_1 - 1)}. \tag{3.70}
\]

This implies that \( p_1 \) scales roughly with \( \sqrt{p_2} \).

### 3.4 Spectrum Sharing in the Downlink

To complete the spectrum sharing scenario between a massive MIMO AP and a small MIMO AP, the downlink (or forward link) is considered in this section. In similarity
with Chapter 3.3, the scope is limited to a ZF precoder at AP 2, and the linear precoders, MRT and ZF, are considered at AP 1. The same analysis principle will also be applied in the downlink. Namely, the SINR distributions of the systems will be used to establish the transmit powers of the two systems and SINR operating point. Approximations to the exact results will also be provided to gain insight into how the number of service antennas and system load affect performance. Note that the underlying reason for the use of this method are the same as for the uplink: the small number of service antennas at AP 2 (particularly in contrast to AP 1) does not seem to justify the use of asymptotic (in the number of antennas) results available in the literature.

Recently, Björnson et al. [157] extended an uplink-downlink duality result to the multi-cell scenario in order to avoid explicit derivation of the downlink side, when the uplink equivalent had already been investigated. This approach however will not be pursued here. The reason is that one of the underlying conditions is not met. The original formulation, showing that the optimum of the downlink optimization problem is obtained by solving an equivalent uplink problem, was shown to hold under a sum power constraint. More specifically, Schubert and Boche [178, 170] showed that if a set of SINRs can be achieved in the uplink, then the same point can be achieved in the downlink with the same set of beamforming vectors. The presented theorem simplifies the design of downlink beamformers by converting it to the conceptually simpler uplink equivalent. The theorem does however hinge on the fact that the sum power of the downlink scenario is equal to the uplink case. For the extension of the uplink-downlink duality to the multi-cell scenario, Björnson et al. [157], the sum power constraint remained. This however implies that the access points share a total maximum power constraint. For the work by Björnson et al. on multiple equal massive MIMO access points, this was not an issue, since the channel hardening effect would guarantee nearly static and equal power allocation. In the present scenario with one conventional multi-user MIMO system, it is likely however that the sum power constraint would mean that AP 2 would routinely break its individual power constraint (i.e., in a sense borrowing from AP 1). It therefore seems prudent not to apply the duality property in the present case and instead carry out the derivation explicitly.

### 3.4.1 ZF Precoding

The application of a linear ZF precoder at AP \( n \) results in a pre-processing function of the form [74]:

\[
f_n(x_n) = \sqrt{N_n p_n} W_n^{zf} x_n,
\]

where

\[
W_n^{zf} = \sqrt{1/\lambda_n} H_n (H_n^H H_n)^{-1}.
\]
Here, the channel matrices are defined as $H_1 = [H_{11} \ H_{12}]$ and $H_2 = H_{22}$. In similarity with the uplink, AP 2 only ensures that its own signals are separated, whereas AP 1 with considerably more antennas also nulls any transmit power to the UTs belonging to AP 2. As mentioned in Chapter 3.2.1, AP 1 is assumed to be capable of either directly estimating the cross-channels to the UTs associated to AP 2, or successively learning their behaviour by analysing available feedback channels. The normalization factor $\lambda_n$ is chosen to satisfy a transmit power constraint. Three common choices for $\lambda_n$, highlighted by Hochwald and Vishwanath [79], are

i) $\lambda_n = x_n^H (H_n^H H_n)^{-1} x_n$: an instantaneous total power constraint keeping the transmit power constant regardless of channel state or information symbols.

ii) $\lambda_n = \text{tr} \left( (H_n^H H_n)^{-1} \right)$: compensates for any channel fluctuations.

iii) $\lambda_n = \mathbb{E} \left( x_n^H (H_n^H H_n)^{-1} x_n \right)$: serves to set an average transmit power, but allows the instantaneous transmit power to vary with symbol amplitude and channel variations.

Note that $\mathbf{x}_1 = [x_1, x_2, \ldots x_{N_1}, \mathbf{0}_{1 \times N_2}]^T$ has been redefined by appending a $\mathbf{0}_{1 \times N_2}$ vector. This is in order to accommodate the interference channels to the UTs of AP 2. The conventional normalization factor in ii) does not seem to lead to any tractable form, and hence following the works of e.g. Hochwald and Vishwanath [79] and Björnson et al. [157] the long-term $\lambda_n$ in iii) will be used for the analysis. It has been pointed out that the three normalization factors converge asymptotically with the number of antennas [79]. Hence, the assumption clearly applies to AP 1. At AP 2 however, the use of the average $\lambda_2$ in iii) applied in the precoding step has the general effect of compensating for deep fades by increasing transmit power (and vice versa).

Hochwald and Vishwanath [79] stated that the random variable $z_n = x_n^H (H_n^H H_n)^{-1} x_n$, used in the normalization factor, is distributed as a scaled $\mathcal{F}$-distribution (this can be shown by using the same methods presented in Appendices A.2 and A.3, with the additional assumption $x_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$):

$$z_1 \sim \frac{N_1}{M_1 - N_1 - N_2 + 1} \mathcal{F}_{2N_1,2(M_1-N_1-N_2+1)},$$

(3.73)

$$z_2 \sim \frac{N_2}{M_2 - N_2 + 1} \mathcal{F}_{2N_2,2(M_2-N_2+1)}.$$  

(3.74)

Accordingly, the chosen normalization factor has the following simple forms at APs 1 and 2:

$$\lambda_1 = \frac{N_1}{M_1 - N_1 - N_2}$$

(3.75)

$$\lambda_2 = \frac{N_2}{M_2 - N_2}$$

(3.76)
The above forms set the restrictions $M_1 > N_1 + N_2$ and $M_2 > N_2$. For the massive MIMO system, the condition is likely to be met. For AP 2, on the other hand, it can at first sight seem very restrictive. However, it also has to be pointed out that the average per UT SINR also decreases significantly with increased number of UTs and for this reason it may not be advisable to fully load a ZF system. In addition, to avoid the situation where the power normalization factors go to infinity, $N_1 > 1$ and $N_2 > 1$.

Focusing on one UT belonging to each AP, the SINRs using the ZF precoder 3.72 and the normalization factors (3.73) and (3.74) can be expressed as:

$$\gamma_1 = \frac{(M_1 - N_1 - N_2) p_1}{h^H W_2^Z F (h^H W_2^Z F)^H (M_2 - N_2) p_2 + \sigma^2}$$

$$\gamma_2 = \frac{(M_2 - N_2) p_2}{\sigma^2}.$$  \hspace{1cm} (3.77)

In (3.77), $h$ denotes the first column (an arbitrary selection) of the interfering channel $H_{21}$. Note that due to the design of the precoding vectors at AP 1, $\gamma_2$ does not suffer interference. Hence, comparing ZF in the uplink and downlink there is a clear difference. In the uplink, a small fraction of receive dimensions at AP 1 are sacrificed to null interference from the small MIMO system. In the downlink instead, a fraction of the transmit dimensions at AP 1 are lost to ensure that the small MIMO system remains interference-free. In addition, the average SINR of a UT associated to AP 2 is simply $\bar{\gamma}_2 = \gamma_2$, since the ZF precoder compensates (or inverts) any channel fluctuations.

The interference term, $I_1$, in (3.77) makes the evaluation of the average SINR, $\bar{\gamma}_1$, involved. It is however possible to show that (in similarity with the average $\lambda_n$), that the random variable $z = h^H W_2^Z F (h^H W_2^Z F)^H$ is also distributed according to a scaled $\mathcal{F}$-distribution:

$$z \sim \frac{N_2}{M_2 - N_2 + 1} \mathcal{F}_{2N_2,2(M_2-N_2+1)}.$$  \hspace{1cm} (3.79)

It is interesting to note that the interference originating from a ZF precoding transmitter has a very similar form to the interference seen after a ZF equalizer (cf. (3.32)). With the characterization of the interference, it is again possible to bound the average SINR as follows:

$$\frac{(M_1 - N_1 - N_2) p_1}{N_2 p_2 \sigma^2 + 1} \leq \gamma_1 < \frac{(M_1 - N_1 - N_2) p_1}{N_2 p_2 \sigma^2}.$$  \hspace{1cm} (3.80)

The lower bound is once more an application of Jensen’s inequality. The upper bound is obtained by ignoring the noise component in the denominator of (3.77) and using the mean of the interference term whose distribution follows a scaled $\mathcal{F}$-distribution, (3.79). Substituting the upper bound in (3.80) into the equality $\bar{\gamma}_1 = \rho \bar{\gamma}_2$, and solving
3.4. Spectrum Sharing in the Downlink

for \( p_1 \) gives:

\[
\frac{p_1}{\sigma^2} \approx \frac{\rho (M_2 - N_2) N_2}{M_1 - N_1 - N_2} \left( \frac{p_2}{\sigma^2} \right)^2.
\]

(3.81)

Although (3.81) is an approximation (asymptotically tight as \( p_2 \to \infty \)), it still stands in stark contrast to (3.45). Specifically, the transmit power at AP 1 is a function of \( p_2^2 \), and the ratio of antenna numbers is not to a fractional power. With this relation it can be noted that there is a cross-over point where \( p_1 \) becomes larger than \( p_2 \). It of course occurs at

\[
\frac{p_2}{\sigma^2} = \frac{M_1 - N_1 - N_2}{\rho (M_2 - N_2) N_2}.
\]

(3.82)

The difference between uplink and downlink can be traced to the way the interference is handled; it appears in the SINR expression for system 2 in the uplink (see (3.29)), but appears only in system 1 in the downlink (see (3.77)).

In Figs. 3.7 to 3.10 the above approximations are compared with the solution to the original problem statement (3.23). Using the fact that either \( p_1 = p_{\text{max}} \) or \( p_2 = p_{\text{max}} \) at the optimal average common SINR, the original problem reduces to solving \( \gamma_1 = \rho \gamma_2 \) for one transmit power while setting the transmit power of the other system to \( p_{\text{max}} \). The solution that satisfies the power constraints is evidently the feasible operating point.

The bounds in (3.80) can also be used to reject an initial power constraint hypothesis.

3.4.2 MF Precoding

Considering the MF precoder at AP 1, the preprocessing filter is given by [74]

\[
f_1(x_1) = \sqrt{N_1 p_1} W_{11}^{\text{MF}} x_1
\]

(3.83)

\[
= \sqrt{N_1 p_1} H_{11} A_1^{-1/2} x_1,
\]

(3.84)

where the normalization factor \( A_1 \) can take several forms, in similarity with the ZF normalization. Joham et al. [74] derived the normalization factor \( A_1 = \text{tr}(H_{11}^H H_{11}) I \) as part of the solution to the maximum average sum signal strength. Hoydis et al. [83] utilized \( A_1 = \mathcal{E}(\text{tr}(H_{11}^H H_{11}^H)) I \), while Björnson et al. [157] applied \( \text{diag} A_1 = \text{diag}(H_{11}^H H_{11}) \). For the present case, however, a fourth alternative is offered for analytic tractability: \( \text{diag} A_1 = \text{diag}(H_{11}^H H_{11}) \). It can be shown that asymptotically as the number of service antennas approaches infinity, all four normalization factors converge.

The SINR at an arbitrary UT associated to AP 1 using the precoding matrix (3.84) is therefore

\[
\gamma_1 = \frac{h_{11}^H p_1 h_1}{I_1 + I_2 + 1}.
\]

(3.85)
where

\[ I_1 = \|h^H \mathbf{W}_2^p\|^2 (M_2 - N_2) \frac{P_2}{\sigma^2}, \]  
\[ (3.86) \]

and

\[ I_2 = \|h^H \mathbf{H} \Lambda^{-\frac{1}{2}}\|^2 \frac{P_1}{\sigma^2}. \]  
\[ (3.87) \]

\( h \) again denotes the first column (an arbitrary selection) of \( \mathbf{H}_{11} \), and \( \mathbf{H} \) the remaining columns. \( \Lambda \) is defined as the matrix \( \Lambda_1 \) with the first column and row removed. As previously shown in Chapter 3.4.1, \( I_1 \) follows a scaled \( F \)-distribution. With the selected power normalization factor, \( I_2 \) takes the form of a Gamma-distributed random variable, \( G\left(\frac{(N_1 - 1)}{p_1/\sigma^2}\right) \). In addition, the numerator in (3.85), and the random variables \( I_1 \) and \( I_2 \), are pair-wise independent. The expansion of the expectation operator in \( \tau_1 = \mathcal{E}(\gamma_1) \) does not seem to lead to any insightful form. However, having characterized the interference terms, it is again possible to bound the average SINR. The lower bound can be expressed as

\[ \tau_1 = \mathcal{E}(h^H h) \frac{P_1}{\sigma^2} \mathcal{E} \left( \frac{1}{I_1 + I_2 + 1} \right), \]  
\[ (3.88) \]

\[ = M_1 \frac{P_1}{\sigma^2} \mathcal{E} \left( \frac{1}{I_1 + I_2 + 1} \right), \]  
\[ (3.89) \]

\[ \geq M_1 \frac{P_1}{\sigma^2} \mathcal{E} \left( I_1 + I_2 + 1 \right), \]  
\[ (3.90) \]

\[ = \frac{M_1 P_2}{\sigma^2} + (N_1 - 1) \frac{P_1}{\sigma^2} + 1. \]  
\[ (3.91) \]

The inequality in (3.90) is due to Jensen’s inequality. Again, note that this lower bound (ultimately based on the statistics of the underlying Rayleigh fading channel assumption), coincides with the asymptotic results present by Tse and Hanly [39].

A concise upper bound for \( \tau_1 \) appears more elusive. Hence for the MF precoder, the inequality derived by Wooff [177] will be used:

\[ \mathcal{E} \left( \frac{1}{z + 1} \right) \leq \frac{b^2}{a^2 + b^2} + \frac{a^3}{(b^2 + a(a + 1))(a^2 + b^2)}. \]  
\[ (3.92) \]

where \( a = \mathcal{E}(z) \) is the mean, and \( b^2 = \mathcal{E}(z - a)^2 \) is the variance. It is worth noting that (3.92) is a particular application of Wooff’s result, and the general inequality applies to \( \mathcal{E}(z + c)^{-k} \) for \( c > 0 \) and integers \( k > 0 \). Due to the independence between \( I_1 \) and \( I_2 \) the mean has the form:

\[ a = \mathcal{E}(I_1) + \mathcal{E}(I_2), \]
\[ = N_2 \frac{P_2}{\sigma^2} + (N_1 - 1) \frac{P_1}{\sigma^2}, \]  
\[ (3.93) \]

\[ (3.94) \]
3.4. Spectrum Sharing in the Downlink

and the variance can be expressed as

\[ b^2 = \mathcal{E}(z^2) - a^2, \]  

(3.95)

with the raw second moment

\[
\mathcal{E}(z^2) = \frac{(N_2 + 2)(N_2 + 1)(M_2 - N_2)^2}{(M_2 - N_2 + 2)(M_2 - N_2 + 1)} \left( \frac{P_2}{\sigma^2} \right)^2 \\
+ 2(N_1 - 1)N_2 \left( \frac{P_1}{\sigma^2} \right) \left( \frac{P_2}{\sigma^2} \right) \\
+ (N_1 + 1)N_1 \left( \frac{P_1}{\sigma^2} \right)^2. 
\]

(3.96)

One can see that as \( N_1 \to \infty \) the standard deviation, \( b \), grows at a much slower pace than the mean, \( a \), and eventually the upper bound in (3.92) approaches \( 1/a \). This again emphasizes the connection between the mean SINR and the asymptotic result used in the massive MIMO literature.

With MF precoding, AP 1 does not take the UTs belonging to system 2 into account. Therefore the SINR of a UT belonging to AP 2 takes the form

\[ \gamma_2 = \frac{(M_2 - N_2) \frac{P_2}{\sigma^2}}{\left\| h^H \mathbf{H}_{12} \mathbf{A}_1^{-1/2} \right\|^2 \frac{P_1}{\sigma^2} + 1}, \]

(3.97)

where \( h \) is used to denote the channel from a UT in system 2 to AP 1, i.e., a column of the matrix \( \mathbf{H}_{12} \). In similarity with \( I_2 \) in (3.85), the interference term in (3.97) is characterized by a Gamma distribution, \( \mathcal{G}(N_1, \frac{P_1}{\sigma^2}) \). Note that \( \gamma_2 \) has the same general form as that of the SINR in (3.33). Hence applying the same steps as in (3.34) to (3.36), the mean SINR can be described as

\[ \overline{\gamma}_2 = (M_2 - N_2) \frac{P_2}{\sigma^2} \left( \frac{\sigma^2}{P_1} \right)^{N_1} e^{\frac{\sigma^2}{P_1}} \Gamma \left( 1 - N_1, \frac{\sigma^2}{P_1} \right). \]

(3.98)

In addition, \( \overline{\gamma}_2 \) can be bounded by

\[ \frac{(M_2 - N_2) \frac{P_2}{\sigma^2}}{N_1 \frac{P_1}{\sigma^2} + 1} \leq \overline{\gamma}_2 < \frac{(M_2 - N_2) \frac{P_2}{\sigma^2}}{N_1 \frac{P_1}{\sigma^2}}. \]

(3.99)

The mean SINR expressions for system 1 and 2, (3.88) and (3.98), can be employed in problem statement (3.23) to find a specific operating point. Once again, the required transmit power is found by equating \( \overline{\gamma}_1 = \rho \overline{\gamma}_2 \), and assuming the power of the other system is set to \( p_{\text{max}} \). If the average SINR bounds cannot be used to reject an initial power constraint assumption, then both cases need to be solved and the feasible solution kept.

An approximate solution can be found by using the derived bounds. Substituting (3.91) and the upper bound in (3.99) in \( \overline{\gamma}_1 = \rho \overline{\gamma}_2 \) and solving for \( P_1 \) gives

\[ P_1 \approx a + \sqrt{a^2 + b}, \]

(3.100)
3.5. Spectrum Sharing Implications

In environments where users have similar path-loss components, and with the use of linear receiver or transmitter techniques, interference can have a significant impact on the mean SINR. The use of a massive number of antennas at one AP provides a way for the interference to be mitigated. In the uplink,

- the MF equalizer amplifies the desired signal power by a factor of $M_1$ above the interference and noise, while
- the ZF receiver sacrifices $N_2 \ll M_1$ receive dimensions to null the interference from AP 2.

In other words, with MF precoding at AP 1, both systems suffer interference and the transmit power at system 1 has a strong likeness to (3.64).
3.5. Spectrum Sharing Implications

Figure 3.8: Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 100$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$.

Figure 3.9: Exact and approximate average common SINR, $\gamma$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 1000$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$. 
3.5. Spectrum Sharing Implications

![Figure 3.10](image)

**Figure 3.10:** Exact and approximate transmit power, $p_1$, versus $\beta_1$ in the downlink (Chapters 3.4.1 and 3.4.2). $M_1 = 1000$, $M_2 = 4$, $p_2 = p_{\text{max}} = 10$ dB and $\rho = 1$.

Likewise in the downlink,

- the MF precoder enhances the signal strength at the targeted UT by $M_1$, and
- the ZF precoder (altruistically) restricts its transmit signals to the null-space of the interfering channel to the UTs of system 2, losing $N_2$ transmit dimensions in the process.

In addition, for a given operating point, the transmit power of a massive MIMO system, and hence the interference it inflicts on system 2, has been shown to be a decreasing function with $M_1$. The interference can hence be scaled back to an arbitrarily low level given a sufficient number of service antennas at AP 1. The value of the results in Chapters 3.3 and 3.4 now becomes clear: scaling laws involving $M_1$ allows insight into how much system 2 will be affected, and can ultimately be used to dimension a massive MIMO system for coexistence.

Is such coexistence, i.e., where two different systems transmit at the same time on the same frequency, beneficial? It can be argued that with the simpler linear receive and transmit strategies applied here, it may be better from a total sum rate perspective to orthogonalize usage (e.g. different time slots). However, this solution implies a central controller of some form, or some other scheduling mechanism (e.g. listen-before-talk).
3.6. Conclusion

The former method requires standardization between systems and additional overhead (something that has to date been avoided when considering unlicensed bands), while the latter solution has proved to be sensitive to for instance congestion. It is precisely the congested scenario that is of particular interest; otherwise the two systems could have operated on different frequency channels. As the results show, and the numerical results highlight, it is possible for AP 1 and AP 2 to coexist at a desired operating point by judiciously selecting the load parameters (i.e., number of UTs) of both systems.

Another interesting feature of the systems under study, is the fact that it could mirror an interference system with a Wi-Fi node and a mobile communications operator. If the operator’s network is also available to the Wi-Fi UTs (i.e., through subscription) then one or more UTs could shift from system 2 to system 1 to benefit from the lower transmit power requirements, and subsequent SINR improvement due to the lighter load at AP 2. In fact, due to the highly asymmetric antenna conditions it would be beneficial from an SINR perspective if all system 2 UTs migrated to system 1. However, one also has to take into account that, in contrast to Wi-Fi, a subscription to the operator’s network usually incurs a monetary charge.

3.6 Conclusion

In this chapter a coexistence scenario between a small and a large MU-MIMO system employing linear (pre-) processing techniques was analysed. The interference system was analysed by framing it in a maximum common mean SINR problem formulation. This allowed the characterization of the transmit power and mean SINR, as well as concise power scaling relationships giving insight into the performance of the system with varying service antennas and system load.

The use of a large number of service antennas at one of the access points, enables lower transmit power and better interference rejection for a given SINR operating point. These general traits have previously been exposed in other massive MIMO or large scale antenna systems studies. However harnessing these characteristics in a coexistence scenario has not been previously explored. Also, by specifically linking the SINRs of the conventional and massive MIMO systems, the behaviour of the systems could be better explored. The used method has the distinct advantage over the much more common and classical approach of underlay in hierarchical access, that instead of a crude interference temperature constraint, here the SINR and hence by extension the rates of the UTs are explicitly studied. This offers an approach where the effect of sharing the channel can be seen in the performance of both parties simultaneously.

The study suggests that compared to conventional multi-user MIMO systems, the co-
existence of a massive MIMO system offers improved SINR performance throughout. It must be pointed out however that the gain, quantified analytically here, is not in general as optimistic as early isolated massive MIMO literature claims. For instance in the uplink, assuming a ZF or MF processing at the massive MIMO access point, the resulting common SINR is a function of the square root of the number of service antennas, as opposed to a linear function of the number of service antennas. The weakness of the MF processing compared to ZF in the face of strong interference was also noticeable, and mirrors the near-far issue observed in CDMA systems. With the use of linear processing, the additional interference due to coexistence generally degrades UTs SINR compared to the isolated environment. Equivalently, higher transmit power is required to maintain a particular operating point.

In the scenario under focus, the very asymmetric nature of the systems, quite naturally led to the smaller conventional MU-MIMO system being the bottleneck. In the uplink, the interference at the conventional MIMO access point, can be mitigated with an increase in the number of massive MIMO service antennas. In the downlink, the use of a large number of antennas can similarly be used for array gain to reduce transmit power and hence interference, or to place nulls in the radiation pattern and eliminate any interference to the UTs belonging to the conventional MIMO system. The derived scaling laws give an indication of the effectiveness of adjusting the number of service antennas. As an alternative, or in combination, the number of UTs at each access point may also be adjusted. One interesting aspect is for instance that by reducing the number of UTs of the conventional MIMO system by one, many more may be added to the massive MIMO system, and still maintain the original SINR point. As a specific example, in the uplink, reducing the number of UTs from three to two at an access point with four service antennas, allows the massive MIMO system to increase the load from five to nine UTs. This is accomplished with 100 service antennas at the massive MIMO system and a 10 dB normalized power constraint. If instead the access point has 1000 service antennas, about 40 additional UTs may be admitted for the same SINR operating point. The scaling laws may be used precisely in this manner to gain an understanding of the system behaviour with the number of service antennas and load.
Spectrum Sharing with Primary Cooperation

4.1 Introduction

In this chapter an opportunistic transmit antenna selection scheme with PU cooperation is investigated. The basic idea is that the PU receivers adapt their linear receive filters and the SU transmitters select a subset of transmit antennas to allow the sharing of the channel while simultaneously guaranteeing the satisfaction of the PU users. In this manner, opportunistic SU transmission is enabled while meeting a minimum rate threshold for all PU users. The scheme is evaluated in terms of SU average achieved rate and the cost to the PU system in the form of an average transmit power sacrifice. The particularly interesting feature with this scheme is that through cooperation with the PU receivers, an additional user is allowed to transmit in a spatial interweave fashion even though the SU transmitters do not necessarily have enough antennas to freely place their interfering signals in the PU receive space.

4.2 System Model

The system under consideration consists of a PU set, \( P_p \in \{1, 2, \ldots, K_p\} \), communicating simultaneously over a narrow-band channel, and a set of SUs, \( P_s \in \{(K_p + 1), (K_p + 2), \ldots, K\} \), opportunistically accessing the same channel. \( K = K_p + K_s \) represents the total number of transceivers in the system. As illustrated in Fig. 4.1, the \( k \)th transmitter possesses \( M_k \) antennas with which to transmit \( d_k \) streams, and the intended recipient, the \( k \)th receiver, collects the signal on its \( N_k \) antennas. The
discrete-time equivalent form of the received signal at the \( k \)th receiver, assuming perfect synchronization, is

\[
y_k = H_{kk} F_k s_k + \sum_{i=1, i \neq k}^{i=K} H_{ki} F_i s_i + v_k \tag{4.1}
\]

where \( H_{ki} \in \mathbb{C}^{N_k \times M_i} \) is the single-tap channel response between transmitter \( i \) and receiver \( k \), \( F_k \in \mathbb{C}^{M_k \times d_k} \) is the precoding matrix at transmitter \( k \), \( s_k \in \mathbb{C}^{d_k} \) is the symbol vector, and \( v_k \) is the additive white Gaussian noise at receiver \( k \) with covariance matrix \( \sigma_v^2 I \). For the analysis, it is assumed that the channel matrices \( H_{ki} \) are each of full rank and mutually independent, the transmitted symbol elements are independently drawn from a circularly symmetric complex Gaussian distribution with zero mean and unit variance, and the noise is statistically independent from the transmitted signals.

The selection of a subset of transmit antennas at the SU is accomplished through the use of an antenna selection matrix \( A_k \), \( k \in \mathcal{P}_s \), and consists of \( L \) columns, each chosen from \( \{ e_i \}_{i=1}^{M_k} \). Here, \( e_i \) represents column \( i \) of the identity matrix. In (4.1) the SU precoding matrix is thus a combination of the antenna selection matrix and an additional transformation representing beamforming and power allocation

\[
F_k = A_k \hat{F}_k, k \in \mathcal{P}_s. \tag{4.2}
\]

The specific design of \( A_k \) and \( \hat{F}_k \) is described in Chapters 4.4 and 4.5. The channel from the SU transmitter, as observed by all receivers, is limited to the SU active antennas

\[\begin{align*}
\mathcal{P}_s & \quad \{ F_1, F_2, \ldots, F_K \} \\
& \quad \{ H_{11}, H_{12}, \ldots, H_{K} \} \\
& \quad \{ G_1, G_2, \ldots, G_K \}
\end{align*}\]

Figure 4.1: \( K \)-user interference channel where the transceivers in set \( \mathcal{P}_s \) are the secondary user pairs.
and is denoted

\[ \tilde{H}_{ki} = H_{ki} A_i, \quad \forall k \in {\mathcal{P}_p} \cup {\mathcal{P}_s}, i \in {\mathcal{P}_s}. \]  

(4.3)

Knowledge of \( \tilde{H}_{ki} \) at the PU receivers may be gained in several ways. For instance, the PU receivers may be aware of other transmitters and use available training sequences (the SU only transmits on active antennas) to estimate the channels to the interferers. In a second scenario, with greater PU cooperation, the PU receivers estimate the entire \( H_{ki} \) channels, inform the SUs of a set of feasible antenna selection matrices, and confirmation of the antenna choice is subsequently given by the SUs.

A critical question with regards to hierarchical access in general is why the incumbent user or license owner should choose to share the channel with a SU, and potentially subject itself to excessive interference? This is still a hotly debated issue with proponents pointing out potential user capacity and/or spectrum efficiency improvements, and opponents highlighting the risks of excessive interference and overall system complexity. One concrete example of hierarchical access is the usage of TV white space in the United States [179, 180]. Here, the sharing between incumbent television broadcasters and unlicensed secondary entities has been mandated. Another approach to spectrum sharing is the idea of dynamic spectrum leasing [181, 182], where the incumbent has the right to lease a part of its spectrum to other users. In this manner the primary has more control over how its allocated spectrum is employed and can balance remuneration versus quality of service for its own users. It is in this category the present scheme might find a use. The dynamic leasing principle already requires a communication channel between the primary and secondary system (although generally on much longer time scales), and by using multiple antenna techniques spectrum sharing may be enabled on a finer geographical scale. In addition, with the active cooperation from the PUs, it also offers a method to avoid harmful interference to the primary system. If the interference levels cannot be tolerated (i.e., the PUs rate constraints are not met), the SUs are not allowed to transmit. It is the break from a passive, unaware PU, that in effect offers a greater level of protection from SU interference, when compared to a spectrum sensing only approach or database-aided sharing. Lastly, the configuration of the present scheme (a \( K \)-user interference channel in its most general form) makes the technique applicable to a wide range of scenarios. Two examples are for instance, multiple antenna device-to-device communication underlaid with an uplink cellular network [183], or simply between two cellular networks. In addition, an interesting interpretation of a multi-tiered network may be by defining priority data (perhaps delay-sensitive applications such as voice or video) over secondary information, as opposed to different entities.
4.3 PU Description

The PU receiver, able to perfectly estimate the channels from all other transmitters (primary or secondary, indiscriminately), designs its receiver decoding matrix to separate the desired symbols and null interference. The ZF filter for PU $k$ takes the following form [184]

$$G_k = \begin{bmatrix} I_{d_k} & 0_{d_k \times (N_k - d_k)} \end{bmatrix} \bar{H}_k^\dagger$$  \hspace{1cm} (4.4)

where \( \bar{H}_k = [H_{kk} F_k C_k \] and the aggregate cross-channel is defined as

$$C_k = \begin{bmatrix} H_{k1} & \cdots & H_{k(k-1)} & H_{k(k+1)} & \cdots & H_{kK_p} & \tilde{H}_{k(K_p+1)} & \cdots & \tilde{H}_{kK} \end{bmatrix}.$$ \hspace{1cm} (4.5)

Here, \((\cdot)\dagger\) denotes the pseudo-inverse and with the assumption of full rank, independent channel matrices the left inverse exists as long as [185]

$$N_k \geq d_k + L + \sum_{i=1,i\neq k}^{K} (M_i), \quad k \in \mathcal{P}_p.$$ \hspace{1cm} (4.6)

The achievable rate of user $k$, with interference treated as noise, is expressed as (cf. [16])

$$R_k = \log \left| I_{d_k} + G_k H_{kk} Q_k H_{kk}^H (G_k R_k G_k^H)^{-1} \right|$$ \hspace{1cm} (4.7)

The interference-plus-noise covariance matrix is represented as

$$R_k = \sigma_v^2 I_{N_k} + \sum_{i=1,i\neq k}^{K} H_{ki} F_i F_i^H H_{ki}^H$$ \hspace{1cm} (4.8)

and the transmit covariance matrix is defined as

$$Q_k = \mathcal{E} \left[ (F_k s_k) (F_k s_k)^H \right] = F_k F_k^H.$$ \hspace{1cm} (4.9)

The transmit power is concisely described by $P_k = \text{tr} (Q_k)$.

Substituting (4.4) into (4.7) and applying the partitioned matrix inverse [186] (also known as Banachiewicz inversion formula) the achievable rate for PU $k$ reduces to

$$R_k = \log \left| I_{d_k} + Q_k H_{kk}^H \tilde{C}_k H_{kk} \right|$$ \hspace{1cm} (4.10)

where \( \tilde{C}_k = I_{N_k} - C_k (C_k^H C_k)^{-1} C_k^H \) can be recognized to be the (idempotent) projection matrix into the interference null-space.

In the scenario under consideration the PUs have specified rate constraints

$$R_k \geq R_{k,t}, \quad \forall k \in \mathcal{P}_p,$$ \hspace{1cm} (4.11)
perhaps reflecting the requirements of a specific application such as a voice call or video streaming. In this case the PUs are perfectly satisfied with any rate equal to or greater than their rate constraints and it seems natural to use this constraint in a minimization problem to limit the use of another finite resource in many wireless communications systems: power. Since the PU receiver applies ZF, the individual user’s rate (4.10) does not depend on the other PUs choice of precoding matrices. In this case, the $k$th PU’s desire to transmit at or above a target rate, $R_{k,t}$, using minimum transmit power simply reduces to the point-to-point margin adaptation algorithm. Following the results in [16] the optimal precoding matrix at transmitter $k$ takes the form

$$F_k = V_k D_k$$

where the unitary matrix $V_k$ is from the singular value decomposition (SVD) of $\bar{C}_k H_{kk} = U_k \Sigma_k V_k^H$ and $D_k$ is a diagonal matrix with element $i$ set according to WF

$$[D_k]_{ii}^2 = \max \left( \mu_k - \frac{1}{[\Sigma_k]_{ii}}, 0 \right).$$

The water level $\mu_k$ is adjusted so that the PU’s rate meets the specified rate constraint. Note that due to the application of the unitary matrix $V_k$ the effective channel in (4.10) has been diagonalized and so decoding each component of $s_k$ independently (e.g., see [68]) does not cause a rate loss.

With both a minimum rate constraint and a maximum power constraint, situations may arise where due to the channel realization a particular PU cannot reach its desired rate constraint even in the absence of a SU. The behaviour of the PU in such a case will be to remain idle for the duration of the channel realization. The set of PUs that the SU must take into consideration reduces to the active PU set, $P_a$.

### 4.4 Single SU Transmit Method

Based on (4.6) and (4.11) the SU can only transmit if it can ensure that the PU receivers can ZF any interference and that with the resulting post-processing filters the achieved rates of the individual PUs remain above the specified constraints. The complete problem may be formulated as

$$\max_{F_K} R_K$$

subject to $R_{k,t} \leq R_k, \ \forall P_a$

$$P_k \leq P_{k,t}, \ \forall k \in P_a \cup P_s.$$ 

The achievable rate of the SU takes the same form as (4.7), and the SU power constraint is, in similarity with the PUs, defined as $P_K = \text{tr} (Q_K)$. The SU precoder, as specified in (4.2), consists of a selection matrix $A_K$ and a beamforming matrix $\hat{F}_K$. 

$$\max_{F_K} R_K$$

subject to $R_{k,t} \leq R_k, \ \forall P_a$

$$P_k \leq P_{k,t}, \ \forall k \in P_a \cup P_s.$$
Algorithm 1. Antenna selection.

1: Find $P_a$
2: Find $L_{\text{max}}$ using (4.16)
3: $\mathcal{A} \leftarrow \text{GenerateCombinations}(L_{\text{max}})$ using exhaustive subset or correlation based selection (Chapter 4.6.1)
4: for $j = 1$ to $|\mathcal{A}|$ do
5: \[\mathbf{A}_j \leftarrow [\mathcal{A}]_j\]
6: for $k = 1$ to $|P_a|$ do
7: \[\text{Find } \mathbf{G}_k \text{ using (4.4)}\]
8: \[\text{Find } \mathbf{F}_k \text{ using (4.12) and (4.13)}\]
9: end for
10: for $n = 1$ to $n_r$ do
11: \[\text{Find } \hat{\mathbf{F}}_i, \forall i \in P_a \text{ using (4.13) and (4.15)}\]
12: end for
13: end for
14: $j \leftarrow \arg\max_j (\sum_i R_{i,j}) \text{ s.t. } R_{k,j} \geq R_{k,t}$
15: $\mathbf{A} \leftarrow \mathbf{A}_j$

The discrete nature of the antenna selection matrix, $\mathbf{A}_K$, makes an exact analytical solution for the optimal SU rate, $R_{K,\text{opt}}$, seem difficult for the problem described in (4.14). In fact, in the related problem of choosing the best $n$ antenna combination out of a total of $m$ transmit antennas to maximize the achievable rate in a point-to-point MIMO system [187], the optimum solution uses a search through the $(m^n)$ possible combinations.

Using a similar approach, the problem in (4.14) is divided into two parts. In the first step the SU achievable rate is calculated for each antenna selection matrix $\mathbf{A}_{K,i}$ in the set of possible combinations $\mathcal{A}$. In the second step $R_{K,\text{opt}}$ is selected as the largest SU rate, subject to the active PUs’ rate constraints.

For a given antenna selection matrix, the maximum SU achievable rate (w.r.t. $Q_K$) can be obtained from

$$\hat{\mathbf{F}}_K = \mathbf{V}_K \mathbf{D}_K.$$  \hfill (4.15)

In similarity with the construction of the PU precoding matrix, $\mathbf{V}_K$ is the unitary matrix from the SVD of $\mathbf{R}_K^{-1/2} \tilde{\mathbf{H}}_{KK} = \mathbf{U}_K \Sigma_K \mathbf{V}_K^H$ and $\mathbf{D}_K$ is a diagonal matrix with elements set according to WF (4.13). The main difference between PU power minimization and SU rate maximization lies in the fact the water level for the SU, $\mu_K$, is
chosen to saturate the power constraint.

Note that in order to investigate the maximum achievable rate of the SU with transmit antenna selection an ideal receiver has been assumed and hence the SU receive filter \( G_K \) has been omitted from the discussion. In fact, following the results in [16], it can be shown that an appropriate \( G_K \) can be applied to diagonalize the effective SU channel and hence each element in the transmitted symbol vector may be decoded individually without incurring a rate loss.

In an attempt to reduce the search space, a limit on the maximum number of SU transmit antennas may be imposed by inspecting the zero-forcing constraint. Solving (4.6) for \( L \) and adding the SU transmit antenna \( M_K \) limit, the largest number of active transmit antennas while keeping ZF feasible at the PU receivers is

\[
L_{ZF} = \min \left( \min_{k \in P_a} \left( N_k - d_k - \sum_{i=1, i \neq k}^{K-1} M_i \right), M_0 \right).
\] (4.16)

Further, by setting \( d_k = 1 \ \forall k \in P_a \), i.e., assuming each active PU achieves its rate constraint with a single stream, the maximum number of SU transmit antennas is limited to \( L_{max} = L_{ZF} |_{(d_k = 1 \ \forall k)} \). Then, \( |A| \) (size of set \( A \)) follows the partial sums of binomial coefficients (elements of a Bernoulli triangle), \( \sum_{i=0}^{L_{max}} \binom{M_K}{i} \). Although no closed form expressions exist for this sum, lower and upper bounds developed by Worsch [188] show that it still grows exponentially with \( M_K \). To cope with a larger number of antennas, two approximate antenna selection algorithms are presented in Chapter 4.6.

To illustrate the method applied by the SU, consider the following example. Expressing the number of receive and transmit antennas as \( (N_k, M_k) \), PU 1 and PU 2 are each \( (4, 2) \), and the SU has a \( (3, 3) \) antenna configuration. Given a particular channel realization, power constraints and PU rate constraints, the active set \( P_a \) contains both PUs. According to (4.16) the SU can use a maximum of 2 transmit antennas. The SU evaluates its own rate and the PU rates for all \( \sum_{i=0}^{2} \binom{3}{i} = 7 \) antenna combinations (remember the null set is also included), determines that only using either its first or third transmit antenna satisfies \( R_{k,t}, \ \forall k \in P_a \) and finally selects the third antenna since it offers the highest SU rate.

The algorithm performed by the SU is summarized in Algorithm 1.

### 4.5 Multiple SUs

The PUs coexist with each other by simply applying a ZF filter, nulling any interference. This requires that the number of transmit streams is limited to ensure that ZF is
feasible. One option is for instance to use a central controller of some form, and it seems reasonable to assume that for primary coexistence such a mechanism is in place. However the SUs are opportunistic, and may not necessarily belong to the same group (e.g., network operator), and hence the question of how more than one SU may share the spectrum becomes critical.

With a single SU present, maximizing the achievable rate seems a reasonable objective since it efficiently utilizes the shared spectrum (albeit at a cost in terms of power to the PUs). With multiple SUs competing for the same spectrum, not only do the SUs cause interference to the PUs, but the SUs of course also interfere with each other. This further complicates the behaviour of the system. Finding a solution to efficiently use the spectrum in an interference channel, i.e., to maximize the sum rate of the SUs, is still an open question in general, and a whole host of techniques have been proposed. One technique may be the use of interference alignment [7], but one of the weaknesses of interference alignment is that it requires the number of antennas to scale with the number of users. Peters and Heath [60] propose several algorithms for the interference channel, including minimizing an interference leakage metric, finding an MMSE solution, and maximizing a sum SINR metric. One of the issues with these algorithms (and that of interference alignment) is that the different objectives are not a direct function of rate, and therefore begs the question: what problem is actually being solved?

Instead, to illustrate the antenna selection with PU cooperation technique, two multi-user techniques are adapted from the literature that are based on maximizing rate directly. The first technique is based on iterative WF [189], where each SU link attempts to maximize its own rate. The second technique draws inspiration from the paper by Nosrat-Makouei et al. [126], where a new arrival seeks to join an existing system. Finally, it must also be noted that the single SU transmit method described in Chapter 4.4 may still be used for multiple SUs by allocating each SU to separate frequency sub-bands.

### 4.5.1 Iterative WF

Given an antenna selection, and resulting PU transmit strategy, the interference between the SUs is a direct result of the choice of pre-processing matrices, $F_k, k \in P_s$. Even for the special SISO case, this interference channel has been proven to be a strongly NP-hard problem [124] for various objectives, such as sum rate or proportional fairness. Since finding a specific point on the Pareto boundary is analytically intractable, an approximation will instead be sought. Using the principle of iterative WF, for a given antenna selection, each SU will repeatedly apply WF to maximize its
own rate. By iteratively finding the pre-processing matrices, the interference caused to the other nodes is updated and can consequently be taken into account.

It must be stated that there is no guarantee that the final operating point is a global optimum, and in addition unless specific interference magnitude constraints are met (see for instance the work on MIMO Nash equilibrium by Scutari et al. [190]) the algorithm may not converge to a Nash equilibrium. For this study a Nash equilibrium is not a strict requirement, however it is attractive from the point of view that it represents a stable operating point. Finally, the iterative WF technique has in general been observed to be a stable method, providing good solutions with short convergence times [134, 191, 189, 190, 60].

For a specific antenna selection matrix, and given the PU rate constraints and the transmit power constraints the problem may hence be summarized as:

\[
\max_{\mathbf{F}_i} R_i \quad \forall i \in \mathcal{P}_s \\
\text{s.t.} \ R_{k,t} \leq R_k, \\
P_k \leq P_{k,t}, \quad \forall k \in \mathcal{P}_a \\
P_i \leq P_{i,t}, \quad \forall i \in \mathcal{P}_s
\]  

(4.17)

Using the iterative method, each SU updates its own pre-processing filter in response to the other SUs’ transmit strategies. This procedure continues until convergence, or until a predefined maximum number of iterations, \( n_r \), has been reached. Hence, for each iteration, the individual SUs calculate their pre-processing matrices, \( \mathbf{F}_k \), in exactly the same manner as for the single SU case in Chapter 4.4. The final antenna selection matrices for each SU transmitter is selected based on the highest achievable sum rate. The whole procedure is summarized in Algorithm 1.

### 4.5.2 Sequential Admission

The second multiple SU technique is based on the idea of user arrival [192, 193, 126]. In particular the work by Nosrat-Makouei et al. [126], describes a method whereby the existing set of users are considered primary, and a new arrival can only join the system if it caused no additional interference (this was enabled with the use of interference alignment). The same principle may in part be used in the present case. Consider that a new SU wishes to transmit, then the existing SU is treated as a PU and gains protection in the form of a minimum rate constraint. If no antenna selection matrix is found to satisfy the PUs and the existing SUs, the new SU may not transmit. In this manner priority is given to existing SUs. In addition, if the PUs cannot achieve their minimum rate constraints, then the number of existing SUs are reduced.
4.6 Reduced Complexity Antenna Selection Algorithms

This simple scheme has a set of advantages. Due to the opportunistic nature of SUs, the admission of a single SU at a time, may better reflect practical circumstances. With the addition of a single SU to the system, the number of antenna combinations is limited to a single transmitter, thereby restricting the computational burden. Finally, some protection is offered to existing SUs, providing a sense of fairness and perhaps stability to the system. Weaknesses of the scheme include an extended set of parameters that is difficult to define in order to achieve fair or efficient usage (in terms of both spectrum and power), and that low spectrum efficiency may result if existing SU have poorer channel conditions than new arrivals.

The scheme is modelled using the same principle as the single SU case in Chapter 4.4. With the addition of a new SU, the existing SU is handed a minimum rate constraint, and for the simulation purposes in Chapter 4.7 is treated as a member of the PU group. Further gains may be expected by judiciously selecting an appropriate minimum rate constraint, maximum transmit power or number of transmit antennas. These optimization issues are left as future research.

4.6 Reduced Complexity Antenna Selection Algorithms

The computational complexity order of ZF and SVD operations are roughly cubic with the number of antennas. However in order to find the pre- and post-processing matrices that maximize sum rate these steps must be repeated over all possible combinations of transmit antennas (see Algorithm 1). The number of combinations alone increases exponentially with the number of antennas [188], and hence the computational burden quickly becomes excessive. It is true that this problem belongs to the category embarrassingly parallel, where each antenna combination may be treated separately. However the number of processing units must also scale at an exponential rate and hence full parallelization is not a sustainable solution either. It is therefore worthwhile to investigate sub-optimum selection algorithms with lower computational complexity.

Antenna selection, i.e., the principle of using a subset of antennas, was first studied in the context of choosing a group of antennas at one end, while only a single antenna is present at the other end [194, 195]. This scheme is commonly known as hybrid selection/maximum ratio combining. Given a subset size, $L$, the optimal antenna selection rule to maximize SNR, is simply to choose the $L$ antennas with largest channel gains. Allowing for more than a single antenna at both transmitter and receiver, i.e., a MIMO system, complicates the analysis and no closed form solutions seem to be available. To find the $L$ antenna subset that maximizes the rate of a MIMO system, all antenna combinations must be investigated [196, 187].
In order to avoid this exponential increase in computational complexity with the number of antennas several algorithms have been proposed. One of the most rudimentary methods is to simply apply the same procedure as in the single stream case, i.e., a norm-based selection maximizing the sum channel gains [196]. This method does not take into consideration, however, that rate is a function of both the number of streams and the eigenvalues of the channel matrix. To improve the rate performance, Molisch et al. [187] proposed a correlation-based algorithm that selects antennas based on both channel strength and orthogonality (prioritizing orthogonality). In the same article, a mutual information based decision metric was also presented with slightly improved performance, but the method also requires explicit knowledge of the noise variance. Gharavi-Alkhansari and Gershman [197] also applied the correlation-based approach and made the observation that it mirrors an ordered Gram-Schmidt orthogonalization procedure. Also, Gharavi-Alkhansari and Gershman noted that the method, incrementally adding one antenna at a time, is the opposite of the method presented by Gorokhov [198], where the starting point is the full antenna set and at each step one antenna element is removed.

The previously described methods for a MIMO system operate by only examining the channel matrix. If additional constraints such as the type of receive filtering is present, it may be the case that algorithm dealing explicitly with the (sum) rate may be more appropriate. Jorswieck and Mochaourab [199] and Lin and Tsai [200] have both applied a form of greedy selection to maximize rate. The underlying principle is to incrementally choose the transmit antenna that maximizes the sum rate. In both the case of a MIMO system [200], and for multi-user MISO [199], excellent results have been found at high SNR, with an increasing gap occurring between the suboptimal algorithm and the exhaustive search at lower SNRs. This last observation may suggest that the greedy rate selection methods may have a link to correlation based selection where orthogonality is emphasised. More complicated procedures have also been used to better incorporate the transceiver design. For instance, Zhai et al. [201] developed an approach using an iterative concave-convex procedure.

The previous works have dealt with a MIMO channel (either directly or through multiple single antenna transmitters and a multiple antenna receiver). However, in the present case, the SUs that need to perform transmit antenna selection, must also take into consideration the existence of PUs. This added constraint has been shown to be difficult to incorporate in the selection algorithms. For instance, Waheed and Cai [202] resorted to the use of a genetic algorithm and particle swarm optimization. Acknowledging that finding an algorithm that simultaneously takes into account the quality of the SUs direct channels and the effect of interference to the PUs, is an unlikely prospect, a simplified method will be formulated. The basic idea will be to reduce the
4.6. Reduced Complexity Antenna Selection Algorithms

of candidate antenna subsets, and investigate the most appropriate solution within this set. In this manner, a certain level of adaptation is kept since for each antenna configuration the pre-processing matrices are explicitly calculated and hence the PUs reaction is incorporated. Two different candidate antenna subset selection methods will be explored. The first is based on orthogonalization, and the second method is based instead on incrementally adding the transmit antenna that contributes the most in terms of rate.

4.6.1 Correlation Based Selection

The first sub-optimum algorithm with lower computational complexity is derived from the principle of correlation based selection by Molisch et al. [187] and Gharavi-Alkhansari and Gershman [197]. For the single SU, the applied procedure is exactly the same. Consider a channel matrix $H_{KK}$. The first transmit antenna (or column of $H_{KK}$) is chosen based on largest channel gain. The second antenna is chosen such that the gain (or 2-norm) of the element orthogonal to the previous matrix column is maximized. This last step is repeated until $L_{\text{max}}$ antennas have been selected. Clearly, the above procedure (except for halting at $L_{\text{max}}$ columns) describes a form of ordered orthogonalization. The above procedure has a complexity proportional to $N_k^2L_{\text{max}}$ [197]. With a derived list of candidate transmit antennas, the pre- and post-processing matrices for all primary and secondary users is calculated for 1 to $L_{\text{max}}$ transmit antennas. The selection that maximizes the rate is chosen. The procedure follows Algorithm 1, and with the reduced antenna candidate set has a computational complexity of roughly $L_{\text{max}}N_k^3$.

For multiple SUs, the same principle can be applied. Since multiple SUs are present, it is the direct channel gains that are of interest, and used in the ordering procedure. Note that, whereas in the single SU case, the method can be described as an ordered QR-decomposition, this does not hold for multiple SUs. Specifically, for the single SU case, the diagonal elements of the $R$ matrix in the QR-decomposition will hold monotonically decreasing absolute values. In the multiple SU case, the decomposition of the compound channel containing all SU channel paths, will not necessarily result in an $R$ matrix with decreasing diagonal elements. This is due entirely to the fact that the ordering is based on a subset of elements in each column of the compound matrix. Nevertheless, through observation, this effect has been noted to be small (e.g. one or two occasionally swapped elements of very similar magnitude). The candidate antenna set (possibly containing transmit antennas from all SUs, or only a single one) is then cycled through to find the precoders and decoders that maximize the sum rate.

The main advantage of the above procedure is that it only requires $L_{\text{max}}$ calculations of the SVD operation, and hence the overall computational order of the algorithm is
bounded from above by \( N^2_K \), since \( L_{\text{max}} \leq N_K \). However the simplicity of the antenna subset generation (i.e., only using the channel state directly) clearly ignores the cross channels to the PUs, the potential reaction of the PUs, and the more fine-grained interaction between the SUs streams. Expounding on the last statement: The antenna candidate set is only selected based on the SU channels. For instance it does not take advantage of exceptionally weak channels to the PUs, that could have enabled spectrum sharing with close to no PU interference. Also, the main premise of this hierarchical access system is that the PUs adapt their receive filtering to accommodate a SU. The candidate set is not based on any information as to how the PUs react. Finally, selecting the candidate set through orthogonalization has been known to be near-optimal at high SNRs, but not for lower ranges.

**4.6.2 Greedy Rate Selection**

In an attempt to capture a larger portion of the sum rate a second antenna selection algorithm is developed. It is based on the principle of greedy selection, i.e., incrementally adding the transmit antenna that contributes the most in terms of additional rate. The same principle has been applied in a multitude of scenarios, but for the present case with a PU constraint, the most pertinent example is by Jorswieck and Mochaourab [199], who applied greedy user selection to choose a subset of SUs that maximize the sum rate. Since the rate is calculated at each step of the algorithm, the state of the other users’ precoders and equalizers is taken into account, and hence it is expected that such an algorithm should perform better than the correlation based selection, where the antenna set is obtained only through the use of the SUs channel paths.

Denoting the sum of all SU transmit antennas as \( M \), i.e., \( M = \sum_{k \in P} M_k \), the largest antenna candidate set is

\[
|A| = M + (M - 1) + \ldots + (M - L_{\text{max}} + 2) + (M - L_{\text{max}} + 1),
\]

(4.18)

\[
= \sum_{i=(M-L_{\text{max}}+1)}^{M} i,
\]

(4.19)

\[
= L_{\text{max}} (2M - L_{\text{max}} + 1) / 2.
\]

(4.20)

Hence, if the greedy selection loop is not terminated at an earlier stage (this could occur if the sum rate is not increased from the previous step), the greedy selection processes requires on the order of \( N^2_K \) iterations of the SVD operation, and hence overall the greedy selection algorithm has a complexity order of roughly \( N^5_K \).

Since the antenna set is incrementally selected based on rate, the complete SU pre-coder algorithm takes on a different form than Algorithm 1. Algorithm 2 describes
Algorithm 2. Greedy Rate selection.
1: Find active PU set $\mathcal{P}_a$
2: Find set of all SU transmit antennas $\mathcal{U}$, empty set $\mathcal{A} = \{\}$
3: for $m = 1$ to $L_{max}$ do
4: for $j = 1$ to $|\mathcal{U}|$ do
5: $\mathcal{A}_j \leftarrow \{j\} \cup \mathcal{A}$
6: Find $G_k$, $\forall k \in \mathcal{P}_a$ using (4.4)
7: Find $F_k$, $\forall k \in \mathcal{P}_a$ using (4.12) and (4.13)
8: for $n = 1$ to $n_r$ do
9: Find $\hat{F}_k$, $\forall k \in \mathcal{P}_a$ using (4.13) and (4.15)
10: end for
11: Find sum rate $SR_{m,j} = \sum_{i \in \mathcal{P}_s} R_i$
12: end for
13: Find $j^* = \arg\max_{j \in \mathcal{U}} SR_{m,j}$
14: if $SR_{m,j^*} > SR$ then
15: $\mathcal{A} \leftarrow \mathcal{A} \cup \{j^*\}$, $\mathcal{U} \leftarrow \mathcal{U} \setminus \{j^*\}$
16: $SR \leftarrow SR_{m,j^*}$
17: else
18: break
19: end if
20: end for

the complete procedure for the greedy rate selection. The hierarchical access problem involves many difficult to characterize interactions, and the greedy selection method is not claimed to achieve a sum rate optimal solution. However since all users’ reactions (i.e., their rates) are calculated for each additional antenna the algorithm incorporates partially the effect of the different precoding matrices and the PUs receive filters.

### 4.7 Simulation Results and Discussion

The behaviour of the proposed scenario and the performance of the SUs is explored through Monte-Carlo simulations. Rayleigh flat fading channels are considered between all terminals, i.e., $[H_{kl}]_{ij} \sim \mathcal{CN}(0,1)$, $\forall k, l, i, j$. For ease of exposition the results are limited to the case of identical PUs and identical SUs, and in all results the PUs have the same rate and power constraints. If not explicitly stated all results are based on $1 \times 10^5$ trials. The simulations have been separated into separate experiments for
4.7. Simulation Results and Discussion

Figure 4.2: Probability that PU $k$ remains idle due to its inability to reach the rate constraint within the power constraint.

certainty. Experiment 1 exposes the behaviour of the PUs, in particular the principle that the PUs remain inactive if unable to reach their rate constraints. In Experiments 2 and 3, the single SU scenario is examined in terms of both PU rate and transmit power constraint. Finally, Experiments 4 and 5 deal with the multi-SU scenario.

**Experiment 1**

Before analysing the performance of the SUs it is worth emphasizing the PUs’ behaviour. In the scenario under consideration the PU remains idle if it cannot reach its rate constraint within its power budget. Fig. 4.2 shows this trait as the probability of a PU being idle with respect to its rate constraint. For all SU simulations, each PU $k$ is equipped with $N_k = 4$ receive antennas, $M_k = 2$ transmit antennas. Evidently, given a fixed power constraint, setting a higher rate constraint reduces the likelihood that the PU will transmit at all. For a channel realization where all PUs remain idle the SU may unrestrictedly make use of the channel. This type of opportunistic use of the channel (commonly known as time interweave [8] in CR), although effective in its own right, detracts from the case with active aware PUs. Hence for the remainder of the discussion only rate constraints that incur less than 1% idle probability will be considered. From Fig. 4.2 this corresponds to rate constraints of about 9, 12 and 15 b/s/Hz for transmit power constraints of 24, 30 and 36 dB, respectively. Similarly for the case with multiple SUs, where $(N_k, M_k) = (8, 2)$ $\forall k \in \mathcal{P}_a$ and $(N_k, M_k) = (8, 8)$ $\forall k \in \mathcal{P}_s$, the 1% idle probability is reached at rate constraints of around 8 and 20 b/s/Hz for
4.7. Simulation Results and Discussion

Figure 4.3: Average achieved SU rate with respect to PU rate constraint. The SU has $M_K = N_K$ antennas and the PUs have $M_k = 2, N_k = 4, \forall k \in P_p$ and both SU and PUs have power constraints of 30 dB.

transmit power constraints of 10 and 30 dB, respectively. In all results the transmit power is normalized by the receiver noise variance, $\sigma_v^2$.

**Experiment 2**

Fig. 4.3 depicts the average single SU rate with respect to the PU’s rate constraint for three different SU antenna configurations. For comparison the OIA scheme [12] with $M_K = N_K = 3$ antennas has also been included. As expected, the average SU rate decreases with increasing PU rate constraint. This trend can be traced back to the main PU trait. At lower rate constraints the PU can satisfy its rate constraint with the use of a single stream. This allows the PU receive filters to ZF a larger receive subspace giving more leeway to the SU to use a larger number of transmit antennas and more freedom to choose the best antenna combination to maximize its own rate. At higher PU rate constraints, although the PUs’ receivers each have four antennas to receive a maximum of two streams (since each PU transmitter has 2 antennas) the PU receivers still have to ZF the interference from the other PU and hence leaves no room for the SU.

The OIA scheme is included as an alternative for the case when the SU has three transmit antennas. Two remarks are necessary, however. OIA, originally used in a scenario where the PU applies a rate maximization method can be applied to this scenario without amendment and in fact works with multiple PUs if its receivers use
4.7. Simulation Results and Discussion

Figure 4.4: Average PU transmit power under same conditions as in Fig. 4.3. Note that the OIA algorithm does not affect PU system and hence the PU transmit power is equivalent to a system with no SU.

ZF. Also, OIA functions in cases where the PUs are oblivious to the SUs. It is still noteworthy, however, that when the SU transmitter only has one or two transmit antennas (in the given scenario) the antenna selection algorithm and cooperative PUs can still achieve a positive average rate while OIA remains infeasible. In addition it can be noted that in the case where the SU has three transmit antennas and OIA may be applied, there is still a substantial SU rate benefit with the antenna selection algorithm.

The acceptance of a SU into the system does impose a cost to the PUs. In Fig. 4.4 the average transmit power is clearly seen to be increasing with PU rate constraint. The PU transmit power with a SU applying OIA is equivalent to a system in absence of a SU, and hence depicts the minimum average required transmit power for a given rate constraint (since the PUs apply margin adaptation). The difference in power between OIA and the antenna selection algorithms with 1, 2 or 3 transmit antennas can be referred to as the power or margin sacrifice carried by the PU for allowing the presence of a SU.

It is interesting to note that for PU rate constraints smaller than around 8 bits/s/Hz the average PU margin sacrifice is virtually the same whether the SU has 1, 2 or 3 transmit antennas. In this region, for the majority of channel realizations, the SU only uses a single transmit antenna. On average the PU receivers reject one additional dimension, which translates into a fixed average margin sacrifice. Above around 8 bits/s/Hz, however, it can be seen that the 1 and 2 transmit antenna configurations impose a smaller margin sacrifice than $M_K = 3$, simply because they remain idle for a
larger proportion of the channel realizations. At 12 bits/s/Hz the SU seldom accesses the channel, and hence the margin sacrifice approaches zero.

**Experiment 3**

The effect of PU power constraint on the SU average achieved rate is illustrated in Fig. 4.5. At lower rate constraints the PU is more likely to be able to accommodate an additional interfering signal from the SU. In general, the application of the receive ZF filter incurs a signal power loss at the PU receiver which in this case is compensated by an increase in transmit power. If the PU power constraint is reached, but the rate constraint is not satisfied, the SU has no choice but to reduce its number of transmit antennas or choose a less favourable (from its own rate perspective) antenna combination. At lower rate constraints it can be seen that the PU power constraint remains largely inactive and the SU average rates are not limited by it. On the other hand at higher PU rate constraints, the higher PU power constraints allow the PU receivers to ZF the interference to a larger extent. Note here that for power constraints of 24 and 30 dB the 1% idle probability is reached at around 9 and 12 bits/s/Hz, respectively. This implies that above these values the SU opportunistically uses the channel in a time interweave fashion. This is clearly seen in the increase of the average SU rates at higher PU rate constraints.

![Figure 4.5: The effect on average SU achievable rate with varying the PUs power constraint. The SU has \( M_K = N_K = 2 \) antennas and its power constraint is set to 30 dB.](image-url)
4.7. Simulation Results and Discussion

Figure 4.6: Multi-user IWF method. SU rate and PU transmit power as a function of PU rate constraint. Plots (a) and (c) are with power constraint 10 dB for all users, and plots (b) and (d) are with 30 dB power constraint.

**Experiment 4**

The IWF procedure is demonstrated in Fig. 4.6. Due fundamentally to the linear receiver type in operation, the number of receive antennas has been increased to incorporate a larger number of transmit antennas. For all PUs, \((N_k, M_k) = (8, 2), k \in \mathcal{P}_p\), and for all SUs, \((N_k, M_k) = (8, 8), k \in \mathcal{P}_s\). The scenario is comprised of 2 PUs and 3 SUs, i.e., 5 transceivers in total. Note that with an increased number of transmit and receive antennas, the rate is of course also presumed to increase. Therefore, two transmit power levels have been plotted: 30 dB has been included as a direct comparison with the smaller single SU antenna case previously explored, and 10 dB as a more reasonable level given the gain of the additional antennas.

On the same plot, Fig. 4.6, four different antenna selection procedures have also been included. The label “exhaustive” refers to the exhaustive search, with an exponential
computational complexity. The greedy rate selection method is marked as “greedy”, and the correlation based selection is labelled “correlation”. A fourth procedure, called “random” in the plots, has been included as a baseline. For this procedure a SU transmit antenna is selected at random (with equal probability). Additional antenna elements are added until the sum rate is no longer increased or the PUs constraints are no longer met. In effect, this procedure has roughly the same computational complexity as the correlation based selection, but since the antenna elements are chosen in a random fashion with no regards to direct or interfering channels it acts as a marker for the efficacy of the other procedures.

As can be seen in Figs. 4.6 (a) and (b), the antenna selection procedures remain ordered with respect to the SU rate over the entire range of PU rate constraints and the two widely different transmit power levels. The random antenna selection method indeed represents a floor on the performance, with the orthogonal, greedy and exhaustive antenna selection methods giving improved SU rate at a cost in computational complexity. Note that only one of the three SUs’ rate is plotted, since all SUs are symmetrical. Also note that due to the prohibitive cost associated to simulating the exhaustive antenna selection algorithm the exhaustive results have been sampled on a coarser scale, and simulations have been halted for PU rates greater than 7.5 and 20 b/s/Hz for the 10 dB and 30 dB transmit power plots, respectively. For rates higher than the above, a non-negligent number of Monte-Carlo trials result in a situation with no active PUs, and hence the SUs may use any combination of the $3 \times 8$ transmit antennas, leading to exorbitant computation times.

The Figs. 4.6 (a) to (d) exhibit three general regions. For 10 dB transmit power constraint ((a) and (c)), the regions may be delineated as smaller than 4 b/s/Hz PU rate constraint, between 4 b/s/Hz to 8 b/s/Hz, and greater than 8 b/s/Hz. In the first region the PUs can generally meet their rate constraints with only one transmit stream each and accommodate 5 (i.e., $L_{\text{max}}$) SU transmit streams. In the second region, the PUs must increasingly employ 2 transmit antennas each in order to reach their rate constraints within their allowed power budget. Thus fewer interference paths may be cancelled at the PU receivers, leaving the SUs with a smaller number of streams. In the third region, the PU active set is regularly reduced (i.e., the PU cannot reach its rate constraint). This has two consequences, the remaining PUs (if any) may employ more of its receive antennas to null interference from the SU transmitters, and since there are fewer PU transmitters, the total interference power at the SU receivers is reduced. At the very extreme PU rate, it can be seen from the PU transmit power, Fig. 4.6 (c), than the SUs have the channels all to themselves.

The same three regions may be observed for the 30 dB transmit power constraint in Figs. 4.6 (b) and (d). Here a switch to 2 streams per PU occurs around 10 b/s/Hz, and
the reduction in the active PU set happens from around 20 b/s/Hz. The PU transmit power, Fig. 4.6 (d), exhibits an interesting trough at around 12 b/s/Hz. This may be explained by the fact that the PUs can no longer sustain their rate constraints with only one transmit stream, but must switch to using two streams. This results in a transmit power saving for the given rate constraint.

**Experiment 5**

Figs. 4.7 and 4.8 summarize the performance of the sequential admission method in Chapter 4.5.2. As in Experiment 4, each PU is equipped with 2 transmit and 8 receive antennas, while the SU has 8 antennas at each end. For sequential admission, it is assumed that with a new SU arrival, the existing SU effectively becomes part of the PU group. For this reason the simulations have included 1 to 4 PUs and one new SU arrival.

The four different antenna selection algorithms have also been employed in Figs. 4.7 and 4.8. Unsurprisingly, the order (in terms of SU rate) of the selection methods remain the same as in Experiment 4. The results of the different antenna selection methods also emphasize another aspect: at the simulated transmit powers and the various PU numbers, the difference between the exhaustive search and the random antenna selection scheme is most pronounced over the transition regions of the system. The transition regions can for instance be seen in Fig. 4.8 at around 11 and 22 b/s/Hz for the $K_p = 1$ case. The first transition is mainly due to the PU switching from one stream to two. The second transition is primarily due to the SU being barred from the channel, ie the PU needs all of its receive antennas. The final increase in SU rate is a result of the PU being incapable of reaching its target rate and remaining silent. It is hypothesised that it is during these transition regions that it is even more critical to choose SU transmit antennas that suit the PUs. This is in effect what the exhaustive and greedy methods do better than the correlation and random based techniques, since these methods iterate over many more SU transmit antenna subsets.

As expected, as the number of PUs increases, the SU rate decreases. This is a direct consequence of a smaller receive space at the PUs to handle SU interference (i.e., SU limited to fewer transmit antennas), and higher levels of interference from the PUs to the SU receivers. Note that for $K_p = 4$, even with a 30 dB transmit power constraint, there exists a region where the PUs cannot sustain any SU interference. The general shape of the plots matches the IWF results in Fig. 4.6, but since the IWF scenario contains 2 PUs and 3 SUs, the individual rates of the SUs are much lower. Note that if the average sum rate is considered for the IWF method (i.e., multiplying the SU rates in Fig. 4.6 by 3) it is seen that these sum rates are generally above the $K_p = 2$ results.
4.7. Simulation Results and Discussion

Figure 4.7: Multi-user sequential admission method. SU rate (a) and PU transmit power (b) as a function of PU rate constraint, for $P_k = 10$ dB, $\forall k \in \mathcal{P}_p \cup \mathcal{P}_s$. 
4.7. Simulation Results and Discussion

Figure 4.8: Multi-user sequential admission method. SU rate (a) and PU transmit power (b) as a function of PU rate constraint, for $P_k = 30$ dB, $\forall k \in \mathcal{P}_p \cup \mathcal{P}_s$. 
in Figs. 4.7 and 4.8. It must be emphasized that the improved rates are not directly due to the 3 SUs competing for the same frequency-time slot through IWF. Rather it is due to the fact that the antenna selection methods tend to select the best SU (from a rate perspective). The improvement in spectral efficiency is hence a user selection gain.

4.8 Conclusion

In this chapter a cooperative interweave spatial-domain spectrum sharing scheme was developed. Specifically, cooperation between the tiers enabled opportunistic transmission, at a controlled cost to the prioritized system. In similarity with previous works (e.g., [125, 12, 11]) all results are based on the assumption that the SU has perfect global CSI. The achieved performance of the proposed scheme must therefore be seen as an upper bound, ignoring not only channel estimation errors but also the potential overhead of communicating or distributing CSI.

One of the fundamental issues with hierarchical sharing is how to protect the incumbent users effectively. The presented scheme breaks with the traditional view of an unaware PU and by fostering active cooperation the PU is better protected against SU interference and can control SU activity. With the use of multiple antennas the PU is better equipped to handle any additional interference, and the SU is given more freedom in how to design its pre-processing.

A transmit antenna selection scheme was used to increase the probability of matching the interference space with a null-space at the PU receiver. The simulation results show an appreciable gain for the opportunistic user, and perhaps more importantly gives weight to the idea that interweave spatial-domain spectrum sharing can be feasible even in scenarios where the SU transmitter cannot freely place its interference. In addition to the single SU scenario, two multiple SU techniques are explored. One method is based on IWF, whereas the second method instead considers sequential admission. Both techniques are explored with three different antenna selection schemes. The first is an exhaustive search, where the size of the antenna subset has been limited. The second procedure is based on a greedy selection, by incrementally choosing the antenna element that results in the largest rate increase. The third method, designed to lighten the computational burden, is based on selecting a subset of nearly orthogonal transmit antennas.

The improvement in spectrum utilization must be weighed against the transmit power sacrifice incurred by the PUs. Although mobile transmitters may not favour this type of spectrum sharing for the increased drain on batteries, mobile operators may find the
solution alluring simply due to the control given in allowing SU access. In addition, if the primary and secondary entities are not strictly users, but perhaps services on the same network, such as delay sensitive voice versus delay tolerant file transfers, the presented spectrum sharing scheme could be employed without modification.
Conclusion and Future Work

5.1 Conclusion

The focus of this thesis has been on the use of the spatial dimension as an enabler for spectrum sharing applications. To set the scene and gain an understanding of how the rate of a system can scale with the minimum number of transmit and receive antennas without increasing the bandwidth requirement or the total transmit power the MIMO channel was reviewed from an information theoretic perspective. This multiple-antenna framework extends naturally to both the MAC and BC models by considering the separate handling of the information streams at either the transmitting or the receiving side. However, in contrast to the case of the MIMO model, where the channel can be decomposed into a set of orthogonal sub-channels with no penalty on the capacity, the case where either the transmitters or receivers act in isolation significantly changes how the interference between the streams or users is handled. For the MAC this is a multi-user detection problem and for the BC it can be tackled using a pre-cancellation technique (i.e., dirty paper coding). For the case of the IC in general, the situation is still unclear, and it must be emphasized that the search for practical algorithms for all three multiple antenna channel types, MAC, BC, and IC, with reasonable computational complexity is still an ongoing research topic.

In an effort to exploit the difference in users’ spatial signatures two different scenarios were studied in this thesis, one is non-cooperative while the other requires cooperation. In the non-cooperative scenario a large MU-MIMO system shares the spectrum with a small MU-MIMO. Although it has been shown using game theory that such an asymmetric system may lead to an operating point with very poor fairness, by instead invoking the attractive traits of the massive MIMO system, namely a robustness to interference and low transmit power, it is demonstrated that such an interference scenario
5.1. Conclusion

can allow fair coexistence if scaled correctly. The competitive multiple antenna policy is well suited to the spectrum sharing form known as general authorization. A concrete example could be the sharing of an unlicensed band between Wi-Fi and a mobile network operator, as has been proposed for LTE unlicensed. The use of an access point equipped with a large number of service antennas may in this way be used to handle the interference from the Wi-Fi network and simultaneously allow a lower transmit power for its own network and hence reduce interference to others. Two previous issues with the majority of spectrum sharing techniques is the fact that since interference is treated as noise its impact on system performance is significant, and that the requirement of some form of cooperation incurs an additional overhead. With the proposed policy, the scenario was framed as a common mean SINR maximization problem where the mean SINR and transmit power are characterized. The study shows that even with no cooperation, by adjusting the number of service antennas and system loads, the interference at both systems can be managed. As a case in point, in the uplink, reducing the number of UTs from three to two at an access point with four service antennas, allows a massive MIMO system with 100 service antennas to increase the load from five to nine UTs, keeping the maximum normalized power constraint to 10 dB. If the massive MIMO access point is instead equipped with 1000 antennas, about 40 additional UTs may be admitted at the same operating point.

In addition to the characterization of the SINRs at the conventional and massive MIMO systems, concise power scaling laws were derived in order to gain a more intuitive understanding of the behaviour of the interference system. These laws can be used to estimate the number of service antennas and system loads required for a specific SINR operating point. The behaviour of the interference system may also be gauged against an isolated massive MIMO network. For instance, in the uplink, assuming a ZF or MF processing at the massive MIMO access point, the resulting common SINR is a function of the square root of the number of service antennas, as opposed to a linear function for the isolated system. The weakness of the MF processing compared to ZF in the face of strong interference was also noticeable, and mirrors the near-far issue observed in CDMA systems. It must be stated that the SINR characterisation and scaling laws are based on a Rayleigh fading assumption and equal large-scale fading coefficients. It remains to be seen if the assumption applied to permit analysis holds in practical scenarios. The Rayleigh fading assumption has been shown to produce representable results [91]. However the near-equal large scale fading assumption may be too simplistic, although it strongly depends on for instance scheduling mechanism, and power control to mention a few additional factors. The method of using a large number of antennas to coexist with a conventional MIMO system does still seem to be a viable policy where the system capacity may be scaled according to the number of service antennas, load and permitted transmit power.
5.1. Conclusion

For the cooperative scenario, the use of multiple antennas at both the PU and SU is explored to allow spectrum access. Here, due to the limited number of antennas at the secondary transmitter, it is demonstrated how cooperation may be used to increase utilization. The investigation focused on a setting where the primary system has additional capabilities (in terms of additional receive antennas) and can use a linear receiver structure to mitigate or cancel interference from the SU. A SU transmit antenna selection scheme was used to increase the probability of matching the interference space with a null-space at the PU receiver. Crucially, it is the PU that allows the SU to access the spectrum and handles the interference. This is a break from the conventional view of hierarchical access, where the PU is unaware of the existence of other systems, but may be more realistic. The policy requires active cooperation from the PU and global channel state information. However it does highlight one method to give the PU more control as to how and when a SU may share the channel. The ability for the PU to have some control over external interference is deemed a useful property, and may be one way to reassure primary spectrum holders and enable spectrum sharing in the first instance.

In its most general form, the studied scenario forms a classical MIMO interference channel. The overall objective is to maximize the (sum)rate of the secondary user(s), subject to a secondary user power constraint and individual primary user rate constraints. Due to the application of a secondary transmit antenna scheme, the optimal solution involves iterating over all possible antenna combinations. For a small number of transmit antennas this is not a concern. However, for larger systems, the exponential increase in computational complexity makes the optimal solution impractical. Hence, two additional algorithms are proposed based on greedy selection and an orthogonality criterion, that show good performance at high SINRs (> 10 dB). For situations where more than one SU is present, two heuristic approaches have been studied to maximize SU rates: the first method is based on IWF and the second method is designed around sequential admission. In all cases, an improvement in spectrum utilization was observed. However the gain in spectrum utilization comes at a cost or sacrifice in terms of transmit power for the PUs. Since the PU receivers modify their receive filters to accommodate the additional SUs, the net result is that the PU transmitters must increase their transmit power. The specific transmit power sacrifice of course depends on the number of nodes in the system, the number of available antennas and the desired operating point (PU desired rate). It is demonstrated that PU cooperation enables SU communication in situations where transmit zero-forcing beamforming or opportunistic interference alignment remain infeasible. The simulation results show an appreciable gain for the opportunistic user, and perhaps more importantly gives weight to the idea that interweave spatial-domain spectrum sharing can be feasible even in scenarios where the SU transmitter cannot freely place its interference.
5.2 Future Work

The field of spectrum sharing is full of interesting further directions worth of investigation. In general, due to the difficulty in dealing with the interference caused by multiple users sharing the same spectrum, precious few results are available that offer decisive conclusions; the results depend heavily on the chosen network scenario and parameter settings. This reason alone motivates further research into how spectrum sharing can be enabled all the way from a regulatory perspective, down to how it is handled on the physical layer. More closely related to the present thesis, however are the following possible paths:

- Specifically dealing with precoding and decoding vectors or matrices, CSI is of crucial importance. Throughout this thesis perfect CSI has been assumed. This assumption, however common, does not hold in practice where the cost in estimation (e.g. in terms of length of training sequence) leads to a compromise in the quality of estimation, and the time varying aspect of the channel (related to the coherence time) may introduce the issue of out-dated CSI. There may also be a difference in the quality of estimation, depending on if the channel is from one’s own or interfering transmitter. The design of robust precoding and decoding techniques that take the estimation errors into account would therefore be an interesting extension.

- The scenarios considered have been limited to the case where all links are subject to the same large scale fading components. This adequately models a scenario where distances between nodes are on the same order of magnitude. It is clear however, that with single-user detection type receivers that treat interference as noise, the individual channel gains play a fundamental role. In this thesis, the large scale fading components were normalized to unity in order to better expose the potential of multiple antennas. However, if the difference in channel gains may also be leveraged, it could have a significant impact on system level performance. One interesting avenue, therefore, is to attempt to frame the spectrum sharing problem using methods from stochastic geometry [203], where the position of the different nodes follows some random distribution.

- The completely distributed, or non-cooperative, solution proposed in Chapter 3 certainly has very attractive features. No communication exchange between the systems is necessary for operation. However, as emphasized in Chapter 4, if systems are allowed to cooperate then spectrum utilization may be increased where it would otherwise not be possible. Cooperation however, is not a binary choice (full-cooperation versus strict competition), it is rather a sliding scale with
a range of trade-offs. An example of such a classification was given by Gesbert
et al. [204] for the multi-cell cooperation scenario. It would be interesting to
investigate whether there are scopes (e.g. limiting to linear transmit and receive
techniques, or specific fading scenarios) that enable a framework or an overview
of the costs involved with different levels or forms of cooperation with respect to
some performance measure.
A.1 Distribution of MF Interference Term

The following facts are used in the characterization of the random variable in (3.13):

**Definition. 1** [205, Definition 2.7] A random matrix $H^{M \times N}$ is called bi-unitarily invariant if the joint distribution of its entries equals that of $UHV^H$ for any unitary matrices $U$ and $V$ independent of $H$.

**Lemma. 1** [205, Lemma 2.1] Let $h \in \mathbb{C}^{M \times 1}$ be a vector with i.i.d. zero-mean complex Gaussian entries, and denote its QR-decomposition by $h = QR$. Then, the unitary matrix $Q \in \mathbb{C}^{M \times M}$ is independent of the vector $R \in \mathbb{C}^{M \times 1}$.

Consider first a vector $h \in \mathbb{C}^{M \times 1}$ with i.i.d. zero-mean complex Gaussian entries. Then

$$
\frac{h}{\sqrt{h^H h}} = \frac{QR}{\sqrt{h^H h}} = \frac{Qe_1}{\sqrt{h^H h}} = Qe_1,
$$

(A.1)

where $h$ was decomposed into a unitary matrix $Q \in \mathbb{C}^{M \times M}$ and the vector $R \in \mathbb{C}^{M \times 1}$ that in this case contains the vector magnitude. $e_1 = [1 \ 0 \ \ldots \ 0]^T$ is the first column of the $M$ by $M$ identity matrix $I$ (note that column choice is arbitrary).

Given a matrix $H \in \mathbb{C}^{M \times N}$ with i.i.d. zero-mean complex Gaussian entries with variance $\sigma^2$ and the above decomposition in (A.1), it can then be shown that

$$
\frac{h^H HH^H h}{h^H h} = e_1^HQ^H HH^HQe_1 = [A A^H]_{11},
$$

(A.2)

where $A = Q^H H$ is, according to Definition. 1, distributed as $H$. Note that since $M \geq N$, $AA^H$ may be less than full rank. These random matrices cannot be characterized.
as Wishart matrices (extensions to the singular case do exist, however, see e.g., [206]). For the present case, it is possible to see (simply through the mechanics of matrix multiplication) that the \(i\)th diagonal element, \([\mathbf{A} \mathbf{A}^H]_{ii}\), is distributed as \(G(N, \sigma^2)\). A slightly different derivation is also given by Larsson and Jorswieck [107], but by explicitly applying the QR-decomposition in (A.1) and referring to Lemma. 1, it is also possible to state that the random variable \([\mathbf{A} \mathbf{A}^H]_{ii}\) is independent of \(\mathbf{h}^H \mathbf{h}\).

### A.2 ZF SINR Distribution

A derivation of the ZF SINR distribution using block matrix inversion and properties of the idempotent matrix (contrast to the work by Gore at al. [68]). Let \(\mathbf{H}_{nj} \in \mathbb{C}^{M \times N}\) be a tall or square matrix \((M \geq N)\) with i.i.d. zero-mean complex Gaussian entries. First note that

\[
\mathbf{H}_{nj}^H \mathbf{H}_{nj} = \begin{bmatrix}
\mathbf{h}^H \mathbf{h} & \mathbf{h}^H \mathbf{H} \\
\mathbf{H}^H \mathbf{h} & \mathbf{H}^H \mathbf{H}
\end{bmatrix},
\]

where \(\mathbf{H}_{nj}\) has been partitioned as \(\begin{bmatrix} \mathbf{h} & \mathbf{H} \end{bmatrix}\), with \(\mathbf{h} \in \mathbb{C}^{M \times 1}\). Using this partitioning and applying the block inverse

\[
(\mathbf{H}_{nj}^H \mathbf{H}_{nj})^{-1} = \begin{bmatrix}
\mathbf{Y}_{11} & \mathbf{Y}_{12} \\
\mathbf{Y}_{21} & \mathbf{Y}_{22}
\end{bmatrix},
\]

with

\[
\mathbf{Y}_{11} = (\mathbf{h}^H \left( \mathbf{I} - \mathbf{H} (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \right) \mathbf{h})^{-1}.
\]

The remaining elements, \(\mathbf{Y}_{12}, \mathbf{Y}_{21},\) and \(\mathbf{Y}_{22}\), are not specified since only a diagonal element is required for the derivation. Hence the ZF SINR of the first UT is

\[
\gamma = \frac{1}{\left( (\mathbf{H}^H \mathbf{H})^{-1} \right)_{11}} = \mathbf{h}^H \left( \mathbf{I} - \mathbf{H} (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \right) \mathbf{h}.
\]

Now note that \(\mathbf{I} - \mathbf{H} (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H\) is an idempotent matrix. This implies that its eigenvalue decomposition \(\mathbf{Q} \mathbf{D} \mathbf{Q}^H\) consists of the usual unitary matrix \(\mathbf{Q} \in \mathbb{C}^{M \times M}\) and the eigenvalues in \(\mathbf{D}\) are either 1 or 0, with \(\text{tr}(\mathbf{D}) = M - (N - 1)\). Applying Definition. 1 it can be seen that \(\mathbf{a} = \mathbf{Q}^H \mathbf{h}\) possesses the same distribution as \(\mathbf{h}\). In addition, the action of the binary diagonal matrix \(\mathbf{D}\) is to select \(M - N + 1\) entries from the vector \(\mathbf{a}\). Therefore denoting \(\mathbf{b} \in \mathbb{C}^{(M-N+1) \times 1}\) with i.i.d. zero-mean complex Gaussian entries with variance \(\sigma^2\), the ZF SINR is distributed as

\[
\gamma \sim \mathbf{b}^H \mathbf{b}.
\]

The squared euclidean norm of an \(M - N + 1\) vector with complex Gaussian entries of course follows a Gamma distribution, \(G((M - N + 1), \sigma^2)\). This completes the proof.
A.3 ZF Distribution with Interferer

To show that (3.31) follows a scaled $\mathcal{F}$-distribution, the matrix block inverse and the bi-unitarily invariance property will again be applied.

First, recognize that (3.31) can be written as

$$
\begin{bmatrix}
W_{zf}^2 H_{21} (W_{zf}^2 H_{21})^H
\end{bmatrix}_{11} = w^H H_{21} H_{21}^H w,
$$

(A.9)

where $w^H$ is the first row of the pseudo-inverse matrix $W_{zf}^2$. Second, the first row of the pseudo-inverse matrix can be written as

$$
w^H = e_1^H (H_{22} H_{22})^{-1} H_{22}^H
$$

(A.10)

$$
= \begin{bmatrix} \Upsilon_{11} & \Upsilon_{12} \end{bmatrix} \begin{bmatrix} h^H \\ H^H \end{bmatrix}
$$

(A.11)

$$
= \frac{h^H (I - H (H^H H)^{-1} H^H)}{h^H (I - H (H^H H)^{-1} H^H)} h.
$$

(A.12)

In (A.11), the same matrix partition scheme as in (A.3) was used, $\Upsilon_{11}$ is defined as (A.5), and

$$
\Upsilon_{12} = - \Upsilon_{11} h^H H (H^H H)^{-1} H^H.
$$

(A.13)

The idempotent matrix $P = I - H (H^H H)^{-1} H^H$, can again be decomposed as $P = QDQ^H$, where the diagonal matrix $D$ is populated by either 1 or 0 values, summing to $M_2 - N_2 + 1$ in this case. Hence leaning on the bi-unitary invariance property (Definition 1), and the selection action of the diagonal matrix $D$, a row of the pseudo-inverse matrix can be said to be distributed as

$$
w^H \sim a^H Q^H a
$$

(A.14)

where $a \in \mathbb{C}^{(M_2 - N_2 + 1) \times 1}$ is a vector with i.i.d. complex Gaussian elements. Substituting (A.14) into (A.9) one immediately sees the similarity with (A.2) (the difference is that the denominator is squared). Hence following the same method as in Appendix A.1, it is possible to show that

$$
w^H H_{21} H_{21}^H w \sim b^H b
$$

(A.15)

where $b \in \mathbb{C}^{N_1 \times 1}$ is a vector with i.i.d. zero-mean complex Gaussian entries. The form of the random variable in (A.15), i.e., a ratio of independent Gamma distributed random variables can be recognized as a scaled $\mathcal{F}$-distributed random variable.
A.4 Solution Model to Problem (3.23)

The problem (3.23) is solved by recognizing that:

1. Given a transmit power $p_n, n \in \{1, 2\}$, then a unique $p_j, j = \{1, 2\} \setminus n$ can be solved for (either in closed form or using a root finding method).

2. The optimal solution, $\gamma^*$, at $\{p_1^*, p_2^*\}$ to problem (3.23) is reached when either (or both) $p_1^* = p_{\text{max}}$ or (and) $p_2^* = p_{\text{max}}$, for all the receiver equalizers and precoding strategies considered.

Point 1 can be seen by observing that $\gamma_n, n \in \{1, 2\}$ is strictly increasing with $p_n \geq 0$ and $\gamma_n = 0$ at $p_n = 0$ for all equalizer and preprocessing filters considered strategies. In addition, $\gamma_j, j \in \{1, 2\} \setminus n$ is either a strictly decreasing function, or independent ($\gamma_1$ for ZF in the uplink and $\gamma_2$ for ZF in the downlink) of $p_n$ and $\gamma_j > 0$ at $p_n = 0$ for all $p_j > 0$. These conditions ensures a unique intersection.

Point 2 can be shown by following the same argumentation as, e.g., Schubert and Boche [170, lemma. 1]. Given an operating point, $\gamma_1 = \rho \gamma_2$ at a specific transmit power combination $\{p_1 < p_{\text{max}}, p_2 < p_{\text{max}}\}$ then increasing both transmit powers by a factor $\alpha > 1$ results in an increase in both systems’ mean SINRs. Now, consider the case where the proportional increase in transmit powers causes $\gamma_1 > \rho \gamma_2$. In this case $p_1$ can be decreased to lower $\gamma_1$ and simultaneously increase $\gamma_2$ (except for the ZF downlink scenario where $\gamma_2$ is independent of $p_1$). The same argument can be used for the case $\gamma_1 < \rho \gamma_2$. The mean SINRs at the new equilibrium point are greater than the previous values. Hence, the mean SINR can always be increased if neither $p_1 = p_{\text{max}}$ or $p_2 = p_{\text{max}}$ has been reached.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


