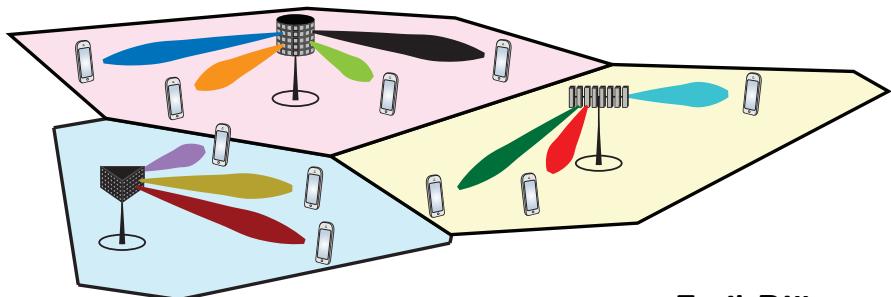


Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency



Emil Björnson
Linköping University
emil.bjornson@liu.se

Jakob Hoydis
Bell Labs, Nokia
jakob.hoydis@nokia.com

Luca Sanguinetti
University of Pisa
luca.sanguinetti@unipi.it

© 2017 E. Björnson, J. Hoydis and L. Sanguinetti

Version of record: Emil Björnson, Jakob Hoydis and Luca Sanguinetti (2017), "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency", Foundations and Trends® in Signal Processing: Vol. 11, No. 3-4, pp 154–655. DOI: 10.1561/2000000093.

Simulation code and supplementary material: <https://massivemimobook.com>

Printed books: Available from now Publishers Inc., <http://www.nowpublishers.com>

This is the authors' version of the manuscript. See the above version of record for the final published manuscript. Date of this version: October 12, 2018.

“Massive MIMO is an essential topic in the field of future cellular networks. I have not seen any other book which can compete at that level of detail and scientific rigor. I liked the didactic style, coming back to root definitions (cellular networks, spectral efficiency, channel models, and so forth) which will be very useful to PhD students and others starting in this area. The models are very well explained and justified as opposed to being imposed out of nowhere. This makes the reading particularly pleasant and rich. Overall, a great tool to researchers and practitioners in the field.”

David Gesbert, EURECOM

“This book provides a modern presentation of the state-of-the-art for Massive MIMO communication. It includes a comprehensive treatment of mathematical tools for analyzing and understanding Massive MIMO networks. The authors provide an enlightening introduction to the topic, suitable for graduate students and professors alike. The book starts with the basic definitions and culminates in a systematic treatment of spectral and energy efficiency. Of particular interest, the book provides an updated assessment of the performance limiting factors, showing for example that pilot contamination is not a fundamental limitation.”

Robert W. Heath Jr., The University of Texas at Austin

Contents

1	Introduction and Motivation	158
1.1	Cellular Networks	160
1.2	Definition of Spectral Efficiency	167
1.3	Ways to Improve the Spectral Efficiency	173
1.4	Summary of Key Points in Section 1	214
2	Massive MIMO Networks	216
2.1	Definition of Massive MIMO	216
2.2	Correlated Rayleigh Fading	222
2.3	System Model for Uplink and Downlink	226
2.4	Basic Impact of Spatial Channel Correlation	228
2.5	Channel Hardening and Favorable Propagation	231
2.6	Local Scattering Spatial Correlation Model	235
2.7	Summary of Key Points in Section 2	243
3	Channel Estimation	244
3.1	Uplink Pilot Transmission	244
3.2	MMSE Channel Estimation	248
3.3	Impact of Spatial Correlation and Pilot Contamination	254
3.4	Computational Complexity and Low-Complexity Estimators	264
3.5	Data-Aided Channel Estimation and Pilot Decontamination	271

3.6	Summary of Key Points in Section 3	274
4	Spectral Efficiency	275
4.1	Uplink Spectral Efficiency and Receive Combining	275
4.2	Alternative UL SE Expressions and Key Properties	301
4.3	Downlink Spectral Efficiency and Transmit Precoding	316
4.4	Asymptotic Analysis	335
4.5	Summary of Key Points in Section 4	351
5	Energy Efficiency	353
5.1	Motivation	354
5.2	Transmit Power Consumption	357
5.3	Definition of Energy Efficiency	362
5.4	Circuit Power Consumption Model	375
5.5	Tradeoff Between Energy Efficiency and Throughput	390
5.6	Network Design for Maximal Energy Efficiency	395
5.7	Summary of Key Points in Section 5	401
6	Hardware Efficiency	403
6.1	Transceiver Hardware Impairments	404
6.2	Channel Estimation with Hardware Impairments	413
6.3	Spectral Efficiency with Hardware Impairments	419
6.4	Hardware-Quality Scaling Law	439
6.5	Summary of Key Points in Section 6	449
7	Practical Deployment Considerations	451
7.1	Power Allocation	452
7.2	Spatial Resource Allocation	468
7.3	Channel Modeling	482
7.4	Array Deployment	500
7.5	Millimeter Wavelength Communications	522
7.6	Heterogeneous Networks	527
7.7	Case Study	537
7.8	Summary of Key Points in Section 7	546
	Acknowledgements	548

Appendices	549
A Notation and Abbreviations	550
B Standard Results	558
B.1 Matrix Analysis	558
B.2 Random Vectors and Matrices	563
B.3 Properties of the Lambert W Function	567
B.4 Basic Estimation Theory	567
B.5 Basic Information Theory	572
B.6 Basic Optimization Theory	575
C Collection of Proofs	579
C.1 Proofs in Section 1	579
C.2 Proofs in Section 3	591
C.3 Proofs in Section 4	593
C.4 Proofs in Section 5	609
C.5 Proofs in Section 6	612
References	621

Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency

Emil Björnson¹, Jakob Hoydis² and Luca Sanguinetti³

¹*Linköping University; emil.bjornson@liu.se*

²*Bell Labs, Nokia; jakob.hoydis@nokia.com*

³*University of Pisa; luca.sanguinetti@unipi.it*

ABSTRACT

Massive multiple-input multiple-output (MIMO) is one of the most promising technologies for the next generation of wireless communication networks because it has the potential to provide game-changing improvements in spectral efficiency (SE) and energy efficiency (EE). This monograph summarizes many years of research insights in a clear and self-contained way and provides the reader with the necessary knowledge and mathematical tools to carry out independent research in this area. Starting from a rigorous definition of Massive MIMO, the monograph covers the important aspects of channel estimation, SE, EE, hardware efficiency (HE), and various practical deployment considerations.

From the beginning, a very general, yet tractable, canonical system model with spatial channel correlation is introduced. This model is used to realistically assess the SE and EE, and is later extended to also include the impact of hardware impairments. Owing to this rigorous modeling approach, a lot of classic “wisdom” about Massive MIMO, based on too simplistic system models, is shown to be questionable.

The monograph contains many numerical examples, which can be reproduced using Matlab code that is available online at https://dx.doi.org/10.1561/2000000093_supp.

Preface

Why We Wrote this Monograph

Massive multiple-input multiple-output (MIMO) is currently a buzzword in the evolution of cellular networks, but there is a great divide between what different people read into it. Some say Massive MIMO was conceived by Thomas Marzetta in a seminal paper from 2010, but the terminology cannot be found in that paper. Some say it is a reincarnation of space-division multiple access (SDMA), but with more antennas than in the field-trials carried out in the 1990s. Some say that any radio technology with at least 64 antennas is Massive MIMO. In this monograph, we explain what Massive MIMO is to us and how the research conducted in the past decades lead to a scalable multiantenna technology that offers great throughput and energy efficiency under practical conditions. We decided to write this monograph to share the insights and know-how that each of us has obtained through ten years of multiuser MIMO research. Two key differences from previous books on this topic are the spatial channel correlation and the rigorous signal processing design considered herein, which uncover fundamental characteristics that are easily overlooked by using more tractable but less realistic models and processing schemes. In our effort to provide a coherent description of the topic, we cover many details that cannot be found in the research literature, but are important to connect the dots.

This monograph is substantially longer than the average monograph published in *Foundations and Trends*, but we did not choose the publisher based on the format but the quality and openness that it offers. We want to reach a broad audience by offering printed books as well as open access to an electronic version. We have made the simulation code available online, to encourage reproducibility and continued research. This monograph is targeted at graduate students, researchers, and professors who want to learn the conceptual and analytical foundations of Massive MIMO, in terms of spectral, energy, and/or hardware efficiency, as well as channel estimation and practical considerations. We also cover some related topics and recent trends, but purposely in less detail, to focus on the unchanging fundamentals and not on the things that current research is targeting. Basic linear algebra, probability theory, estimation theory, and information theory are sufficient to read this monograph. The appendices contain detailed proofs of the analytical results and, for completeness, the basic theory is also summarized.

Structure of the Monograph

Section 1 introduces the basic concepts that lay the foundation for the definition and design of Massive MIMO. Section 2 provides a rigorous definition of the Massive MIMO technology and introduces the system and channel models that are used in the remainder of the monograph. Section 3 describes the signal processing used for channel estimation on the basis of uplink (UL) pilots. Receive combining and transmit precoding are considered in Section 4 wherein expressions for the spectral efficiency (SE) achieved in the UL and downlink (DL) are derived and the key insights are described and exemplified. Section 5 shows that Massive MIMO also plays a key role when designing highly energy-efficient cellular networks. Section 6 analyzes how transceiver hardware impairments affect the SE and shows that Massive MIMO makes more efficient use of the hardware. This opens the door for using components with lower resolution (e.g., fewer quantization bits) to save energy and cost. Section 7 provides an overview of important practical aspects, such as spatial resource allocation, channel modeling, array deployment, and the role of Massive MIMO in heterogeneous networks.

How to Use this Monograph

Researchers who want to delve into the field of Massive MIMO (e.g., for the purpose of performing independent research) can basically read the monograph from cover to cover. However, we stress that Sections 5, 6, and 7 can be read in any order, based on personal preferences.

Each section ends with a summary of key points. A professor who is familiar with the broad field of MIMO can read these summaries to become acquainted with the content, and then decide what to read in detail.

A graduate-level course can cover Sections 1–4 in depth or partially. Selected parts of the remaining sections may also be included in the course, depending on the background and interest of the students. An extensive slide set and homework exercises are made available for teachers who would like to give a course based on this monograph.

The authors, October 2017

1

Introduction and Motivation

Wireless communication technology has fundamentally changed the way we communicate. The time when telephones, computers, and Internet connections were bound to be wired, and only used at predefined locations, has passed. These communications services are nowadays wirelessly accessible almost everywhere on Earth, thanks to the deployment of cellular wide area networks (e.g., based on the GSM¹, UMTS², and LTE³ standards), local area networks (based on different versions of the WiFi standard IEEE 802.11), and satellite services. Wireless connectivity has become an essential part of the society—as vital as electricity—and as such the technology itself spurs new applications and services. We have already witnessed the streaming media revolution, where music and video are delivered on demand over the Internet. The first steps towards a fully networked society with augmented reality applications, connected homes and cars, and machine-to-machine communications have also been taken. Looking 15 years into the future, we will find new innovative wireless services that we cannot predict today.

¹Global System for Mobile Communications (GSM).

²Universal Mobile Telecommunications System (UMTS).

³Long Term Evolution (LTE).

The amount of wireless voice and data communications has grown at an exponential pace for many decades. This trend is referred to as *Cooper's law* because the wireless researcher Martin Cooper [91] noticed in the 1990s that the number of voice and data connections has doubled every two-and-a-half years, since Guglielmo Marconi's first wireless transmissions in 1895. This corresponds to a 32% annual growth rate. Looking ahead, the Ericsson Mobility Report forecasts a compound annual growth rate of 42% in mobile data traffic from 2016 to 2022 [109], which is even faster than Cooper's law. The demand for wireless data connectivity will definitely continue to increase in the foreseeable future; for example, since the video fidelity is constantly growing, since new must-have services are likely to arise, and because we are moving into a networked society, where all electronic devices connect to the Internet. An important question is how to evolve the current wireless communications technologies to meet the continuously increasing demand, and thereby avoid an imminent data traffic crunch. An equally important question is how to satisfy the rising expectations of service quality. Customers will expect the wireless services to work equally well anywhere and at any time, just as they expect the electricity grid to be robust and constantly available. To keep up with an exponential traffic growth rate and simultaneously provide ubiquitous connectivity, industrial and academic researchers need to turn every stone to design new revolutionary wireless network technologies. This monograph explains what the Massive multiple-input multiple-output (MIMO) technology is and why it is a promising solution to handle several orders-of-magnitude⁴ more wireless data traffic than today's technologies.

The cellular concept for wireless communication networks is defined in Section 1.1, which also discusses how to evolve current network technology to accommodate more traffic. Section 1.2 defines the spectral efficiency (SE) notion and provides basic information-theoretic results that will serve as a foundation for later analysis. Different ways to improve the SE are compared in Section 1.3, which motivates the design of Massive MIMO. The key points are summarized in Section 1.4.

⁴In communications, a factor ten is called one order-of-magnitude, while a factor 100 stands for two orders-of-magnitude and so on.

1.1 Cellular Networks

Wireless communication is based on radio, meaning that electromagnetic (EM) waves are designed to carry information from a transmitter to one or multiple receivers. Since the EM waves propagate in all possible directions from the transmitter, the signal energy spreads out and less energy reaches a desired receiver as the distance increases. To deliver wireless services with sufficiently high received signal energy over wide coverage areas, researchers at Bell Labs postulated in 1947 that a cellular network topology is needed [277]. According to this idea, the coverage area is divided into cells that operate individually using a fixed-location base station; that is, a piece of network equipment that facilitates wireless communication between a device and the network. The cellular concept was further developed and analyzed over the subsequent decades [291, 116, 204, 364] and later deployed in practice. Without any doubt, the cellular concept was a major breakthrough and has been the main driver to deliver wireless services in the last forty years (since the “first generation” of mobile phone systems emerged in the 1980s). In this monograph, a cellular network is defined as follows.

Definition 1.1 (Cellular network). A cellular network consists of a set of base stations (BSs) and a set of user equipments (UEs).⁵ Each UE is connected to one of the BSs, which provides service to it. The downlink (DL) refers to signals sent from the BSs to their respective UEs, while the uplink (UL) refers to transmissions from the UEs to their respective BSs.⁶

While this definition specifies the setup that we will study, it does not cover every aspect of cellular networks; for example, to enable efficient handover between cells, a UE can momentarily be connected to multiple BSs.

⁵The terms BS and UE stem from GSM and LTE standards, respectively, but are used in this monograph without any reference to particular standards.

⁶In a fully cooperative cellular network, called network MIMO [126] or cell-free system [240], all BSs are connected to a central processing site and are used to jointly serve all UEs in the network. In this case, the DL (UL) refers to signals transmitted from (to) all the BSs to (from) each UE. Such a cellular network is not the focus of this monograph, but cell-free systems are briefly described in Section 7.4.3 on p. 509.

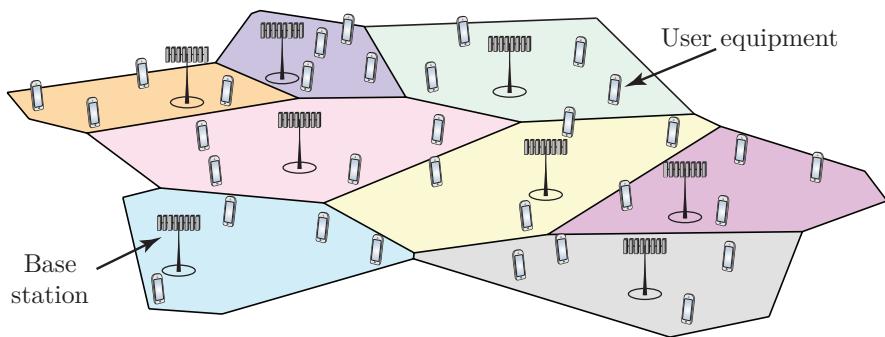


Figure 1.1: A basic cellular network, where each BS covers a distinct geographical area and provides service to all UEs in it. The area is called a “cell” and is illustrated with a distinct color. The cell may consist of all geographic locations where this BS provides the strongest DL signal.

An illustration of a cellular network is provided in Figure 1.1. This monograph focuses on the wireless communication links between BSs and UEs, while the remaining network infrastructure (e.g., fronthaul, backhaul, and core network) is assumed to function perfectly. There are several branches of wireless technologies that are currently in use, such as the IEEE 802.11 family for WiFi wireless local area networks (WLANs), the 3rd Generation Partnership Project (3GPP) family with GSM/UMTS/LTE for mobile communications [128], and the competing 3GPP2 family with IS-95/CDMA2000/EV-DO. Some standards within these families are evolutions of each other, optimized for the same use case, while others are designed for different use cases. Together they form a *heterogeneous network* consisting of two main tiers:

1. Coverage tier: Consisting of outdoor cellular BSs that provide wide-area coverage, mobility support, and are shared between many UEs;
2. Hotspot tier: Consisting of (mainly) indoor BSs that offer high throughput in small local areas to a few UEs.

The term “heterogeneous” implies that these two tiers coexist in the same area. In particular, the hotspot BSs are deployed to create *small cells (SCs)* within the coverage area of the cellular BSs, as illustrated in

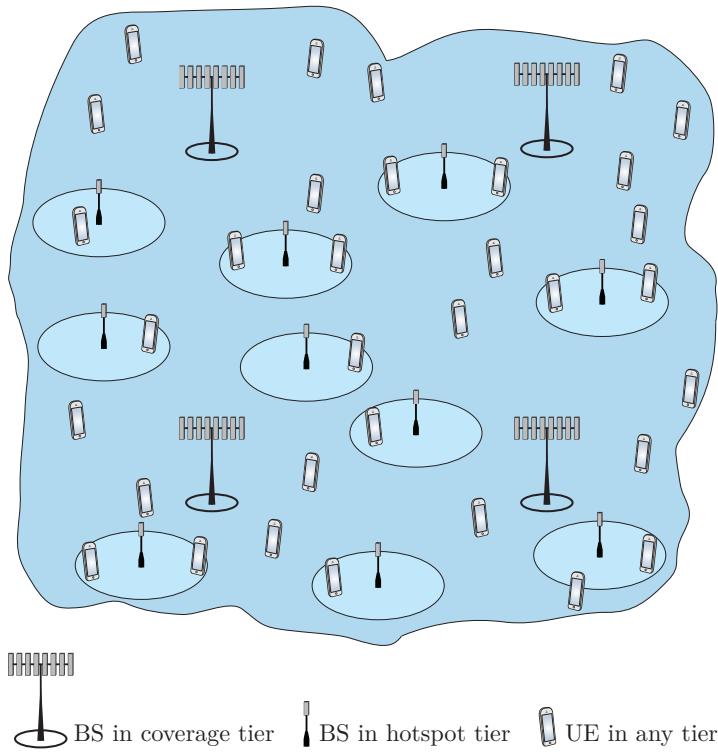


Figure 1.2: Current wireless networks are heterogeneous since a tier of SCs is deployed to offload traffic from the coverage tier. BSs in the coverage tier and in the hotspot tier are depicted differently, as shown in the figure. To improve the area throughput of the coverage tier, it is particularly important to increase the SE, because densification and the use of additional bandwidth at higher frequencies would degrade mobility support and coverage.

Figure 1.2. The two tiers may utilize the same frequency spectrum, but, in practice, it is common to use different spectrum to avoid inter-tier coordination; for example, the coverage tier might use LTE and operate in the 2.1 GHz band, while the hotspot tier might use WiFi in the 5 GHz band.

Cellular networks were originally designed for wireless voice communications, but it is wireless data transmissions that dominate nowadays [109]. Video on-demand accounts for the majority of traffic in wireless networks and is also the main driver of the predicted increase in traffic

demand [86]. The *area throughput* is thus a highly relevant performance metric of contemporary and future cellular networks. It is measured in bit/s/km² and can be modeled using the following high-level formula:

$$\text{Area throughput [bit/s/km}^2\text{]} = B \text{ [Hz]} \cdot D \text{ [cells/km}^2\text{]} \cdot \text{SE [bit/s/Hz/cell]} \quad (1.1)$$

where B is the bandwidth, D is the average cell density, and SE is the SE per cell. The SE is the amount of information that can be transferred per second over one Hz of bandwidth, and it is later defined in detail in Section 1.2.

These are the three main components that determine the area throughput, and that need to be increased in order to achieve higher area throughput in future cellular networks. This principle applies to the coverage tier as well as to the hotspot tier. Based on (1.1), one can think of the area throughput as being the volume of a rectangular box with sides B , D , and SE ; see Figure 1.3. There is an inherent dependence between these three components in the sense that the choice of frequency band and cell density affects the propagation conditions; for example, the probability of having a line-of-sight (LoS) channel between the transmitter and receiver (and between out-of-cell interferers and the receiver), the average propagation losses, etc. However, one can treat these three components as independent as a first-order approximation to gain basic insights. Consequently, there are three main ways to improve the area throughput of cellular networks:

1. Allocate more bandwidth;
2. Densify the network by deploying more BSs;
3. Improve the SE per cell.

The main goal of this section is to demonstrate how we can achieve major improvements in SE. These insights are then utilized in Section 2 on p. 216 to define the Massive MIMO technology.

1.1.1 Evolving Cellular Networks for Higher Area Throughput

Suppose, for the matter of argument, that we want to design a new cellular network that improves the area throughput by a factor of 1000

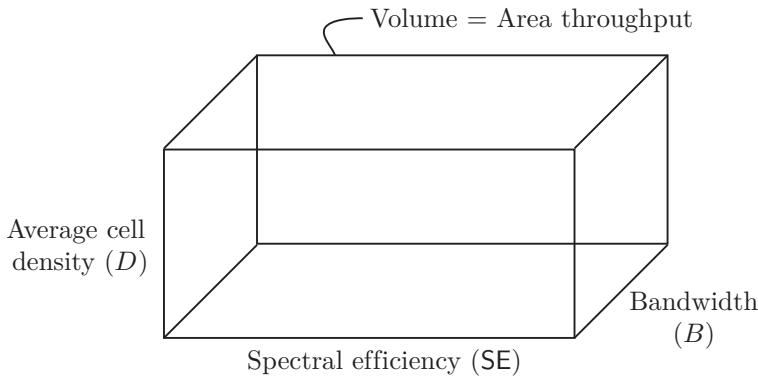


Figure 1.3: The area throughput can be computed according to (1.1) as the volume of a rectangular box where the bandwidth, average cell density, and SE are the length of each side.

over existing networks; that is, to solve “the $1000\times$ data challenge” posed by Qualcomm [271]. Note that such a network can handle the three orders-of-magnitude increase in wireless data traffic that will occur over the next 15–20 years, if the annual traffic growth rate continues to be in the range of 41%–59%. How can we handle such a huge traffic growth according to the formula in (1.1)?

One potential solution would be to increase the bandwidth B by $1000\times$. Current cellular networks utilize collectively more than 1 GHz of bandwidth in the frequency range below 6 GHz. For example, the telecom operators in Sweden have licenses for more than 1 GHz of spectrum [65], while the corresponding number in USA is around 650 MHz [30]. An additional 500 MHz of spectrum is available for WiFi [65]. This means that a $1000\times$ increase corresponds to using more than 1 THz of bandwidth in the future. This is physically impractical since the frequency spectrum is a global resource that is shared among many different services, and also because it entails using much higher frequency bands than in the past, which physically limits the range and service reliability. There are, however, substantial bandwidths in the millimeter wavelength (mmWave) bands (e.g., in the range 30–300 GHz) that can be used for short-range applications. These mmWave bands are attractive in the hotspot tier, but less so in the coverage tier since the signals at

those frequencies are easily blocked by objects and human bodies and thus cannot provide robust coverage.

Another potential solution would be to densify the cellular network by deploying $1000\times$ more BSs per km^2 . The inter-BS distances in the coverage tier are currently a few hundred meters in urban areas and the BSs are deployed at elevated locations to avoid being shadowed by large objects and buildings. This limits the number of locations where BSs can be deployed in the coverage tier. It is hard to densify without moving BSs closer to UEs, which leads to increased risks of being in deep shadow, thereby reducing coverage. Deploying additional hotspots is a more viable solution. Although WiFi is available almost everywhere in urban areas, the average inter-BS distance in the hotspot tier can certainly shrink down to tens of meters in the future. Reusing the spectrum from the coverage tier or using mmWave bands in these SCs can also bring substantial improvements to the area throughput [197]. Nevertheless, this solution is associated with high deployment costs, inter-cell interference issues [19], and is not suitable for mobile UEs, which would have to switch BS very often. Note that even under a substantial densification of the hotspot tier, the coverage tier is still required to support mobility and avoid coverage holes.

Higher cell density and larger bandwidth have historically dominated the evolution of the coverage tier, which explains why we are approaching a saturation point where further improvements are increasingly complicated and expensive. However, it might be possible to dramatically improve the SE of future cellular networks. This is particularly important for BSs in the coverage tier that, as explained above, can neither use mmWave bands nor rely on network densification. Increasing the SE corresponds to using the BSs and bandwidth that are already in place more efficiently by virtue of new modulation and multiplexing techniques. The principal goal is to select a rectangular box, as illustrated in Figure 1.4, where each side represents the multiplicative improvement in either B , D , or SE. As shown in the figure, there are different ways to choose these factors in order to achieve $1000\times$ higher area throughput. A pragmatic approach is to first investigate how much the SE can be improved towards the $1000\times$ goal and then jointly

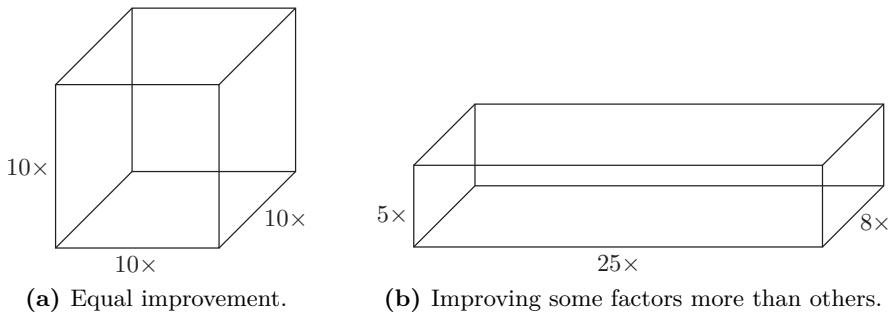


Figure 1.4: Examples of different ways to achieve a $1000\times$ improvement in area throughput. Each side of the rectangular box represents an improvement factor in either B , D , or SE in (1.1), and their multiplication (i.e., the volume) equals $1000\times$.

increase B and D to take care of the remaining part of the ambitious final goal. Section 4 on p. 275 shows why Massive MIMO is considered the most promising technology for improving the SE in future cellular networks.

Remark 1.1 (Massive MIMO versus SCs in mmWave bands). This monograph focuses on the coverage tier, which will remain the most challenging tier in the future since it should provide ubiquitous coverage, support mobility, and simultaneously deliver a uniform service quality within each cell. All of this must be achieved without any substantial densification or use of mmWave spectrum because that would inevitably result in patchy coverage. This is why major improvements in SE are needed. We will demonstrate that Massive MIMO can deliver that. In contrast, the main purpose of the hotspot tier is to reduce the pressure on the coverage tier by offloading a large portion of the traffic from low-mobility UEs. Since only short-range best-effort communications must be supported, this tier can be enhanced by straightforward cell densification and by using the large bandwidths available in mmWave bands. The use of Massive MIMO in mmWave bands will be discussed in Section 7.5 on p. 522, while the combination of Massive MIMO and SCs is considered in Section 7.6 on p. 527.

1.2 Definition of Spectral Efficiency

We now provide a definition of SE for a communication channel with a bandwidth of B Hz. The Nyquist-Shannon sampling theorem implies that the band-limited communication signal that is sent over this channel is completely determined by $2B$ real-valued equal-spaced samples per second [298]. When considering the complex-baseband representation of the signal, B complex-valued samples per second is the more natural quantity [314]. These B samples are the degrees of freedom available for designing the communication signal. The SE is the amount of information that can be transferred reliably per complex-valued sample.

Definition 1.2 (Spectral efficiency). The SE of an encoding/decoding scheme is the average number of bits of information, per complex-valued sample, that it can reliably transmit over the channel under consideration.

From this definition, it is clear that the SE is a deterministic number that can be measured in bit per complex-valued sample. Since there are B samples per second, an equivalent unit of the SE is bit per second per Hertz, often written in short-form as bit/s/Hz. For fading channels, which change over time, the SE can be viewed as the average number of bit/s/Hz over the fading realizations, as will be defined below. In this monograph, we often consider the SE of a channel between a UE and a BS, which for simplicity we refer to as the “SE of the UE”. A related metric is the *information rate* [bit/s], which is defined as the product of the SE and the bandwidth B . In addition, we commonly consider the sum SE of the channels from all UEs in a cell to the respective BS, which is measured in bit/s/Hz/cell.

The channel between a transmitter and a receiver at given locations can support many different SEs (depending on the chosen encoding/decoding scheme), but the largest achievable SE is of key importance when designing communication systems. The maximum SE is determined by the channel capacity, which was defined by Claude Shannon in his seminal paper [297] from 1948. The following theorem provides the capacity for the channel illustrated in Figure 1.5.



Figure 1.5: A general discrete memoryless channel with input x and output y .

Theorem 1.1 (Channel capacity). Consider a discrete memoryless channel with input x and output y , which are two random variables. Any SE smaller or equal to the channel capacity

$$C = \sup_{f(x)} (\mathcal{H}(y) - \mathcal{H}(y|x)) \quad (1.2)$$

is achievable with arbitrarily low error probability, while larger values cannot be achieved. The supremum is taken with respect to all feasible input distributions $f(x)$, while $\mathcal{H}(y)$ is the differential entropy of the output and $\mathcal{H}(y|x)$ is the conditional differential entropy of the output given the input.

The terminology of discrete memoryless channels and entropy is defined in Appendix B.5 on p. 572. We refer to [297] and textbooks on information theory, such as [94], for the proof of Theorem 1.1. The set of feasible input distributions depends on the application, but it is common to consider all distributions that satisfy a constraint on the input power. In wireless communications, we are particularly interested in channels where the received signal is the superposition of a scaled version of the desired signal and additive Gaussian noise. These channels are commonly referred to as additive white Gaussian noise (AWGN) channels. The channel capacity in Theorem 1.1 can be computed in closed form in the following canonical case from [298], which is also illustrated in Figure 1.6.

Corollary 1.2. Consider a discrete memoryless channel with input $x \in \mathbb{C}$ and output $y \in \mathbb{C}$ given by

$$y = hx + n \quad (1.3)$$

where $n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent noise. The input distribution is power-limited as $\mathbb{E}\{|x|^2\} \leq p$ and the channel response $h \in \mathbb{C}$ is known at the output.

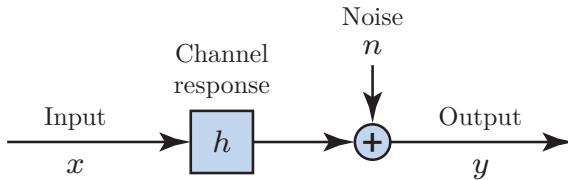


Figure 1.6: A discrete memoryless channel with input x and output $y = hx + n$, where h is the channel response and n is independent Gaussian noise.

If h is deterministic, then the channel capacity is

$$C = \log_2 \left(1 + \frac{p|h|^2}{\sigma^2} \right) \quad (1.4)$$

and is achieved by the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

If h is a realization of a random variable \mathbb{H} that is independent of the signal and noise, then the ergodic⁷ channel capacity is

$$C = \mathbb{E} \left\{ \log_2 \left(1 + \frac{p|h|^2}{\sigma^2} \right) \right\} \quad (1.5)$$

where the expectation is with respect to h . This is called a fading channel and the capacity is achieved by the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Proof. The proof is available in Appendix C.1.1 on p. 579. □

The channel considered in Corollary 1.2 is called a single-input single-output (SISO) channel because one input signal is sent and results in one output signal. An average power constraint is assumed in the corollary and throughout this monograph, but other constraints also exist in practice; see Remark 7.1 on p. 460 for a further discussion. The practical meaning of the channel capacity can be described by considering the transmission of an information sequence with N scalar inputs, generated by an ergodic stochastic process, over the discrete

⁷The capacity of a fading channel requires that the transmission spans asymptotically many realizations of the random variable that describes the channel. This is referred to as the ergodic capacity since a stationary ergodic random fading process is required if the statistical properties shall be deducible from a single sequence of channel realizations. Each channel realization is used for a predetermined and finite number of input signals, then a new realization is taken from the random process.

memoryless channel in Corollary 1.2. If the scalar input has an SE smaller or equal to the capacity, the information sequence can be encoded such that the receiver can decode it with arbitrarily low error probability as $N \rightarrow \infty$. In other words, an infinite decoding delay is required to achieve the capacity. The seminal work in [267] quantifies how closely the capacity can be approached at a finite length of the information sequence. The SE is generally a good performance metric whenever data blocks of thousands of bits are transmitted [50].

The capacity expressions in (1.4) and (1.5) have a form that is typical for communications: the base-two logarithm of one plus the signal-to-noise ratio (SNR)-like expression

$$\frac{\text{Received signal power}}{\text{Noise power}} = \frac{\overbrace{p|h|^2}^{\text{Received signal power}}}{\underbrace{\sigma^2}_{\text{Noise power}}} \quad (1.6)$$

This is the actual measurable SNR for a deterministic channel response h , while it is the instantaneous SNR for a given channel realization when h is random. Since the SNR fluctuates in the latter case, it is more convenient to consider the average SNR when describing the quality of a communication channel. We define the average SNR as

$$\text{SNR} = \frac{p\mathbb{E}\{|h|^2\}}{\sigma^2} \quad (1.7)$$

where the expectation is computed with respect to the channel realizations. We call $\mathbb{E}\{|h|^2\}$ the average *channel gain* since it is the average scaling of the signal power incurred by the channel.

Transmissions in cellular networks are in general corrupted by interference from simultaneous transmissions in the same and other cells. By adding such interference to the channel in Figure 1.6, we obtain the discrete memoryless interference channel in Figure 1.7. The interference is not necessarily independent of the input x and the channel h . The exact channel capacity of interference channels is generally unknown, but convenient lower bounds can be obtained. Inspired by [36, 214], the following corollary provides the lower capacity bounds that will be used repeatedly in this monograph.

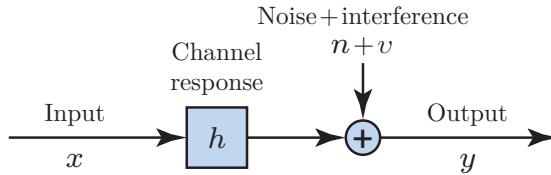


Figure 1.7: A discrete memoryless interference channel with input x and output $y = hx + v + n$, where h is the channel response, n is independent Gaussian noise, and v is the interference, which is uncorrelated with the input and the channel.

Corollary 1.3. Consider a discrete memoryless interference channel with input $x \in \mathbb{C}$ and output $y \in \mathbb{C}$ given by

$$y = hx + v + n \quad (1.8)$$

where $n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent noise, the channel response $h \in \mathbb{C}$ is known at the output, and $v \in \mathbb{C}$ is random interference. The input is power-limited as $\mathbb{E}\{|x|^2\} \leq p$.

If h is deterministic and the interference v has zero mean, a known variance $p_v \in \mathbb{R}_+$, and is uncorrelated with the input (i.e., $\mathbb{E}\{x^*v\} = 0$), then the channel capacity C is lower bounded as

$$C \geq \log_2 \left(1 + \frac{p|h|^2}{p_v + \sigma^2} \right) \quad (1.9)$$

where the bound is achieved using the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Suppose $h \in \mathbb{C}$ is instead a realization of the random variable \mathbb{H} and that \mathbb{U} is a random variable with realization u that affects the interference variance. The realizations of these random variables are known at the output. If the noise n is conditionally independent of v given h and u , the interference v has conditional zero mean (i.e., $\mathbb{E}\{v|h, u\} = 0$) and conditional variance denoted by $p_v(h, u) = \mathbb{E}\{|v|^2|h, u\}$, and the interference is conditionally uncorrelated with the input (i.e., $\mathbb{E}\{x^*v|h, u\} = 0$), then the ergodic⁸ channel capacity C is lower bounded as

$$C \geq \mathbb{E} \left\{ \log_2 \left(1 + \frac{p|h|^2}{p_v(h, u) + \sigma^2} \right) \right\} \quad (1.10)$$

⁸When transmitting an information sequence over this fading channel, a sequence of realizations of \mathbb{H} and \mathbb{U} is created, forming stationary ergodic random processes. Each set of realizations (h, u) is used for a predetermined and finite number of input signals, then a new set of realizations is taken from the random processes.

where the expectation is taken with respect to h and u , and the bound is achieved using the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Proof. The proof is available in Appendix C.1.2 on p. 580. \square

Note that in Corollary 1.3, we use the shorthand notation $\mathbb{E}\{v|h, u\}$ for the conditional expectation $\mathbb{E}\{v|\mathbb{H} = h, \mathbb{U} = u\}$. For notational convenience, we will from now on omit the random variables in similar expressions and only write out the realizations.

The lower bounds on the channel capacity in Corollary 1.3 are obtained by treating the interference as an additional source of noise in the decoder, which might not be optimal from an information-theoretic point of view. For example, if an interfering signal is very strong, then one can potentially decode it and subtract the interference from the received signal, before decoding the desired signal. This is conceptually simple, but harder to perform in a practical cellular network, where the interfering signals change over time and the cells are not fully cooperating. In fact, there should not be any strongly interfering signal in a well-designed cellular network. In the low-interference regime, it is optimal (i.e., capacity-achieving) to treat interference as additional noise, as shown in [230, 296, 20, 21, 295].

We utilize SE expressions of the type in Corollary 1.3 throughout this monograph and stress that these might not be the highest achievable SEs, but SEs that can be achieved by low-complexity signal processing in the receiver, where interference is treated as noise. The SE expressions in (1.9) and (1.10) have a form typical for wireless communications: the base-two logarithm of one plus the expression

$$\text{SINR} = \frac{\text{Received signal power}}{\underbrace{p_v}_{\text{Interference power}} + \underbrace{\sigma^2}_{\text{Noise power}}} \quad (1.11)$$

$$\qquad \qquad \qquad \overbrace{p|h|^2}$$

that can be interpreted as the signal-to-interference-plus-noise ratio (SINR). Formally, this is only an SINR when h and p_v are deterministic; the expression is otherwise random. For simplicity, we will refer to any term a that appears as $\mathbb{E}\{\log_2(1 + a)\}$ in an SE expression as an *instantaneous SINR* (with slight abuse of terminology).

The SE expressions presented in this section are the fundamental building blocks for the theory developed in later sections. The capacity results consider discrete memoryless channels, which are different from practical continuous wireless channels. However, the bandwidth B can be divided into narrow subchannels (e.g., using orthogonal frequency-division multiplexing (OFDM)) that are essentially memoryless if the symbol time is much longer than the delay spread of the propagation environment [314].

1.3 Ways to Improve the Spectral Efficiency

There are different ways to improve the per-cell SE in cellular networks. In this section, we will compare different approaches to showcase which ones are the most promising. For simplicity, we consider a two-cell network where the average channel gain between a BS and every UE in a cell is identical, as illustrated in Figure 1.8. This is a tractable model for studying the basic properties of cellular communications, due to the small number of system parameters. It is an instance of the *Wyner model*, initially proposed by Aaron Wyner in [353] and studied for fading channels in [304]. It has been used extensively to study the fundamental information-theoretic properties of cellular networks; see the monograph [303] and references therein. More realistic, but less tractable, network models will be considered in later sections.

In the UL scenario shown in Figure 1.8, the UEs in cell 0 transmit to their serving BS, while the UL signals from the UEs in cell 1 leak into cell 0 as interference. The average channel gain from a UE in cell 0 to its serving BS is denoted by β_0^0 , while the interfering signals from UEs in cell 1 have an average channel gain of β_1^0 . Similarly, the average channel gain from a UE in cell 1 to its serving BS is denoted by β_1^1 , while the interfering signals from UEs in cell 0 have an average channel gain of β_0^1 . Notice that the superscript indicates the cell of the receiving BS and the subscript indicates the cell that the transmitting UE resides in. The average channel gains are positive dimensionless quantities that are often very small since the signal energy decays quickly with the propagation distance; values in the range from -70 dB to -120 dB are common within the serving cell, while even smaller values appear for

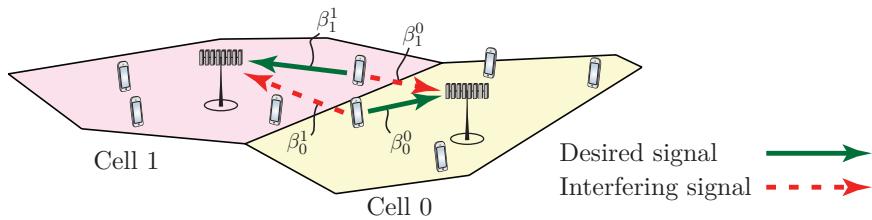


Figure 1.8: Illustration of the notion of desired and interfering UL signals in a two-cell network. In the Wyner model, every UE in cell 0 has the same value of the average channel gain β_0^0 from its serving BS and of the average channel gain β_1^1 to the other-cell BS, while every UE in cell 1 has the same value of β_1^0 and β_1^1 .

interfering signals. As shown later, it is not the absolute values that are of main importance when computing the SE, but the relative strength of the interference as compared to the desired signals. For simplicity, we assume that the intra-cell channel gains are equal (i.e., $\beta_0^0 = \beta_1^1$) and that the inter-cell channel gains are equal as well (i.e., $\beta_1^0 = \beta_1^1$); this is commonly assumed in the Wyner model. We can then define the ratio $\bar{\beta}$ between the inter-cell and intra-cell channel gains as

$$\bar{\beta} = \frac{\beta_1^0}{\beta_0^0} = \frac{\beta_1^1}{\beta_0^0} = \frac{\beta_1^0}{\beta_1^1} = \frac{\beta_1^1}{\beta_1^1}. \quad (1.12)$$

This ratio will be used in the analysis of both UL and DL. We typically have $0 \leq \bar{\beta} \leq 1$, where $\bar{\beta} \approx 0$ corresponds to a negligibly weak inter-cell interference and $\bar{\beta} \approx 1$ means that the inter-cell interference is as strong as the desired signals (which may happen for UEs at the cell edge). We will use this model in the remainder of Section 1, to discuss different ways to improve the SE per cell.

1.3.1 Increase the Transmit Power

The SE naturally depends on the strength of the received desired signal, represented by the average SNR, defined in (1.7). Using the Wyner model described above, the average SNR of a UE in cell 0 is

$$\text{SNR}_0 = \frac{p}{\sigma^2} \beta_0^0 \quad (1.13)$$

where p denotes the UE's transmit power and σ^2 is the noise power.

These power quantities are measured in Joule per time interval. Any type of time interval can be utilized as long as it is the same for both the signal and the noise, but common choices are “one second” or “one sample”. The parameter SNR_0 plays a key role in many of the expressions computed in this section.

Assume that there is one active UE per cell and that each BS and UE is equipped with a single antenna. Notice that with “antenna” we refer to a component with a size that is smaller than the wavelength (e.g., a patch antenna) and not the type of large high-gain antennas that are used at the BSs in conventional cellular networks. Antennas and antenna arrays are further discussed in Section 7.4 on p. 500.

Focusing on a flat-fading⁹ wireless channel, the symbol-sampled complex-baseband signal $y_0 \in \mathbb{C}$ received at the BS in cell 0 is

$$y_0 = \underbrace{h_0^0 s_0}_{\text{Desired signal}} + \underbrace{h_1^0 s_1}_{\text{Interfering signal}} + \underbrace{n_0}_{\text{Noise}} \quad (1.14)$$

where the additive receiver noise is modeled as $n_0 \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$. The scalars $s_0, s_1 \sim \mathcal{N}_{\mathbb{C}}(0, p)$ in (1.14) represent the information signals¹⁰ transmitted by the desired and interfering UEs, respectively. Moreover, their channel responses are denoted by $h_0^0 \in \mathbb{C}$ and $h_1^0 \in \mathbb{C}$, respectively. The properties of these channel responses depend on the propagation environment. In this section, we consider one model of LoS propagation and one model of non-line-of-sight (NLoS) propagation. In single-antenna LoS propagation, h_0^0 and h_1^0 are deterministic scalars corresponding to the square-root of the (average) channel gains:

$$h_i^0 = \sqrt{\beta_i^0} \quad \text{for } i = 0, 1. \quad (1.15)$$

In general, the channel response will also have a phase rotation, but it is neglected here since it does not affect the SE. The channel gain

⁹In flat-fading channels, the coherence bandwidth of the channel is larger than the signal bandwidth [314]. Therefore, all frequency components of the signal will experience the same magnitude of fading, resulting in a scalar channel response.

¹⁰The information signals are assumed to be complex Gaussian distributed since this maximizes the differential entropy of the signal (see Lemma B.21 on p. 574) and achieves the capacity in interference-free scenarios (see Corollary 1.2). In practice, quadrature amplitude modulation (QAM) schemes with finite number of constellation points are commonly used, which leads to a small shaping-loss as compared to having infinitely many constellation points from a Gaussian distribution.

β_i^0 can be interpreted as the macroscopic *large-scale fading* in LoS propagation, caused by distance-dependent pathloss. The impact of the transceiver hardware, including the antenna gains, is also absorbed into this parameter. The parameter is constant if the transmitter and receiver are fixed, while it changes if the transmitter and/or receiver move. Microscopic movements (at the order of the wavelength) can be modeled as phase-rotations in h_i^0 , while large movements (at the order of meters) lead to substantial changes in β_i^0 . We consider a fixed value of h_i^0 in order to apply the SE expression in Corollary 1.3 for deterministic channels.

In NLoS propagation environments, the channel responses are random variables that change over time and frequency. If there is sufficient scattering between the UEs and the BS, then h_0^0 and h_1^0 are well-modeled as

$$h_i^0 \sim \mathcal{N}_{\mathbb{C}}(0, \beta_i^0) \quad \text{for } i = 0, 1 \quad (1.16)$$

as validated by the channel measurements reported in [337, 177, 83, 365]. The transmitted signal reaches the receiver through many different paths and the superimposed received signals can either reinforce or cancel each other. When the number of paths is large, the central limit theorem motivates the use of a Gaussian distribution. This phenomenon is known as small-scale fading and is a microscopic effect caused by small variations in the propagation environment (e.g., movement of the transmitter, receiver, or other objects). In contrast, the variance β_i^0 is interpreted as the macroscopic large-scale fading, which includes distance-dependent pathloss, shadowing, antenna gains, and penetration losses in NLoS propagation. The channel model in (1.16) is called Rayleigh fading, because the magnitude $|h_i^0|$ is a Rayleigh distributed random variable.

Notice that the average channel gain is $\mathbb{E}\{|h_i^0|^2\} = \beta_i^0$, for $i = 0, 1$, in both propagation cases in order to make them easily comparable. Practical channels can contain a mix of a deterministic LoS component and a random NLoS component, but, by studying the differences between the two extreme cases, we can predict what will happen in the mixed cases as well. The following lemma provides closed-form SE expressions for the LoS and NLoS cases.

Lemma 1.4. Suppose the BS in cell 0 knows the channel responses. An achievable¹¹ UL SE for the desired UE in the LoS case is

$$\text{SE}_0^{\text{LoS}} = \log_2 \left(1 + \frac{1}{\bar{\beta} + \frac{1}{\text{SNR}_0}} \right) \quad (1.17)$$

with $\bar{\beta}$ and SNR_0 given by (1.12) and (1.13), respectively. In the NLoS case (with $\bar{\beta} \neq 1$), an achievable UL SE is

$$\begin{aligned} \text{SE}_0^{\text{NLoS}} &= \mathbb{E} \left\{ \log_2 \left(1 + \frac{p|h_0^0|^2}{p|h_1^0|^2 + \sigma^2} \right) \right\} \\ &= \frac{e^{\frac{1}{\text{SNR}_0}} E_1 \left(\frac{1}{\text{SNR}_0} \right) - e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1 \left(\frac{1}{\text{SNR}_0 \bar{\beta}} \right)}{\log_e(2) \left(1 - \bar{\beta} \right)} \end{aligned} \quad (1.18)$$

where $E_1(x) = \int_1^\infty \frac{e^{-xu}}{u} du$ denotes the exponential integral and $\log_e(\cdot)$ denotes the natural logarithm.

Proof. The proof is available in Appendix C.1.3 on p. 582. \square

This lemma shows that the SE is fully characterized by the SNR of the desired signal, SNR_0 , and the relative strength of the inter-cell interference, $\bar{\beta}$. Note that the closed-form NLoS expression in (1.18) only applies for $\bar{\beta} \neq 1$. Recall that $0 \leq \bar{\beta} \leq 1$ is the typical range of $\bar{\beta}$. The pathological case $\bar{\beta} = 1$ represents a cell-edge scenario where the desired and interfering signals are equally strong. An alternative expression can be derived for $\bar{\beta} = 1$, using the same methodology as in the proof of Lemma 1.4, but it does not provide any further insights and is therefore omitted.

The SE is naturally an increasing function of the SNR, which is most easily seen from the LoS expression in (1.17), where the SE is the logarithm of the following SINR expression:

$$\frac{1}{\bar{\beta} + \frac{1}{\text{SNR}_0}} = \frac{\overbrace{p\beta_0^0}^{\text{Signal power}}}{\underbrace{p\beta_1^0}_{\text{Interference power}} + \underbrace{\sigma^2}_{\text{Noise power}}}. \quad (1.19)$$

¹¹Recall that an SE is achievable if there exists a sequence of codes such that the maximum probability of error in transmission for any message of length N converges to zero as $N \rightarrow \infty$ [94]. Any SE smaller or equal to the capacity is thus achievable.

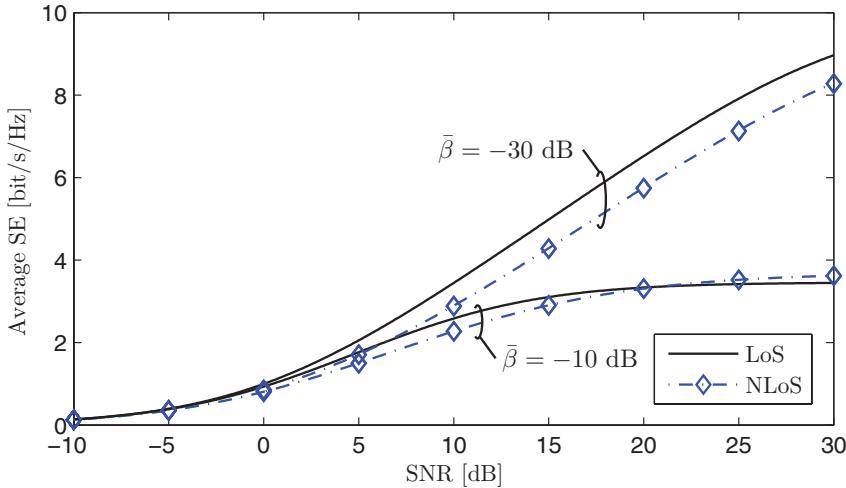


Figure 1.9: Average UL SE as a function of the SNR for different cases of inter-cell interference strength, $\bar{\beta} \in \{-10, -30\}$ dB, and different channel models.

One can improve the SE by increasing the transmit power p . However, the SE will not increase indefinitely with p . In the LoS case, we have

$$\text{SE}_0^{\text{LoS}} \rightarrow \log_2 \left(1 + \frac{1}{\bar{\beta}} \right) \quad \text{as } p \rightarrow \infty \quad (1.20)$$

where the limit is completely determined by the strength of the interference. This is due to the fact that the desired UE and the interfering UE both increase their transmit powers, which is the case of interest in cellular networks since good service quality should be guaranteed in all cells. The corresponding limit in the NLoS case is

$$\text{SE}_0^{\text{NLoS}} \rightarrow \frac{1}{1 - \bar{\beta}} \log_2 \left(\frac{1}{\bar{\beta}} \right) \quad \text{as } p \rightarrow \infty \quad (1.21)$$

which can be proved by expanding the exponential integrals in (1.18) using the identity in [3, Eq. (5.1.11)] and then taking the limit $p \rightarrow \infty$.

To exemplify these behaviors, Figure 1.9 shows the SE as a function of the SNR, where an SNR increase is interpreted as increasing the transmit power p . We consider two different strengths of the inter-cell interference: $\bar{\beta} = -10$ dB and $\bar{\beta} = -30$ dB. The SE converges quickly to the LoS limit $\log_2(1 + 1/\bar{\beta}) \approx 3.46$ bit/s/Hz and the NLoS

limit $\log_2(1/\bar{\beta})/(1 - \bar{\beta}) \approx 3.69$ bit/s/Hz in the former case, since the interference is only 10 dB weaker than the desired signal. In the case of $\bar{\beta} = -30$ dB, the convergence to the LoS limit 9.97 bit/s/Hz and NLoS limit 9.98 bit/s/Hz is less visible in the considered SNR range, since the interference is weaker and the logarithm makes the SE grow slowly. Nevertheless, we notice that going from $\text{SNR}_0 = 10$ dB to $\text{SNR}_0 = 30$ dB only doubles the SE, though 100 times more transmit power is required. The NLoS case provides slightly lower SE than the LoS case for most SNRs, due to the random fluctuations of the squared magnitude $|h_0^0|^2$ of the channel. However, the randomness turns into a small advantage at high SNR, where the limit is slightly higher in NLoS because the interference can be much weaker than the signal for some channel realizations. This behavior is seen for $\bar{\beta} = -10$ dB in Figure 1.9, while it occurs at higher SNRs for $\bar{\beta} = -30$ dB.

In summary, increasing the SNR by using more transmit power improves the SE, but the positive effect quickly pushes the network into an interference-limited regime where no extraordinary SEs can be obtained. This is basically because of the lack of *degrees of freedom* at the BS, which cannot separate the desired signal from the interference from a single observation.¹² This interference-limited regime is where the coverage tier operates in current networks, while the situation for the hotspot tier depends on how the BSs are deployed. For example, the signals at mmWave frequencies are greatly attenuated by walls and other objects. A mmWave SC will typically cover a very limited area, but on the other hand the cell might be noise-limited since the interfering signals from SCs in other rooms are also attenuated by walls. The SE range in Figure 1.9 is comparable to what contemporary networks deliver (e.g., 0–5 bit/s/Hz in LTE [144]). Hence, a simple power-scaling approach cannot contribute much to achieving higher SE in cellular networks.

Remark 1.2 (Increasing cell density). Another way to increase the SNR is to keep the transmit power fixed and increase the cell density D

¹²The transmission scheme considered in this example is not optimal. The UEs could take turns in transmitting, thereby achieving an SE that grows without bound, but with a pre-log factor of 1/2 if each UE is active 50% of the time. More generally, interference alignment methods can be used to handle the interference [70].

instead. It is commonly assumed in channel modeling that the average channel gain is inversely proportional to the propagation distance to some fixed “pathloss” exponent. Under such a basic propagation model, the power of the received desired signal and the inter-cell interference increase at roughly the same pace when D is increased, since both the distance to the desired BS and the interfering BSs are reduced. This implies that the interference-limited SE limit is obtained also when D increases. While D cannot be much increased in the coverage tier, cell densification is a suitable way to improve the hotspot tier [198]; the area throughput in (1.1) increases linearly with D as long as the basic propagation model holds true. At some point, this model will, however, become invalid since the pathloss exponent will also reduce with the distance and approach the free-space propagation scenario with an exponent of two [19]. Cell densification is no longer desired in this extreme short-range scenario since the sum power of the interfering signals increase faster than the desired signal power.

1.3.2 Obtain an Array Gain

Instead of increasing the UL transmit power, the BS can deploy multiple receive antennas to collect more energy from the EM waves. This concept has at least been around since the 1930s [257, 117], with the particular focus on achieving spatial diversity; that is, to combat the channel fading in NLoS propagation by deploying multiple receive antennas that observe different fading realizations. The related idea of using multiple transmit antennas to increase the received signal power was described as early as 1919 [10]. Having multiple receive antennas also allows the receiver to distinguish between signals with different spatial directivity by using spatial filtering/processing [324]. Implementations of these methods have been referred to as “adaptive” or “smart” antennas [16, 350]. In general, it is more convenient to equip the BSs with multiple antennas than the UEs, because the latter are typically compact commercial end-user products powered by batteries and relying on low-cost components.

Suppose the BS in cell 0 is equipped with an array of M antennas. The channel responses from the desired and interfering UEs can then be represented by the vectors $\mathbf{h}_0^0 \in \mathbb{C}^M$ and $\mathbf{h}_1^0 \in \mathbb{C}^M$, respectively. The

m th element of each vector is the channel response observed at the m th BS antenna, for $m = 1, \dots, M$. The scalar received UL signal in (1.14) is then extended to a received vector $\mathbf{y}_0 \in \mathbb{C}^M$, modeled as

$$\mathbf{y}_0 = \underbrace{\mathbf{h}_0^0 s_0}_{\text{Desired signal}} + \underbrace{\mathbf{h}_1^0 s_1}_{\text{Interfering signal}} + \underbrace{\mathbf{n}_0}_{\text{Noise}} \quad (1.22)$$

where $\mathbf{n}_0 \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \sigma^2 \mathbf{I}_M)$ is the receiver noise over the BS array and the transmit signals s_0 and s_1 are defined as in (1.14).

To analyze the SE of this UL single-input multiple-output (SIMO) channel with inter-cell interference, we need to extend the propagation models to the multiple antenna case. In the LoS case, we consider a horizontal uniform linear array (ULA) with antenna spacing $d_H \in (0, 0.5]$, which is measured in the number of wavelengths between adjacent antennas. Hence, if λ denotes the wavelength at the carrier frequency, then the antenna spacing is λd_H meters. Channel models for other array geometries are considered in Section 7.3 on p. 482. We further assume that the UEs are located at fixed locations in the far-field of the BS array, which leads to the following deterministic channel response [254]:

$$\mathbf{h}_i^0 = \sqrt{\beta_i^0} \left[1 \ e^{2\pi j d_H \sin(\varphi_i^0)} \dots e^{2\pi j d_H (M-1) \sin(\varphi_i^0)} \right]^T \text{ for } i = 0, 1 \quad (1.23)$$

where $\varphi_i^0 \in [0, 2\pi)$ is the azimuth angle to the UE, relative to the boresight of the array at the BS in cell 0, and β_i^0 describes the macroscopic large-scale fading. The channel response in (1.23) can also have a common phase rotation of all elements, but it is neglected here since it does not affect the SE. The LoS propagation model is illustrated in Figure 1.10, where a plane wave reaches the array from a generic azimuth angle φ . When comparing two adjacent antennas, one of them observes a signal that has traveled $d_H \sin(\varphi)$ longer than the other one. This leads to the array response in (1.23) with phase rotations that are multiples of $d_H \sin(\varphi)$, as also illustrated in Figure 1.10.

In the NLoS case, we assume for now that the channel response is spatially uncorrelated over the array. This yields

$$\mathbf{h}_i^0 \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_M, \beta_i^0 \mathbf{I}_M \right) \text{ for } i = 0, 1 \quad (1.24)$$

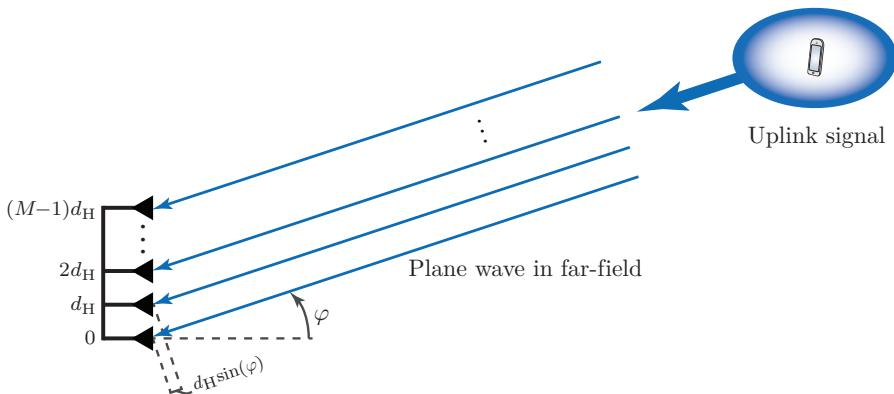


Figure 1.10: LoS propagation between a transmitting single-antenna UE and a BS equipped with a ULA with M antennas. The antenna spacing is d_H wavelengths, the azimuth angle to the UE is φ , and the UE is located in the far-field of the array, so that a plane wave reaches it. Note that the setup is illustrated from above.

where β_i^0 describes the macroscopic large-scale fading, while the randomness and Gaussian distribution account for the small-scale fading. This channel model is called *uncorrelated Rayleigh fading* or independent and identically distributed (i.i.d.) Rayleigh fading, since the elements in \mathbf{h}_i^0 are uncorrelated (and also independent) and have Rayleigh distributed magnitudes. Uncorrelated Rayleigh fading is a tractable model for rich scattering conditions, where the BS array is surrounded by many scattering objects, as compared to the number of antennas. We will use it to describe the basic properties in this section, while a more general and realistic model is introduced in Section 2.2 on p. 222 and then used in the remainder of the monograph. Channel modeling is further discussed in Section 7.3 on p. 482. The NLoS propagation model with uncorrelated Rayleigh fading is illustrated in Figure 1.11. Notice that the average channel gain β_i^0 is, for simplicity, assumed to be the same for all BS antennas. This is a reasonable approximation when the distance between the BS and UE is much larger than the distance between the BS antennas. However, in practice, there can be several decibels of channel gain variations between the antennas [122]. This fact is neglected in this section, but has a strong impact on the SE when M is large; see Section 4.4 on p. 335 for further details.

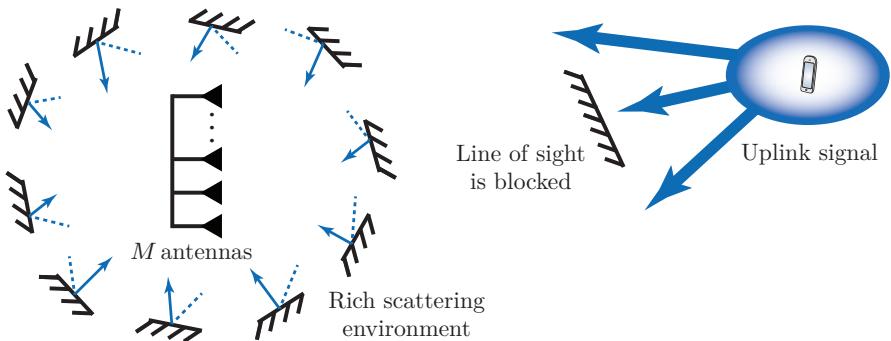


Figure 1.11: NLoS propagation with uncorrelated Rayleigh fading between a transmitting single-antenna UE and a BS equipped with an array of M antennas. The LoS path is blocked, but the signal finds multiple other paths via scattering objects. The BS is surrounded by many scattering objects so that the UE location has no impact on the spatial directivity of the received signal.

The benefits of having multiple antennas at the BS appear when the BS knows the channel response of the desired UE. This knowledge enables the BS to coherently combine the received signals from all antennas. Estimation of the channel response is thus a key aspect in multiantenna systems and will be further discussed in Section 1.3.5 and later analyzed in detail in Section 3 on p. 244. For now, we assume that the channel responses are known at the BS and can be used to select a *receive combining vector* $\mathbf{v}_0 \in \mathbb{C}^M$. This vector is multiplied with the received signal in (1.22) to obtain

$$\mathbf{v}_0^H \mathbf{y}_0 = \underbrace{\mathbf{v}_0^H \mathbf{h}_0^0 s_0}_{\text{Desired signal}} + \underbrace{\mathbf{v}_0^H \mathbf{h}_1^0 s_1}_{\text{Interfering signal}} + \underbrace{\mathbf{v}_0^H \mathbf{n}_0}_{\text{Noise}}. \quad (1.25)$$

Receive combining is a linear projection, which transforms the SIMO channel into an effective SISO channel that may support higher SEs than in the single-antenna case, if the combining vector is selected judiciously. There are many different combining schemes, but a simple and popular one is *maximum ratio (MR) combining*, defined as

$$\mathbf{v}_0 = \mathbf{h}_0^0. \quad (1.26)$$

This is a vector that maximizes the ratio $|\mathbf{v}_0^H \mathbf{h}_0^0|^2 / \|\mathbf{v}_0\|^2$ between the power of the desired signal and the squared norm of the combining

vector [172, 68].¹³ The following lemma gives closed-form SE expressions for the case of MR combining.

Lemma 1.5. Suppose the BS in cell 0 knows the channel responses and applies MR combining to the received signal in (1.22). An achievable UL SE for the desired UE in the LoS case is

$$\text{SE}_0^{\text{LoS}} = \log_2 \left(1 + \frac{M}{\bar{\beta} g(\varphi_0^0, \varphi_1^0) + \frac{1}{\text{SNR}_0}} \right) \quad (1.27)$$

where the function $g(\varphi, \psi)$ is defined as

$$g(\varphi, \psi) = \begin{cases} \frac{\sin^2(\pi d_H M (\sin(\varphi) - \sin(\psi)))}{M \sin^2(\pi d_H (\sin(\varphi) - \sin(\psi)))} & \text{if } \sin(\varphi) \neq \sin(\psi) \\ M & \text{if } \sin(\varphi) = \sin(\psi). \end{cases} \quad (1.28)$$

Similarly, an achievable UL SE for the desired UE in the NLoS case (with $\bar{\beta} \neq 1$) is

$$\begin{aligned} \text{SE}_0^{\text{NLoS}} = & \left(\frac{1}{\left(1 - \frac{1}{\bar{\beta}}\right)^M} - 1 \right) \frac{e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1\left(\frac{1}{\text{SNR}_0 \bar{\beta}}\right)}{\log_e(2)} \\ & + \sum_{m=1}^M \sum_{l=0}^{M-m} \frac{(-1)^{M-m-l+1}}{\left(1 - \frac{1}{\bar{\beta}}\right)^m} \frac{\left(e^{\frac{1}{\text{SNR}_0}} E_1\left(\frac{1}{\text{SNR}_0}\right) + \sum_{n=1}^l \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{j! \text{SNR}_0^j} \right)}{(M-m-l)! \text{SNR}_0^{M-m-l} \bar{\beta} \log_e(2)} \end{aligned} \quad (1.29)$$

where $n!$ denotes the factorial function and $E_1(x) = \int_1^\infty \frac{e^{-xu}}{u} du$ denotes the exponential integral.

Proof. The proof is available in Appendix C.1.4 on p. 583. \square

This lemma shows that the SE is characterized by the SNR of the desired signal, SNR_0 , the strength of the inter-cell interference, $\bar{\beta}$, and the number of BS antennas, M . Notice that by having M receive antennas, the array collects M times more energy from the desired

¹³The Cauchy-Schwartz inequality can be used to prove that $\mathbf{v}_0 = \mathbf{h}_0^0$ maximizes the ratio $|\mathbf{v}_0^H \mathbf{h}_0^0|^2 / \|\mathbf{v}_0\|^2$.

and interfering signals, and also from the noise. In the LoS case in (1.27), the gain of the desired signal scales as M . The linear scaling with the number of antennas is called *array gain*. It shows that MR coherently combines all the received energy from the desired signal, because the combining vector is matched to the channel response of the desired UE. In contrast, MR combines the noise and the interfering signal components non-coherently over the array since \mathbf{v}_0 is independent of \mathbf{h}_1^0 and \mathbf{n}_0 . As a consequence, the interference power $\bar{\beta}g(\varphi_0^0, \varphi_1^0)$ in (1.27) can be upper bounded as

$$\bar{\beta}g(\varphi_0^0, \varphi_1^0) \leq \frac{\bar{\beta}}{M} \frac{1}{\sin^2(\pi d_H(\sin(\varphi_0^0) - \sin(\varphi_1^0)))} \quad (1.30)$$

when $\sin(\varphi_0^0) \neq \sin(\varphi_1^0)$, which decreases as $1/M$ when more receive antennas are added. The basic reason that MR combining rejects the interfering signal is that the M antennas provide the BS with M spatial degrees of freedom, which can be used to separate the desired signal from the interfering signal. In particular, the directions of the LoS channel responses \mathbf{h}_0^0 and \mathbf{h}_1^0 gradually become orthogonal as M increases. This property is called (asymptotically) *favorable propagation* [245], since UEs with orthogonal channels can communicate with the BS simultaneously without causing mutual interference. We will further discuss this property in Section 1.3.3 and also in Section 2.5.2 on p. 233.

The equation $\sin(\varphi_0^0) = \sin(\varphi_1^0)$ has two unique solutions: $\varphi_0^0 = \varphi_1^0$ and the mirror reflection $\varphi_0^0 = \pi - \varphi_1^0$. Hence, the ULA can only uniquely resolve angles either in the interval $[-\pi/2, \pi/2]$ or in the interval $[\pi/2, 3\pi/2]$ at the other side of the array. The discussion above does not apply when $\sin(\varphi_0^0) = \sin(\varphi_1^0)$, because then $g(\varphi_0^0, \varphi_1^0) = M$ instead. It is natural that both the desired and the interfering signal scale linearly with M in this case, because the two signals arrive from exactly the same angle (or its mirror reflection). This will most likely never happen in practice, but we can infer from (1.28) that the interference is stronger when the UEs' angles are similar to each other. For example,

we can utilize the fact that $\sin(\pi z) \approx \pi z$ for $|z| < 0.2$ to show that

$$\begin{aligned} g(\varphi, \psi) &= \frac{\sin^2(\pi d_H M (\sin(\varphi) - \sin(\psi)))}{M \sin^2(\pi d_H (\sin(\varphi) - \sin(\psi)))} \\ &\approx \frac{(\pi d_H M (\sin(\varphi) - \sin(\psi)))^2}{M (\pi d_H (\sin(\varphi) - \sin(\psi)))^2} = M \end{aligned} \quad (1.31)$$

if $d_H M |\sin(\varphi) - \sin(\psi)| < 0.2$. The angular interval for which this holds becomes smaller as the aperture $d_H M$ of the ULA increases, but it exists for any finite-sized array. Since it is $d_H M$ that determines the angular resolution, the interference is reduced by either increasing the number of antennas M and/or using a larger antenna spacing d_H . This is in contrast to the signal term, which is proportional only to the number of antennas. For a given array aperture, it is therefore beneficial to have many antennas rather than widely separated antennas. Note that we have considered a two-dimensional LoS model in this section where only the azimuth angle can differ between the UEs. In practice, UEs can also have different elevation angles to the BS array and this can be exploited to separate the UEs. These aspects will be discussed in more detail in Section 7.4.2 on p. 503.

To illustrate these behaviors, the function $g(\varphi_0^0, \varphi_1^0)$ is shown in Figure 1.12 for a desired UE at the fixed angle $\varphi_0^0 = 30^\circ$, while the angle of the interfering UE is varied between -180° and 180° . The antenna-spacing is $d_H = 1/2$. In the single-antenna case, we have $g(\varphi_0^0, \varphi_1^0) = 1$ irrespective of the angles, which is in line with Lemma 1.4. When the BS has multiple antennas, $g(\varphi_0^0, \varphi_1^0)$ depends strongly on the individual UE angles. There are interference peaks when the two UEs have the same angle (i.e., $\varphi_1^0 = 30^\circ$) and when the angles are each others' mirror reflections (i.e., $\varphi_1^0 = 180^\circ - 30^\circ = 150^\circ$). The function is equal to M at these peaks, because the interfering signal is coherently combined by the MR combining (just as the desired signal). When the ULA can resolve the individual UEs, the interference level instead decreases rapidly (notice the logarithmic vertical scale) and gets generally smaller as M increases. In these cases, the interference level oscillates as the interfering UE's angle is varied, but is approximately $1/M$ times weaker than in the single-antenna case. Hence, the multiple BS antennas help to

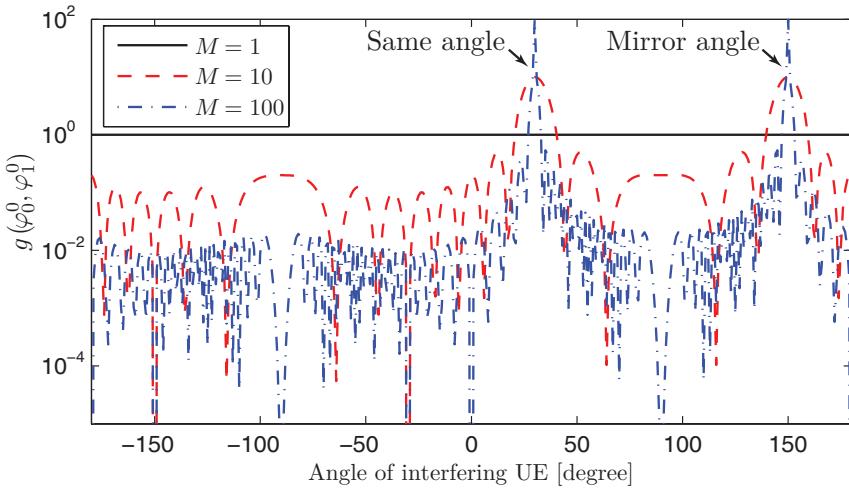


Figure 1.12: The function $g(\varphi_0^0, \varphi_1^0)$ in (1.28) that determines the interference level in an LoS scenario. The desired UE is at the fixed angle $\varphi_0^0 = 30^\circ$ and the interfering UE has a varying angle $\varphi_1^0 \in [-180^\circ, 180^\circ]$.

suppress interference, as long as the UE angles are sufficiently different.

The SE in the NLoS case is harder to interpret since the closed-form expression in (1.29) has a complicated structure with several summations and special functions. Fortunately, we can obtain the following convenient lower bound that is very tight for $M \gg 1$ (see Figure 1.14 for a comparison).

Corollary 1.6. A lower bound on the UL SE in (1.29) for NLoS channels is

$$\text{SE}_0^{\text{NLoS}} = \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_0^0\|^2}{p \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} + \sigma^2} \right) \right\} \geq \log_2 \left(1 + \frac{M-1}{\bar{\beta} + \frac{1}{\text{SNR}_0}} \right). \quad (1.32)$$

Proof. The proof is available in Appendix C.1.5 on p. 586. \square

The SE expression above can be interpreted similarly to the LoS expression in (1.27); it is the logarithm of one plus an SINR expression where the signal power increases as $(M-1)$. A linear array gain is thus obtained for both LoS and NLoS channels. It is the lower-bounding

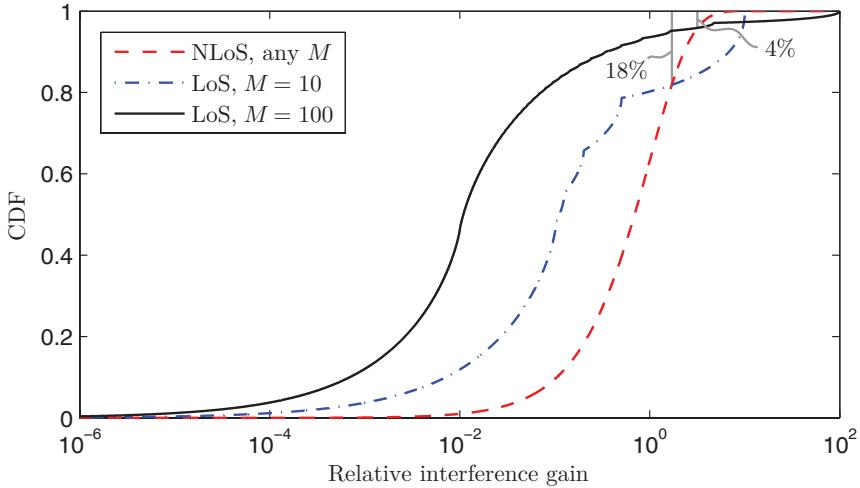


Figure 1.13: CDF of the relative interference gain in (1.33), using logarithmic scale on the horizontal axis. The randomness in the NLoS case is due to Rayleigh fading, while it is due to random UE angles in the LoS cases. The percentages of realizations when LoS gives higher interference gain than NLoS are indicated.

technique used in Corollary 1.6 that made the desired signal scale as $(M - 1)$, instead of M which is the natural array gain obtained with MR combining. However, the difference is negligible when M is large. The interference power in (1.32) is independent of M , in contrast to the LoS case in (1.27) where it decays as $1/M$. This scaling behavior suggests that NLoS channels provide less favorable propagation than LoS channels, but the reality is more complicated. To exemplify this, Figure 1.13 shows the cumulative distribution function (CDF) of the relative interference gain

$$\frac{1}{\beta_1^0} \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} \quad (1.33)$$

which determines how well interference is suppressed by MR combining.

For NLoS channels, (1.33) can be shown to have an $\text{Exp}(1)$ distribution, irrespectively of the value of M . In contrast, (1.33) equals $g(\varphi_0^0, \varphi_1^0)$ in (1.28) for LoS channels, which is a function of M and the UE angles. Figure 1.13 considers the LoS case with $M = 10$ and $M = 100$, and shows the CDF over different uniformly distributed UE angles between

0 and 2π (with $d_H = 1/2$). The CDF of the small-scale fading with NLoS channels is also shown. Figure 1.13 shows that LoS channels often provide several orders-of-magnitude lower interference gains than NLoS channels, but this only applies to the majority of random angle realizations. There is a small probability that the interference gain is larger in LoS than in NLoS; it happens in 18% of the realizations with $M = 10$ and 4% of the realizations with $M = 100$. This corresponds to cases when $\sin(\varphi_0^0) \approx \sin(\varphi_1^0)$ so that the array cannot resolve and separate the UE angles. As discussed earlier, this occurs approximately when $d_H M |\sin(\varphi_0^0) - \sin(\varphi_1^0)| < 0.2$. This happens less frequently for random angles as M increases (for fixed d_H), since the array aperture grows and thus obtains a better spatial resolution. Nevertheless, for any finite M , there will be a small angular interval around φ_0^0 where incoming interference will be amplified just as the desired signal. Since the array is unable to separate UEs with such small angle differences, time-frequency scheduling might be needed to separate them; see Section 7.2.2 on p. 474 for further guidelines for scheduling.

The favorable propagation concept provides a way to quantify the ability to separate UE channels at a BS with many antennas [245]. The channels \mathbf{h}_i^0 and \mathbf{h}_k^0 are said to provide asymptotically favorable propagation if

$$\frac{(\mathbf{h}_i^0)^H \mathbf{h}_k^0}{\sqrt{\mathbb{E}\{\|\mathbf{h}_i^0\|^2\} \mathbb{E}\{\|\mathbf{h}_k^0\|^2\}}} \rightarrow 0 \quad \text{as } M \rightarrow \infty. \quad (1.34)$$

For fading channels, different types of convergence can be considered in (1.34). Herein, we consider almost sure convergence, also known as convergence with probability one, but the literature also contains definitions that build on weaker types of convergence (e.g., convergence in probability). The interpretation of (1.34) is that the channel directions $\mathbf{h}_i^0 / \sqrt{\mathbb{E}\{\|\mathbf{h}_i^0\|^2\}}$ and $\mathbf{h}_k^0 / \sqrt{\mathbb{E}\{\|\mathbf{h}_k^0\|^2\}}$ becomes asymptotically orthogonal. The condition in (1.34) is satisfied for LoS channels as well as for NLoS channels with uncorrelated Rayleigh fading [245]. One can show that the superposition of LoS and NLoS components also satisfies (1.34). Channel measurements with large BS arrays have also confirmed that the UE channels decorrelate as more antennas are added [120, 150]; see

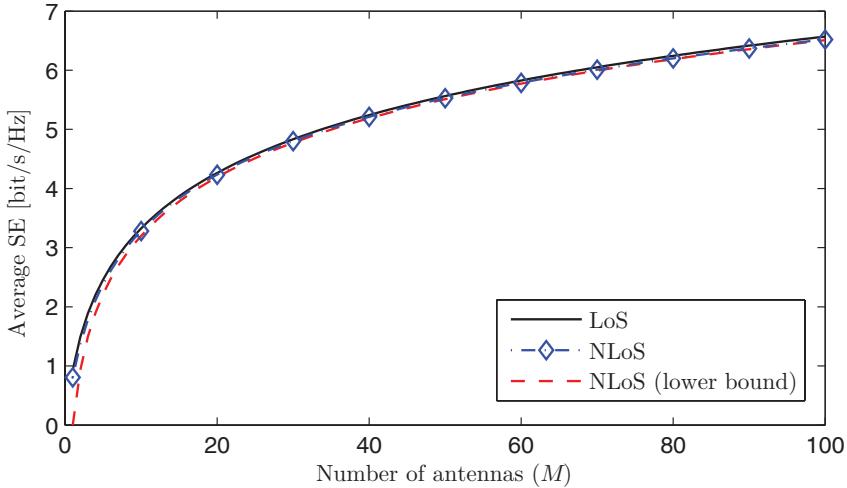


Figure 1.14: Average UL SE as a function of the number of BS antennas M for different channel models. The SNR is $\text{SNR}_0 = 0 \text{ dB}$ and the strength of the inter-cell interference is $\bar{\beta} = -10 \text{ dB}$.

Section 7.3.4 on p. 495 for further details on channel measurements. Note that (1.34) does not imply that channel responses become orthogonal, in the sense that $(\mathbf{h}_i^0)^H \mathbf{h}_k^0 \rightarrow 0$. We later provide a general definition of asymptotically favorable propagation in Section 2.5.2 on p. 233.

Figure 1.14 shows the average SE as a function of the number of BS antennas when the SNR of the desired UE is fixed at $\text{SNR}_0 = 0 \text{ dB}$ and the strength of the inter-cell interference is $\bar{\beta} = -10 \text{ dB}$. The LoS case considers a ULA with $d_H = 1/2$ and the results are averaged over different independent UE angles, all being uniformly distributed from 0 to 2π . Despite the rather poor SNR and interference conditions, Figure 1.14 shows that, by going from $M = 1$ to $M = 10$ antennas, one can improve the SE from 0.8 bit/s/Hz to 3.3 bit/s/Hz. This is achieved thanks to the array gain provided by MR combining. We notice that the lower bound on the SE with NLoS propagation in Corollary 1.6 is very tight for $M > 10$. The SE is a monotonically increasing function of M and grows without limit as $M \rightarrow \infty$, in contrast to the power-scaling case analyzed in Section 1.3.2 where the SE saturated in the high-SNR regime. This is once again due to MR combining, which selectively collects more signal energy from the array, without collecting

more interference energy. The difference between LoS and NLoS is negligible in Figure 1.14 because the channel fading has a gradually smaller impact on the mutual information between the transmitted and received signal as more antennas are added [142]. This is attributed to the spatial diversity from having multiple receive antennas that observe independent fading realizations, which are unlikely to all be nearly zero simultaneously. This phenomenon has been known for a long time; in fact, the early works [257, 117] on multiantenna reception focused on combating channel fading. The term *channel hardening* was used in [142] to describe a fading channel that behaves almost deterministically due to spatial diversity.

In the Massive MIMO literature [243], a channel \mathbf{h}_i^0 is said to provide asymptotic channel hardening if

$$\frac{\|\mathbf{h}_i^0\|^2}{\mathbb{E}\{\|\mathbf{h}_i^0\|^2\}} \rightarrow 1 \quad (1.35)$$

almost surely as $M \rightarrow \infty$. The essence of this result is that the channel variations reduce as more antennas are added, in the sense that the normalized instantaneous channel gain converges to the deterministic average channel gain. It is no surprise that deterministic LoS channels provide channel hardening. More importantly, in NLoS propagation,

$$\frac{\|\mathbf{h}_i^0\|^2}{\mathbb{E}\{\|\mathbf{h}_i^0\|^2\}} = \frac{\|\mathbf{h}_i^0\|^2}{M\beta_i^0} \rightarrow 1 \quad (1.36)$$

almost surely as $M \rightarrow \infty$. This is an example of the strong law of large numbers (see Lemma B.12 on p. 564) and can be interpreted as the variations of $\|\mathbf{h}_i^0\|^2/M$ becoming increasingly concentrated around its mean value $\mathbb{E}\{\|\mathbf{h}_i^0\|^2\}/M = \beta_i^0$ as more antennas are added. This does not mean that $\|\mathbf{h}_i^0\|^2$ becomes deterministic; in fact, its standard deviation grows as \sqrt{M} , while the standard deviation of $\|\mathbf{h}_i^0\|^2/M$ goes asymptotically to zero as $1/\sqrt{M}$. Asymptotic channel hardening can be also proved for other channel distributions, as will be further discussed in Section 2.5.1 on p. 231.

The channel hardening effect for the M -dimensional channel $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{I}_M)$ is illustrated in Figure 1.15. The mean value of the normalized instantaneous channel gain $\|\mathbf{h}\|^2/\mathbb{E}\{\|\mathbf{h}\|^2\}$ and the 10% and 90%

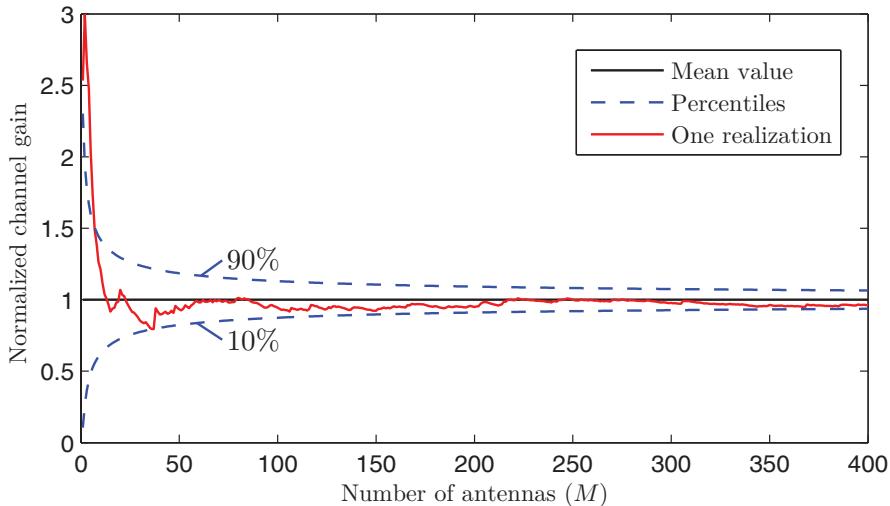


Figure 1.15: Illustration of the channel hardening phenomenon for an M -dimensional channel $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{I}_M)$. The normalized instantaneous channel gain $\|\mathbf{h}\|^2 / \mathbb{E}\{\|\mathbf{h}\|^2\}$ approaches its average value 1 and the standard deviations reduces as $1/\sqrt{M}$.

percentiles are shown for different numbers of antennas. One random realization is also shown. As expected, we have $\|\mathbf{h}\|^2 / \mathbb{E}\{\|\mathbf{h}\|^2\} \approx 1$ when M is large. The convergence towards this limit is gradual, but the approximation is reasonably tight for $M \geq 50$.

In summary, increasing the number of BS antennas improves the SE, which even grows without bound when $M \rightarrow \infty$. This is because the BS can process its received signal over the array to selectively increase the signal gain without collecting more interference. In contrast, increasing the transmit power will increase both the signal and interference equally much and give an upper SE limit. Nevertheless, the SE grows only logarithmically with the number of antennas, as $\log_2(M)$, which does not provide sufficient scalability to achieve any order-of-magnitude improvement in SE in future cellular networks.

Remark 1.3 (Physical limits of large arrays). The scaling behavior obtained by the asymptotic analysis above has been validated experimentally for practical antenna numbers [120, 150]. However, it is important to note that the physics prevent us from letting the size of the array grow

indefinitely as $M \rightarrow \infty$, since the propagation environment is enclosed by a finite volume [281]. Ideally, we can cover the surface of this volume with antennas, and neglect any absorption, to collect all signal energy, but we can never collect more energy than was transmitted. This is not an issue when we deal with hundreds or thousands of antennas since a “large” channel gain of -60 dB in cellular communications implies that we need one million antennas to collect all the transmitted energy. In conclusion, the limit $M \rightarrow \infty$ is not physically achievable, but asymptotic analysis can still be suitable for investigating the system behavior at practically large antenna numbers. Other channel distributions than uncorrelated Rayleigh fading are, however, needed to get reliable results; see Section 2.2 on p. 222 and Section 7.3 on p. 482 for further details.

1.3.3 Uplink Space-Division Multiple Access

Increasing the transmit power or using multiple BS antennas can only bring modest improvements to the UL SE, as previously shown. This is because these methods improve the SINR, which appears inside the logarithm of the SE expression, thus the SE increases slowly. We would like to identify a way that improves the SE at the outside of the logarithm instead. Since the logarithmic expressions in Lemmas 1.4 and 1.5 describe the SE of the channel between a particular UE and its serving BS, we can potentially serve multiple UEs, say K UEs, simultaneously in each cell and achieve a sum SE that is the summation of K SE expressions of the types in Lemmas 1.4 and 1.5. An obvious bottleneck of such multiplexing of UEs is the co-user interference that increases with K and now appears also within each cell. The intra-cell interference can be much stronger than the inter-cell interference and needs to be suppressed if a K -fold increase in SE is actually to be achieved.

Space-division multiple access (SDMA) was conceived in the late 1980s and early 1990s [349, 308, 17, 280, 125, 373] to handle the co-user interference in a cell by using multiple antennas at the BS to reject interference by spatial processing. Multiple field-trials were carried out in the 1990s, using (at least) up to ten antennas [15, 96, 16]. The

information-theoretic capacity¹⁴ of these systems was characterized in the early 2000s and described in [74, 129, 335, 342, 366, 127] for single-cell systems, where the terminology “multiuser MIMO” was used. Note that the K UEs are the multiple inputs and the M BS antennas are the multiple outputs, thus the MIMO terminology is used irrespective of how many antennas each UE is equipped with.¹⁵ Extensions of multiuser MIMO to cellular networks have been developed and surveyed in papers such as [276, 33, 294, 46, 126, 208], but the exact capacity is hard to obtain in this case.

We will now analyze a cellular network with UL SDMA transmission by assuming that there are K active UEs in each cell, as previously illustrated in Figure 1.8. The channel response between the k th desired UE in cell 0 and the serving BS is denoted by $\mathbf{h}_{0k}^0 \in \mathbb{C}^M$ for $k = 1, \dots, K$, while the channel responses from the interfering UEs in cell 1 to the BS in cell 0 are denoted by $\mathbf{h}_{1i}^0 \in \mathbb{C}^M$ for $i = 1, \dots, K$. Notice that the subscript still indicates the identity of the UE, while the superscript is the index of the receiving BS. The received multiantenna UL signal in (1.22) is then generalized to

$$\mathbf{y}_0 = \underbrace{\sum_{k=1}^K \mathbf{h}_{0k}^0 s_{0k}}_{\text{Desired signals}} + \underbrace{\sum_{k=1}^K \mathbf{h}_{1k}^0 s_{1k}}_{\text{Interfering signals}} + \underbrace{\mathbf{n}_0}_{\text{Noise}} \quad (1.37)$$

where $s_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, p)$ is the signal transmitted by the k th UE in cell j and the receiver noise $\mathbf{n}_0 \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \sigma^2 \mathbf{I}_M)$ is the same as before.

We consider the same LoS and NLoS propagation models as before. More precisely, the LoS channel response for UE k in cell j is

$$\mathbf{h}_{jk}^0 = \sqrt{\beta_j^0} \begin{bmatrix} 1 & e^{2\pi j d_H \sin(\varphi_{jk}^0)} & \dots & e^{2\pi j d_H (M-1) \sin(\varphi_{jk}^0)} \end{bmatrix}^T \quad (1.38)$$

¹⁴When there are K UEs in the network, the conventional one-dimensional capacity notion generalizes to a K -dimensional capacity region that represents the set of capacities that the K UEs can achieve simultaneously. The sum capacity represents one point in this region and has gained particular traction since it is the one-dimensional metric that describes the aggregate capacity of the network. This and other operating points are further described in Section 7.1 on p. 452.

¹⁵The terminology “multiuser SIMO” was used in the 1990s for the case of SDMA with single-antenna UEs [254], but nowadays the information-theoretic multiuser MIMO terminology dominates and it is adopted in this monograph.

where $\varphi_{jk}^0 \in [0, 2\pi)$ is the azimuth angle relative to the boresight of the BS array in cell 0. In the NLoS case, the corresponding channel response between UE k in cell j and the BS in cell 0 is defined as

$$\mathbf{h}_{jk}^0 \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_M, \beta_j^0 \mathbf{I}_M \right) \quad (1.39)$$

and assumed to be statistically independent between UEs. Recall that we use the Wyner model in which, for simplicity, the average channel gain β_j^0 is assumed to be the same for all UEs in cell j .

Since the BS in cell 0 receives a superposition of the signals transmitted by its K desired UEs, it needs to process the received signal in (1.37) to separate the UEs in the spatial domain—simply speaking, by directing its hearing towards the location of each desired UE. The separation of UEs is more demanding in SDMA than in conventional time-frequency multiplexing of UEs, because it requires the BS to have knowledge of the channel responses [127]. For example, the BS in cell 0 can use knowledge of its k th UE's channel response to tailor a receive combining vector $\mathbf{v}_{0k} \in \mathbb{C}^M$ to this UE channel. This vector is multiplied with the received signal in (1.37) to obtain

$$\mathbf{v}_{0k}^H \mathbf{y}_0 = \underbrace{\mathbf{v}_{0k}^H \mathbf{h}_{0k}^0 s_{0k}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^K \mathbf{v}_{0k}^H \mathbf{h}_{0i}^0 s_{0i}}_{\text{Intra-cell interference}} + \underbrace{\sum_{i=1}^K \mathbf{v}_{0k}^H \mathbf{h}_{1i}^0 s_{1i}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{v}_{0k}^H \mathbf{n}_0}_{\text{Noise}}. \quad (1.40)$$

The purpose of the receive combining is to make the desired signal much stronger than the sum of interfering signals and noise. MR combined with

$$\mathbf{v}_{0k} = \mathbf{h}_{0k}^0 \quad (1.41)$$

is a popular suboptimal choice since it maximizes the relative power $|\mathbf{v}_{0k}^H \mathbf{h}_{0k}^0|^2 / \|\mathbf{v}_{0k}\|^2$ of the desired signal, but it is not the optimal choice when there are interfering signals [28, 348, 349]. The receive combining design for multiuser MIMO is analytically similar to multiuser detection in code-division multiple access (CDMA) [202, 205, 106] and the key methods were developed at roughly the same time. In Section 4.1 on p. 275, we will show that it is the *multicell minimum mean-squared*

error (*M-MMSE*) combining vector

$$\mathbf{v}_{0k} = p \left(p \sum_{i=1}^K \mathbf{h}_{0i}^0 (\mathbf{h}_{0i}^0)^H + p \sum_{i=1}^K \mathbf{h}_{1i}^0 (\mathbf{h}_{1i}^0)^H + \sigma^2 \mathbf{I}_M \right)^{-1} \mathbf{h}_{0k}^0 \quad (1.42)$$

that maximizes the UL SE in cellular networks. This combining scheme has received its name from the fact that it also minimizes the mean-squared error (MSE) $\mathbb{E}\{|s_{0k} - \mathbf{v}_{0k}^H \mathbf{y}_0|^2\}$ between the desired signal s_{0k} and the receive combined signal $\mathbf{v}_{0k}^H \mathbf{y}_0$, where the expectation is with respect to the transmit signals (while the channels are considered deterministic). Interfering signals from all cells are taken into account in M-MMSE combining and the matrix inverse in (1.42) has a role similar to that of a whitening filter in classic signal processing [175]. M-MMSE combining maximizes the SINR by finding the best balance between amplifying the desired signal and suppressing interference in the spatial domain. The price to pay is the increased computational complexity from inverting a matrix and the need to learn the matrix that is inverted in (1.42).

The next lemma provides closed-form SE expressions for the case of MR combining. M-MMSE combining will be studied by simulations.

Lemma 1.7. If the BS in cell 0 knows the channel responses of all UEs and applies MR combining to detect the signals from each of its K desired UEs, then an achievable UL sum SE [bit/s/Hz/cell] in the LoS case is

$$\text{SE}_0^{\text{LoS}} = \sum_{k=1}^K \log_2 \left(1 + \frac{M}{\sum_{\substack{i=1 \\ i \neq k}}^K g(\varphi_{0k}^0, \varphi_{0i}^0) + \bar{\beta} \sum_{i=1}^K g(\varphi_{0k}^0, \varphi_{1i}^0) + \frac{1}{\text{SNR}_0}} \right) \quad (1.43)$$

with $g(\cdot, \cdot)$ being defined in (1.28).

With NLoS channels, an achievable UL sum SE [bit/s/Hz/cell] and

a closed-form lower bound are

$$\begin{aligned} \text{SE}_0^{\text{NLoS}} &= \sum_{k=1}^K \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_{0k}^0\|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sum_{i=1}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{1i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sigma^2} \right) \right\} \\ &\geq K \log_2 \left(1 + \frac{M - 1}{(K - 1) + K \bar{\beta} + \frac{1}{\text{SNR}_0}} \right). \end{aligned} \quad (1.44)$$

Proof. The proof is available in Appendix C.1.6 on p. 587. \square

The sum SE expressions in Lemma 1.7 have similar forms as the ones in Lemma 1.5 and Corollary 1.6, but are more complicated due to the addition of intra-cell interference and the greater amount of inter-cell interference. In the LoS case, SDMA results in the summation of K SE expressions, one per desired UE. The desired signal gains inside the logarithms increase linearly with M and thus every UE experiences the full array gain when using MR combining. The drawback of SDMA is seen from the denominator, where the interference terms contain contributions from $K - 1$ intra-cell UEs and K inter-cell UEs. Each interference term has the same form as in the single-user case in Lemma 1.5, thus one can expect the interference to be the lowest when the UEs have well-separated angles (to avoid the worst cases illustrated in Figure 1.12). Recall from (1.30) that the function $g(\varphi, \psi)$ decreases as $1/M$ for any $\sin(\varphi) \approx \sin(\psi)$. In conjunction with the array gain of the desired signal, we can thus serve multiple UEs and still maintain roughly the same SINR per UE if M is increased proportionally to \sqrt{K} to counteract the increased interference.¹⁶

The NLoS case in Lemma 1.7 generalizes the lower bound in Corollary 1.6 to $K \geq 1$ and the bound is tight for $M \gg 1$. An exact closed-form expression similar to (1.29) can also be obtained, but it contains many summations and is omitted since it does not provide

¹⁶To obtain this scaling behavior, we notice that the desired signal power grows as M and the interference power is proportional to K/M , due to the bound in (1.30). The signal-to-interference ratio becomes M^2/K and thus it is sufficient to scale M proportionally to \sqrt{K} to achieve a constant signal-to-interference ratio as K grows.

any additional insight. The gain from SDMA is easily seen from (1.44); there is a factor K in front of the logarithm that shows that the sum SE increases proportionally to the number of UEs. This multiplicative factor is known as the *multiplexing gain* and achieving this gain is the main point with SDMA. Inside the logarithm, the desired signal power increases linearly with M , while the intra-cell interference power $K - 1$ and the inter-cell interference power $K \bar{\beta}$ increase linearly with K . This means that, as we add more UEs, we can counteract the increasing interference by adding a proportional amount of additional BS antennas; more precisely, we can maintain roughly the same SINR per UE by increasing M jointly with K to keep the antenna-UE ratio M/K fixed. Interestingly, this means that more antennas are needed to suppress interference with MR combining in the NLoS case than in the LoS case, where M only needs to increase as \sqrt{K} . The explanation is that all interfering UEs cause substantial interference in the NLoS case, while only the ones with sufficiently similar angles to the desired UE does that in the LoS case (and the angular interval where this happens decreases with M).

To exemplify these behaviors, Figure 1.16 shows the average sum SE as a function of the number of UEs per cell, for either $M = 10$ or $M = 100$ antennas. The sum SE with MR combining is shown in Figure 1.16a based on the analytic formulas from Lemma 1.7, while Monte-Carlo simulations are used for M-MMSE combining in Figure 1.16b. In both cases, the SNR is fixed at $\text{SNR}_0 = 0$ dB and the strength of the inter-cell interference is $\bar{\beta} = -10$ dB. The antenna spacing is $d_H = 1/2$ in the LoS case and the results are averaged over different independent UE angles, all being uniformly distributed from 0 to 2π .

Figure 1.16 shows that the sum SE is a slowly increasing function of K in the case of $M = 10$, because the BS does not have enough spatial degrees of freedom to separate the UEs—neither by MR nor by M-MMSE combining. The behavior is completely different when $M = 100$ antennas are used since the channel response of each UE is then a 100-dimensional vector but there are only up to 20 UEs per cell so the UE channels only span a small portion of the spatial dimensions that the BS can resolve. Consequently, the sum SE increases almost

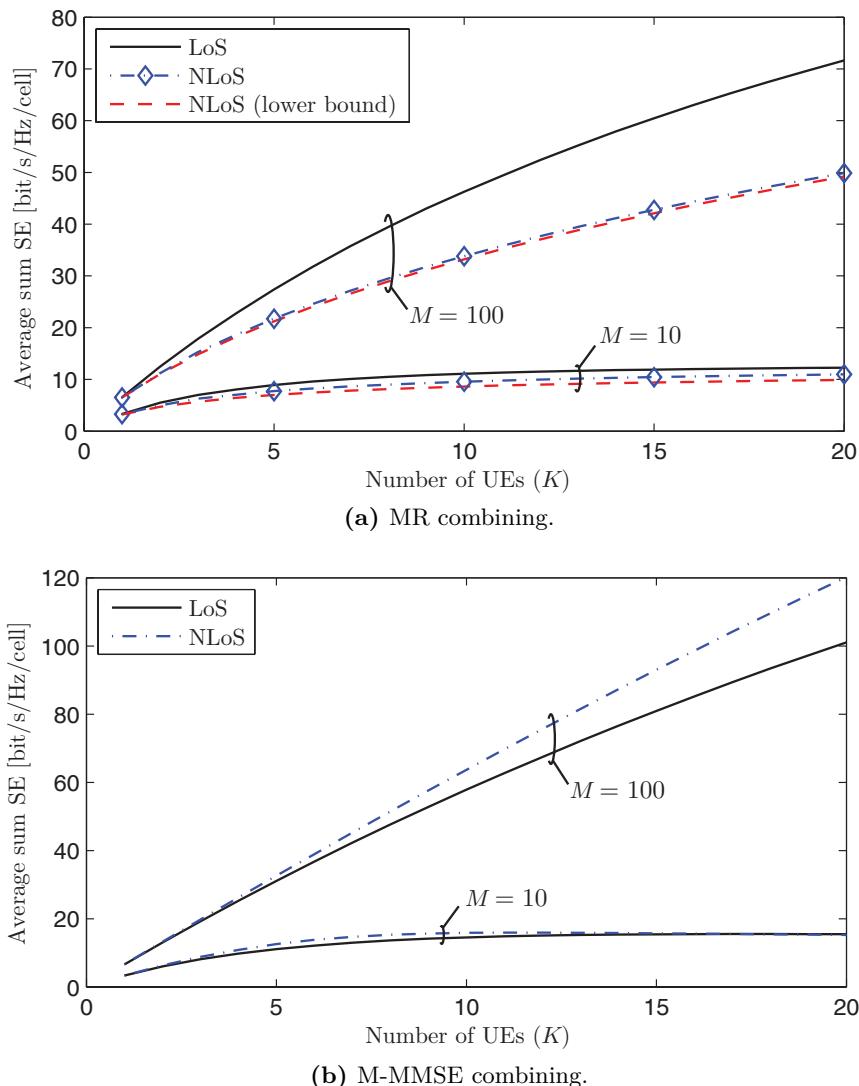


Figure 1.16: Average UL sum SE as a function of the number of UEs per cell for different combining schemes, different channel models, and either $M = 10$ or $M = 100$ BS antennas. The SNR is $\text{SNR}_0 = 0$ dB and the strength of the inter-cell interference is $\bar{\beta} = -10$ dB. The sum SE grows linearly with K as long as M/K remains large. M-MMSE rejects interference more efficiently than MR.

linearly with the number of UEs and we can achieve a roughly K -fold improvement in sum SE over a single-user scenario. For example, we achieve an SE of 3.3 bit/s/Hz/cell with $(M, K) = (10, 1)$ using MR/M-MMSE combining and can increase it to 71.6 bit/s/Hz/cell with MR and 101 bit/s/Hz/cell with M-MMSE for $(M, K) = (100, 20)$. This corresponds to $21\times$ and $31\times$ gains in SE, respectively. These numbers were selected from the LoS curves, because the NLoS case shows some interesting behaviors that deserve further discussion. The sum SE is considerably lower with NLoS than with LoS when using MR combining, while we get the opposite result when using M-MMSE combining. The reason for this is that each UE is affected by interference from many UEs in the NLoS case, while only a few UEs with similar angles cause strong interference in the LoS case. If the interference is ignored, as with MR combining, the SE is lower in the NLoS case due to the larger sum interference power. However, it is easier for M-MMSE combining to reject interference in NLoS than in LoS, where there might be a few UEs with channels that are nearly parallel to the desired UE's channel. That is why the SE is higher in the NLoS when using M-MMSE.

We now consider cases wherein M is increased proportionally to K , to suppress the inter-user interference that increases with K . The proportionality constant M/K is called *antenna-UE ratio*. Figure 1.17 shows the sum SE obtained by M-MMSE combining, as a function of K for different antenna-UE ratios: $M/K \in \{1, 2, 4, 8\}$. The SE grows almost linearly with K in all four cases, as expected from Lemma 1.7. The steepness of the curves increases as M/K increases, since it becomes easier to suppress the interference when $M \gg K$. Looking at the NLoS case with $K = 10$, the first doubling of the number of antennas (from $M/K = 1$ to $M/K = 2$) gives a 94% gain in SE, while the second doubling gives another 51% gain and the third doubling gives yet another 29% gain. Since the relative improvements are decaying, we say that $M/K \geq 4$ is the preferred operating regime for multiuser MIMO.¹⁷ The LoS and NLoS cases once again provide comparable results.

¹⁷We will revisit this statement in Section 7.2.2 on p. 474, where scheduling is discussed. By taking the channel estimation overhead into account, we will show that for a given M there is a particular K that maximizes the sum SE.

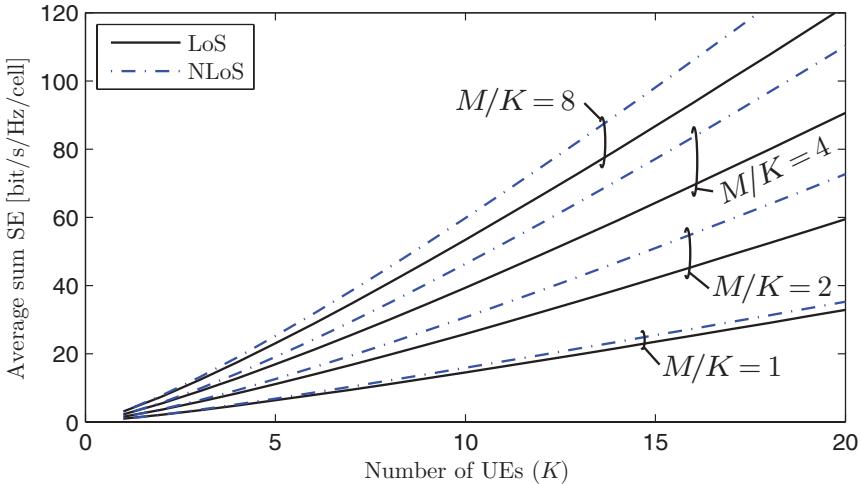


Figure 1.17: Average UL sum SE with M-MMSE combining as a function of the number of UEs per cell, when the number of antennas increases with K with different fixed antenna-UE ratios M/K . The SNR is $\text{SNR}_0 = 0 \text{ dB}$ and the strength of the inter-cell interference is $\bar{\beta} = -10 \text{ dB}$. The sum SE grows as M/K increases.

M-MMSE is the linear receive combining scheme that maximizes the SE. The basic characteristic of linear schemes is that they treat interference as spatially colored noise. From a channel capacity perspective, this is only optimal when the interference between each pair of UEs is sufficiently small [230, 296, 20, 21, 295]. The information theory for interference channels proves that strong interference sources should be canceled using non-linear receiver processing schemes, such as successive interference cancellation, before the desired signals are decoded [314]. However, such schemes are rather impractical, since one needs to store large blocks of received signals and then decode the UEs' data sequentially, leading to high complexity, large memory requirements, and latency issues. If we would limit ourselves to linear receiver processing schemes, how large is the performance loss?

Figure 1.18 quantifies the performance loss of linear receiver processing as compared to non-linear receiver processing, as a function of the number of UEs. The figure shows the ratio between the average UL sum SE achieved by M-MMSE combining and by successive interfer-

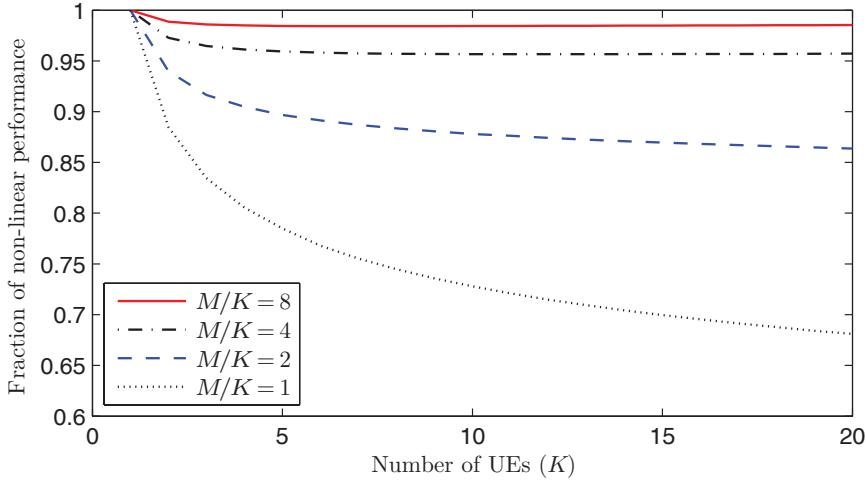


Figure 1.18: Ratio between the average UL sum SE achieved with M-MMSE combining and with non-linear receiver processing, as a function of the number of UEs per cell. The number of antennas M increases with K for different fixed antenna-UE ratios: $M/K \in \{1, 2, 4, 8\}$. The SNR is $\text{SNR}_0 = 0 \text{ dB}$ and the strength of the inter-cell interference is $\bar{\beta} = -10 \text{ dB}$.

ence cancelation, where the intra-cell signals are decoded sequentially while treating inter-cell interference as noise [314]. The setup is the same as in the previous figure, but we only consider NLoS propagation for simplicity. The non-linear scheme performs much better for $M/K = 1$, in which case M-MMSE only achieves 70%–80% of its sum SE. The performance difference reduces quickly as M/K increases. For $M/K = 4$ and $M/K = 8$, we only lose a few percentages in sum SE by using M-MMSE instead of the non-linear scheme, even if there is as much as 20 UEs. The interpretation is that the favorable propagation, achieved by having many BS antennas, makes the interference between each pair of UEs sufficiently small to make linear receiver processing nearly optimal. When there are many active UEs, the total interference caused to a UE can indeed be large, but nevertheless, linear processing performs well since the interference between each pair of UEs is small. Similar observations have been made in the overview articles [50, 209, 210].

In summary, UL SDMA transmission can increase the sum SE per cell by more than one order-of-magnitude. This is achieved by serving K UEs simultaneously and increasing the number of BS antennas to achieve an array gain that counteracts the increased interference. This leads to an operating regime with antenna-UE ratio $M/K \geq c$, for some preferably large value c , where we can provide K -fold gains in sum SE. This is the type of highly scalable SE improvements that are needed to handle much higher data volumes in the coverage tier of future cellular networks. Note that the SE per UE is not dramatically changed, thus the use of more spectrum is still key to improving the throughput per UE. The sum SE gains are achievable with both LoS and NLoS channels, using either MR combining that maximizes the array gain or M-MMSE combining that also suppresses interference to maximize the SE. Non-linear processing schemes can only bring minor performance improvements in the preferable operating regime and are therefore not considered in the remainder of this monograph.

Remark 1.4 (Multiantenna UEs). We have shown above that SDMA transmission with many single-antenna UEs and an even larger number of BS antennas achieves high sum SE. What would happen if the UEs were also equipped with multiple antennas? The cost, size, and complexity of each UE will certainly increase. The positive effect is that a UE with N_{UE} antennas can transmit up to N_{UE} simultaneous data streams to its serving BS. From the BS's perspective, each stream can be treated as a signal from a separate “virtual” UE and the signal can only be distinguished if it has a different spatial directivity than the other signals. This means that the vector that describes the channel response from the BS to the n th antenna of a particular UE should be nearly orthogonal to the other antennas' channel responses (for $n = 1, \dots, N_{\text{UE}}$). In NLoS propagation, this is achieved when the UE antennas observe nearly uncorrelated random channel realizations, which is possible in a rich scattering environment with an adequate antenna spacing. Channel orthogonality is much harder to achieve in LoS propagation since the angle between the BS and a UE in the far-field is roughly the same for all the antennas at the UE; recall from (1.28) that the inner product $g(\varphi, \psi)$ between LoS channel responses with angles φ

and ψ is large whenever $\varphi \approx \psi$. Hence, the benefit of sending multiple data signals cannot be exploited in propagation environments with only a dominating LoS path. The UE can, however, achieve an additional array gain proportional to N_{UE} by coherently combining the signals over N_{UE} antennas, if it knows the channel responses. This monograph focuses on single-antenna UEs, but the results can be readily applied to N_{UE} -antenna UEs by viewing them as N_{UE} virtual UEs that transmit N_{UE} separate signals, representing different data streams. The paper [194] considers multiantenna UEs and shows that the SE is maximized when a particular number of data streams are received/transmitted per cell (see Section 7.2.2 on p. 474 for a further discussion). Suppose this number of streams is K^*_{stream} and that each UE is allocated as many streams as it has antennas. The analysis in [194] indicates that roughly the same sum SE is achieved when having K UEs that are equipped with N_{UE} antennas and when having $N_{\text{UE}}K$ single-antenna UEs. Hence, the distinct advantage of having multiple UE antennas occurs at low user load, $K < K^*_{\text{stream}}$, where the only way to send all K^*_{stream} streams is to allocate multiple streams per UE.

1.3.4 Downlink Space-Division Multiple Access

This section has so far focused on the UL, where we have identified SDMA as a suitable way to improve the SE by an order-of-magnitude or more. We will now describe how SDMA is applied in the DL. We continue to use the Wyner model, which is illustrated in Figure 1.19 for the DL. The main difference from the UL in Figure 1.8 is that the signals are transmitted from BSs instead of from UEs. There are K active UEs in each cell and the serving BS sends a separate signal to each of them using linear transmit precoding from an array of M antennas. Precoding means that each data signal is sent from all antennas, but with different amplitude and phase to direct the signal spatially. This is also called beamforming, but we refrain from using this terminology since it can give the misleading impression that a signal beam is always formed in a particular angular direction and that analog phase-shifters are used. In contrast, precoding means that each antenna's transmit signal is generated separately in the digital baseband, which gives full

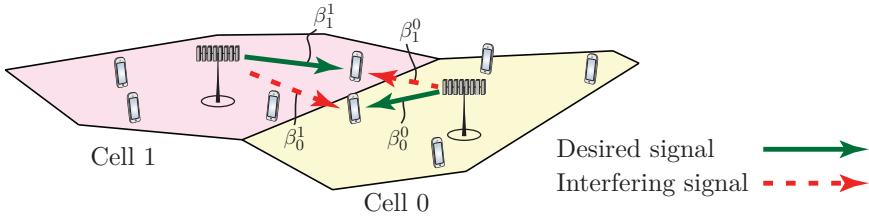


Figure 1.19: Illustration of the notion of desired and interfering DL signals in a two-cell network. In the Wyner model, every UE in cell 0 has the same value of the average channel gains β_0^0 and β_1^0 , while every UE in cell 1 has the same value of β_0^1 and β_1^1 .

flexibility in the signal generation.¹⁸ Angular beams are a special case of precoding that is useful in LoS propagation, but for NLoS channels the transmitted signal might not have a distinct angular directivity, but can still be precoded such that the multipath components are received coherently at the UE.

Similar to the UL, the DL channel response between the BS in cell 0 and its k th desired UE is denoted by $(\mathbf{h}_{0k}^0)^H$ for $k = 1, \dots, K$. The DL channel response between the BS in cell 1 and the k th UE in cell 0 is denoted by $(\mathbf{h}_{0k}^1)^H$. The transpose represents the fact that we are now looking at the channel from the opposite direction, while the complex conjugate is added for notational convenience. There is no such conjugation in practice, but it simplifies the notation and does not change the SE.

The received DL signal $z_{0k} \in \mathbb{C}$ at UE k in cell 0 is modeled as

$$\begin{aligned}
 z_{0k} = & \underbrace{(\mathbf{h}_{0k}^0)^H \mathbf{w}_{0k} \varsigma_{0k}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^K (\mathbf{h}_{0k}^0)^H \mathbf{w}_{0i} \varsigma_{0i}}_{\text{Intra-cell interference}} \\
 & + \underbrace{\sum_{i=1}^K (\mathbf{h}_{0k}^1)^H \mathbf{w}_{1i} \varsigma_{1i}}_{\text{Inter-cell interference}} + n_{0k} \quad (1.45)
 \end{aligned}$$

where $\varsigma_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, p)$ is the signal transmitted to the k th UE in cell j

¹⁸An animation of precoding is found at <https://youtu.be/XBb481RNqGw>.

and $\mathbf{w}_{jk} \in \mathbb{C}^M$ is the corresponding unit-norm precoding vector (i.e., $\|\mathbf{w}_{jk}\| = 1$) that determines the spatial directivity of the signal. The receiver noise at this UE is denoted by $n_{0k} \sim \mathcal{N}_\mathbb{C}(0, \sigma^2)$.

We consider the same LoS and NLoS propagation models as before. In the LoS case, we have the multiple-input single-output (MISO) channel response

$$\mathbf{h}_{jk}^l = \sqrt{\beta_j^l} \begin{bmatrix} 1 & e^{2\pi j d_H \sin(\varphi_{jk}^l)} & \dots & e^{2\pi j d_H (M-1) \sin(\varphi_{jk}^l)} \end{bmatrix}^\top \quad (1.46)$$

between UE k in cell j and the BS in cell l , where $\varphi_{jk}^l \in [0, 2\pi)$ is the azimuth angle relative to the boresight of the transmitting BS array. In the NLoS case, the corresponding channel response is

$$\mathbf{h}_{jk}^l \sim \mathcal{N}_\mathbb{C}(\mathbf{0}_M, \beta_j^l \mathbf{I}_M) \quad (1.47)$$

and is assumed to be independent between UEs. Recall from (1.12) that we use the same notation, $\bar{\beta} = \beta_0^1 / \beta_0^0$, for the relative strength of inter-cell interference in the DL as in the UL.

The precoding vectors \mathbf{w}_{jk} , for $k = 1, \dots, K$ and $j = 0, 1$, can be selected in a variety of ways. As seen from the received signal in (1.45), each UE is affected by all the precoding vectors; the own precoding vector is multiplied with the channel response from the serving BS, while the other ones cause interference and are multiplied with the channel response from the corresponding transmitting BSs. Hence, the precoding vectors should be selected carefully in the DL, based on knowledge of the channel responses. We will study this in detail in Section 4.3 on p. 316, but for now we consider MR precoding with

$$\mathbf{w}_{jk} = \frac{\mathbf{h}_{jk}^j}{\|\mathbf{h}_{jk}^j\|}. \quad (1.48)$$

This precoding vector focuses the DL signal at the desired UE to achieve the maximum array gain, similar to MR combining in the UL. Note that $\|\mathbf{w}_{jk}\|^2 = 1$, which implies that the total transmit power of the BS is constant, irrespective of the number of antennas. Consequently, the transmit power per BS antenna decreases roughly as $1/M$. The following lemma provides SE expressions for MR precoding.

Lemma 1.8. If the BSs use MR precoding and the UEs in cell 0 know their respective effective channels $(\mathbf{h}_{0k}^0)^H \mathbf{w}_{0k}$ and the interference variance, then an achievable DL sum SE [bit/s/Hz/cell] in the LoS case is

$$\text{SE}_0^{\text{LoS}} = \sum_{k=1}^K \log_2 \left(1 + \frac{M}{\sum_{\substack{i=1 \\ i \neq k}}^K g(\varphi_{0i}^0, \varphi_{0k}^0) + \bar{\beta} \sum_{i=1}^K g(\varphi_{1i}^1, \varphi_{0k}^1) + \frac{1}{\text{SNR}_0}} \right). \quad (1.49)$$

With NLoS channels, a DL sum SE [bit/s/Hz/cell] and a closed-form lower bound are

$$\begin{aligned} \text{SE}_0^{\text{NLoS}} &= \sum_{k=1}^K \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_{0k}^0\|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0i}^0\|^2} + \sum_{i=1}^K p \frac{|(\mathbf{h}_{0k}^1)^H \mathbf{h}_{1i}^1|^2}{\|\mathbf{h}_{1i}^1\|^2} + \sigma^2} \right) \right\} \\ &\geq K \log_2 \left(1 + \frac{(M-1)}{(K-1) \frac{M-1}{M} + K \bar{\beta} + \frac{1}{\text{SNR}_0}} \right). \end{aligned} \quad (1.50)$$

Proof. The proof is available in Appendix C.1.7 on p. 588. \square

The DL sum SE in this lemma is very similar to the UL sum SE in Lemma 1.7. The NLoS case only differs in the extra multiplicative term $\frac{M-1}{M}$ in the denominator of (1.50), which is almost one for large M . The LoS case only differs in the angles that appear in each expression; all angles in the UL are from UEs to the BS in cell 0, while the DL includes both the angles from the desired UE to all transmitting BSs and the angles from the other UEs that these BSs are transmitting to (representing the directivity of each DL signal). Some of the similarities are induced by the Wyner model since we have assumed that the inter-cell interference is equally strong in the UL and DL (i.e., $\beta_0^1 = \beta_1^0$); in general, there are also differences in the average channel gains, as we elaborate on in Section 4.3.2 on p. 320. Nonetheless, when using the Wyner model, the UL simulations in Figures 1.16–1.17 are representative

for the DL performance as well—no additional simulations are needed to uncover the basic behaviors.

The array gain is M with MR processing in both UL and DL, but it is obtained differently. In the UL, the BS makes M observations of the desired signal over its M receive antennas, each being corrupted by an independent noise term. By coherently combining the M signal components, the signal power grows proportionally to M while the noise realizations add incoherently so that the noise variance is unchanged. In the DL, the M transmit antennas have different channels to the receiving UE. Since the total transmit power is fixed, the signal power per antenna is reduced as $1/M$ and the signal amplitude as $1/\sqrt{M}$. With precoding that makes the M transmitted signal components add coherently at the UE, the received signal's amplitude grows as $M/\sqrt{M} = \sqrt{M}$ and the received signal power therefore grows as M .

1.3.5 Acquiring Channel State Information

The channel responses, \mathbf{h}_{jk}^j , are utilized by BS j to process the UL and DL signals. We have assumed so far that the channel responses are known perfectly, but in practice, these vectors need to be estimated regularly. More precisely, the channel responses are typically only constant for a few milliseconds and over a bandwidth of a few hundred kHz. A random distribution is commonly used to model the channel variations. The current set of channel response realizations is called the *channel state* and the knowledge that the BSs have of them is referred to as the *channel state information (CSI)*. Full statistical CSI regarding the distributions¹⁹ of random variables is assumed to be available anywhere in the network, while instantaneous CSI regarding the current channel realizations need to be acquired at the same pace as the channels change. The main method for CSI acquisition is pilot signaling, where a predefined pilot signal is transmitted from an antenna. As illustrated in Figure 1.20, any other antenna in the network can simultaneously receive the transmission and compare it with the known pilot signal to

¹⁹It is in many cases sufficient to know the first- and second-order moments of the random variables, but for simplicity we assume that the full distributions are available.

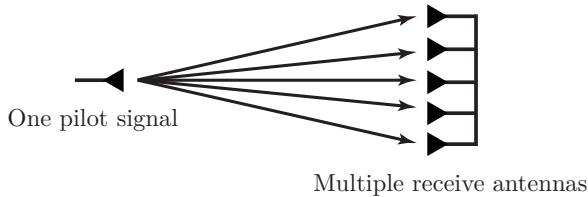


Figure 1.20: When an antenna is transmitting a pilot signal, any number of receive antennas can simultaneously receive the pilot signal and use it to estimate their respective channels to the transmitter.

estimate the channel from the transmitting antenna. If we instead need to estimate the channel response from two transmitting antennas, two orthogonal pilot signals are generally required to separate the signals from the two antennas [182, 195, 38]. The orthogonality is achieved by spending two samples on the transmission, as further explained in Section 3.1 on p. 244. The number of orthogonal pilot signals is proportional to the number of transmit antennas, while any number of receive antennas can “listen” to the pilots simultaneously and estimate their individual channels to the transmitters.

Every pilot signal that is transmitted could have been a signal that carried payload data, thus we want to minimize this overhead caused by pilot signaling. In SDMA, there are key differences between UL and DL in terms of the overhead for channel acquisition. There are K single-antenna UEs per cell and thus K pilot signals are required to estimate the channels in the UL. Similarly, there are M antennas at the BS and thus M pilot signals are required to estimate the channels in the DL. Since having an antenna-UE ratio $M/K \geq 4$ is the preferable operating regime in SDMA, the overhead from sending DL pilots is typically much larger than that from UL pilots. A BS antenna is only useful if we know the channel response, which limits the number of BS antennas that we can utilize in practice, unless we can find a workaround.

The UL and DL can be separated in either time or frequency; see Figure 1.21. If the UL and DL are separated in time, using a time-division duplex (TDD) protocol, then the channel responses are *reciprocal*²⁰ [254].

²⁰The physical propagation channels are reciprocal, but the transceiver chains are

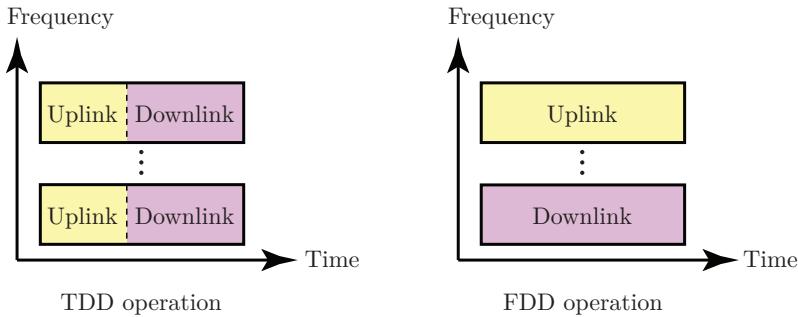


Figure 1.21: Illustration of two ways to divide a block of time/frequency resources between UL and DL. Each solid box represents a time-frequency block where the channel responses are constant and need to be estimated.

This means that the channel response is the same in both directions and can be estimated at the BS using only K UL pilots. Only the BS in cell j needs to know the complete channel response \mathbf{h}_{jk}^j to its k th UE, while the corresponding UE only needs to know the effective scalar channel $g_{jk} = (\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}$ that is obtained after precoding. Since the value of g_{jk} is constant as long as the channels are constant, it can be estimated blindly from the DL payload data signals, irrespective of the channel distribution [243].²¹ For example, the BS can use its CSI to adjust the phase of \mathbf{w}_{jk} so that the phase of g_{jk} becomes (nearly) deterministic, thereby mainly the magnitude $|g_{jk}|$ needs to be estimated. Channel hardening improves the estimation quality since the relative variations in $|g_{jk}|/\mathbb{E}\{|g_{jk}|\}$ becomes smaller. Consequently, a TDD protocol only requires K pilots, independently of the number of antennas, M .

If the UL and DL are instead separated in frequency, using a frequency-division duplex (FDD) protocol, then the UL and DL channels are always different and we cannot rely on reciprocity. Hence, we need to send pilots in both UL and DL. In addition, the estimates of the DL channel responses need to be fed back to the BS, to enable DL precoding computation. The feedback overhead is approximately

generally not fully reciprocal. This is further discussed in Section 6.4.4 on p. 445.

²¹DL pilot signals can be utilized to improve the estimation quality, but this does not necessarily improve the SE since the overhead for channel estimation increases [243].

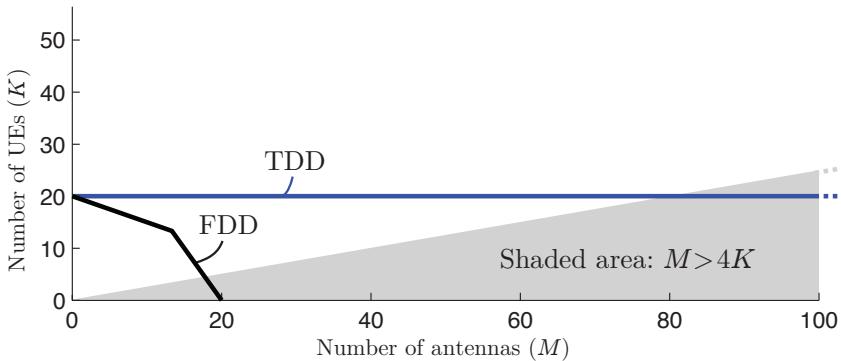


Figure 1.22: Illustration of the operating points (M, K) that are supported by using $\tau_p = 20$ pilots, for TDD and FDD protocols. The shaded area corresponds to the preferable operating points for SDMA systems. The TDD protocol is scalable with respect to the number of antennas and the number of UEs that can be supported is only limited by τ_p .

the same as that of sending $\max(M, K)$ additional UL pilot signals.²² The precoded channels $g_{jk} = (\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}$ can be estimated from the DL signals, as described for TDD above. Hence, an FDD protocol has a pilot/feedback overhead that is equivalent to sending $M + K$ pilots in the UL and M pilots in the DL. To compare this with TDD, suppose the frequency resources in FDD are divided equally between UL and DL. The average pilot overhead of the FDD protocol is then $\frac{M+K+\max(M,K)}{2}$.

We will now illustrate the important difference in pilot dimensionality between TDD and FDD operation. Consider an SDMA system that can afford τ_p pilots. This value determines the combinations of M and K that can be supported. The TDD protocol supports up to $K = \tau_p$ UEs and an arbitrary M . The FDD protocol supports any M and K such that $\frac{M+K+\max(M,K)}{2} \leq \tau_p$. The operating points supported by these protocols are illustrated in Figure 1.22 for $\tau_p = 20$. The shaded area indicates $M \geq 4K$, which are the operating points attractive for

²²This approximation assumes analog CSI feedback, where UE k sends the value of each element in \mathbf{h}_{jk}^j as a real-valued data symbol and this feedback is multiplexed using SDMA. More precisely, with the multiplexing gain $\min(M, K)$ of SDMA, we need $\max(M, K)$ symbol transmissions to feed back the MK channel coefficients. Quantized digital feedback is another option, but it gives roughly the same overhead if the feedback accuracy should be the same [73].

SDMA as discussed in Section 1.3.3 (see for example the results in Figure 1.17). The tradeoff between antennas and UEs caused by the FDD protocol leads to a very limited intersection with the shaded area. In contrast, the TDD protocol is entirely scalable with respect to M and the number of pilots only limits how many UEs can be supported. Any number of antennas can be used, but preferably we select one of the many operating points that lie in the shaded area.

In summary, SDMA systems should ideally be combined with TDD, by exploiting the reciprocity between UL and DL channels. This is because the required channel acquisition overhead in TDD is K , while it is $\frac{M+K+\max(M,K)}{2}$ in FDD. The FDD overhead is around 50% larger when $M \approx K$, while it is much larger for $M \gg K$, which is the preferable operating regime for SDMA. Note that it is the channel acquisition needed for DL precoding that differs between TDD and FDD, while the UL works essentially the same.

Remark 1.5 (Channel parameterizations). In some propagation scenarios, the set of possible M -dimensional channel responses can be parameterized using much less than M parameters. A key example is LoS propagation where the model that we used in (1.38) mainly depends on the angle φ_{jk}^0 between the BS and the UE. Instead of transmitting M DL pilots, we can in the LoS case select a set of equally spaced angles between 0 and π and send precoded DL pilot signals only in these directions. If the number of such angles is much smaller than M , then this method can enable FDD operation with reduced pilot overhead and can still give good estimation quality [50]. However, LoS channel parameterizations require the array geometry to be predefined and that the antennas are phase-calibrated, in the sense that the phase drifts incurred by the radio frequency (RF) hardware are known and can be compensated for. In particular, the model in (1.38) is only valid for phase-calibrated ULAs. There are several drawbacks with building a system that strictly relies on channel parameterizations. One is that even if some UE channels can be parameterized efficiently, there might not exist a single low-dimensional parameterization model that applies to all channels—it is sufficient that one part of the cell provides approximately uncorrelated Rayleigh fading to discourage the use of channel

parametrization for simplified DL estimation. Another drawback is that practical channels are not bound to follow a particular channel model. NLoS channels can consist of various multipath components that arrive from different angles and with different phase-rotations, while practical LoS channels contain random reflections and scattering, in addition to the deterministic LoS path. TDD operation is generally preferred because we want to design a network that can operate efficiently in any kind of propagation environment, with any array geometry, and without inter-antenna phase-calibration. However, TDD also has its own specific challenges: *i*) the SNR is slightly lower than in FDD since the power amplifier is only turned on part of the time; *ii*) the transmitter and receiver hardware of an antenna must be calibrated to maintain channel reciprocity (see Section 6.4.4 on p. 445 for a further discussion).

1.4 Summary of Key Points in Section 1

- Users of future networks will demand wireless connectivity with uniform service quality, anywhere at any time.
- The demand for data traffic increases rapidly and calls for higher area throughput in future cellular networks. This can be achieved by cell densification, allocating more frequency spectrum, and/or improving the SE [bit/s/Hz/cell].
- Current and future network infrastructure consists of two key parts: the coverage tier and the hotspot tier. The area throughput needs to be improved in both tiers.
- The coverage tier takes care of coverage, mobility, and guarantees a minimum service quality. To increase the area throughput of this tier, it is preferred to increase the SE, since densification or the use of spectrum at higher frequencies degrade the mobility support and coverage.
- The hotspot tier offloads traffic from the coverage tier, for example, from low-mobility indoor UEs. Densification and the use of new spectrum at higher frequencies are attractive ways to increase the area throughput of this tier, but the SE can be also improved by an array gain.
- The SE of a single UE is a slowly increasing, logarithmic function of the SINR. Only modest SE gains are possible by increasing the SINR (e.g., by using higher transmit power or deploying multiple antennas at the BS).
- A K -fold SE gain is achievable by serving K UEs per cell, on the same time/frequency resources, using SDMA. The number of BS antennas is preferably increased with K to get an array gain that compensates for the increased interference.

- Each BS should have more antennas, M , than UEs, leading to an antenna-UE ratio $M/K > 1$. This makes linear UL receive combining and DL transmit precoding nearly optimal since each interfering UE contributes with relatively little interference.
- When the number of BS antennas is large, the effective channels to the desired UEs are almost deterministic after combining/precoding, although the channel responses are random. This phenomenon is called channel hardening.
- CSI is used by the BS to spatially separate the UEs in UL and DL. The channels are most efficiently estimated with a TDD protocol that utilizes channel reciprocity, since only UL pilot signals are required and no feedback is needed.

2

Massive MIMO Networks

This section defines many of the basic concepts related to Massive MIMO, which will be used in later sections. A formal definition of Massive MIMO networks is provided in Section 2.1, along with a description of the considered coherence block structure. Spatial channel correlation is introduced and the correlated Rayleigh fading channel model is defined in Section 2.2. The UL and DL system models that will be used in the remainder of this monograph are provided in Section 2.3. In Section 2.4, we exemplify how spatial channel correlation can affect the system performance. The properties of channel hardening and favorable propagation are then defined in Section 2.5 and analyzed for spatially correlated channels. Section 2.6 introduces the local scattering channel model, which will be used in later sections to provide qualitative insights into the impact of spatial channel correlation. The key points are summarized in Section 2.7.

2.1 Definition of Massive MIMO

Based on the discussion in Section 1, a highly spectrally efficient coverage tier in a cellular network can be characterized as follows:

- It uses SDMA to achieve a multiplexing gain by serving multiple UEs on the same time-frequency resources.
- It has more BS antennas than UEs per cell to achieve efficient interference suppression. If the anticipated number of UEs grows in a cell, the BS should be upgraded so that the number of antennas increases proportionally.
- It operates in TDD mode to limit the CSI acquisition overhead, due to the multiple antennas, and to not rely on parametrizable channel models.

The Massive MIMO technology from [208, 212] embraces these design guidelines, making it an efficient way to achieve high SE in the coverage tier of future wireless networks. It is hard to find a concise definition of Massive MIMO in prior literature, but the following is the definition considered in this monograph.

Definition 2.1 (Canonical Massive MIMO network). A Massive MIMO network is a multicarrier cellular network with L cells that operate according to a synchronous TDD protocol.¹ BS j is equipped with $M_j \gg 1$ antennas, to achieve channel hardening. BS j communicates with K_j single-antenna UEs simultaneously on each time/frequency sample, with antenna-UE ratio $M_j/K_j > 1$. Each BS operates individually and processes its signals using linear receive combining and linear transmit precoding.

We consider this as the canonical form of Massive MIMO because it has the characteristics listed above and is in line with Marzetta's seminal work. It also represents the technology that has been demonstrated in real-time Massive MIMO testbeds [329, 139]. However, there are important research efforts that deviate from the canonical form (or attempt to broaden it). In particular, finding an efficient FDD protocol for Massive

¹A synchronous TDD protocol refers to a protocol in which UL and DL transmissions within different cells are synchronized. As discussed in [208], this constitutes a worst-case scenario from the standpoint of inter-cell interference. In Section 4.2.4 on p. 315, we briefly discuss the potential impact of asynchronous pilot transmission.

MIMO is highly desirable, since there are vast amounts of spectrum reserved for FDD operation. In mobile scenarios, the estimation/feedback overhead of FDD operation is prohibitive, unless something is done to reduce it. The predominant approach is to parameterize the channel (as discussed in Remark 1.5 on p. 212) and utilize the parametrization to reduce the channel estimation and feedback overhead. This principle was analyzed for small-scale MIMO in the 1990s [125, 254], while some early results for FDD Massive MIMO can be found in [7, 84, 273, 77]. These works are based on the hypothesis that the channels can be parameterized in a particular way, which is then utilized to achieve a more efficient estimation and feedback procedure. However, this line of research is still in its infancy since the underlying hypothesis has not been proved experimentally. This is why FDD operation is not considered in this monograph, but we stress that designing and demonstrating an efficient FDD Massive MIMO implementation is a great challenge that needs to be tackled [50].

Two other deviations from the canonical form of Massive MIMO are the use of multiantenna UEs [32, 194, 31] and single-carrier transmission [264]. The former was discussed in Remark 1.4 on p. 203, while the multicarrier assumption in Definition 2.1 deserves further explanation.

The propagation channels change over time and frequency. The bandwidth B equals the number of complex-valued samples that describe the signal per second. The time interval between two samples thus decreases as the bandwidth increases. Wireless channels are dispersive, meaning that the signal energy that is transmitted over a given time interval spreads out and is received over a longer time interval. If the sample interval is short, as compared to the dispersiveness of the channel, there will be a substantial overlap between adjacent transmitted samples at the receiver. The channel then has memory, which makes it harder to estimate it and to process the transmitted and received signal to combat inter-sample interference. A classic solution is to divide the bandwidth into many subcarriers, each having a sufficiently narrow bandwidth so that the effective time interval between samples is much longer than the channel dispersion. The subcarrier channels are then essentially memoryless and we can apply the information-theoretic

results described in Section 1.2 on p. 167 on each subcarrier. There are different multicarrier modulation schemes, whereof both conventional OFDM [357] and filter bank multi-carrier (FBMC) modulation [111] have been analyzed in the context of Massive MIMO.

The important thing from the Massive MIMO perspective is not which multicarrier modulation scheme is used, but that the frequency resources are divided into flat-fading subcarriers. The *coherence bandwidth* B_c describes the frequency interval over which the channel responses are approximately constant. One or multiple subcarriers fit(s) into the coherence bandwidth, thus the channel observed on adjacent subcarriers are either approximately equal or closely related through a deterministic transformation. Hence, there is generally no need to estimate the channel on every subcarrier. Similarly, the time variations of the channels are small between adjacent samples and the *coherence time* T_c describes the time interval over which the channel responses are approximately constant.

Definition 2.2 (Coherence block). A *coherence block* consists of a number of subcarriers and time samples over which the channel response can be approximated as constant and flat-fading. If the coherence bandwidth is B_c and the coherence time is T_c , then each coherence block contains $\tau_c = B_c T_c$ complex-valued samples.

The number of practically useful samples per coherence block can be smaller than $B_c T_c$. For example, if the cyclic prefix in an OFDM implementation adds 5% to the OFDM symbol time, then the number of useful samples is $B_c T_c / 1.05$.

The concepts of multicarrier modulation and coherence block are illustrated in Figure 2.1. The random channel responses in one coherence block are statistically identical to the ones in any other coherence block, irrespective of whether they are separated in time and/or frequency. Hence, the channel fading is described by a stationary ergodic random process. The performance analysis is therefore carried out by studying a single statistically representative coherence block. We assume that the channel realizations are independent between any pair of blocks, which is known as a block fading assumption.²

²The independence assumption is not strictly necessary, since we consider ergodic

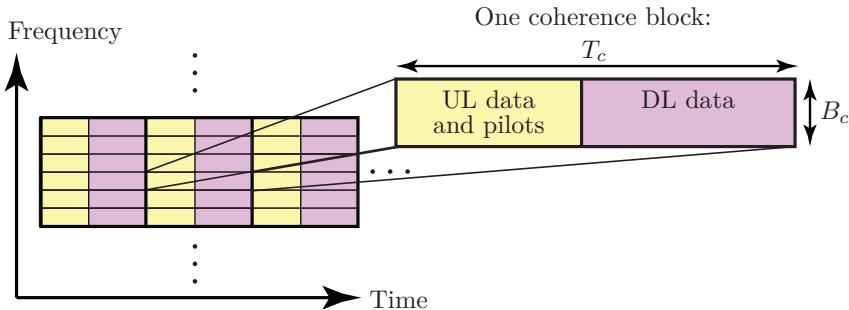
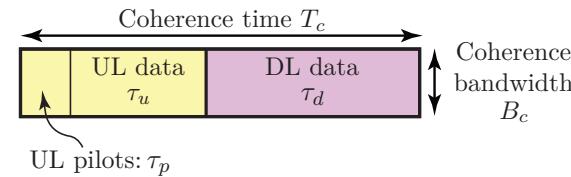
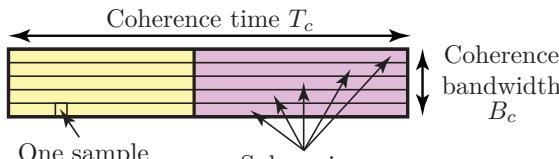


Figure 2.1: The TDD multicarrier modulation scheme of a canonical Massive MIMO network. The time-frequency plane is divided into coherence blocks in which each channel is time-invariant and frequency-flat.



(a) The samples are used for UL pilots, UL data, and DL data.



(b) The samples can belong to different subcarriers.

Figure 2.2: Each coherence block contains $\tau_c = B_c T_c$ complex-valued samples.

Each coherence block is operated in TDD mode and Figure 2.2 illustrates how the τ_c samples are located in the time and frequency plane. The samples are used for three different things:

- τ_p UL pilot signals;

SEs where all possible channel realizations are observed during the communication. Hence, there can be some correlation between the channel realizations in different blocks, but the key assumption is that this correlation is not utilized. The processing in the transmitter and receiver is carried out using only long-term statistics and measurements made in the current block.

- τ_u UL data signals;
- τ_d DL data signals.

Clearly, we need $\tau_p + \tau_u + \tau_d = \tau_c$. The fraction of UL and DL data can be selected based on the network traffic characteristics, while the number of pilots per coherence block is a design parameter. Many user applications (e.g., video streaming and web browsing) mainly generate DL traffic, which can be dealt with by selecting $\tau_d > \tau_u$.

The size of a coherence block is determined by the propagation environment, UE mobility, and carrier frequency. Each UE has an individual coherence bandwidth and coherence time, but it is hard to dynamically adapt the network to these values since the same protocol should apply to all UEs. A practical solution is to dimension the coherence block for the worst-case propagation scenario that the network should support. If a UE has a much larger coherence time/bandwidth, then it does not have to send pilots in every block.³

Remark 2.1 (Rule-of-thumb for channel coherence). It is hard to give a precise dimensionality of the coherence block since it depends on many physical factors, but there is a common rule-of-thumb [314]. The coherence time is the time interval over which the phase and amplitude variations in the channel due to UE mobility are negligible. This can be approximated as the time it takes to move a substantial fraction of the wavelength λ , say, a quarter of the wavelength: $T_c = \lambda/(4v)$ where v is the velocity of the UE. Hence, the coherence time is inversely proportional to the carrier frequency and the channels need to be estimated less frequently in the conventional cellular frequency range of 1–6 GHz as compared to the mmWave frequency range of 30–300 GHz.⁴ The coherence bandwidth is determined by phase differences in the multipath propagation. It can be approximated as $B_c = 1/(2T_d)$ where T_d is the delay spread (i.e., the time difference between the shortest

³More precisely, suppose a particular UE has a coherence bandwidth of $\check{B}_c \geq B_c$ and a coherence time of $\check{T}_c \geq T_c$. Let $k_1 = \lfloor \check{B}_c/B_c \rfloor$ and $k_2 = \lfloor \check{T}_c/T_c \rfloor$, then the UE only needs to send pilots in every k_1 th coherence block in the frequency dimension and every k_2 th coherence block in the time dimension.

⁴The antenna radiation pattern also affects the (effective) coherence time and it might change depending on the antenna design and carrier frequency [321].

and longest path). To give quantitative numbers, suppose the carrier frequency is 2 GHz, which gives the wavelength $\lambda = 15$ cm. In an outdoor scenario with $T_c = 1$ ms and $B_c = 200$ kHz, we can support mobility of $v = 37.5$ m/s = 135 km/h and delay spread of $2.5\ \mu\text{s}$ (i.e., 750 m path differences). The coherence block contains $\tau_c = 200$ samples in this scenario that supports high mobility and high channel dispersion. In an indoor scenario with $T_c = 50$ ms and $B_c = 1$ MHz, we can instead support mobility of $v = 0.75$ m/s = 2.7 km/h and delay spread of $0.5\ \mu\text{s}$ (or 150 m path differences). The coherence block contains $\tau_c = 50\,000$ samples in this scenario with low mobility and low channel dispersion.

2.2 Correlated Rayleigh Fading

The channel response between UE k in cell l and the BS in cell j is denoted by $\mathbf{h}_{lk}^j \in \mathbb{C}^{M_j}$, where each of the elements corresponds to the channel response from the UE to one of the BS's M_j antennas. Notice that the superscript of \mathbf{h}_{lk}^j is the BS index and the subscript identifies the cell and index of the UE. The channel response is the same in both UL and DL of a coherence block. For notational convenience, we use \mathbf{h}_{lk}^j for the UL channel and $(\mathbf{h}_{lk}^j)^\text{H}$ for the DL channel, although there is only a transpose and not any complex conjugate in practice. The additional conjugation does not change the SE or any other performance metric, but simplifies the notation.

Since the channel response is a vector, it is characterized by its norm and its direction in the vector space. Both are random variables in a fading channel. The channel model characterizes their respective distribution and statistical independence/dependence.

Definition 2.3 (Spatial channel correlation). A fading channel $\mathbf{h} \in \mathbb{C}^M$ is *spatially uncorrelated* if the channel gain $\|\mathbf{h}\|^2$ and the channel direction $\mathbf{h}/\|\mathbf{h}\|$ are independent random variables, and the channel direction is uniformly distributed over the unit-sphere in \mathbb{C}^M . The channel is otherwise *spatially correlated*.

An example of a spatially uncorrelated channel model is the uncorrelated Rayleigh fading, which was defined in (1.24). Practical channels

are generally spatially correlated, also known as having space-selective fading [254], since the antennas have non-uniform radiation patterns and the physical propagation environment makes some spatial directions more probable to carry strong signals from the transmitter to the receiver than other directions. The spatial channel correlation is particularly important for large arrays since these have a good spatial resolution as compared to the number of scattering clusters (see Section 7.3 on p. 482 for further details). Therefore, in the remainder of this monograph, we concentrate on *correlated Rayleigh fading* channels such that

$$\mathbf{h}_{lk}^j \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_{M_j}, \mathbf{R}_{lk}^j \right) \quad (2.1)$$

where $\mathbf{R}_{lk}^j \in \mathbb{C}^{M_j \times M_j}$ is the positive semi-definite⁵ spatial correlation matrix (and it is also the covariance matrix due to the zero mean). This matrix is assumed to be known at the BS and the estimation of such matrices is discussed in Section 3.3.3 on p. 260. The Gaussian distribution is used to model the small-scale fading variations. The channel response is assumed to take a new independent realization from this distribution in every coherence block, as a stationary ergodic random process. The spatial correlation matrix, on the other hand, describes the macroscopic propagation effects, including the antenna gains and radiation patterns at the transmitter and receiver. The normalized trace

$$\beta_{lk}^j = \frac{1}{M_j} \text{tr} \left(\mathbf{R}_{lk}^j \right) \quad (2.2)$$

determines the average channel gain from one of the antennas at BS j to UE k in cell l . Uncorrelated Rayleigh fading with $\mathbf{R}_{lk}^j = \beta_{lk}^j \mathbf{I}_{M_j}$ is a special case of this model, but the spatial correlation matrix is in general not diagonal. The parameter β_{lk}^j is also referred to as the large-scale fading coefficient and is often modeled in decibels as

$$\beta_{lk}^j = \Upsilon - 10\alpha \log_{10} \left(\frac{d_{lk}^j}{1 \text{ km}} \right) + F_{lk}^j \quad (2.3)$$

where d_{lk}^j [km] is the distance between the transmitter and the receiver, the *pathloss exponent* α determines how fast the signal power decays

⁵A Hermitian matrix is positive definite or positive semi-definite if and only if all of its eigenvalues are positive or non-negative, respectively.

with the distance, and Υ determines the median channel gain at a reference distance of 1 km. In theoretical studies, the parameters Υ and α can be computed according to one of the many established propagation models; see for example [287]. These parameters are functions of the carrier frequency, antenna gains, and vertical heights of the antennas, which are derived from fitting (2.3) to measurements. The only non-deterministic term in (2.3) is $F_{lk}^j \sim \mathcal{N}(0, \sigma_{sf}^2)$. This term is called the *shadow fading* and creates log-normal random variations around the nominal value $\Upsilon - 10\alpha \log_{10}(d_{lk}^j/(1 \text{ km}))$ [dB]. The shadow fading can either be viewed as a model of physical blockage from large obstacles or simply as a random correction term to obtain a model that better fits practical channel measurements. The variance σ_{sf}^2 of the shadow fading determines how large the random variations are, and is often reported in terms of the standard deviation σ_{sf} . The latter is considered a constant here, but it can also depend on the cell indices and other parameters.

The eigenstructure of \mathbf{R}_{lk}^j determines the spatial channel correlation of the channel \mathbf{h}_{lk}^j ; that is, which spatial directions are statistically more likely to contain strong signal components than others. Strong spatial correlation is characterized by large eigenvalue variations. An example of how to generate \mathbf{R}_{lk}^j is provided in Section 2.6, while detailed modeling is considered in Section 7.3 on p. 482.

Remark 2.2 (A generative model for channel vectors). We can generate a random channel vector $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R})$ as follows. Let the eigenvalue decomposition of $\mathbf{R} \in \mathbb{C}^{M \times M}$ be given as $\mathbf{R} = \mathbf{UDU}^H$, where $\mathbf{D} \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing the $r = \text{rank}(\mathbf{R})$ positive non-zero eigenvalues of \mathbf{R} and $\mathbf{U} \in \mathbb{C}^{M \times r}$ consists of the associated eigenvectors, such that $\mathbf{U}^H \mathbf{U} = \mathbf{I}_r$. Then, \mathbf{h} can be generated as

$$\mathbf{h} = \mathbf{R}^{\frac{1}{2}} \check{\mathbf{e}} = \mathbf{UD}^{\frac{1}{2}} \mathbf{U}^H \check{\mathbf{e}} \sim \mathbf{UD}^{\frac{1}{2}} \mathbf{e} \quad (2.4)$$

where $\check{\mathbf{e}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{I}_M)$, $\mathbf{e} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_r, \mathbf{I}_r)$, and the last step implies that the distributions of \mathbf{h} and $\mathbf{UD}^{\frac{1}{2}} \mathbf{e}$ are identical. It is straightforward to verify that \mathbf{h} is a complex Gaussian vector with zero mean and spatial correlation matrix $\mathbb{E}\{\mathbf{h}\mathbf{h}^H\} = \mathbf{R}$. Moreover, we can clearly see that the generative model is driven by a random vector with $r \leq M$.

degrees of freedom. This seemingly negative impact of spatial channel correlation is further discussed in Section 2.4. The expression in (2.4) is also referred to as the *Karhunen-Loeve expansion* of \mathbf{h} .

By studying correlated Rayleigh fading channels, we can capture some important aspects of practical Massive MIMO channels and yet analyze the performance in a tractable way. What are the limiting assumptions behind this model? First, the model assumes that the mean value is zero. Suppose a particular channel response has a non-zero mean $\bar{\mathbf{h}}_{lk}^j$, in the sense that $\mathbf{h}_{lk}^j \sim \mathcal{N}_{\mathbb{C}}(\bar{\mathbf{h}}_{lk}^j, \mathbf{R}_{lk}^j)$. The communication performance over such a channel is typically better than the performance over the corresponding zero-mean channel with the same correlation matrix $\mathbf{R}_{lk}^j + \bar{\mathbf{h}}_{lk}^j(\bar{\mathbf{h}}_{lk}^j)^H$, since the average power $\mathbb{E}\{\|\mathbf{h}_{lk}^j\|^2\}$ is the same but there is more randomness in the zero-mean case. Hence, it is a pessimistic assumption to consider zero-mean channels. Second, the model assumes that the channel is Gaussian distributed, which is not completely true in practice. However, as explained later in Section 2.5, the channel hardening and favorable propagation phenomena make the communication performance almost independent of the small-scale fading realizations; it mainly depends on the first and second order moments of the channels, which represent the large-scale fading. Hence, most of the results in this monograph hold for other channel distributions as well (as long as some technical conditions on the higher-order moments are satisfied).

Remark 2.3 (Mobility). The channel fading model describes random variations caused by microscopic movements that affect the multipath propagation, while the spatial correlation matrix describes macroscopic effects such as pathloss, shadowing, and spatial channel correlation. The capacity analysis assumes stationary ergodic fading channels with fixed statistics, which limits the scope to microscopic mobility.⁶ Macroscopic mobility (e.g., for UEs in moving cars) can be handled by dividing the time axis into segments where the channel statistics are approximately

⁶If a UE's mobility path is known a priori, its long-term statistics can be defined and used for ergodic capacity analysis. However, if the mobility is considered random, then the ergodic approach would require the UE to visit all possible locations before the signal can be decoded, making the results practically questionable.

fixed and then computing a separate ergodic SE for each segment. This makes practical sense in Massive MIMO if large communication bandwidths are used, so that we can transmit sufficiently long codewords to approach the ergodic capacity also in short time segments. The channel hardening also improves the convergence to the ergodic capacity (or SE), since the fading variations are smaller than in single-antenna systems. Under extremely high mobility or short-packet transmissions, other performance metrics such as bit error rate (BER) and outage capacity are more suitable [314].

2.3 System Model for Uplink and Downlink

Having defined Massive MIMO, we will now define the UL and DL system models that are used in the remainder of this monograph.

2.3.1 Uplink

The UL transmission in Massive MIMO is illustrated in Figure 2.3. The received UL signal $\mathbf{y}_j \in \mathbb{C}^{M_j}$ at BS j is modeled as

$$\begin{aligned} \mathbf{y}_j &= \sum_{l=1}^L \sum_{k=1}^{K_l} \mathbf{h}_{lk}^j s_{lk} + \mathbf{n}_j \\ &= \underbrace{\sum_{k=1}^{K_j} \mathbf{h}_{jk}^j s_{jk}}_{\text{Desired signals}} + \underbrace{\sum_{l=1, l \neq j}^L \sum_{i=1}^{K_l} \mathbf{h}_{li}^j s_{li}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{n}_j}_{\text{Noise}} \end{aligned} \quad (2.5)$$

where $\mathbf{n}_j \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \sigma_{\text{UL}}^2 \mathbf{I}_{M_j})$ is independent additive receiver noise with zero mean and variance σ_{UL}^2 . The UL signal from UE k in cell l is denoted by $s_{lk} \in \mathbb{C}$ and has power $p_{lk} = \mathbb{E}\{|s_{lk}|^2\}$, irrespective of whether it is a random payload data signal $s_{lk} \sim \mathcal{N}_{\mathbb{C}}(0, p_{lk})$ or a deterministic pilot signal with $p_{lk} = |s_{lk}|^2$. The channels are constant within a coherence block, while the signals and noise take new realization at every sample. During data transmission, the BS in cell j selects the receive combining vector $\mathbf{v}_{jk} \in \mathbb{C}^{M_j}$ to separate the signal from its k th

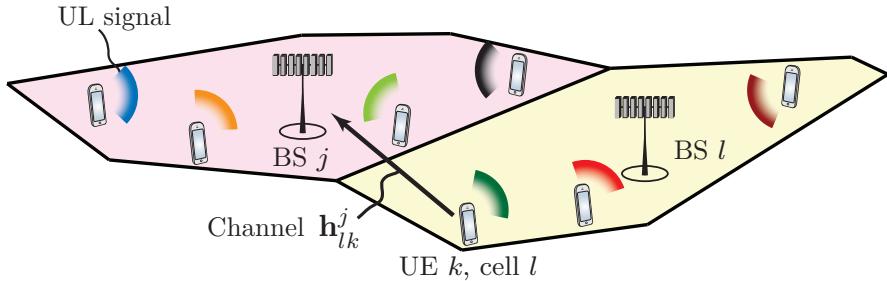


Figure 2.3: Illustration of the UL Massive MIMO transmission in cell j and cell l . The channel vector between BS j and UE k in cell l is called \mathbf{h}_{lk}^j .

desired UE from the interference as

$$\mathbf{v}_{jk}^H \mathbf{y}_j = \underbrace{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j s_{jk}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} \mathbf{v}_{jk}^H \mathbf{h}_{ji}^j s_{ji}}_{\text{Intra-cell signals}} + \underbrace{\sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j s_{li}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{v}_{jk}^H \mathbf{n}_j}_{\text{Noise}}. \quad (2.6)$$

The selection of combining vectors, based on estimated channels, and the corresponding UL SEs will be studied in Section 4.1 on p. 275. Note that receive combining is linear processing scheme that is also known as linear detection. Recall from Figure 1.18 that linear schemes provides nearly the same performance as non-linear schemes when the antenna-UE ratio is large.

2.3.2 Downlink

The DL transmission in Massive MIMO is illustrated in Figure 2.4. The BS in cell l transmits the DL signal

$$\mathbf{x}_l = \sum_{i=1}^{K_l} \mathbf{w}_{li} \varsigma_{li} \quad (2.7)$$

where $\varsigma_{lk} \sim \mathcal{N}_{\mathbb{C}}(0, \rho_{lk})$ is the DL data signal intended for UE k in the cell and ρ_{lk} is the signal power. This signal is assigned to a transmit precoding vector $\mathbf{w}_{lk} \in \mathbb{C}^{M_l}$ that determines the spatial directivity of the transmission. The precoding vector satisfies $\mathbb{E}\{\|\mathbf{w}_{lk}\|^2\} = 1$, such that $\mathbb{E}\{\|\mathbf{w}_{lk} \varsigma_{lk}\|^2\} = \rho_{lk}$ is the transmit power allocated to this UE.

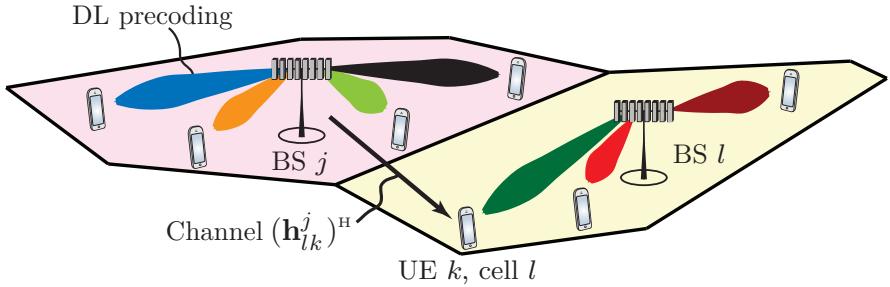


Figure 2.4: Illustration of the DL Massive MIMO transmission in cell j and cell l . The channel vector between BS j and UE k in cell l is called \mathbf{h}_{lk}^j .

The received signal $y_{jk} \in \mathbb{C}$ at UE k in cell j is modeled as

$$\begin{aligned}
 y_{jk} &= \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \mathbf{x}_l + n_{jk} \\
 &= \sum_{l=1}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li} + n_{jk} \\
 &= \underbrace{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} \varsigma_{jk}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} (\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji} \varsigma_{ji}}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{l=1 \\ l \neq j}}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li}}_{\text{Inter-cell interference}} + n_{jk} \\
 &\quad \text{Noise}
 \end{aligned} \tag{2.8}$$

where $n_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{DL}}^2)$ is independent additive receiver noise with variance σ_{DL}^2 . The channels are constant within a coherence block, while the signals and noise take new realization at every sample. The selection of transmit precoding vectors and the corresponding DL SEs will be studied in Section 4.3 on p. 316.

2.4 Basic Impact of Spatial Channel Correlation

There is a wide-spread belief that spatial channel correlation is detrimental for MIMO communications. This is indeed the case for single-user point-to-point MIMO channels with multiple antennas at both transmitter and receiver [168, 234]. However, for multiuser communications with single-antenna UEs the picture changes because it is the collection

of the UEs' spatial correlation matrices that determines the network performance. The UEs are generally physically separated by multiple wavelengths so that their channels are well modeled as statistically uncorrelated. In addition, although the channel of each UE can exhibit high spatial correlation at the BS, the spatial correlation matrices can be highly different between UEs. These are two fundamental differences from a point-to-point MIMO channel, where spatial channel correlation is seen from both the transmitter and the receiver and where the channel from the transmit antennas exhibits almost the same spatial correlation to each of the receive antennas (and vice versa).

To better understand the impact that spatial channel correlation can have on multiuser MIMO, let us consider the UL of a single-cell scenario and assume that the channels are perfectly known. The UEs' channels are distributed as $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_k)$, for $k = 1, \dots, K$, and we make the slightly artificial assumption that

$$\mathbf{R}_k = K \mathbf{U}_k \mathbf{U}_k^H \quad (2.9)$$

where $\mathbf{U}_k \in \mathbb{C}^{M \times M/K}$ are tall unitary matrices (i.e., $\mathbf{U}_k^H \mathbf{U}_k = \mathbf{I}_{M/K}$) and we assume that $\mathbf{U}_k^H \mathbf{U}_j = \mathbf{0}_{M/K \times M/K}$ for all $k \neq j$. The factor K in (2.9) normalizes the average channel gain such that $\beta_k = \frac{1}{M} \text{tr}(\mathbf{R}_k) = 1$. The channel model described by (2.9) implies that each UE has a strongly spatially correlated channel with only M/K rather than M degrees of freedom, which refers to the number of non-zero eigenvalues of the correlation matrix. However, at the same time the eigenspaces of the individual correlation matrices are all orthogonal. This means that, although the UEs' channels are random, they "live" in mutually orthogonal subspaces. This fact can be more easily seen from the Karhunen-Loeve expansion in (2.4) of the channels:

$$\mathbf{h}_k = \sqrt{K} \mathbf{U}_k \mathbf{e}_k \quad (2.10)$$

where $\mathbf{e}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M/K}, \mathbf{I}_{M/K})$. To understand the impact of spatial channel correlation, let us consider the received UL signal $\mathbf{y} \in \mathbb{C}^M$ at the BS, which is given by

$$\mathbf{y} = \sum_{i=1}^K \mathbf{h}_i s_i + \mathbf{n} \quad (2.11)$$

where $s_i \in \mathbb{C}$, for $i = 1, \dots, K$, are the UL signals of power p_i , and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \sigma_{\text{UL}}^2 \mathbf{I}_M)$ is receiver noise. Multiplying \mathbf{y} by the correlation-eigenspace \mathbf{U}_k of UE k , we obtain

$$\begin{aligned}\mathbf{U}_k^H \mathbf{y} &= \mathbf{U}_k^H \left(\sum_{i=1}^K \mathbf{h}_i s_i + \mathbf{n} \right) \\ &= \sum_{i=1}^K \sqrt{K} \mathbf{U}_k^H \mathbf{U}_i \mathbf{e}_i s_i + \mathbf{U}_k^H \mathbf{n} \\ &= \sqrt{K} \mathbf{e}_k s_k + \check{\mathbf{n}}_k\end{aligned}\tag{2.12}$$

where $\check{\mathbf{n}}_k = \mathbf{U}_k^H \mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M/K}, \sigma_{\text{UL}}^2 \mathbf{I}_{M/K})$. Thanks to the structure of the spatial correlation matrices, the multiuser channel is divided into K orthogonal single-user channels with M/K effective antennas, which is advantageous because there is no interference. Based on (2.12), the average SNR/SINR of UE k is

$$\mathbb{E} \{ \text{SNR}_k \} = \mathbb{E} \left\{ \frac{K p_k \| \mathbf{e}_k \|^2}{\sigma_{\text{UL}}^2} \right\} = \frac{M p_k}{\sigma_{\text{UL}}^2}\tag{2.13}$$

which indicates that each UE gets the full array gain of M . One way of interpreting this result is that the antenna array captures the same amount of energy, but this energy is concentrated on a subset of the spatial directions or degrees of freedom.

The scenario above has to be compared to the case where all UEs share the same correlation matrix $\mathbf{R} = K \mathbf{U} \mathbf{U}^H$, where $\mathbf{U} \in \mathbb{C}^{M \times M/K}$ is a tall unitary matrix, and thus $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R})$, for $k = 1, \dots, K$. The detrimental effect of a common correlation matrix is apparent when we multiply the received UL signal \mathbf{y} by the correlation-eigenspace \mathbf{U} :

$$\begin{aligned}\mathbf{U}^H \mathbf{y} &= \sum_{i=1}^K \sqrt{K} \mathbf{U}^H \mathbf{U} \mathbf{e}_i s_i + \mathbf{U}^H \mathbf{n} \\ &= \sum_{i=1}^K \sqrt{K} \mathbf{e}_i s_i + \check{\mathbf{n}}\end{aligned}\tag{2.14}$$

where $\check{\mathbf{n}} = \mathbf{U}^H \mathbf{n}$. In contrast to (2.12), this is not a single-user channel, but a K -user channel with M/K effective uncorrelated antennas. The common spatial correlation matrix essentially reduces the degrees of

freedom that all UEs share. In this scenario, spatial channel correlation has a clear negative impact.

In summary, it is not the individual spatial correlation matrices that manifest the system behavior, but the collection of all UEs' correlation matrices. Spatial channel correlation can be very beneficial in Massive MIMO if the UEs have sufficiently different spatial correlation matrices. This applies also to small-scale multiuser MIMO systems, as demonstrated in [373, 97, 87, 340, 52]. The case of totally orthogonal correlation matrices does hardly occur in practice and simply served as an extreme example to explain the basic impact that spatial channel correlation can have on multiuser communications.

2.5 Channel Hardening and Favorable Propagation

Two important properties of multiantenna channels were uncovered in Section 1: channel hardening and favorable propagation. We will now provide formal definitions of these properties and interpret them using the correlated fading model that was introduced in Section 2.2.

2.5.1 Channel Hardening

Channel hardening makes a fading channel behave as deterministic. This property alleviates the need for combating small-scale fading (e.g., by adapting the transmit powers) and improves the DL channel gain estimation. In Section 4, we will also show that channel hardening can be utilized to obtain simpler and more intuitive SE expressions.

Definition 2.4 (Channel hardening). A propagation channel \mathbf{h}_{jk}^j provides asymptotic channel hardening if

$$\frac{\|\mathbf{h}_{jk}^j\|^2}{\mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\}} \rightarrow 1 \quad (2.15)$$

almost surely as $M_j \rightarrow \infty$.

This definition says that the gain $\|\mathbf{h}_{jk}^j\|^2$ of an arbitrary fading channel \mathbf{h}_{jk}^j is close to its mean value when there are many antennas. This should be interpreted in the sense that the relative deviation from

the average channel gain $\mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\} = \text{tr}(\mathbf{R}_{jk}^j)$ vanishes asymptotically. This does not mean that $\|\mathbf{h}_{jk}^j\|^2 \rightarrow \text{tr}(\mathbf{R}_{jk}^j)$, because both these terms generally diverge as $M_j \rightarrow \infty$, but one can interpret the result as

$$\frac{1}{M_j} \|\mathbf{h}_{jk}^j\|^2 - \frac{1}{M_j} \text{tr}(\mathbf{R}_{jk}^j) \rightarrow 0 \quad (2.16)$$

almost surely as $M_j \rightarrow \infty$. With correlated Rayleigh fading, $\mathbf{h}_{jk}^j \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \mathbf{R}_{jk}^j)$, a sufficient condition⁷ for asymptotic channel hardening is that the spectral norm $\|\mathbf{R}_{jk}^j\|_2$ of the spatial correlation matrix is bounded and $\beta_{jk}^j = \frac{1}{M_j} \text{tr}(\mathbf{R}_{jk}^j)$ remains strictly positive as $M_j \rightarrow \infty$. The interpretation of these asymptotic properties is further discussed in Section 4.4 on p. 335.

What is important for practical purposes is not the asymptotic result, but how close to asymptotic channel hardening we are with a practical number of antennas. This can be quantified by considering

$$\mathbb{V} \left\{ \frac{\|\mathbf{h}_{jk}^j\|^2}{\mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\}} \right\} = \frac{\mathbb{V}\{\|\mathbf{h}_{jk}^j\|^2\}}{(\mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\})^2} \stackrel{(a)}{=} \frac{\text{tr}((\mathbf{R}_{jk}^j)^2)}{(\text{tr}(\mathbf{R}_{jk}^j))^2} = \frac{\text{tr}((\mathbf{R}_{jk}^j)^2)}{(M_j \beta_{jk}^j)^2} \quad (2.17)$$

where (a) follows from applying Lemma B.14 on p. 564. This is the variance of the expression in (2.15) and it should be close to zero if channel hardening is to be observed.⁸ Note that the numerator of (2.17) is the sum of the squared eigenvalues of \mathbf{R}_{jk}^j , while $M_j \beta_{jk}^j$ in the denominator is the sum of the eigenvalues. In the special case of uncorrelated fading, we have $\mathbf{R}_{jk}^j = \beta_{jk}^j \mathbf{I}_{M_j}$ and hence (2.17) becomes $1/M_j$. In this special case, $M_j = 100$ is typically sufficient to benefit from channel hardening. Hence, we note that (2.17) should be in order of 10^{-2} or smaller to obtain hardening.

The numerator $\text{tr}((\mathbf{R}_{jk}^j)^2)$ of (2.17) is a so-called Schur-convex function of the eigenvalues [166, Example 2.5]. For a given M_j and average eigenvalue β_{jk}^j , this implies that the variance is maximized

⁷This can be proved by substituting $\beta_{jk}^j = \text{tr}(\mathbf{R}_{jk}^j)/M_j$ into the left-hand side of (2.15) and then applying Lemma B.13 on p. 564.

⁸A necessary but not sufficient condition for channel hardening is that the variance goes to zero. This condition implies convergence in (2.15) in probability, but not almost sure convergence.

when one eigenvalue is $M_j \beta_{jk}^j$ and the remaining ones are zero, while it is minimized when all the eigenvalues are equal to β_{jk}^j . This suggests that spatial channel correlation, which is characterized by eigenvalue variations in \mathbf{R}_{jk}^j , increases (2.17) and thereby reduces the level of channel hardening that is observed for a given number of antennas. Another way to view it is that more antennas are required to achieve a certain value in (2.17) under spatially correlated fading than with uncorrelated fading.

2.5.2 Favorable Propagation

Favorable propagation makes the directions of two UE channels asymptotically orthogonal. This property makes it easier for the BS to mitigate interference between these UEs, which generally improves the SE and makes it sufficient to use linear combining and precoding.

Definition 2.5 (Favorable propagation). The pair of channels \mathbf{h}_{li}^j and \mathbf{h}_{jk}^j to BS j provide asymptotically favorable propagation if

$$\frac{(\mathbf{h}_{li}^j)^H \mathbf{h}_{jk}^j}{\sqrt{\mathbb{E}\{\|\mathbf{h}_{li}^j\|^2\} \mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\}}} \rightarrow 0 \quad (2.18)$$

almost surely as $M_j \rightarrow \infty$.

This definition says that the inner product of the normalized channels $\mathbf{h}_{li}^j / \sqrt{\mathbb{E}\{\|\mathbf{h}_{li}^j\|^2\}}$ and $\mathbf{h}_{jk}^j / \sqrt{\mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\}}$ goes asymptotically to zero. Since the norms of the channels grow with M_j , favorable propagation does not imply that the inner product of \mathbf{h}_{li}^j and \mathbf{h}_{jk}^j goes to zero; that is, the channel directions become orthogonal, but not the channel responses. For correlated Rayleigh fading channels, a sufficient condition for (2.18) is that the spatial correlation matrices \mathbf{R}_{li}^j and \mathbf{R}_{jk}^j have spectral norms that are bounded and the average channel gains $\beta_{li}^j = \frac{1}{M_j} \text{tr}(\mathbf{R}_{li}^j)$ and $\beta_{jk}^j = \frac{1}{M_j} \text{tr}(\mathbf{R}_{jk}^j)$ remain strictly positive as $M_j \rightarrow \infty$. Notice that under this condition, the two channels will also exhibit asymptotic channel hardening.

One way to quantify how close to asymptotic favorable propagation we are with a practical number of antennas is to consider

$$\mathbb{V} \left\{ \frac{(\mathbf{h}_{li}^j)^H \mathbf{h}_{jk}^j}{\sqrt{\mathbb{E}\{\|\mathbf{h}_{li}^j\|^2\} \mathbb{E}\{\|\mathbf{h}_{jk}^j\|^2\}}} \right\} = \frac{\text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j)}{\text{tr}(\mathbf{R}_{li}^j) \text{tr}(\mathbf{R}_{jk}^j)} = \frac{\text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j)}{M_j^2 \beta_{li}^j \beta_{jk}^j} \quad (2.19)$$

which is the variance of the expression in (2.18). This is a measure of how orthogonal the channel directions are, which determines how much interference the UEs cause to each other. The connection is particularly strong when using MR combining/precoding, where the inner product between the channels appear directly in the received signals (see (1.40) and (1.45)). Ideally, the variance in (2.19) should be zero.⁹ In practice, the variance is non-zero and therefore we can benefit from using combining/precoding schemes that mitigate inter-user interference. If both channels have uncorrelated fading, the variance becomes $1/M_j$ and thus decreases with an increasing number of antennas. In general, it is the spatial channel correlation that determines the variance in (2.19). It is zero if the UEs have orthogonal correlation-eigenspaces, while the worst-case appears when the UEs have identical eigenspaces and only a few strong eigenvalues. This result is in line with the observations made in Section 2.4.

Note that channel hardening and favorable propagation are two related, but different properties. We described a sufficient condition under which both properties hold, but it is not a necessary condition. Generally speaking, a channel model can have both properties, one of them, or none. The keyhole channel provides favorable propagation, but not channel hardening [243]. In contrast, two LoS channels (e.g., of the type in (1.23)) that have the same azimuth angle provide channel hardening, but not favorable propagation.

Finally, we stress that Massive MIMO does not require or formally rely on any of these properties, but any multiuser MIMO system performs better when the two properties are satisfied.

⁹A necessary but not sufficient condition for favorable propagation is that the variance goes to zero. This condition implies convergence in (2.18) in probability, but not almost sure convergence.

2.6 Local Scattering Spatial Correlation Model

Since spatial channel correlation is an important property of multiuser MIMO, we will now develop a spatial correlation model that will be used in the numerical examples of subsequent sections. The model is rather simple, as compared to the state-of-the-art channel models later described in Section 7.3 on p. 482, but captures some key characteristics and has an intuitive structure. The subspaces of the correlation matrices will be parameterized by the azimuth angles to the UEs, making it easy to determine if two UEs are spatially separable by comparing their respective angles.

Our goal is to develop a model for the spatial correlation matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$ for a NLoS channel between a UE and a BS equipped with a ULA. The UE and BS indices are dropped for simplicity. The received signal at the BS is the superposition of N_{path} multipath components, where N_{path} is a large number. Suppose the scattering is localized around the UE, while the BS is elevated and thus has no scatterers in its near-field. Each of the multipath components thus results in a plane wave that reaches the array from a particular angle $\bar{\varphi}_n$ and gives an array response $\mathbf{a}_n \in \mathbb{C}^M$ similar to the LoS case in (1.23):

$$\mathbf{a}_n = g_n \begin{bmatrix} 1 & e^{2\pi j d_H \sin(\bar{\varphi}_n)} & \dots & e^{2\pi j d_H (M-1) \sin(\bar{\varphi}_n)} \end{bmatrix}^T \quad (2.20)$$

where $g_n \in \mathbb{C}$ accounts for the gain and phase-rotation for this path and d_H is the antenna spacing in the array (measured in number of wavelengths). The channel response \mathbf{h} is the superposition

$$\mathbf{h} = \sum_{n=1}^{N_{\text{path}}} \mathbf{a}_n \quad (2.21)$$

of the array responses of the N_{path} components. Suppose the angles $\bar{\varphi}_n$ are i.i.d. random variables with angular probability density function (PDF) $f(\bar{\varphi})$ and g_n are i.i.d. random variables with zero-mean and variance $\mathbb{E}\{|g_n|^2\}$. The variance represents the average gain of the n th path and the total average gain of the multipath components is denoted by $\beta = \sum_{n=1}^{N_{\text{path}}} \mathbb{E}\{|g_n|^2\}$. The multidimensional central limit theorem then implies that

$$\mathbf{h} \rightarrow \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}), \quad N_{\text{path}} \rightarrow \infty \quad (2.22)$$

where the convergence is in distribution and the correlation matrix is $\mathbf{R} = \mathbb{E}\{\sum_n \mathbf{a}_n \mathbf{a}_n^H\}$. This is the general motivation behind the correlated Rayleigh fading model. Note that in our particular setup the (l, m) th element of \mathbf{R} is

$$\begin{aligned} [\mathbf{R}]_{l,m} &= \sum_{n=1}^{N_{\text{path}}} \mathbb{E}\left\{|g_n|^2\right\} \mathbb{E}\left\{e^{2\pi j d_H(l-1) \sin(\bar{\varphi}_n)} e^{-2\pi j d_H(m-1) \sin(\bar{\varphi}_n)}\right\} \\ &= \beta \int e^{2\pi j d_H(l-m) \sin(\bar{\varphi})} f(\bar{\varphi}) d\bar{\varphi} \end{aligned} \quad (2.23)$$

where we used the definition of β and let $\bar{\varphi}$ denote the angle of an arbitrary multipath component. The integral expression in (2.23) can be computed numerically for any angular distribution. Since $[\mathbf{R}]_{l,m}$ depends on the difference $l - m$, but not on the individual values of l and m , \mathbf{R} is Toeplitz matrix. Due to the assumed lack of scattering around the BS, it is reasonable to further assume that all the multipath components originate from a scattering cluster around the UE; that is, $\bar{\varphi} = \varphi + \delta$, where φ is a deterministic nominal angle and δ is a random deviation from the nominal angle with standard deviation σ_φ . We refer to this as the *local scattering model* and notice that Gaussian distributed deviations $\delta \sim \mathcal{N}(0, \sigma_\varphi^2)$ [4, 313, 363, 373], Laplace distributed deviations $\delta \sim \text{Lap}(0, \sigma_\varphi/\sqrt{2})$ [225, Section 7.4.2], [161, 256], as well as uniformly distributed deviations $\delta \sim U[-\sqrt{3}\sigma_\varphi, \sqrt{3}\sigma_\varphi]$ [7, 284, 301, 363] can be found in the literature.¹⁰ The latter case is also known as the one-ring model, since all the scatterers can be assumed to lie on a circle centered at the UE. This setup is illustrated in Figure 2.5. We stress that the correlated fading is caused by the scattering being localized around the UE, in contrast to the uncorrelated fading case illustrated in Figure 1.11 that also contained rich scattering in the vicinity of the BS.

The standard deviation $\sigma_\varphi \geq 0$ is measured in radians and is called the *angular standard deviation (ASD)*, since it determines how large the deviations from the nominal angle are. A reasonable value of σ_φ in urban cellular networks is 10° [256], while smaller values are expected

¹⁰Since only $\varphi \in [-\pi, \pi]$ is of interest in the angular domain, the Gaussian and Laplace distributions can either be truncated to this interval (and scaled to maintain a PDF that integrates to one) or applied as they are, letting the periodicity of $\sin(\varphi)$ wrap the distribution into the interval of interest.

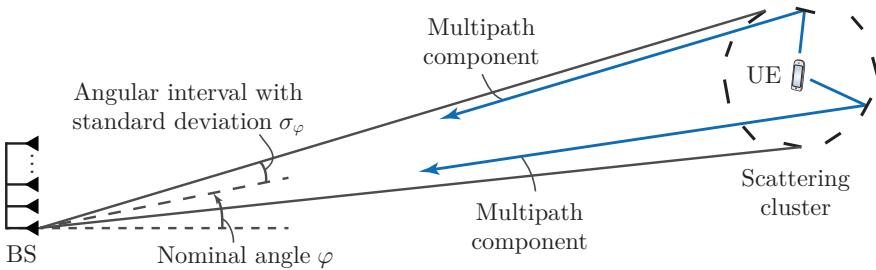


Figure 2.5: Illustration of NLoS propagation under the local scattering model, where the scattering is localized around the UE. Two of the many multipath components are shown. The nominal angle φ and the angular standard deviation (ASD) σ_φ of the multipath components are key parameters to model the spatial correlation matrix.

in flat rural areas and larger values in hilly areas [254].

To illustrate the effect of spatial channel correlation, Figure 2.6 shows the eigenvalues of \mathbf{R} in decreasing order, when using the local scattering model with $M = 100$ antennas, the nominal angle $\varphi = 30^\circ$, and the ASD $\sigma_\varphi = 10^\circ$. The correlation matrices are normalized such that $\text{tr}(\mathbf{R}) = M$. The aforementioned three distributions of the angular deviations are compared with the reference case of uncorrelated fading: $\mathbf{R} = \mathbf{I}_M$. The figure shows that the spatial channel correlation makes around 30 of the 100 eigenvalues larger than in the uncorrelated case, while the remaining eigenvalues are substantially smaller. In fact, a uniform angular distribution makes 68% of the eigenvalues 30 dB smaller than in the reference case, while this happens for 40% of the eigenvalues with Gaussian distribution and 19% with Laplace distribution. These percentages remain roughly the same if M is increased.

Clearly, a 10° ASD leads to high spatial channel correlation with a low-rank correlation matrix where many eigenvalues are negligibly small. One should be careful when interpreting this result since it is based on the rather simple local scattering model. Large eigenvalue variations are expected in practice, but the angular distribution might be less smooth (e.g., non-Gaussian with multiple peaks) and also vary between BS antennas due to near-field scattering; see the measurements in [121, Figure 4] for an example. Hence, despite the low-rank behavior, one should not expect the spatial correlation matrices to be parametriz-

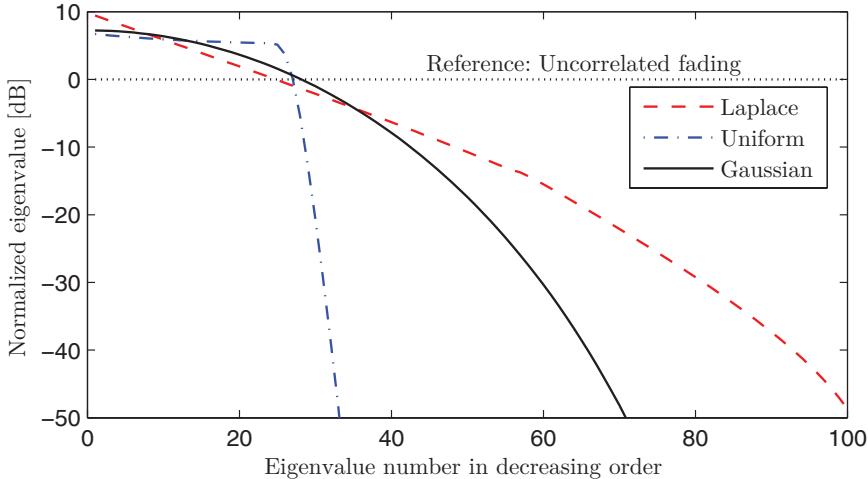


Figure 2.6: Eigenvalues of the spatial correlation matrix \mathbf{R} when using the local scattering model with $M = 100$, nominal angle $\varphi = 30^\circ$, and either Laplace, uniform, or Gaussian angular distribution with standard deviation $\sigma_\varphi = 10^\circ$. Uncorrelated fading is shown as a reference case.

able using a nominal angle and a Gaussian/Laplace/uniform angular distribution in practice.

The spatial channel correlation reduces as σ_φ increases. Since the angles are wrapped in the angular domain, the scatterers become asymptotically uniformly distributed between $-\pi$ and $+\pi$ as $\sigma_\varphi \rightarrow \infty$ (for any of the three distributions mentioned above). This does, however, not lead to fully uncorrelated fading since a ULA has better resolution in some angular directions than others. The fully uncorrelated fading case is instead obtained from (2.23) in the pathological case of $\sin(\delta) \sim U[-1, 1]$.

2.6.1 Impact on Channel Hardening and Favorable Propagation

In addition to affecting the rank of the spatial correlation matrix, the nominal angle and ASD also affect how many antennas are needed to approach asymptotic channel hardening and favorable propagation.

Recall from Section 2.5.1 that a channel \mathbf{h} hardens if $\|\mathbf{h}\|^2 / \mathbb{E}\{\|\mathbf{h}\|^2\} \rightarrow 1$ as $M \rightarrow \infty$. Figure 2.7 shows the “variance” of the channel hardening,

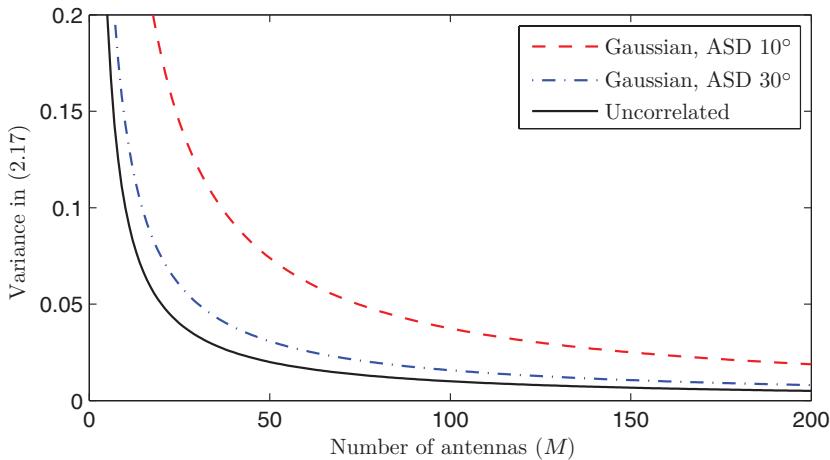


Figure 2.7: Variance of the channel hardening, defined in (2.17), as a function of the number of antennas. Uncorrelated fading is compared with the local scattering model, using $\varphi = 30^\circ$ and Gaussian angular distribution.

as defined in (2.17), for different numbers of antennas. The smaller the variance, the more the channel has hardened. We compare uncorrelated Rayleigh fading with the local scattering model, using $\varphi = 30^\circ$ and a Gaussian angular distribution with $\sigma_\varphi \in \{10^\circ, 30^\circ\}$. The smallest variance is achieved with uncorrelated fading, while spatial channel correlation basically shifts the curve to the right. With $\sigma_\varphi = 30^\circ$, which represents moderate spatial correlation, the difference from uncorrelated fading is rather small. However, with $\sigma_\varphi = 10^\circ$, the strong spatial correlation leads to a large loss in channel hardening. For example, $M = 200$ with $\sigma_\varphi = 10^\circ$ gives the same variance as $M = 53$ with uncorrelated fading.

The “variance” of the favorable propagation, as defined in (2.19), is illustrated in Figure 2.8 for $M = 100$ antennas. We consider a desired UE with a fixed nominal angle of 30° and an interfering UE with a nominal angle that is varied between -180° and 180° . A smaller variance implies that the UEs’ channel directions are closer to be orthogonal. We once again compare uncorrelated fading with the local scattering model using a Gaussian angular distribution with $\sigma_\varphi \in \{10^\circ, 30^\circ\}$. With uncorrelated fading, the variance is independent of the UEs’ angles,

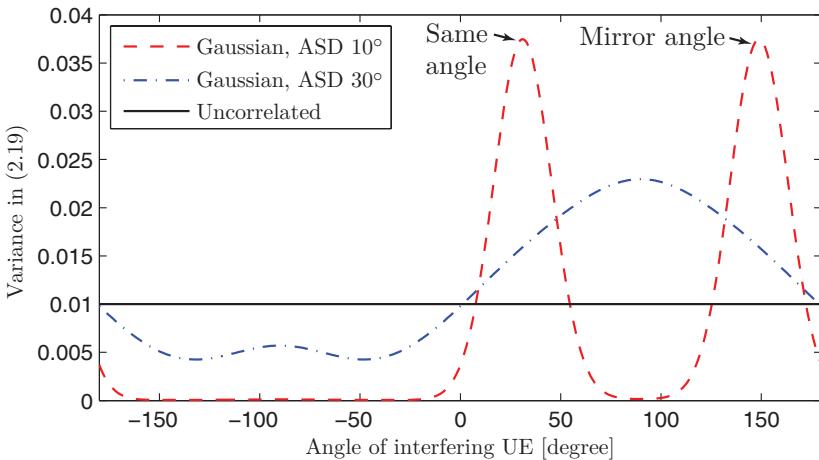


Figure 2.8: Variance of the favorable propagation, defined in (2.19), for $M = 100$. The desired UE has a nominal angle of 30° , while the angle of the interfering UE is varied between -180° and 180° . Uncorrelated fading is compared with the local scattering model, using Gaussian angular distribution.

as expected. In contrast, the variance depends strongly on the UE angles when there is spatial channel correlation. When the interfering UE has around the same nominal angle as the desired UE (or is close to the mirror reflection angle $180^\circ - 30^\circ = 150^\circ$), the variance is substantially larger than with uncorrelated fading. This represents the case when the UEs have similar spatial correlation matrices. When the ASD is small, there are visible peaks at 30° and 150° . As the ASD increases, these peaks widen and eventually merge into a single peak, as in the case of $\sigma_\varphi = 30^\circ$. In that case, the largest variance actually occurs when the UEs have different angles. When the UEs have well-separated angles, the variance is substantially smaller than with uncorrelated fading. This is the same basic behavior as exemplified in Section 2.4; spatial channel correlation is good if the UEs have very different correlation-eigenspaces, while it is bad if the UEs have similar correlation-eigenspaces. If we compute the average variance over different angles of the interfering UE, it becomes 0.01 with uncorrelated fading, 0.0076 for $\sigma_\varphi = 10^\circ$ and 0.012 for $\sigma_\varphi = 30^\circ$. This suggests that we will, on the average, observe slightly more favorable propagation

under strong spatial channel correlation, than when the correlation is weaker.

In summary, spatial channel correlation reduces the level of channel hardening observed for a given number of antennas. Spatial correlation can also improve the level of favorable propagation if the UEs have different spatial characteristics. How these behaviors affect the communication performance will be investigated in later sections.

2.6.2 Approximate Expression with Gaussian Angular Distribution

In the case of $\delta \sim \mathcal{N}(0, \sigma_\varphi^2)$, we can compute an approximate closed-form expression for \mathbf{R} when the ASD is small (e.g., below 15°), such that $\sin(\delta) \approx \delta$ and $\cos(\delta) \approx 1$. We can then approximate (2.23) as

$$\begin{aligned}
[\mathbf{R}]_{l,m} &= \beta \int_{-\infty}^{\infty} e^{2\pi j d_H(l-m) \sin(\varphi+\delta)} \frac{1}{\sqrt{2\pi}\sigma_\varphi} e^{-\frac{\delta^2}{2\sigma_\varphi^2}} d\delta \\
&\approx \beta \int_{-\infty}^{\infty} e^{2\pi j d_H(l-m) \sin(\varphi)} e^{2\pi j d_H(l-m) \cos(\varphi)\delta} \frac{1}{\sqrt{2\pi}\sigma_\varphi} e^{-\frac{\delta^2}{2\sigma_\varphi^2}} d\delta \\
&= \beta e^{2\pi j d_H(l-m) \sin(\varphi)} e^{-\frac{\sigma_\varphi^2}{2}(2\pi d_H(l-m) \cos(\varphi))^2} \\
&\quad \cdot \underbrace{\frac{1}{\sqrt{2\pi}\sigma_\varphi} \int_{-\infty}^{\infty} e^{-\frac{(\delta-2\pi j \sigma_\varphi^2 d_H(l-m) \cos(\varphi))^2}{2\sigma_\varphi^2}} d\delta}_{=1} \\
&= \beta e^{2\pi j d_H(l-m) \sin(\varphi)} e^{-\frac{\sigma_\varphi^2}{2}(2\pi d_H(l-m) \cos(\varphi))^2} \tag{2.24}
\end{aligned}$$

where the approximation is based on $\sin(\varphi + \delta) = \sin(\varphi) \cos(\delta) + \cos(\varphi) \sin(\delta) \approx \sin(\varphi) + \cos(\varphi)\delta$. The last equality identifies an integral over the entire PDF of a Gaussian distribution (which is equal to one).

The approximate closed-form expression in (2.24) can be utilized to reduce the computational complexity in simulations. The expression also offers some insights into the structure of the correlation matrix. Notice that $[\mathbf{R}]_{l,m} = \beta e^{2\pi j d_H(l-m) \sin(\varphi)}$ for $\sigma_\varphi = 0$. In this extreme case, all multipath components arrive from the angle φ and give the rank-one

correlation matrix

$$\mathbf{R} = \beta \begin{bmatrix} e^{2\pi j d_H \sin(\varphi)} & \dots & e^{2\pi j d_H (M-1) \sin(\varphi)} \end{bmatrix}^T \\ \cdot \begin{bmatrix} e^{-2\pi j d_H \sin(\varphi)} & \dots & e^{-2\pi j d_H (M-1) \sin(\varphi)} \end{bmatrix}. \quad (2.25)$$

This matrix is formed based on the array response vector in (1.23) of a ULA. For $\sigma_\varphi > 0$, the diagonal elements are the same, but the off-diagonal elements decay as $e^{-\frac{\sigma_\varphi^2}{2}(2\pi d_H(l-m)\cos(\varphi)\delta)^2}$ and thus go to zero as σ_φ grows. When the off-diagonal elements reduce, the rank of the matrix increases. Although the small-angle approximation is not valid when σ_φ is large, the decay indicates that the correlation matrix becomes increasingly similar to a scaled identity matrix for large ASDs.

We will utilize the local scattering model with Gaussian angular distribution in the remainder of this monograph to illustrate how spatially correlated channels behave, as compared to uncorrelated channels. However, one should bear in mind that the local scattering model dates back to the 1990s, when BSs with relatively few antennas and elevated deployment at masts were the norm. Practical Massive MIMO channels are likely to experience scattering in the near-field of the BS, multiple scattering clusters, and shadowing over the array [121, 122], which are three key effects not captured by the local scattering model. We return to channel modeling in Section 7.3 on p. 482.

2.7 Summary of Key Points in Section 2

- Massive MIMO builds on decades of research insights into how to design efficient SDMA systems.
- A canonical Massive MIMO network uses a multicarrier TDD protocol. BS j is equipped with M_j antennas and serves K_j single-antenna UEs in each channel coherence block.
- By having an antenna-UE ratio $M_j/K_j > 1$, the BS benefits from favorable propagation that makes the UEs' channel directions almost orthogonal when M_j is large. It is therefore sufficient to use linear receive combining and transmit precoding in Massive MIMO.
- By having $M_j \gg 1$, the BS also benefits from channel hardening that makes the effective channels after combining/precoding almost immune to small-scale fading.
- The propagation channels and antenna arrays create spatial channel correlation, which has a non-negligible impact on the channel hardening and favorable propagation. That is why the correlated Rayleigh fading model, where the correlation is represented by the spatial correlation matrices, must be adopted in the analysis.
- The local scattering correlation model captures the basic characteristics of spatial channel correlation, in terms of a nominal angle and ASD. It will be used to exemplify the impact of spatial channel correlation in later sections.

3

Channel Estimation

This section describes how channel estimation is carried out at the BSs based on UL pilot transmission. The system model for pilot transmission is provided in Section 3.1 along with the basic pilot sequence design. The minimum mean-squared error (MMSE) estimator is derived and analyzed in Section 3.2. The impacts of spatial channel correlation and pilot contamination are exemplified in Section 3.3. The computational complexity is quantified in Section 3.4.1 and two low-complexity channel estimators are described and compared. Data-aided channel estimation is briefly discussed in Section 3.5. The key points are summarized in Section 3.6.

3.1 Uplink Pilot Transmission

To make efficient use of the massive number of antennas, each BS needs to estimate the channel responses from the UEs that are active in the current coherence block. It is particularly important for BS j to have estimates of the channels from the UEs in cell j . Channel estimates from interfering UEs in other cells can also be useful to perform interference suppression during data transmission. Recall from Section 2.1 on p. 216

that τ_p samples are reserved for UL pilot signaling in each coherence block. Each UE transmits a *pilot sequence* that spans these τ_p samples. The pilot sequence of UE k in cell j is denoted by $\phi_{jk} \in \mathbb{C}^{\tau_p}$. It is assumed to have unit-magnitude elements, to obtain a constant power level, and this implies that $\|\phi_{jk}\|^2 = \phi_{jk}^H \phi_{jk} = \tau_p$. The elements of ϕ_{jk} are scaled by the UL transmit power as $\sqrt{p_{jk}}$ and then transmitted as the signal s_{jk} in (2.5) over τ_p UL samples, leading to the received UL signal $\mathbf{Y}_j^p \in \mathbb{C}^{M_j \times \tau_p}$ at BS j . This signal is given by

$$\mathbf{Y}_j^p = \underbrace{\sum_{k=1}^{K_j} \sqrt{p_{jk}} \mathbf{h}_{jk}^j \phi_{jk}^T}_{\text{Desired pilots}} + \underbrace{\sum_{l=1, l \neq j}^L \sum_{i=1}^{K_l} \sqrt{p_{li}} \mathbf{h}_{li}^j \phi_{li}^T}_{\text{Inter-cell pilots}} + \underbrace{\mathbf{N}_j^p}_{\text{Noise}} \quad (3.1)$$

where $\mathbf{N}_j^p \in \mathbb{C}^{M_j \times \tau_p}$ is the independent additive receiver noise with i.i.d. elements distributed as $\mathcal{N}_{\mathbb{C}}(0, \sigma_{UL}^2)$. \mathbf{Y}_j^p is the observation that BS j can utilize to estimate the channel responses. To estimate the channel of a particular UE, the BS needs to know which pilot sequence this UE has transmitted. This is why the pilots are deterministic sequences and the pilot assignment is typically made when the UE connects to the BS; for example, using a random access procedure. The pilot assignment and random access are further discussed in Section 7.2.1 on p. 468.

Suppose, for the sake of argument, that BS j wants to estimate the channel \mathbf{h}_{li}^j from an arbitrary UE i in cell l . The BS can then multiply/correlate \mathbf{Y}_j^p with the pilot sequence ϕ_{li} of this UE, leading to the processed received pilot signal $\mathbf{y}_{jli}^p \in \mathbb{C}^{M_j}$, given as

$$\mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^* = \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \sqrt{p_{l'i'}} \mathbf{h}_{l'i'}^j \phi_{l'i'}^T \phi_{li}^* + \mathbf{N}_j^p \phi_{li}^* \quad (3.2)$$

which has the same dimension as \mathbf{h}_{li}^j . For the k th UE in the BS's own

cell, (3.2) can be expressed as

$$\begin{aligned} \mathbf{y}_{jjk}^p &= \mathbf{Y}_j^p \phi_{jk}^* \\ &= \underbrace{\sqrt{p_{jk}} \mathbf{h}_{jk}^j \phi_{jk}^* \phi_{jk}^*}_{\text{Desired pilot}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} \sqrt{p_{ji}} \mathbf{h}_{ji}^j \phi_{ji}^* \phi_{jk}^*}_{\text{Intra-cell pilots}} + \underbrace{\sum_{\substack{l=1 \\ l \neq j}}^L \sum_{i=1}^{K_l} \sqrt{p_{li}} \mathbf{h}_{li}^j \phi_{li}^* \phi_{jk}^*}_{\text{Inter-cell pilots}} + \underbrace{\mathbf{N}_j^p \phi_{jk}^*}_{\text{Noise}}. \end{aligned} \quad (3.3)$$

The second and third terms in (3.3) represent interference and contain inner products of the form $\phi_{li}^* \phi_{jk}^*$ between the pilot of the desired UE and the pilot of another UE i in cell l . If the pilot sequences of two UEs are orthogonal (i.e., $\phi_{li}^* \phi_{jk}^* = 0$), then the corresponding interference term in (3.3) vanishes and does not affect the estimation. Ideally, we would like all pilot sequences to be orthogonal, but since the pilots are τ_p -dimensional vectors, for a given τ_p , we can only find a set of at most τ_p mutually orthogonal sequences. The finite length of the coherence blocks imposes the constraint $\tau_p \leq \tau_c$ that makes it impossible to assign mutually orthogonal pilots to all UEs in practice. Since longer pilots come at the price of having fewer samples for data transmission, it is non-trivial to optimize the pilot length; however, a rule-of-thumb is that τ_p should always be smaller than $\tau_c/2$ [49].

We assume that the network utilizes a set of τ_p mutually orthogonal pilot sequences. These can be gathered as the columns of the UL *pilot book* $\Phi \in \mathbb{C}^{\tau_p \times \tau_p}$, which satisfies $\Phi^H \Phi = \tau_p \mathbf{I}_{\tau_p}$. It is recommended to have $\tau_p \geq \max_l K_l$ pilots so that each BS can allocate different UL pilot sequences among its UEs, but this is not mandatory. The reason for making such an assumption is that the strongest interference usually originates from within the own cell. The coordination of pilot assignment across cells is also important and is further discussed in Section 7.2.1 on p. 468. We define the set

$$\mathcal{P}_{jk} = \left\{ (l, i) : \phi_{li} = \phi_{jk}, \quad l = 1, \dots, L, i = 1, \dots, K_l \right\} \quad (3.4)$$

with the indices of all UEs that utilize the same pilot sequence as UE k in cell j . Hence, $(l, i) \in \mathcal{P}_{jk}$ implies that UE i in cell l uses the same pilot as UE k in cell j . Note that $(j, k) \in \mathcal{P}_{jk}$ by definition.

Using the notation in (3.4), the expression in (3.3) simplifies to

$$\mathbf{y}_{j,jk}^p = \underbrace{\sqrt{p_{jk}}\tau_p \mathbf{h}_{jk}^j}_{\text{Desired pilot}} + \underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \sqrt{p_{li}}\tau_p \mathbf{h}_{li}^j}_{\text{Interfering pilots}} + \underbrace{\mathbf{N}_j^p \phi_{jk}^*}_{\text{Noise}}. \quad (3.5)$$

Note that $\mathbf{y}_{j,jk}^p = \mathbf{y}_{jli}^p$ for all $(l, i) \in \mathcal{P}_{jk}$, since these UEs use the same pilot. We also note that $\mathbf{N}_j^p \phi_{jk}^* \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \sigma_{\text{UL}}^2 \tau_p \mathbf{I}_{M_j})$, since the pilot sequences are deterministic and $\|\phi_{jk}\|^2 = \tau_p$.

The processed received signal $\mathbf{y}_{j,jk}^p$ in (3.5) is a sufficient statistic for estimating \mathbf{h}_{jk}^j since there is no loss in useful information as compared to using the originally received signal \mathbf{Y}_j^p [175]. The reason is that the desired component $\mathbf{h}_{jk}^j \phi_{jk}^T$ in \mathbf{Y}_j^p can be brought back from $\mathbf{y}_{j,jk}^p$ by multiplying with ϕ_{jk}^T from the right and the interfering terms are either zero or can be brought back in the same way. Similarly, \mathbf{y}_{jli}^p is a sufficient statistic for estimating \mathbf{h}_{li}^j . The processed received signal is used in Section 3.2 for channel estimation.

3.1.1 Design of Mutually Orthogonal Pilot Sequences

The pilot book Φ is designed under the conditions that all elements have unit magnitude (i.e., $|[\Phi]_{i_1, i_2}| = 1$ for $i_1 = 1, \dots, \tau_p$, $i_2 = 1, \dots, \tau_p$) and that all columns are mutually orthogonal (i.e., $\Phi^H \Phi = \tau_p \mathbf{I}_{\tau_p}$). All pilot books that satisfy these constraints are equivalent in terms of estimation performance, but the choice can have an impact on the practical implementation. In fact, only the mutual orthogonality and the norms $\|\phi_{jk}\|$ determine the estimation accuracy, while the unit magnitude assumption was made to keep a constant power level per sample. We will exemplify two explicit ways to design the pilot books.

A Walsh-Hadamard matrix $\Phi = \mathbf{A}_{\tau_p}$ is a $\tau_p \times \tau_p$ matrix that satisfies the two conditions for being a pilot book and whose elements are either $+1$ or -1 . Since each element is a point in a binary phase-shift keying (BPSK) constellation, these pilot sequences are easily implemented in any system that supports BPSK modulated data transmission. Walsh-Hadamard matrices only exist for some matrix dimensions [339]; for example, matrices with dimensions of a power of two: $\tau_p = 2^n$ for

$n = 0, 1, \dots$. These matrices can be generated recursively as follows [29]:

$$\mathbf{A}_1 = 1 \quad (3.6)$$

$$\mathbf{A}_{2^n} = \begin{bmatrix} \mathbf{A}_{2^{n-1}} & \mathbf{A}_{2^{n-1}} \\ \mathbf{A}_{2^{n-1}} & -\mathbf{A}_{2^{n-1}} \end{bmatrix} \quad n = 1, 2, \dots \quad (3.7)$$

To generate a pilot book of arbitrary dimension (e.g., not a power of two), the discrete Fourier transform (DFT) matrix

$$\Phi = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_{\tau_p} & \omega_{\tau_p}^2 & \dots & \omega_{\tau_p}^{\tau_p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_{\tau_p}^{\tau_p-1} & \omega_{\tau_p}^{2(\tau_p-1)} & \dots & \omega_{\tau_p}^{(\tau_p-1)(\tau_p-1)} \end{bmatrix} \quad (3.8)$$

can be utilized [37], where $\omega_{\tau_p} = e^{-j2\pi/\tau_p}$ is a τ_p th primitive root of 1. Note that the elements in (3.8) are located at τ_p different equally spaced points on the unit circle, thus they correspond to a τ_p -ary phase-shift keying (PSK) constellation.

These two types of sequences are used as spreading codes in UMTS [2] and in LTE [199]. The UL pilots (called reference signals) in LTE are, however, based on Zadoff-Chu sequences, which have unit-norm elements but also the additional feature that each sequence is the cyclic shift of another sequence [309]. This property is particularly useful to mitigate inter-symbol interference in single-carrier transmission. See [309] for algorithms that generate Zadoff-Chu sequences.

3.2 MMSE Channel Estimation

We will now derive an estimator of the channel response \mathbf{h}_{li}^j , based on the received pilot signal \mathbf{Y}_j^p in (3.1) and a pilot book with mutually orthogonal sequences. The channel is a realization of a random variable, thus Bayesian estimators are desirable since they take the statistical distributions of the variables into account; see Appendix B.4 on p. 567 for an introduction to estimation theory. Bayesian estimators require that the distributions are known. Recall from (2.1) that $\mathbf{h}_{li}^j \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \mathbf{R}_{li}^j)$. The minimum mean-squared error (MMSE) estimator of \mathbf{h}_{li}^j is the

vector $\hat{\mathbf{h}}_{li}^j$ that minimizes the MSE $\mathbb{E}\{\|\mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j\|^2\}$. It is provided in the following theorem.

Theorem 3.1. Using a pilot book with mutually orthogonal sequences, the MMSE estimate of the channel \mathbf{h}_{li}^j based on the observation \mathbf{Y}_j^p in (3.1) is

$$\hat{\mathbf{h}}_{li}^j = \sqrt{p_{li}} \mathbf{R}_{li}^j \Psi_{li}^j \mathbf{y}_{jli}^p \quad (3.9)$$

where

$$\Psi_{li}^j = \left(\sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'} \tau_p \mathbf{R}_{l'i'}^j + \sigma_{UL}^2 \mathbf{I}_{M_j} \right)^{-1}. \quad (3.10)$$

The estimation error $\tilde{\mathbf{h}}_{li}^j = \mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j$ has correlation matrix $\mathbf{C}_{li}^j = \mathbb{E}\{\tilde{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H\}$, given by

$$\mathbf{C}_{li}^j = \mathbf{R}_{li}^j - p_{li} \tau_p \mathbf{R}_{li}^j \Psi_{li}^j \mathbf{R}_{li}^j. \quad (3.11)$$

Proof. The proof is available in Appendix C.2.1 on p. 591. \square

This theorem provides the mechanism to compute the MMSE estimate of the channel from any UE in the network to BS j . The estimation quality is represented by the MSE, which is $\mathbb{E}\{\|\mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j\|^2\} = \text{tr}(\mathbf{C}_{li}^j)$ for the MMSE estimator. A good estimation quality is represented by a small MSE.

To estimate \mathbf{h}_{li}^j based on (3.9), the BS should correlate the received pilot signal with the pilot sequence used by UE i in cell l , as $\mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^*$, and then multiply this observation with the two matrices Ψ_{li}^j and \mathbf{R}_{li}^j . The matrix Ψ_{li}^j is the inverse of the normalized correlation matrix $\mathbb{E}\{\mathbf{y}_{jli}^p (\mathbf{y}_{jli}^p)^H\}/\tau_p$ of the processed received signal, while \mathbf{R}_{li}^j is the spatial correlation matrix of the channel to be estimated. These multiplications suppress interference and noise that do not share the same second-order statistics as \mathbf{h}_{li}^j . Note that the MMSE estimator in (3.9) is linear, in the sense that $\hat{\mathbf{h}}_{li}^j$ is formed by multiplying the processed received signal \mathbf{y}_{jli}^p with matrices. The estimator in Theorem 3.1 is therefore sometimes called the linear MMSE (LMMSE) estimator. However, we prefer to use the MMSE notion to make it clear that one cannot further reduce the MSE by using a non-linear estimator.

For notational convenience, we define

$$\hat{\mathbf{H}}_l^j = \begin{bmatrix} \hat{\mathbf{h}}_{l1}^j & \dots & \hat{\mathbf{h}}_{lK_l}^j \end{bmatrix} \quad (3.12)$$

as the $M_j \times K_l$ matrix with the estimates of all channels from UEs in cell l to BS j .

Note that the transmit power appears in the estimation error correlation matrix in (3.11) only as a product with the pilot length: $p_{li}\tau_p$. We define the effective SNR during pilot signaling from UE k in cell j to its serving BS j as

$$\text{SNR}_{jk}^p = \frac{p_{jk}\tau_p\beta_{jk}^j}{\sigma_{\text{UL}}^2} \quad (3.13)$$

where we recall that $\beta_{jk}^j = \frac{1}{M_j} \text{tr}(\mathbf{R}_{jk}^j)$ was defined in (2.2) as the average channel gain to the antennas in the BS array. The terminology *effective SNR* implies that the pilot *processing gain* τ_p is included in the SNR. The processing gain is obtained from the fact that the pilot sequence spans τ_p samples. If the pilot sequences are 10 samples long, then the effective SNR is 10 dB larger than the nominal SNR at a single sample. This gain is highly desirable for achieving good estimation quality also for UEs with limited transmit power and/or weak channel conditions.

If we consider the random realizations of the MMSE channel estimate and the corresponding estimation error in an arbitrary coherence block, the following statistical properties hold.

Corollary 3.2. The MMSE estimate $\hat{\mathbf{h}}_{li}^j$ and the estimation error $\tilde{\mathbf{h}}_{li}^j$ are independent random variables, distributed as follows:

$$\hat{\mathbf{h}}_{li}^j \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_{M_j}, \mathbf{R}_{li}^j - \mathbf{C}_{li}^j \right) \quad (3.14)$$

$$\tilde{\mathbf{h}}_{li}^j \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_{M_j}, \mathbf{C}_{li}^j \right). \quad (3.15)$$

Proof. The proof is available in Appendix C.2.1 on p. 591. \square

The statistical distributions stated in Corollary 3.2 are useful when we later compute the SE of each UE. We can also observe that the average squared norm $\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^j\|^2\} = \text{tr}(\mathbf{R}_{li}^j) - \text{tr}(\mathbf{C}_{li}^j)$ of the estimated channel is smaller than that of the true channel, but it increases when

the MSE $\text{tr}(\mathbf{C}_{li}^j)$ decreases. In the special case of $\text{tr}(\mathbf{C}_{li}^j) = 0$, we have $\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^j\|^2\} = \mathbb{E}\{\|\mathbf{h}_{li}^j\|^2\} = \text{tr}(\mathbf{R}_{li}^j)$, since the estimate is perfect.

In practice, Theorem 3.1 is particularly important for estimating the intra-cell channels. However, also the inter-cell channels from any UE in the entire network to BS j can be estimated. An important observation can be made by comparing the MMSE estimate in (3.9) of an intra-cell channel $\hat{\mathbf{h}}_{jk}^j$ with the estimate $\hat{\mathbf{h}}_{li}^j$ of a UE in another cell that utilizes the same pilot sequence (i.e., $(l, i) \in \mathcal{P}_{jk}$ which implies $\phi_{li} = \phi_{jk}$ and $\mathcal{P}_{li} = \mathcal{P}_{jk}$). In this case, we have $\Psi_{jk}^j = \Psi_{li}^j$ and $\mathbf{y}_{jjk}^p = \mathbf{y}_{jli}^p$, thus the same matrix inverse is multiplied with the same processed received signal. It is only the scalar and the first matrix in (3.9) that are different. If \mathbf{R}_{jk}^j is invertible, we can write the relation as

$$\hat{\mathbf{h}}_{li}^j = \frac{\sqrt{p_{li}}}{\sqrt{p_{jk}}} \mathbf{R}_{li}^j (\mathbf{R}_{jk}^j)^{-1} \hat{\mathbf{h}}_{jk}^j. \quad (3.16)$$

This implies that the two estimates are strongly correlated, but generally the vectors are linearly independent (i.e., non-parallel) since one cannot write $\hat{\mathbf{h}}_{li}^j$ as a scalar times $\hat{\mathbf{h}}_{jk}^j$ unless \mathbf{R}_{li}^j and \mathbf{R}_{jk}^j are equal up to a scaling factor. In the special case of spatially uncorrelated channels with $\mathbf{R}_{jk}^j = \beta_{jk}^j \mathbf{I}_{M_j}$ and $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{I}_{M_j}$, the two channel estimates are parallel vectors that only differ in scaling. This is an unwanted property caused by the inability of BS j to separate UEs that have transmitted the same pilot sequence and have the same spatial characteristics. This situation is illustrated in Figure 3.1. The following corollary highlights a key consequence of the colliding pilot transmissions.

Corollary 3.3. Consider UE k in cell j and UE i in cell l . The correlation matrix of the respective channel estimates at BS j is

$$\mathbb{E}\left\{\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{li}^j)^H\right\} = \begin{cases} \sqrt{p_{li} p_{jk}} \tau_p \mathbf{R}_{jk}^j \Psi_{li}^j \mathbf{R}_{li}^j & (l, i) \in \mathcal{P}_{jk} \\ \mathbf{0}_{M_j \times M_j} & (l, i) \notin \mathcal{P}_{jk}. \end{cases} \quad (3.17)$$

The antenna-averaged correlation coefficient is

$$\frac{\mathbb{E}\left\{(\hat{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\right\}}{\sqrt{\mathbb{E}\left\{\|\hat{\mathbf{h}}_{jk}^j\|^2\right\} \mathbb{E}\left\{\|\hat{\mathbf{h}}_{li}^j\|^2\right\}}} = \begin{cases} \frac{\text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{li}^j)}{\sqrt{\text{tr}(\mathbf{R}_{jk}^j \mathbf{R}_{jk}^j \Psi_{li}^j) \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{li}^j \Psi_{li}^j)}} & (l, i) \in \mathcal{P}_{jk} \\ 0 & (l, i) \notin \mathcal{P}_{jk} \end{cases} \quad (3.18)$$

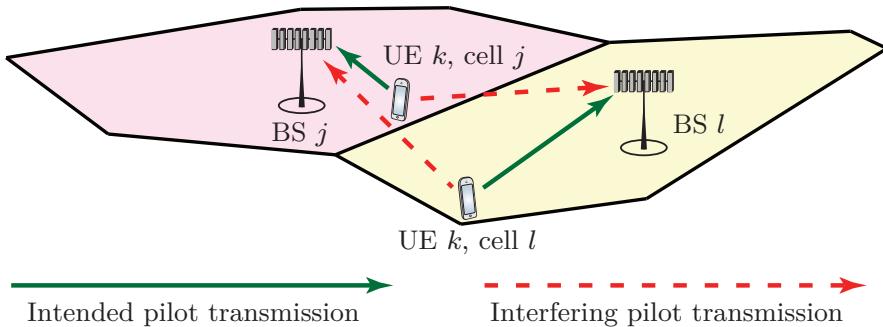


Figure 3.1: When two UEs transmit the same pilot sequence, their respective BSs receive a superposition of their signals—they contaminate each others' pilot transmissions. Since it is challenging for the BSs to separate the UEs, the estimates of their respective channels will be correlated.

despite the fact that $\mathbb{E}\left\{(\mathbf{h}_{li}^j)^H \mathbf{h}_{jk}^j\right\} = 0$ for all UE combinations with $(l, i) \neq (j, k)$.

Proof. The expression in (3.17) follows from taking the expressions in (3.9) for the UEs' channel estimates and then computing the expectation of their outer products. If $(l, i) \in \mathcal{P}_{jk}$, we have $\mathbf{y}_{jjk}^p = \mathbf{y}_{jli}^p$ and then the non-zero expectation is obtained from direct computation, utilizing $\mathbb{E}\{\mathbf{y}_{jli}^p (\mathbf{y}_{jli}^p)^H\} = \tau_p (\Psi_{li}^j)^{-1}$. If $(l, i) \notin \mathcal{P}_{jk}$, then \mathbf{y}_{jjk}^p and \mathbf{y}_{jli}^p are independent, since these vectors contain different channels and independent noise variables. The expectation is then zero. Finally, (3.18) is obtained from (3.17) by exploiting the fact that $(\hat{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j = \text{tr}(\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{li}^j)^H)$ and simplifying the expression by utilizing that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any matrices \mathbf{A} and \mathbf{B} such that \mathbf{A} and \mathbf{B}^T have the same dimensions. \square

This corollary describes one of the key characteristics of the *pilot contamination* phenomenon: UEs that transmit the same pilot sequence contaminate each others' channel estimates. The interference not only reduces the estimation quality (i.e., increases the MSE) but also makes the channel estimates statistically dependent—although the true channels are statistically independent. Pilot contamination has an important impact beyond channel estimation, since the contamination makes it particularly hard for the BS to mitigate interference between UEs that use the same pilot. Pilot contamination is often described as a main

characteristic and limiting factor of Massive MIMO. It was the key focus of some of the early works on the topic [208, 131, 169], but the phenomenon is not unique to Massive MIMO. It exists in most cellular networks because of the practical necessity to reuse the time-frequency resources across cells. Pilot contamination can, however, have a greater impact on Massive MIMO than on conventional networks. This is partially because the large number of UEs requires the pilot sequences to be reused more frequently in space and partially because the signal processing in Massive MIMO is particularly good at suppressing interference between UEs with orthogonal pilots. We return to pilot contamination in Section 3.3.2 and in the SE analysis in Section 4 on p. 275.

Recall that the MMSE estimator minimizes the MSE of the channel estimate, which is defined as

$$\mathbb{E}\{\|\mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j\|^2\} = \mathbb{E}\{\|\tilde{\mathbf{h}}_{li}^j\|^2\} = \mathbb{E}\{\text{tr}(\tilde{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H)\} = \text{tr}(\mathbf{C}_{li}^j). \quad (3.19)$$

To compare the estimation quality obtained with different estimation schemes in different scenarios, the normalized MSE (NMSE) defined as

$$\text{NMSE}_{li}^j = \frac{\text{tr}(\mathbf{C}_{li}^j)}{\text{tr}(\mathbf{R}_{li}^j)} \quad (3.20)$$

is a suitable metric, since it measures the relative estimation error per antenna. This is a value between 0 (perfect estimation) and 1 (achieved by using the mean value of the variable, $\mathbb{E}\{\mathbf{h}_{li}^j\}$, as the estimate).

Remark 3.1 (Other channel distributions). The MMSE estimator in Theorem 3.1 utilizes moments of the channel (i.e., the zero mean and the correlation matrix) as well as the fact that the channel is complex Gaussian distributed. In practice, the mean value and correlation matrix are rather easy to estimate, while it is hard to validate how close to Gaussian the channel distribution is. This is fortunately not a big deal because the estimator in (3.9) is also the LMMSE estimator for non-Gaussian channels with zero mean and the same known correlation matrix (see Appendix B.4 on p. 567). Hence, the same estimation expression can be used for other types of channels, but the estimate and estimation error are only uncorrelated in this case (not independent), which affects the performance analysis.

3.3 Impact of Spatial Correlation and Pilot Contamination

To understand the basic properties of the MMSE estimator, we will exemplify how spatial channel correlation and pilot contamination affects its performance. We will also outline how to acquire the channel statistics in practice.

3.3.1 Impact of Spatial Correlation on Channel Estimation

The basic properties of channel estimation are best described when we consider the estimation of the channel response of a UE that has a unique pilot sequence. The estimation is then only affected by noise and not by interference. Consider an arbitrary channel $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R})$, where the UE and BS indices are dropped for brevity. Let $\mathbf{R} = \mathbf{U}\Lambda\mathbf{U}^H$ denote the eigenvalue decomposition of the correlation matrix, where the unitary matrix $\mathbf{U} \in \mathbb{C}^{M \times M}$ contains the eigenvectors, also called eigendirections, and the diagonal matrix $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_M)$ contains the corresponding eigenvalues. The estimation error correlation matrix in (3.11) becomes

$$\begin{aligned} \mathbf{C} &= \mathbf{R} - p\tau_p \mathbf{R} \left(p\tau_p \mathbf{R} + \sigma_{\text{UL}}^2 \mathbf{I}_M \right)^{-1} \mathbf{R} \\ &= \mathbf{U} \left(\Lambda - p\tau_p \Lambda \left(p\tau_p \Lambda + \sigma_{\text{UL}}^2 \mathbf{I}_M \right)^{-1} \Lambda \right) \mathbf{U}^H \\ &= \mathbf{U} \text{diag} \left(\lambda_1 - \frac{p\tau_p \lambda_1^2}{p\tau_p \lambda_1 + \sigma_{\text{UL}}^2}, \dots, \lambda_M - \frac{p\tau_p \lambda_M^2}{p\tau_p \lambda_M + \sigma_{\text{UL}}^2} \right) \mathbf{U}^H \quad (3.21) \end{aligned}$$

where the second equality follows from the fact that $\mathbf{I}_M = \mathbf{U}\mathbf{U}^H$ and $\mathbf{U}^{-1}\mathbf{U} = \mathbf{I}_M$. The last expression in (3.21) is identified as an eigenvalue decomposition with eigenvectors in \mathbf{U} and the m th eigenvalue given by

$$\lambda_m - \frac{p\tau_p \lambda_m^2}{p\tau_p \lambda_m + \sigma_{\text{UL}}^2} = \frac{\sigma_{\text{UL}}^2 \lambda_m}{p\tau_p \lambda_m + \sigma_{\text{UL}}^2} = \frac{\lambda_m}{\text{SNR}^p \frac{\lambda_m}{\beta} + 1} \quad (3.22)$$

where SNR^p is the effective SNR defined in (3.13) and $\beta = \frac{1}{M} \sum_{n=1}^M \lambda_n$. Hence, the estimation error correlation matrix \mathbf{C} has the same eigenvectors as the spatial correlation matrix \mathbf{R} , but the eigenvalues are different and generally smaller due to the subtraction in (3.22). The eigenvalues of \mathbf{C} in (3.22) represent the estimation error variance in each

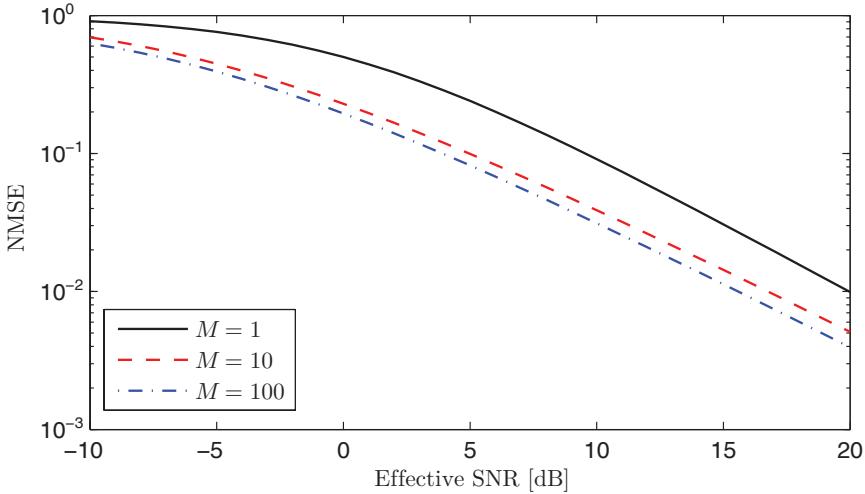


Figure 3.2: NMSE in MMSE estimation of a spatially correlated channel, based on the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$. The results are averaged over different nominal UE angles.

eigendirection. As the effective SNR increases, all these error variances reduce and approach zero as $\text{SNR}^p \rightarrow \infty$, thus showing that error-free estimation is possible in this asymptotic regime. Another important observation from (3.22) is that an eigendirection of \mathbf{R} with a large eigenvalue λ_m has a smaller normalized error variance

$$\frac{\frac{\lambda_m}{\text{SNR}^{p\frac{\lambda_m}{\beta}} + 1}}{\lambda_m} = \frac{1}{\text{SNR}^{p\frac{\lambda_m}{\beta}} + 1} \quad (3.23)$$

than an eigendirection with a smaller eigenvalue. The intuition is that the eigendirections are estimated independently and strong eigendirections are easier to estimate since the SNR is higher.

These properties are illustrated numerically in Figure 3.2 for spatial correlation matrices generated by the local scattering model, defined in (2.23), with Gaussian angular distribution. Figure 3.2 shows the NMSE, defined in (3.20), as a function of SNR^p (the effective SNR) with either $M = 1$, $M = 10$, or $M = 100$ antennas. The results are averaged over different uniformly distributed nominal angles between 0° and 360° , while the ASD is $\sigma_\varphi = 10^\circ$. Figure 3.2 shows that the NMSE

is monotonically decreasing with the SNR, as expected from (3.22). An NMSE of around 10^{-2} is achieved at an SNR of 20 dB, which means that the estimation error variance is only 1% of the original variance of the channel. Note that this effective SNR can be achieved by having a nominal SNR of 10 dB and pilot sequences with $\tau_p = 10$, thus it is not particularly high.

Interestingly, the NMSE in Figure 3.2 also reduces as more antennas are added. This property is due to the spatial channel correlation, as seen from the fact that a spatially uncorrelated channel with $\mathbf{R} = \beta\mathbf{I}$ gives the NMSE $1/(\text{SNR}^p + 1)$ which is independent of M . Hence, it is easier to estimate spatially correlated channels due to the structure in their statistics. This also implies that the average gain $\mathbb{E}\{\|\hat{\mathbf{h}}\|^2\} = \text{tr}(\mathbf{R} - \mathbf{C})$ of the estimated channel is larger under spatial correlation.

The impact of spatial channel correlation is further studied in Figure 3.3, where the NMSE is shown as a function of the ASD σ_φ . The effective SNR is 10 dB and there are $M = 100$ antennas. Figure 3.3 shows that the error is smaller when the ASD is small (i.e., with high spatial correlation). This is explained by the fact that most of the channel's variance lies in a few eigenvalues when σ_φ is small (cf. Figure 2.6). As concluded from (3.23), it is easier to estimate strong eigendirections than weaker ones. The NMSE for uncorrelated channels is shown in Figure 3.3 as a reference. For strongly spatially correlated channels, the estimation error can be two orders of magnitude smaller than in the uncorrelated case, while this benefit is basically lost when σ_φ reaches around 40° .

3.3.2 Impact of Pilot Contamination on Channel Estimation

We will now illustrate the basics of pilot contamination by considering a scenario where two UEs use the same pilot sequence. BS j estimates the channel of UE k in its own cell, while UE i in cell l transmits the same pilot. The mutual interference that these UEs cause during pilot transmission has two main consequences:

- The channel estimates become correlated;
- The estimation quality is reduced.

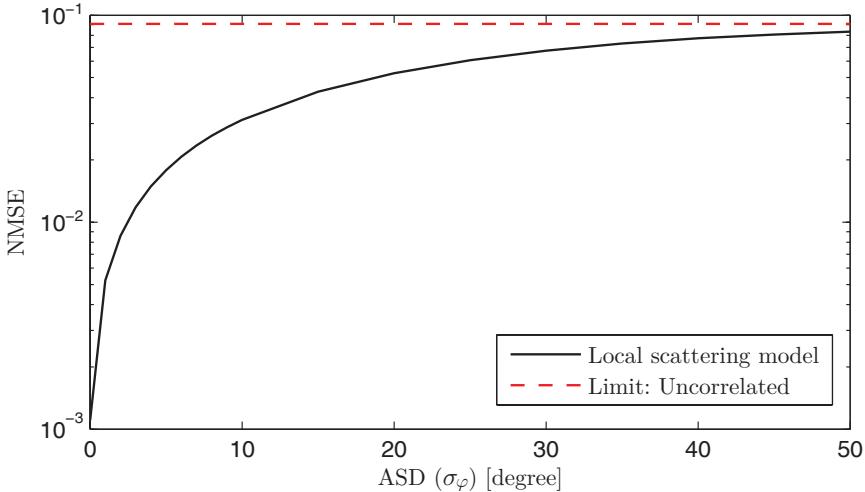


Figure 3.3: NMSE in estimation of a spatially correlated channel, as a function of the ASD in the local scattering model, defined in (2.23), with Gaussian angular distribution. The effective SNR is 10 dB and $M = 100$.

Starting with the former, Figure 3.4 shows the antenna-averaged correlation coefficient between the channel estimates, as defined in (3.18), when the effective SNR from the desired UE is 10 dB and the interfering signal is 10 dB weaker than that. Both correlation matrices are generated using the local scattering model with Gaussian angular distribution and $\text{ASD } \sigma_\varphi = 10^\circ$, but using different nominal angles at BS j . The desired UE has a fixed angle of 30° (measured as described in Figure 2.5), while the angle of the interfering UE is varied between -180° and 180° .

The first observation from Figure 3.4 is that the UE angles play a key role when the BS is equipped with multiple antennas. If the UEs have the same angle, the correlation coefficient is one, meaning that the estimates are identical (up to a scaling factor). If the UE angles are well separated, the correlation coefficient is instead nearly zero. This indicates that not only the average channel gains but also the eigenstructure of the spatial correlation matrices determine the impact of pilot contamination. This is different from the single-antenna case (and multiantenna case with uncorrelated fading), in which the correlation coefficient is equal to one, irrespective of the UE angles. The

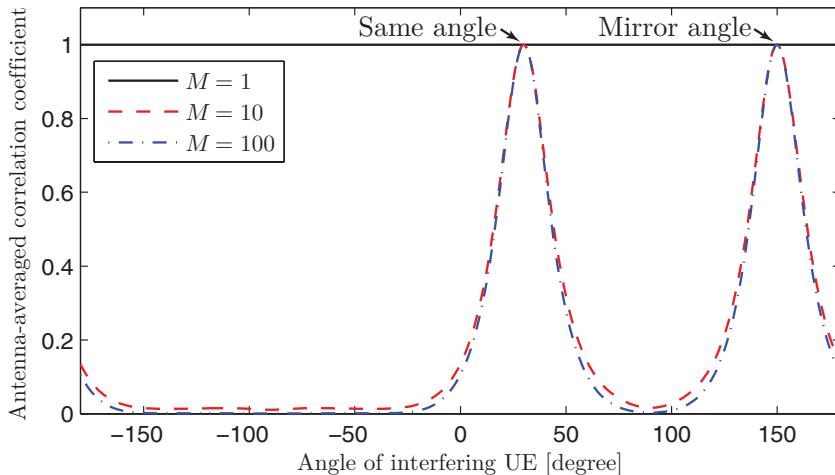


Figure 3.4: Absolute value of the antenna-averaged correlation coefficient in (3.18) between the channel estimates of the desired UE and an interfering UE that uses the same pilot. The local scattering model with Gaussian angular distribution is used and the desired UE has a nominal angle of 30° , while the angle of the interfering UE is varied between -180° and 180° .

conclusion is that spatial channel correlation can mitigate the impact of pilot contamination and we expect this to happen also with other spatially correlated channel models. Depending on the array geometry there might be certain angle pairs that give a resonance behavior in the multiantenna case. Since we consider a horizontal ULA in this simulation, the array cannot separate signals arriving from 30° and from the mirror reflection angle $180^\circ - 30^\circ = 150^\circ$.

The second main consequence of pilot contamination is the reduced estimation quality. We will study this impact in the same scenario as above. Figure 3.5 shows the NMSE of the estimate of the desired channel with $M = 100$ antennas and either uncorrelated fading or the local scattering model with ASD $\sigma_\varphi = 10^\circ$. The effective SNR from the desired UE is 10 dB and the interfering signal is either equally strong, 10 dB weaker, or 20 dB weaker. In the spatially correlated case, when the UE angles are well separated, the NMSE is around 0.04 irrespective of how strong the interfering pilot signal is. This implies that the pilot contamination has a negligible impact on the estimation quality when

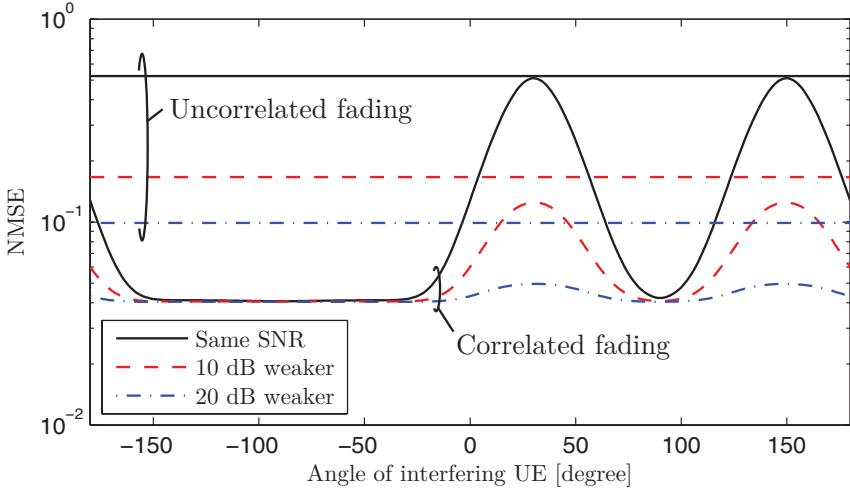


Figure 3.5: NMSE in estimation of the desired UE's channel when there is an interfering UE, which uses the same pilot. There are $M = 100$ antennas. The local scattering model with Gaussian angular distribution is used and the desired UE has a nominal angle of 30° , while the angle of the interfering UE is varied between -180° and 180° . The NMSE with uncorrelated fading is shown as reference.

the UEs have nearly orthogonal correlation-eigenspaces. The NMSE increases when the UEs have similar angles, particularly when the interfering UE has a strong channel to the BS. If the UEs' channels instead exhibit uncorrelated fading, the NMSEs are consistently larger than under spatial correlation and also angle-independent. Hence, spatial channel correlation is helpful in practice to improve the estimation quality under pilot contamination.

In the extreme case of $\mathbf{R}_{jk}^j \mathbf{R}_{li}^j = \mathbf{0}_{M_j \times M_j}$, the UE channels have orthogonal correlation-eigenspaces. The antenna-averaged correlation coefficient between the channel estimates, defined in (3.18), is then zero. Furthermore, the estimation error correlation matrix in (3.11) simplifies as

$$\begin{aligned} \mathbf{C}_{jk}^j &= \mathbf{R}_{jk}^j - p_{jk}\tau_p \mathbf{R}_{jk}^j \left(p_{jk}\tau_p \mathbf{R}_{jk}^j + p_{li}\tau_p \mathbf{R}_{li}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \mathbf{R}_{jk}^j \\ &= \mathbf{R}_{jk}^j - p_{jk}\tau_p \mathbf{R}_{jk}^j \left(p_{jk}\tau_p \mathbf{R}_{jk}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \mathbf{R}_{jk}^j \end{aligned} \quad (3.24)$$

which does not depend on the interfering UE. This property is easily proved by using Lemma B.6 on p. 560. Consequently, it is theoretically

possible to let two UEs share a pilot sequence, without causing pilot contamination, if their spatial correlation matrices satisfy the orthogonality condition $\mathbf{R}_{jk}^j \mathbf{R}_{li}^j = \mathbf{0}_{M_j \times M_j}$. This can theoretically happen under strong spatial channel correlation, while it will never happen for spatially uncorrelated channels. However, $\mathbf{R}_{jk}^j \mathbf{R}_{li}^j \approx \mathbf{0}_{M_j \times M_j}$ can happen when the interfering UE has a very weak channel. These behaviors have been utilized in [154, 363], among others, to guide the pilot assignment and UE scheduling in Massive MIMO.

3.3.3 Imperfect Statistical Knowledge

The MMSE estimator utilizes the channel statistics. For example, if BS j wants to estimate the channel to UE i in cell j , it can only apply the estimator in Theorem 3.1 if it knows the correlation matrix \mathbf{R}_{li}^j and the sum of the correlation matrices, $(\Psi_{li}^j)^{-1}$, of the UEs that utilize the same pilot sequence. We will exemplify how BS j can estimate the correlation matrix \mathbf{R}_{li}^j of the channel $\mathbf{h}_{li}^j \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \mathbf{R}_{li}^j)$. We will then describe how $(\Psi_{li}^j)^{-1}$ can be obtained in a similar manner. The UE and BS indices are dropped for simplicity.

In general, the BS observes many realizations of $\mathbf{h} = [h_1 \dots h_M]^T$ in different coherence blocks, distributed over time and frequency. Suppose the BS has made N independent observations $\mathbf{h}[1], \dots, \mathbf{h}[N]$, where $\mathbf{h}[n] = [h_1[n] \dots h_M[n]]^T$ is the n th observation. For a particular antenna index m , the law of large numbers (see Lemma B.12 on p. 564) implies that the sample variance $\sum_{n=1}^N \frac{1}{N} |h_m[n]|^2$ converges (almost surely) to the true variance $\mathbb{E}\{|h_m|^2\}$ as $N \rightarrow \infty$. The standard deviation of the sample variance decays as $1/\sqrt{N}$ [175], thus a small number of observations is sufficient to get a good variance estimate. The corresponding approach to estimate the $M \times M$ correlation matrix \mathbf{R} is to form the sample correlation matrix

$$\hat{\mathbf{R}}_{\text{sample}} = \frac{1}{N} \sum_{n=1}^N \mathbf{h}[n] (\mathbf{h}[n])^H. \quad (3.25)$$

Each element of $\hat{\mathbf{R}}_{\text{sample}}$ converges to the corresponding element of \mathbf{R} as described above. However, it is more challenging to obtain a sample correlation matrix whose eigenvalues and eigenvectors are well aligned

with those of \mathbf{R} , because the estimation errors in all the M^2 elements of $\hat{\mathbf{R}}_{\text{sample}}$ affect the eigenstructure. This might be important for channel estimation, since the MMSE estimator exploits the eigenstructure of \mathbf{R} to obtain a better estimate. Fortunately, there are techniques to make the channel estimation in Massive MIMO robust to imperfect knowledge of the spatial correlation matrix [57, 189, 299]. Note that only the diagonal elements of \mathbf{R} are essential for Bayesian estimation, because they describe the variance of the unknown variables, while the off-diagonal elements only describe the correlation between variables. Hence, we can alternatively form the diagonalized sample correlation matrix

$$\hat{\mathbf{R}}_{\text{diagonal}} = \begin{bmatrix} \frac{1}{N} \sum_{n=1}^N |h_1[n]|^2 & & \\ & \ddots & \\ & & \frac{1}{N} \sum_{n=1}^N |h_M[n]|^2 \end{bmatrix} \quad (3.26)$$

by ignoring the correlation between the elements in \mathbf{h} . If $\hat{\mathbf{R}}_{\text{diagonal}}$ is used for channel estimation instead of \mathbf{R} , we will effectively estimate each element h_m separately from the other elements of \mathbf{h} , as if we only have one BS antenna. In other words, we are not exploiting the spatial channel correlation.

It was proposed in [299] to estimate the spatial correlation matrix as the convex combination

$$\hat{\mathbf{R}}(c) = c\hat{\mathbf{R}}_{\text{sample}} + (1 - c)\hat{\mathbf{R}}_{\text{diagonal}} \quad (3.27)$$

between the conventional sample correlation matrix and the diagonalized sample correlation matrix. The diagonal elements of $\hat{\mathbf{R}}(c)$ are the same as in $\hat{\mathbf{R}}_{\text{diagonal}}$, while the off-diagonal elements are proportional to $c \in [0, 1]$. A small value of c reduces the off-diagonal elements and thus can be used to purposely underestimate the correlation between the channel coefficients. This can be viewed as a regularization of $\hat{\mathbf{R}}_{\text{sample}}$. The parameter c can be optimized experimentally to achieve robust estimation under imperfect correlation matrix knowledge (i.e., for finite N). In this single-user example, the NMSE of any linear estimator

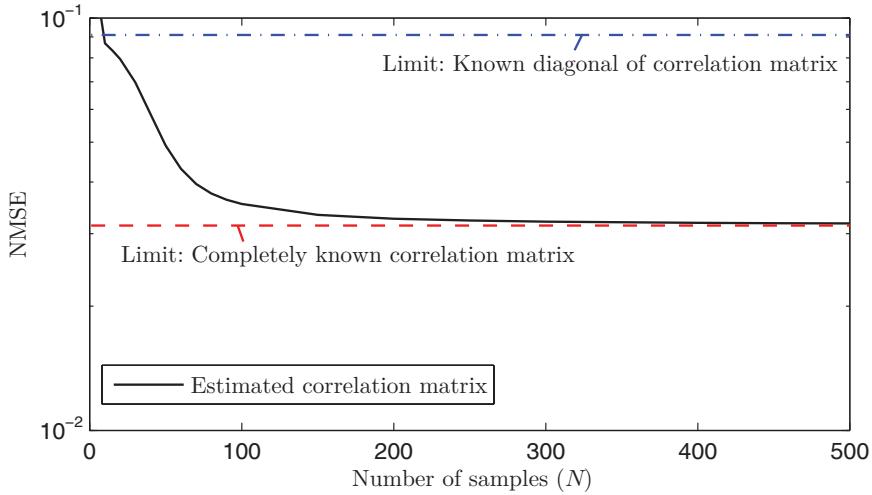


Figure 3.6: NMSE in the estimation of a spatially correlated channel with ASD $\sigma_\varphi = 10^\circ$, when having imperfect spatial correlation matrix knowledge, as a function of the number of samples used to compute the correlation matrix estimate.

$\hat{\mathbf{h}} = \mathbf{A}\mathbf{Y}^p\boldsymbol{\phi}^*$ can be computed as

$$\text{NMSE}(\mathbf{A}) = 1 - \frac{2\sqrt{p}\tau_p \Re(\text{tr}(\mathbf{R}\mathbf{A})) - \tau_p \text{tr}(\mathbf{A}(p\tau_p\mathbf{R} + \sigma_{UL}^2 \mathbf{I}_M)\mathbf{A}^H)}{\text{tr}(\mathbf{R})} \quad (3.28)$$

where the matrix \mathbf{A} specifies which linear estimator is used. The true MMSE estimator in Theorem 3.1 is given by $\mathbf{A} = \sqrt{p}\mathbf{R}(p\tau_p\mathbf{R} + \sigma_{UL}^2 \mathbf{I}_M)^{-1}$, while the estimated correlation matrix in (3.27) can be used to select $\mathbf{A}(c) = \sqrt{p}\hat{\mathbf{R}}(c)(p\tau_p\hat{\mathbf{R}}(c) + \sigma_{UL}^2 \mathbf{I}_M)^{-1}$ instead. This is a heuristic estimator, but c can be optimized to get a small $\text{NMSE}(\mathbf{A}(c))$.

The average NMSE with imperfect correlation matrix knowledge is shown in Figure 3.6. We consider the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$, $M = 100$ antennas, and an effective SNR of 10 dB. The NMSE is averaged over different nominal angles (from 0° to 360°) and different sample realizations, and c is numerically optimized for each N to achieve a low average NMSE. Figure 3.6 shows the NMSE as a function of the number of samples. The first few samples are essential to get a reasonable estimate of $\hat{\mathbf{R}}_{\text{diagonal}}$. With as little as $N = 10$ samples, we can exploit some of the spatial

channel correlation to achieve an NMSE that is smaller than for an uncorrelated channel (with a known correlation matrix that equals the main diagonal of \mathbf{R}). The NMSE decreases with N and asymptotically approaches the lower bound where \mathbf{R} is completely known. Interestingly, not more than 200 samples are needed to achieve an NMSE close to the lower bound. This equals $2M$, which indicates that the channel estimation is rather insensitive to having imperfect channel statistics.

A similar approach as the one detailed above can be taken to estimate Ψ_{li}^j [57]. This matrix is given by $\Psi_{li}^j = \tau_p(\mathbb{E}\{\mathbf{y}_{jli}^p(\mathbf{y}_{jli}^p)^H\})^{-1}$ and thus particularly convenient to estimate since we can use the received signals from the existing pilot transmissions to form a sample correlation matrix (which is then regularized as described above). In contrast, the estimation of the individual correlation matrices requires additional pilot signals that are designed for correlation matrix estimation; see [57] for a comparison of two approaches.

In practice, the spatial correlation matrix evolves over time, due to mobility that creates variations in the large-scale fading. It is necessary to track the changes, which can be accomplished by computing the sample correlation matrix over a sliding time window that contains N samples. The number of samples is selected to achieve a sufficiently good estimate, while the time-frequency intervals between samples can be selected based on the UE mobility. The measurements in [332] suggest that the large-scale fading is constant for a time interval around 100 times longer than the coherence time, thus it is possible to obtain hundreds of samples for correlation matrix estimation if necessary. The overhead for correlation matrix estimation under mobility is quantified in [57].

Remark 3.2 (Correlation matrix estimation from a small number of samples). In addition to the regularization approach described above, there are alternative ways to generate correlation matrix estimates from a relatively small number of samples N . The conventional sample correlation matrix is a consistent estimator when $N \rightarrow \infty$ for a fixed M , which is a limit that is hard to approach if M is large. If N is comparable in size to M , one can instead use G-estimation methods [217, 93], which provide consistent estimators when $M, N \rightarrow \infty$ with a fixed ratio. In

addition, the paper [211] considers the case $N < M$, in which $\hat{\mathbf{R}}_{\text{sample}}$ is rank-deficient, and generates a full-rank correlation estimate that retains the eigenvectors of $\hat{\mathbf{R}}_{\text{sample}}$ (in contrast to the regularization approach above that changes the eigenvectors). The correlation estimation can be further improved if the channels have a special structure that is known a priori. For example, [137] provides algorithms for estimating the correlation matrices of channels that have a limited angle-delay support that is also separable between UEs. There are also methods to track how a low-rank subspace in \mathbf{R} evolves over time [105].

3.4 Computational Complexity and Low-Complexity Estimators

The downside of having many antennas is that there are many signal observations to process in the digital baseband. We will now assess the computational complexity of MMSE estimation, using the methodology described in Appendix B.1.1 on p. 558, where only the numbers of complex multiplications and divisions are counted. The MMSE channel estimates in Theorem 3.1 are computed at BS j once per coherence block for each of the K_j intra-cell UEs. The inter-cell channels are optional to estimate, but if they are utilized in the precoding/combining, they also need to be computed once per coherence block. The processed received pilot signal at BS j is multiplied in (3.9) with two $M_j \times M_j$ matrices. Since these matrices only depend on the spatial correlation matrices, the matrix product can be precomputed and only updated when the channel statistics have changed substantially (e.g., due to UE mobility or new UE scheduling decisions). Note that the channel statistics typically are the same over all subcarriers, so only one matrix is precomputed per UE. The precomputation generally requires $(4M_j^3 - M_j)/3$ complex multiplications and M_j complex divisions per UE (see Lemma B.2 on p. 559 for details). If we estimate the channels from multiple UEs that use the same pilot sequence, we only need to compute the Ψ -matrix once and thus only spend M_j^3 multiplications per additional UE (except the first one).

In contrast to the cubic complexity of the precomputations, the estimation in every coherence block only entails correlating the received signal matrix with the pilot sequence as $\mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^*$ and then mul-

tiplying it with the precomputed statistical matrix $\sqrt{p_{li}} \mathbf{R}_{li}^j \boldsymbol{\Psi}_{li}^j$. These operations require $M_j \tau_p + M_j^2$ complex multiplications per UE (see Lemma B.1 on p. 559) and can be parallelized by computing $\tau_p + M_j$ complex multiplications separately for each of the M_j antennas. Hence, one can imagine a hardware implementation with very efficient computations in every coherence block, based on precomputed statistical matrices, and a less time-critical outer process that updates the precomputed matrices at regular intervals. If we need to estimate the channels to another UE that uses the same pilot, the additional cost is only M_j^2 multiplications since \mathbf{y}_{jli}^p is already known.

Remark 3.3 (Polynomial matrix expansion). In case the hardware implementation cannot handle the computational complexity of exact MMSE estimation, one can resort to approximations. For example, [299] proposed a method to rewrite the matrix inverse in the MMSE estimation expression as an equivalent polynomial matrix expansion, which can then be truncated since the lower-order polynomial terms have the most significant impact on the estimate. A similar approach can be taken for receive combining and transmit precoding, as explained later in Remark 4.2 on p. 296, where the main principle is also outlined. The complexity of this method is quadratic in M_j and linear in the number of terms that are used in the truncated polynomial. Another option is to utilize an estimator that from the beginning does not require matrix-matrix multiplications or large-dimensional inversions. Some options are described next.

3.4.1 Alternative Channel Estimation Schemes

If BS j cannot manage the computational complexity of MMSE channel estimation, there are alternative estimation schemes. An arbitrary linear estimator of \mathbf{h}_{li}^j , based on \mathbf{y}_{jli}^p in (3.2), can be written as $\mathbf{A}_{li}^j \mathbf{y}_{jli}^p$, for some deterministic matrix $\mathbf{A}_{li}^j \in \mathbb{C}^{M_j \times M_j}$ that specifies the estimation scheme. The corresponding MSE $\mathbb{E}\{\|\mathbf{h}_{li}^j - \mathbf{A}_{li}^j \mathbf{y}_{jli}^p\|^2\}$ can be computed

as

$$\begin{aligned} \text{MSE}(\mathbf{A}_{li}^j) &= \text{tr}(\mathbf{R}_{li}^j) - 2\sqrt{p_{li}}\tau_p \Re \left(\text{tr}(\mathbf{R}_{li}^j \mathbf{A}_{li}^j) \right) \\ &\quad + \tau_p \text{tr} \left(\mathbf{A}_{li}^j \left(\mathbf{\Psi}_{li}^j \right)^{-1} (\mathbf{A}_{li}^j)^H \right) \end{aligned} \quad (3.29)$$

with $\mathbf{\Psi}_{li}^j$ given by (3.10). The MMSE estimator is obtained for $\mathbf{A}_{li}^j = \sqrt{p_{li}}\mathbf{R}_{li}^j \mathbf{\Psi}_{li}^j$, but we can alternatively choose an \mathbf{A}_{li}^j that makes the estimate easier to compute. Diagonal matrices are particularly useful to reduce the computational complexity since each element of \mathbf{y}_{jli}^p can then be multiplied with only one scalar instead of M_j non-zero scalars from \mathbf{A}_{li}^j . For any deterministic \mathbf{A}_{li}^j , the estimate $\mathbf{A}_{li}^j \mathbf{y}_{jli}^p$ and estimation error $\tilde{\mathbf{h}}_{li}^j = \mathbf{h}_{li}^j - \mathbf{A}_{li}^j \mathbf{y}_{jli}^p$ are Gaussian distributed, but they are generally correlated random variables—an important difference from the MMSE estimator. In particular, we have that

$$\mathbb{E} \left\{ \hat{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H \right\} = \sqrt{p_{li}}\tau_p \mathbf{A}_{li}^j \mathbf{R}_{li}^j - \tau_p \mathbf{A}_{li}^j \left(\mathbf{\Psi}_{li}^j \right)^{-1} (\mathbf{A}_{li}^j)^H. \quad (3.30)$$

We will now provide two examples for selecting the matrix \mathbf{A}_{li}^j .

Element-wise MMSE Channel Estimator

Based on the discussion in Section 3.3.3, one obvious alternative is to estimate each element of \mathbf{h}_{li}^j separately and thereby ignore the correlation between the elements. More precisely, we can look at the processed received signal in (3.2) and only consider one of the M_j elements at a time. The following corollary provides the resulting element-wise MMSE (EW-MMSE) estimator.

Corollary 3.4. Based on the observation $[\mathbf{y}_{jli}^p]_m$, BS j can compute the MMSE estimate of the m th element $[\mathbf{h}_{li}^j]_m$ of the channel from UE i in cell l as

$$[\hat{\mathbf{h}}_{li}^j]_m = \frac{\sqrt{p_{li}}[\mathbf{R}_{li}^j]_{mm}}{\sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'} \tau_p [\mathbf{R}_{l'i'}^j]_{mm} + \sigma_{\text{UL}}^2} [\mathbf{y}_{jli}^p]_m. \quad (3.31)$$

The estimation error variance of this element is

$$[\mathbf{R}_{li}^j]_{mm} - \frac{p_{li}\tau_p \left([\mathbf{R}_{li}^j]_{mm}\right)^2}{\sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'}\tau_p [\mathbf{R}_{l'i'}^j]_{mm} + \sigma_{UL}^2}. \quad (3.32)$$

Proof. The proof is identical to that of Theorem 3.1, except that we only consider one of the elements in $\hat{\mathbf{h}}_{li}^j$ and the corresponding element in \mathbf{y}_{jli}^p to perform the estimation. \square

The EW-MMSE estimator corresponds to letting \mathbf{A}_{li}^j be diagonal with

$$[\mathbf{A}_{li}^j]_{mm} = \frac{\sqrt{p_{li}}[\mathbf{R}_{li}^j]_{mm}}{\sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'}\tau_p [\mathbf{R}_{l'i'}^j]_{mm} + \sigma_{UL}^2} \quad m = 1, \dots, M. \quad (3.33)$$

The computational complexity per UE is proportional to M_j , both when precomputing the fractional expression in (3.31) (at the slow time scale that the large-scale fading changes) and when multiplying it with the processed received pilot signal once per coherence block. This is substantially lower than the complexity of the original MMSE estimator, except in the special case when all the spatial correlation matrices are diagonal so that one can estimate each channel element separately without performance loss. Note that the main complexity saving comes from the fact that \mathbf{A}_{li}^j is diagonal.

The MSE achieved by the EW-MMSE estimator is obtained by summing up the estimation error variances from (3.32), which can be expressed as

$$\text{MSE} = \text{tr}(\mathbf{R}_{li}^j) - \sum_{m=1}^M \frac{p_{li}\tau_p \left([\mathbf{R}_{li}^j]_{mm}\right)^2}{\sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'}\tau_p [\mathbf{R}_{l'i'}^j]_{mm} + \sigma_{UL}^2}. \quad (3.34)$$

Although each element is estimated using the MMSE principle, the vector with estimates and the vector with estimation errors are correlated when using the EW-MMSE estimator as it follows by inserting (3.33) into (3.30).

Least-square Channel Estimator

The EW-MMSE estimator does not utilize the full spatial correlation matrices, but only the elements on the main diagonals (which can be estimated as described in (3.26)). In case these partial statistics are unknown or unreliable (e.g., due to rapid changes in the UE scheduling in other cells), it might be necessary to consider estimators that require no prior statistical information. The least-squares (LS) estimator has been used for this purpose since the beginning of SDMA [125, 37]. In our setup, we have the observation \mathbf{y}_{jli}^p in (3.2), which contains the desired channel in the form of $\sqrt{p_{li}}\tau_p \mathbf{h}_{li}^j$. The LS estimate of \mathbf{h}_{li}^j is defined as the vector $\hat{\mathbf{h}}_{li}^j$ that minimizes the squared deviation $\|\mathbf{y}_{jli}^p - \sqrt{p_{li}}\tau_p \hat{\mathbf{h}}_{li}^j\|^2$. The smallest value is zero and is attained by

$$\hat{\mathbf{h}}_{li}^j = \frac{1}{\sqrt{p_{li}}\tau_p} \mathbf{y}_{jli}^p. \quad (3.35)$$

This is a linear estimator with

$$\mathbf{A}_{li}^j = \frac{1}{\sqrt{p_{li}}\tau_p} \mathbf{I}_{M_j} \quad (3.36)$$

and since the matrix is diagonal, the computational complexity per coherence block is proportional to M_j . The matrix \mathbf{A}_{li}^j has no explicit dependence on the channel statistics, but it depends on the transmit power, which the UE might change when the statistics change.

The MSE achieved by the LS estimator in (3.35) cannot be computed unless the channel statistics are actually known, but it can be obtained by substituting $\mathbf{A}_{li}^j = \frac{1}{\sqrt{p_{li}}\tau_p} \mathbf{I}_{M_j}$ into (3.29) and simplifying:

$$\text{MSE} = \text{tr} \left(\sum_{(l',i') \in \mathcal{P}_{li} \setminus (l,i)} \frac{p_{l'i'}}{p_{li}} \mathbf{R}_{l'i'}^j + \frac{\sigma_{\text{UL}}^2}{p_{li}\tau_p} \mathbf{I}_{M_j} \right). \quad (3.37)$$

Note that, since the LS estimator is suboptimal, the estimate and the estimation error are correlated:

$$\mathbb{E} \left\{ \hat{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H \right\} = \mathbf{R}_{li}^j - \frac{1}{p_{li}\tau_p} \left(\Psi_{li}^j \right)^{-1}. \quad (3.38)$$

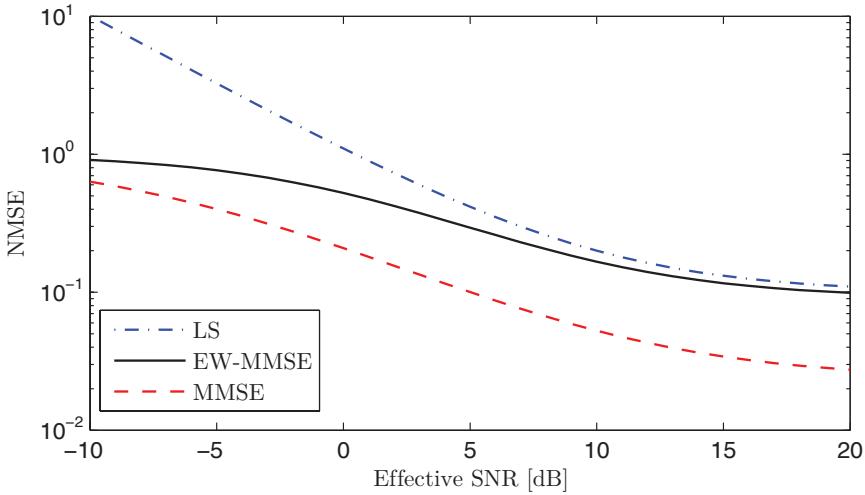


Figure 3.7: NMSE in the estimation of a spatially correlated channel, based on the local scattering model with Gaussian angular distribution, for different estimators. The results are averaged over different nominal angles and $\sigma_\varphi = 10^\circ$.

3.4.2 Comparison of Complexity and Estimation Quality

The estimation quality of the MMSE, EW-MMSE, and LS estimators are compared in Figure 3.7, in terms of NMSE. We consider a scenario where BS j estimates the channel of its UE k , while a UE in another cell transmits the same pilot sequence. The effective SNR of the desired UE is varied from -10 dB to 20 dB, while the interfering signal is assumed to always be 10 dB weaker. The local scattering model is considered with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$, and the results are averaged over different nominal angles between 0° and 360° . Figure 3.7 shows that the three estimators provide rather different NMSEs. The MMSE estimator is systematically the best estimator since it fully exploits the spatial channel correlation. The EW-MMSE estimator provides decent estimation performance (equivalent to MMSE estimation of an uncorrelated channel), but there is a substantial gap from the MMSE estimator—even at high SNR where the error floor (caused by pilot contamination) has a higher value. The LS estimator performs very poorly at low SNR where the NMSE is above 1, while the trivial all-zero estimate $\hat{\mathbf{h}}_{jk}^j = \mathbf{0}_{M_j}$ gives an NMSE of 1. At higher

Scheme	Correlating with pilot	Per UE	Precomputation
MMSE	$M_j \tau_p$	M_j^2	$\frac{4M_j^3 - M_j}{3}$
EW-MMSE	$M_j \tau_p$	M_j	M_j
LS	$M_j \tau_p$	—	—

Table 3.1: Computational complexity per coherence block for channel estimation. The first column is the number of multiplications when correlating the received signal with a pilot sequence and the second column is the multiplications required for estimating the channel of a UE using that pilot sequence. The third column is the complexity for precomputation per UE.

SNRs, the LS estimator is comparable to the EW-MMSE estimator, but their respective error floors are different (if there is pilot contamination). The LS estimator can provide decent estimates of the channel direction, $\mathbf{h}_{jk}^j / \|\mathbf{h}_{jk}^j\|$, while the lack of statistical information makes it harder to get the right scaling of the channel norm $\|\mathbf{h}_{jk}^j\|$.

The computational complexities of the MMSE, EW-MMSE, and LS estimators are summarized in Table 3.1, in terms of complex multiplications. The complexity is divided into three parts: The complexity of correlating the received signal \mathbf{Y}_j^p with a pilot sequence, the complexity of estimating the channel of a UE (after correlating with its pilot), and the complexity of precomputing the statistical coefficients. The complexities of these operations were computed in Section 3.4 for the MMSE estimator, while the complexities of EW-MMSE and LS are obtained analogously, based on Appendix B.1.1 on p. 558. The MMSE estimator has the highest complexity, both when computing an estimate and when precomputing the statistical coefficients. EW-MMSE is more complex than LS, but the difference is small when comparing the sums of the first two columns: $M_j \tau_p + M_j \approx M_j \tau_p$ for practical values of τ_p .

Figure 3.8 illustrates the number of complex multiplications per coherence block as a function of M in a scenario with $K = \tau_p = 10$ UEs in each cell. We have neglected the complexity for precomputation of matrices that only depend on the channel statistics, as they typically are constant for a large number of coherence blocks. MMSE has the highest complexity, followed by EW-MMSE, which reduces the complexity by 45%–90% since the correlation between antennas is not exploited in the

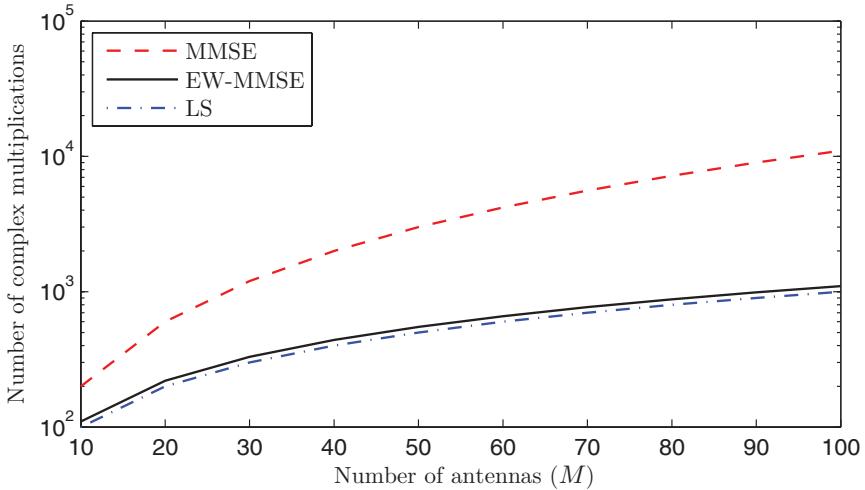


Figure 3.8: Number of complex multiplications per coherence block with 10 UEs, when using different channel estimation schemes. The complexity of precomputing statistical matrices is not accounted for.

channel estimation. The complexity reduction of using LS instead of EW-MMSE is marginal: for $M = 100$ we only save an additional 1% by using LS.

3.5 Data-Aided Channel Estimation and Pilot Decontamination

In scenarios with much pilot contamination, the pilot-based MMSE estimator in Theorem 3.1 might not be sufficient to get a good estimation quality. The amount of pilot contamination can be reduced by increasing τ_p so that each pilot sequence is reused less frequently in space, but at the price of using fewer samples for data transmission per coherence block. Alternatively, the UL data sequences can be utilized for channel estimation, so that the UEs' channels are discriminated based on transmitted sequences of length $\tau_p + \tau_u$ instead of τ_p . The data sequences are not known to the BS in advance, but can nonetheless be used for data-aided estimation, which is classically known as semi-blind estimation [75] and more recently called pilot decontamination [232].

This approach has been taken in a series of papers on Massive MIMO [242, 232, 362, 203, 238, 152, 333, 334]. The main principle is to form an $M_j \times M_j$ sample correlation matrix of the received block

of UL signals in a coherence block. Since the channels are constant, as $\tau_p + \tau_u \rightarrow \infty$, each of the strongest eigenvalues of the sample correlation matrix corresponds to one of the UEs and the respective eigenvector is an estimate of the UE channel (up to a phase ambiguity). BS j can typically infer that the K_j strongest eigenvalues correspond to its K_j UEs, while the weaker eigenvalues correspond to interfering UEs in other cells or receiver noise. By projecting the received signal onto the eigenspace of the K_j strongest eigenvalues, interference and noise can be rejected. The eigenvalue-based separation between signal and interference subspaces can be performed blindly (i.e., without pilot sequences) [232, 238, 334] or by also exploiting spatial channel correlation [362], but it is still desired to transmit orthogonal pilot sequences within each cell to identify which UE corresponds to which eigenvalue and to resolve the phase ambiguity in the channel estimates [242, 238]. Since the limit $\tau_p + \tau_u \rightarrow \infty$ requires the channel coherence time to grow indefinitely, which will not happen in practice, an exact separation of signal and interference subspaces is not possible with data-aided channel estimation and some pilot contamination will remain.¹ However, if implemented judiciously, data-aided channel estimation is always better or equally good as pilot-only MMSE estimation since it uses more observations in the estimation process. The largest benefits are observed when the SNR is low (because noise is also mitigated by the subspace projections) and when there are strong sources of interference [333]. The downside of data-aided channel estimation is the increased computational complexity.

Remark 3.4 (Alternative pilot structures). We have considered the coherence block structure in Figure 2.2, where all UEs send their pilots simultaneously, followed by simultaneous transmission of UL data, and then DL data transmission. There are alternative approaches. One approach is to time-shift the coherence blocks between cells, such that some cells send DL data when their neighboring cells send UL pilots and vice versa [114]. With this approach, there is still inter-cell interference

¹In fact, pilot contamination is not an issue if $\tau_p + \tau_u \rightarrow \infty$, because then we can allow for $\tau_p = \sum_{j=1}^L K_j$ so that each UE can get its own orthogonal pilot while keeping the channel estimation overhead negligible.

that reduces the estimation quality, but it is now the BSs in neighboring cells that cause the largest contamination and not the UEs in those cells. Hence, the channel estimates of the desired UEs in a given cell will now be correlated with the channels from neighboring BSs, which is less critical since these channels are irrelevant during data transmission.

Another approach is to superimpose the pilot sequences on the UL data transmission, which allows for setting $\tau_p = 0$ and nonetheless having τ_u mutually orthogonal pilot sequences at disposal [320, 328]. The benefit of this approach is that long pilot sequences can be transmitted (and reduced infrequently in the network) without sacrificing the number of samples available for data transmission. The price to pay is additional interference between pilot and data transmissions, which can be rather large since, in contrast to the pilot design in this section, there is also interference between intra-cell UEs during channel estimation. It is therefore important to divide the transmit power between pilots and data in a judicious way. A hybrid pilot solution, where some users have superimposed pilots and some others have conventional pilots, may bring the best of both paradigms [319].

3.6 Summary of Key Points in Section 3

- Channel estimation at the BS is key to achieve the full potential of Massive MIMO. This is typically accomplished using UL pilot transmission.
- The MMSE estimator exploits the statistical characteristics to obtain good estimates. Spatial correlation makes it easier to estimate channels to large antenna arrays. The gains are robust to imperfect knowledge of the statistics.
- The computational complexity of the MMSE estimator grows quadratically with the number of antennas. The alternative EW-MMSE estimator can greatly reduce complexity by neglecting the spatial correlation between antennas. If the channel statistics are unknown, the LS estimator can be used instead.
- Since the channel coherence blocks are of limited size, it is necessary to reuse pilot sequences across cells. The inter-cell interference increases the estimation errors and also makes the channel estimates of two UEs that use the same pilot are correlated. This phenomenon is called pilot contamination. The correlation is low when the channel gain of the interfering UE is weak, as compared to that of the desired UE, or when the correlation matrices are sufficiently different.

4

Spectral Efficiency

In this section, we analyze the achievable UL and DL SEs, based on the channel estimation framework developed in the previous section. Expressions for the SE in the UL are derived in Section 4.1. Different receive combining schemes are evaluated and the impacts of spatial channel correlation and pilot contamination are revisited. In Section 4.3, achievable SE expressions for the DL are derived with different DL channel estimation schemes. The key differences and similarities between the UL and DL expressions are described and the performance of different precoding schemes is evaluated. The asymptotic behavior of the SE, when the number of BS antennas grows infinitely large, is considered in Section 4.4. The key points are summarized in Section 4.5.

4.1 Uplink Spectral Efficiency and Receive Combining

We will now study the achievable SE of the UL payload data transmission with different receive combining schemes. Each BS detects the desired signals by using linear receive combining. Recall that UE k in cell j transmits a random data signal $s_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, p_{jk})$ for $j = 1, \dots, L$ and

$k = 1, \dots, K_j$. The variance p_{jk} is the transmit power (i.e., the average energy per sample).

The receiving BS j selects the combining vector $\mathbf{v}_{jk} \in \mathbb{C}^{M_j}$ for its k th UE, as a function of the channel estimates obtained from the pilot transmission. The combining vector should depend on $\hat{\mathbf{h}}_{jk}^j$, in order to coherently combine the desired signal components received over the M_j antennas, but it can also depend on the estimates of other channels, if the BS wishes to suppress interference (from the own and/or other cells). During data transmission, BS j correlates the received signal \mathbf{y}_j from (2.5) with the combining vector to obtain

$$\begin{aligned} \mathbf{v}_{jk}^H \mathbf{y}_j = & \underbrace{\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j s_{jk}}_{\text{Desired signal over estimated channel}} + \underbrace{\mathbf{v}_{jk}^H \tilde{\mathbf{h}}_{jk}^j s_{jk}}_{\text{Desired signal over unknown channel}} \\ & + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} \mathbf{v}_{jk}^H \mathbf{h}_{ji}^j s_{ji}}_{\text{Intra-cell interference}} + \underbrace{\sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j s_{li}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{v}_{jk}^H \mathbf{n}_j}_{\text{Noise}}. \end{aligned} \quad (4.1)$$

A similar expression was given in (2.6), but the key difference in (4.1) is that the desired signal term has been divided into two parts: one that is received over the known estimated channel $\hat{\mathbf{h}}_{jk}^j$ from UE k in the cell and one that is received over the unknown estimation error $\tilde{\mathbf{h}}_{jk}^j$ of the channel. The former part can be utilized straight away for signal detection, while the latter part is less useful since only the distribution of the estimation error is known (see Corollary 3.2 on p. 250). The SE in Massive MIMO is generally computed by treating the latter part as additional interference in the signal detection, by utilizing Corollary 1.3 on p. 171. In doing so, we obtain the following result.

Theorem 4.1. If MMSE channel estimation is used, then the UL ergodic channel capacity of UE k in cell j is lower bounded by $\text{SE}_{jk}^{\text{UL}}$ [bit/s/Hz] given by

$$\text{SE}_{jk}^{\text{UL}} = \frac{\tau_u}{\tau_c} \mathbb{E} \left\{ \log_2 \left(1 + \text{SINR}_{jk}^{\text{UL}} \right) \right\} \quad (4.2)$$

with

$$\text{SINR}_{jk}^{\text{UL}} = \frac{p_{jk} |\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j|^2}{\sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} p_{li} |\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{li}^j|^2 + \mathbf{v}_{jk}^H \left(\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right) \mathbf{v}_{jk}} \quad (4.3)$$

and where the expectation is with respect to the channel estimates.

Proof. The proof is available in Appendix C.3.1 on p. 593. \square

The capacity lower bound in Theorem 4.1 represents an achievable SE for the UL. Later in this section, we will provide an alternative lower bound, which is less tight but commonly used in research papers since it can lead to closed-form expressions. We refer to $\text{SINR}_{jk}^{\text{UL}}$ in (4.3) as the UL instantaneous SINR since it appears as $\frac{\tau_u}{\tau_c} \mathbb{E}\{\log_2(1 + \text{SINR}_{jk}^{\text{UL}})\}$ in the SE expression. However, it is not an SINR in the conventional sense, because it involves both instantaneous channel estimates and averages over channel estimation errors—this implies that we cannot measure $\text{SINR}_{jk}^{\text{UL}}$ in a given coherence block. Note that $\text{SINR}_{jk}^{\text{UL}}$ is a random variable that takes a new independent realization in each coherence block. The pre-log factor $\frac{\tau_u}{\tau_c}$ in (4.2) is the fraction of samples per coherence block that are used for UL data. Since $\tau_u = \tau_c - \tau_p - \tau_d$, the pre-log factor increases if we shorten the length τ_p of the pilot sequences (i.e., reduce the pilot overhead) and/or reduce the number of samples τ_d used for DL data.

The SE expression provided in Theorem 4.1 holds for any choice of the receive combining vector, under the assumption that the MMSE estimator is used for channel estimation. MR combining with $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j$ is commonly considered in the Massive MIMO literature, often motivated by asymptotic arguments that only apply for uncorrelated Rayleigh fading channels with very many antennas [208, 49]. We will show in Section 4.4 that MR is generally not asymptotically optimal. For this reason, we will not assume the use of MR here. Instead, we will optimize the combining vector and compare the result with MR and other alternative schemes. Note that $\text{SINR}_{jk}^{\text{UL}}$ only depends on \mathbf{v}_{jk} , thus each combining vector can be tailored to its associated UE

without taking the SE achieved by other UEs into account. The following corollary finds the “optimal” receive combining vector, in the sense of maximizing the SE expression provided in Theorem 4.1.

Corollary 4.2. The instantaneous UL SINR in (4.3) for UE k in cell j is maximized by the multicell minimum mean-squared error (M-MMSE) combining vector

$$\mathbf{v}_{jk} = p_{jk} \left(\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} (\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \mathbf{C}_{li}^j) + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \hat{\mathbf{h}}_{jk}^j \quad (4.4)$$

which leads to

$$\begin{aligned} \text{SINR}_{jk}^{\text{UL}} &= \\ p_{jk} (\hat{\mathbf{h}}_{jk}^j)^H &\left(\sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} p_{li} \hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \hat{\mathbf{h}}_{jk}^j. \end{aligned} \quad (4.5)$$

Proof. The proof is available in Appendix C.3.2 on p. 594. \square

We mentioned M-MMSE receive combining already in (1.42), but Corollary 4.2 derives its expression for the practical case when only the estimated channels are known. It is called M-MMSE combining since (4.4) not only maximizes the instantaneous SINR but also minimizes the MSE in the data detection; that is, the average squared distance between the desired signal and the processed received signal.

Corollary 4.3. The M-MMSE combining vector in (4.4) is the vector \mathbf{v}_{jk} that minimizes the conditional MSE

$$\mathbb{E} \left\{ |s_{jk} - \mathbf{v}_{jk}^H \mathbf{y}_j|^2 \mid \{\hat{\mathbf{h}}_{li}^j\} \right\} \quad (4.6)$$

where the expectation is conditioned on the current set $\{\hat{\mathbf{h}}_{li}^j\}$ of all channel estimate realizations (for $l = 1, \dots, L$ and $i = 1, \dots, K_l$).

Proof. The proof is available in Appendix C.3.3 on p. 595. \square

Note that the combining vector in (4.4) is the only one that minimizes the MSE, while the instantaneous SINR in (4.3) does not change if we multiply \mathbf{v}_{jk} with an arbitrary non-zero scalar (i.e., we can normalize the vector arbitrarily). The latter can be seen as an artifact from the mutual information definition, which disregards non-destructive signal processing because it does not reduce the information content. For practical discrete signal constellations, such as QAM, the scaling is important in the detection; the received signals are equalized to match the given decision regions for the constellation. M-MMSE combining for Massive MIMO has previously been studied in [246, 134, 193, 43, 239].

The structure of M-MMSE combining is quite intuitive. The matrix that is inverted in (4.4) is the conditional correlation matrix $\mathbf{C}_{\mathbf{y}_j} = \mathbb{E}\{\mathbf{y}_j\mathbf{y}_j^H | \{\hat{\mathbf{h}}_{li}^j\}\}$ of the received signal in (4.1), given the current set of channel estimates. The multiplication $\mathbf{C}_{\mathbf{y}_j}^{-1/2}\mathbf{y}_j$ corresponds to whitening of the received signal; that is, $\mathbb{E}\{\mathbf{C}_{\mathbf{y}_j}^{-1/2}\mathbf{y}_j(\mathbf{C}_{\mathbf{y}_j}^{-1/2}\mathbf{y}_j)^H | \{\hat{\mathbf{h}}_{li}^j\}\} = \mathbf{I}_M$. The whitened received signal has spatially uncorrelated elements, which means that the total received power is equally strong in all directions. If we denote the whitened combining vector as \mathbf{u}_{jk} , it is related to the original combining vector as $\mathbf{v}_{jk} = \mathbf{C}_{\mathbf{y}_j}^{-1/2}\mathbf{u}_{jk}$. The highest desired signal power is now received from the spatial direction $\mathbf{C}_{\mathbf{y}_j}^{-1/2}\hat{\mathbf{h}}_{jk}^j$ and due to the whitening, which makes the total power equal in all directions, the interference plus noise power is lowest in this direction. Hence, the optimal whitened combining vector can be selected as $\mathbf{u}_{jk} = \mathbf{C}_{\mathbf{y}_j}^{-1/2}\hat{\mathbf{h}}_{jk}^j$. This results into $\mathbf{v}_{jk} = \mathbf{C}_{\mathbf{y}_j}^{-1/2}\mathbf{u}_{jk} = \mathbf{C}_{\mathbf{y}_j}^{-1}\hat{\mathbf{h}}_{jk}^j$, which is equal to (4.4) up to a scaling factor. In other words, M-MMSE combining is obtained by whitening followed by MR combining. The whitening process is illustrated in Figures 4.1 and 4.2, where we observe that the choice of combining vector clearly affects the powers of the desired and interfering signals. The M-MMSE combining vector is easily identified from the whitened signal, while this is not the case when inspecting the original signal.

The “multicell” notion is not strictly necessary in the name M-MMSE, but we will use it to differentiate the true MMSE combining in (4.4) from the single-cell variant described in Section 4.1.1.

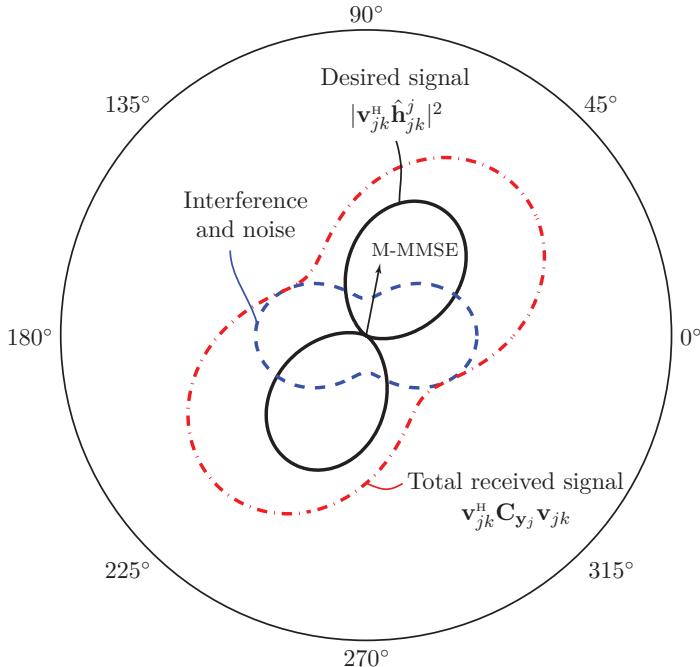


Figure 4.1: The total received signal power $\mathbf{v}_{jk}^H \mathbf{C}_{y_j}^{-1} \mathbf{v}_{jk}$ depends on the combining vector \mathbf{v}_{jk} . It is shown in this figure as the distance from the origin for different angles of a unit-norm combining vector, assuming $M = 2$ and that all vectors are real-valued. The total received power is divided into desired signal power $|\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j|^2$ and interference plus noise power $\mathbf{v}_{jk}^H \mathbf{C}_{y_j}^{-1} \mathbf{v}_{jk} - |\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j|^2$. M-MMSE combining finds a non-trivial tradeoff between high signal power and low interference/noise, which maximizes the instantaneous SINR.

By defining the diagonal matrix $\mathbf{P}_l = \text{diag}(p_{l1}, \dots, p_{lK_l}) \in \mathbb{R}^{K_l \times K_l}$ with the transmit powers of all UEs in cell l , we can collect the M-MMSE combining vectors for all UEs in cell j in a compact matrix form:

$$\begin{aligned} \mathbf{V}_j^{\text{M-MMSE}} &= \left[\mathbf{v}_{j1} \dots \mathbf{v}_{jK_j} \right] \\ &= \left(\sum_{l=1}^L \hat{\mathbf{H}}_l^j \mathbf{P}_l (\hat{\mathbf{H}}_l^j)^H + \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \hat{\mathbf{H}}_j^j \mathbf{P}_j \end{aligned} \quad (4.7)$$

where $\hat{\mathbf{H}}_l^j$ was defined in (3.12) as a matrix containing the estimates of all channels from UEs in cell l to BS j .

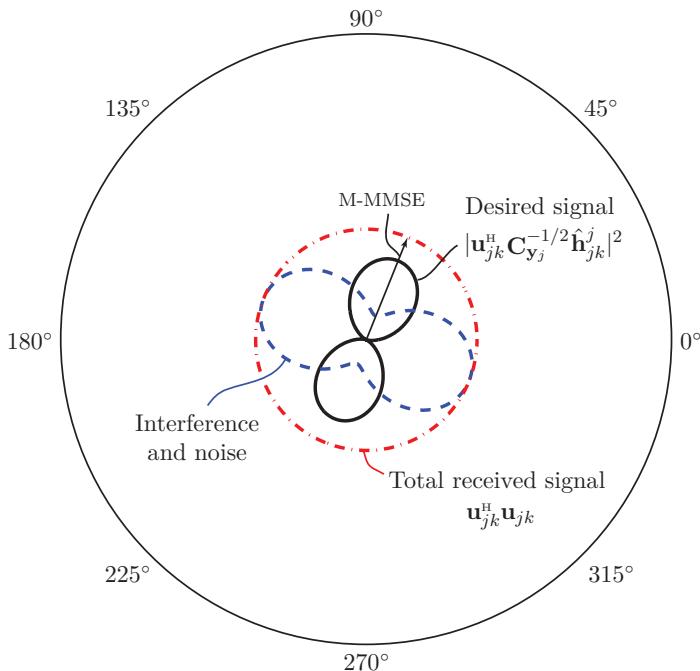


Figure 4.2: This figure continues the example from Figure 4.1. It shows how the total received signal power $\mathbf{u}_{jk}^H \mathbf{u}_{jk}$ of the whitened signal depends on the combining vector $\mathbf{u}_{jk} = \mathbf{C}_{y_j}^{1/2} \mathbf{v}_{jk}$. The total received power is divided into desired signal power $|\mathbf{u}_{jk}^H \mathbf{C}_{y_j}^{-1/2} \hat{\mathbf{h}}_{jk}^j|^2$ and interference plus noise power $\mathbf{u}_{jk}^H \mathbf{u}_{jk} - |\mathbf{u}_{jk}^H \mathbf{C}_{y_j}^{-1/2} \hat{\mathbf{h}}_{jk}^j|^2$. M-MMSE combining jointly maximizes the desired signal power and minimizes the interference/noise of the whitened signal.

4.1.1 Alternative Receive Combining Schemes

Although M-MMSE combining is optimal, it is not so frequently used in the research literature. There are several reasons for this. One is the high computational complexity of computing the $M_j \times M_j$ matrix inverse in (4.7) when M_j is large. The complexity is also affected by the need to estimate the channels and acquiring the channel statistics of all UEs. Another reason is that the performance of M-MMSE is hard to analyze mathematically, while there are alternative schemes that can give more insightful closed-form SE expressions. A third reason is that receive combining schemes often are developed for single-cell scenarios and then applied heuristically in multicell scenarios.

We will now present the alternative receive combining schemes that are most common in the literature and explain how these are obtained as simplifications of M-MMSE combining. The alternative schemes are generally suboptimal and the conditions under which they are nearly optimal are generally not satisfied in practice. Hence, the alternative schemes provide lower SEs but are practically useful to reduce the computational complexity and/or the amount of channel estimates and channel statistics that are needed to compute the combining matrix \mathbf{V}_j .

If BS j only estimates the channels from its own UEs [148, 135, 184], we obtain the single-cell minimum mean-squared error (S-MMSE) combining scheme with¹

$$\begin{aligned} \mathbf{V}_j^{\text{S-MMSE}} &= \left(\hat{\mathbf{H}}_j^j \mathbf{P}_j (\hat{\mathbf{H}}_j^j)^H + \sum_{i=1}^{K_j} p_{ji} \mathbf{C}_{ji}^j + \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} p_{li} \mathbf{R}_{li}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \hat{\mathbf{H}}_j^j \mathbf{P}_j. \end{aligned} \quad (4.8)$$

This combining matrix is obtained from (4.7) by replacing the term $\hat{\mathbf{H}}_l^j \mathbf{P}_l (\hat{\mathbf{H}}_l^j)^H + \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j$ with its average $\mathbb{E}\{\hat{\mathbf{H}}_l^j \mathbf{P}_l (\hat{\mathbf{H}}_l^j)^H + \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j\} = \sum_{i=1}^{K_l} p_{li} \mathbf{R}_{li}^j$ for all $l \neq j$, using Corollary 3.2 on p. 250. This scheme coincides with M-MMSE when there is only one isolated cell, but is generally different and has a substantially weaker ability to suppress interference from interfering UEs in other cells. This can be a major drawback since a few strong interfering UEs in other cells might be located near the cell edge and thus cause as much interference as the intra-cell UEs.

If the channel conditions are good and the interfering signals from other cells are weak, we can neglect all the correlation matrices in (4.8)

¹Strictly speaking, if there is no pilot contamination, S-MMSE minimizes the MSE $\mathbb{E}\{|s_{jk} - \mathbf{v}_{jk}^H \mathbf{y}_j|^2 | \{\hat{\mathbf{h}}_{ji}^j\}\}$ given only the set $\{\hat{\mathbf{h}}_{ji}^j\}$ of intra-cell channel estimates.

and obtain

$$\begin{aligned}\mathbf{V}_j^{\text{RZF}} &= \left(\hat{\mathbf{H}}_j^j \mathbf{P}_j (\hat{\mathbf{H}}_j^j)^H + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \hat{\mathbf{H}}_j^j \mathbf{P}_j \\ &= \hat{\mathbf{H}}_j^j \mathbf{P}_j^{\frac{1}{2}} \left(\mathbf{P}_j^{\frac{1}{2}} (\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j \mathbf{P}_j^{\frac{1}{2}} + \sigma_{\text{UL}}^2 \mathbf{I}_{K_j} \right)^{-1} \mathbf{P}_j^{\frac{1}{2}} \\ &= \hat{\mathbf{H}}_j^j \left((\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j + \sigma_{\text{UL}}^2 \mathbf{P}_j^{-1} \right)^{-1}\end{aligned}\quad (4.9)$$

where the second equality follows from the first matrix identity in Lemma B.5 on p. 560. We call this regularized zero-forcing (RZF) combining. The main benefit over S-MMSE is that a $K_j \times K_j$ matrix is inverted in (4.9) instead of an $M_j \times M_j$ matrix, which can substantially reduce the complexity since $M_j \gg K_j$ is typical in Massive MIMO. This benefit comes with a SE loss since, in general, the channel conditions will not be good to every UE and the interfering signals from other cells are non-negligible. The *regularization* terminology refers to the fact that (4.9) is a pseudo-inverse of the estimated channel matrix $\hat{\mathbf{H}}_j^j$ where the matrix that is inverted has been regularized by the diagonal matrix $\sigma_{\text{UL}}^2 \mathbf{P}_j^{-1}$. Regularization, a classic signal processing technique, improves the numerical stability of an inverse. In our case, it provides weighting between interference suppression (for small regularization terms) and maximizing the desired signals (for large regularization terms).

The combining expression in (4.9) can be further approximated when the SNR is high, in the sense that the regularization term $\sigma_{\text{UL}}^2 \mathbf{P}_j^{-1} \rightarrow 0 \mathbf{I}_{K_j}$. The same approximation can be applied in the regime of many antennas where $(\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j + \sigma_{\text{UL}}^2 \mathbf{P}_j^{-1} \approx (\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j$ since the diagonal of $(\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j$ increases with M_j while the regularization term remains constant. In both cases, we can neglect the regularization term and obtain the zero-forcing (ZF) combining matrix

$$\mathbf{V}_j^{\text{ZF}} = \hat{\mathbf{H}}_j^j \left((\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j \right)^{-1} \quad (4.10)$$

which is the pseudo-inverse of $(\hat{\mathbf{H}}_j^j)^H$. If we compute $(\hat{\mathbf{H}}_j^j)^H \mathbf{V}_j$ for any combining scheme, the k th diagonal matrix is the desired signal gain of the k th UE in cell j and the (k, i) th element represents the interference that UE k causes to UE i in the same cell (for $k \neq i$). The combining vector in (4.10) is called ZF because $(\hat{\mathbf{H}}_j^j)^H \mathbf{V}_j^{\text{ZF}} = (\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j ((\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j)^{-1} =$

\mathbf{I}_{K_j} which implies that (on average) all the interference from intra-cell UEs is canceled, while the desired signals remain non-zero. Since the true channel matrix is \mathbf{H}_j^j and not $\hat{\mathbf{H}}_j^j$, there will be residual interference also with ZF. Note that \mathbf{V}_j^{ZF} only exists if the $K_j \times K_j$ matrix $(\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j$ has full rank, which is typically the case when $M_j \geq K_j$. Since not every UE exhibits a high SNR in practice, it is expected that ZF will provide lower SEs than RZF.

In low SNR conditions, we instead have $(\hat{\mathbf{H}}_j^j)^H \hat{\mathbf{H}}_j^j + \sigma_{\text{UL}}^2 \mathbf{P}_j^{-1} \approx \sigma_{\text{UL}}^2 \mathbf{P}_j^{-1}$ and RZF in (4.9) is approximately equal to $\frac{1}{\sigma_{\text{UL}}^2} \hat{\mathbf{H}}_j^j \mathbf{P}_j$. If we further remove the diagonal matrix $\frac{1}{\sigma_{\text{UL}}^2} \mathbf{P}_j$ (recall that the normalization of a combining vector does not affect the instantaneous UL SINR), we obtain

$$\mathbf{V}_j^{\text{MR}} = \hat{\mathbf{H}}_j^j \quad (4.11)$$

which is known as MR combining. This scheme was considered already in Section 1, but the main difference is that we now use estimated channels instead of the exact ones (which are unknown in practice). Note that MR does not require any matrix inversion, in contrast to the previously mentioned schemes. Since not every UE exhibits a low SNR in practice, it is expected that MR will provide lower SEs than RZF.

4.1.2 Computational Complexity of Receive Combining

The computational complexity of the aforementioned receive combining schemes can be evaluated in detail using the framework provided in Appendix B.1.1 on p. 558. The basic complexity in the signal reception comes from computing $\mathbf{v}_{jk}^H \mathbf{y}_j$ for every received UL signal \mathbf{y}_j and every UE in the cell. This complexity is the same for every combining scheme. Each inner product requires M_j complex multiplications, which gives a total of $\tau_u M_j K_j$ complex multiplications per coherence block.

In addition, we need to account for the complexity of computing the combining matrix \mathbf{V}_j once per coherence block. The combining schemes in (4.7)–(4.11) are all computed using elementary matrix operations, such as matrix-matrix multiplication and matrix inversion. The computational complexity can be computed using the framework

described in Appendix B.1.1 on p. 558, where it is concluded that complex multiplications and divisions dominate the complexity, while additions and subtractions can be neglected. Table 4.1 summarizes the total complexity of each combining scheme, using Lemmas B.1 and B.2 on p. 559.² In these computations, it has been assumed that the intra-cell channel estimates $\hat{\mathbf{H}}_j^j$ and the scaled estimates $\hat{\mathbf{H}}_j^j \mathbf{P}_l^{\frac{1}{2}}$, as well as the statistical matrices $\sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j$, $\sum_{l \neq j} \sum_{i=1}^{K_l} p_{li} \mathbf{R}_{li}^j$, and $\sigma_{\text{UL}}^2 \mathbf{I}_{M_j}$ are available for free at BS j . This is because the complexity of the channel estimation was previously quantified in Table 3.1. However, since M-MMSE is the only scheme that utilizes the inter-cell channel estimates $\hat{\mathbf{H}}_l^j$, for $l \neq j$, we have included the complexity of computing these estimates in Table 4.1. This is the reason that the complexity of M-MMSE depends on which estimator that is used.

The key benefit of using another combining scheme than “optimal” M-MMSE is the reduced computational complexity. Figure 4.3 illustrates the number of complex multiplications per coherence block as a function of either the number of UEs or the number of BS antennas. We consider a scenario with $L = 9$ cells and $\tau_u = 200 - K$ samples for data transmissions. In particular, in Figure 4.3a we assume that $K \in [1, 40]$ and $M = 100$ in every cell (i.e., $K_j = K$ and $M_j = M$ for $j = 1, \dots, L$). On the other hand, in Figure 4.3b we consider $K = 10$ and let M vary from 10 to 100. Note that the vertical axes use a logarithmic scale. The complexity increases with the number of UEs and BSs antennas for all combining schemes. M-MMSE has clearly the highest complexity, followed by S-MMSE. As shown in Figure 4.3a, the use of S-MMSE reduces the complexity by 10%–50% over M-MMSE, since the inter-cell channel estimates are not utilized in the computation. The complexity reduction for $K = 10$ is 17%–37%, as illustrated in Figure 4.3b. RZF and ZF provide even lower complexity since these schemes invert

²MR combining, as defined in (4.11), is given directly from the channel estimates, without the need for any further multiplications or divisions. However, in practical implementations, we typically normalize the combining vector such that the factor $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j$ in front of the desired signal s_{jk} is close to one (or another constant). The complexity of this normalization is accounted for by K_j complex divisions in Table 4.1. The other combining schemes considered in this section provides this normalization automatically.

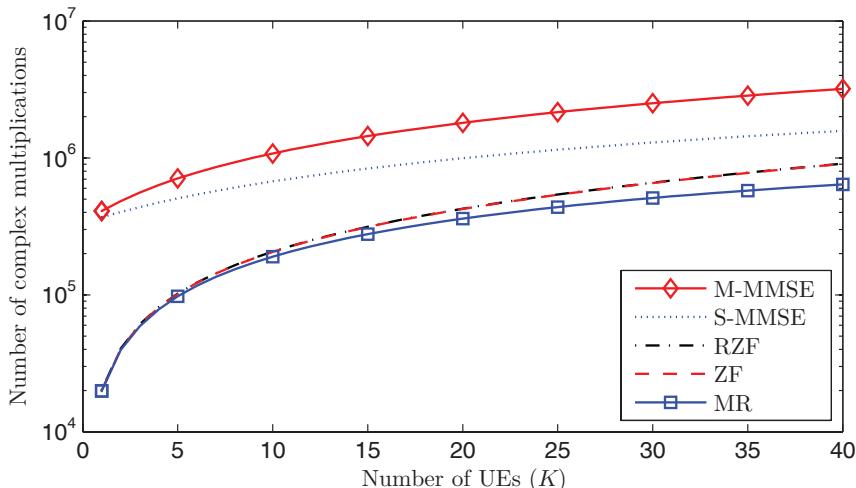
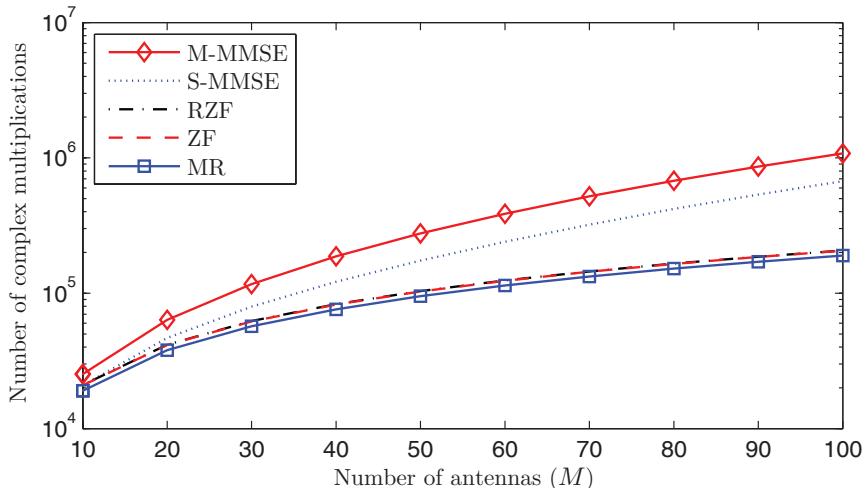
(a) Complexity for $M = 100$ and varying K .(b) Complexity for $K = 10$ and varying M .

Figure 4.3: Number of complex multiplications per coherence block when using different combining schemes. The computation of combining matrices and the inner products with received signals are both accounted for.

Scheme	Reception Multiplic.	Computing combining vectors Multiplic.	Div.
M-MMSE (MMSE est.)	$\tau_u M_j K_j$	$\sum_{l=1}^L \frac{(3M_j^2 + M_j)K_l}{2} + \frac{M_j^3 - M_j}{3} + M_j \tau_p (\tau_p - K_j)$	M_j
M-MMSE (EW-MMSE est.)	$\tau_u M_j K_j$	$\sum_{l=1}^L \frac{(M_j^2 + 3M_j)K_l}{2} + (M_j^2 - M_j)K_j + \frac{M_j^3 - M_j}{3} + M_j \tau_p (\tau_p - K_j)$	M_j
S-MMSE	$\tau_u M_j K_j$	$\frac{3M_j^2 K_j}{2} + \frac{M_j K_j}{2} + \frac{M_j^3 - M_j}{3}$	M_j
RZF	$\tau_u M_j K_j$	$\frac{3K_j^2 M_j}{2} + \frac{3K_j M_j}{2} + \frac{K_j^3 - K_j}{3}$	K_j
ZF	$\tau_u M_j K_j$	$\frac{3K_j^2 M_j}{2} + \frac{K_j M_j}{2} + \frac{K_j^3 - K_j}{3}$	K_j
MR	$\tau_u M_j K_j$	—	K_j

Table 4.1: Computational complexity per coherence block of different receive combining schemes. Only complex multiplications (Multiplic.) and complex divisions (Div.) are considered, while additions/subtractions are neglected; see Appendix B.1.1 on p. 558 for details.

substantially smaller $K_j \times K_j$ matrices (compared to the $M_j \times M_j$ matrices that are inverted by M-MMSE and S-MMSE). As it follows from Figure 4.3a, this property reduces the complexity by 72%–95% as compared to M-MMSE. Finally, MR provides the lowest computational complexity since no matrix inverses are computed, which also means that all computations can be parallelized in the implementation (a separate processing core can be used per antenna and UE). The complexity reduction compared to RZF and ZF, in number of multiplications, is only substantial when the number of UEs is large; Figure 4.3a shows that with $K = 10$, we only save 8% in complexity by using MR instead of RZF.

The price to pay for reduced complexity is a reduction in SE. Before illustrating the performance-complexity tradeoff, we will define a simulation scenario that will be repeatedly used throughout the monograph.

Parameter	Value
Network layout	Square pattern (wrap-around)
Number of cells	$L = 16$
Cell area	$0.25 \text{ km} \times 0.25 \text{ km}$
Number of antennas per BS	M
Number of UEs per cell	K
Channel gain at 1 km	$\Upsilon = -148.1 \text{ dB}$
Pathloss exponent	$\alpha = 3.76$
Shadow fading (standard deviation)	$\sigma_{sf} = 10$
Bandwidth	$B = 20 \text{ MHz}$
Receiver noise power	-94 dBm
UL transmit power	20 dBm
DL transmit power	20 dBm
Samples per coherence block	$\tau_c = 200$
Pilot reuse factor	$f = 1, 2 \text{ or } 4$
Number of UL pilot sequences	$\tau_p = fK$

Table 4.2: System parameters of the running example. Each cell covers a square area of $0.25 \text{ km} \times 0.25 \text{ km}$ and is deployed on a grid of 4×4 cells. A wrap-around topology is used, as illustrated in Figure 4.4. The UEs are uniformly and independently distributed in each cell, at distances larger than 35 m from the BS.

4.1.3 Definition of the Running Example

To exemplify the performance of Massive MIMO under somewhat realistic conditions, we will now define a 16-cell setup that will be used as a *running example* in the remainder of Section 4 and also in later sections. The key parameters are given in Table 4.2 and explained below. The purpose of the simulation examples is to qualitatively describe the basic phenomena and characteristics of Massive MIMO and to enable direct comparison between different simulation results. However, most simulations are based on rather simple channel models and power allocation schemes, so we cannot draw general quantitative conclusions. Optimized power allocation is considered in Section 7.1 on p. 452 and a case study using a realistic channel model and optimized power allocation is provided in Section 7.7 on p. 537.

In the running example, each cell covers a square of $0.25 \text{ km} \times 0.25 \text{ km}$ and is deployed on a square grid of 4×4 cells.³ A wrap-around topology is used to simulate that all BSs receive equally much interference from all directions; see Figure 4.4 for an illustration. More precisely, for each combination of UE and BS, we consider eight alternative locations of the BS and determine which one has the shortest distance to the UE. Only this location is used when computing the large-scale fading and nominal angle between the UE and BS. The large-scale fading model in (2.3) is used with the median channel gain $\Upsilon = -148.1 \text{ dB}$ at 1 km, $\alpha = 3.76$ as the pathloss exponent, and $\sigma_{\text{sf}} = 10$ as the standard deviation of the shadow fading. These propagation parameters are inspired by the NLoS macro cell 3GPP model for 2 GHz carriers, which are described in [119, A.2.1.1.2-3]. The UEs are uniformly and independently distributed in each cell, at distances larger than 35 m from the BS.⁴

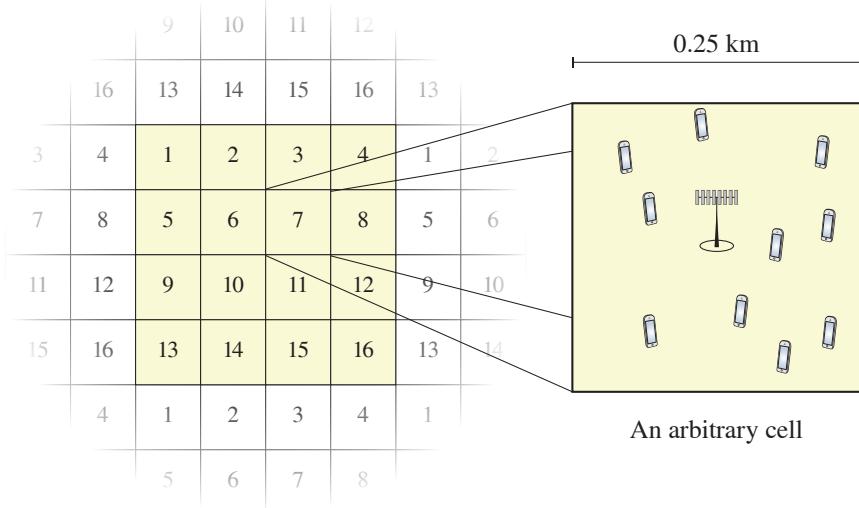
We consider communication over a 20 MHz bandwidth with a total receiver noise power of -94 dBm (consisting of thermal noise and a noise figure of 7 dB in the receiver hardware). Unless stated otherwise, we consider a UL transmit power of 20 dBm per UE and, when needed, each BS allocates 20 dBm of DL transmit power per UE. With these parameters, the median SNR of a UE at 35 m from its serving BS is 20.6 dB, while a UE in any of the corners of a square cell gets -5.8 dB . Note that the median removes the impact of the shadow fading, thus larger SNR variations are obtained in the simulations.

Two Rayleigh fading channel models with different spatial characteristics will be used along with the running example:

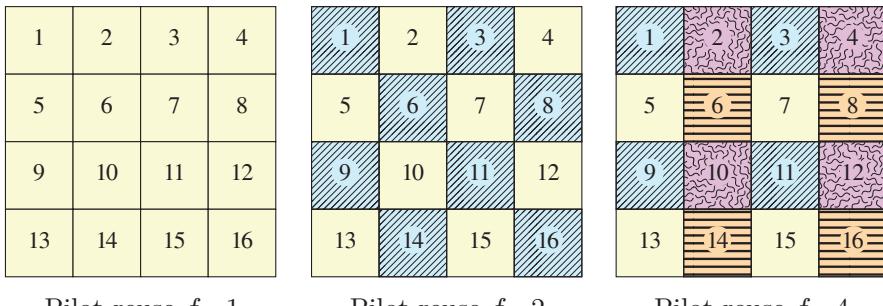
- **Gaussian local scattering with ASD σ_φ :** The spatial correlation matrices are generated using the local scattering model,

³Practical BSs are not deployed in such a regular pattern, but rather seem stochastically deployed [18, 200] because many external constraints affect the deployment. However, when studying the achievable performance of a network, it is common practice to consider an easily reproducible regular deployment.

⁴Due to the random shadow fading, it can happen that a UE gets a better channel to another BS than the one in its own square. In practice, such a UE may exploit this macro-diversity to connect to the other BS. In the running example, we disregard these situations by only considering collections of shadow fading realizations for UE k in cell j for which $\beta_{jk}^j \geq \beta_{jk}^l$ for $l = 1, \dots, L$. This makes sure that the UE connects to the BS in its own square, while retaining macro-diversity towards shadowing.



(a) Basic simulation setup



(b) Examples of pilot reuse factors

Figure 4.4: Illustration of the running example. 16 square cells are located on a 4×4 grid. (a) A wrap-around topology is considered where each cell has multiple locations and the distance between any arbitrary UE and arbitrary BS is always the shortest among all nine combinations. The angular properties are extracted from the shortest path. (b) Four different ways to reuse pilot sequences across the cells are shown, where a larger pilot reuse factor implies that more orthogonal pilot sequences are required.

defined in (2.23). The nominal angles are computed as the LoS angles between the UEs and the BSs. The angular distribution around the nominal angle is Gaussian with zero mean and standard deviation σ_φ , whose value is specified each time we use this model.

- **Uncorrelated fading:** The spatial correlation matrices are generated as $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{I}_{M_j}$ for $l, j = 1, \dots, L$ and $i = 1, \dots, K_l$.

The former model provides strong spatial channel correlation, while the latter provides no spatial channel correlation.

Each coherence block consists of $\tau_c = 200$ samples. This dimensionality supports high mobility and large channel dispersion at 2 GHz carriers, as exemplified in Remark 2.1 on p. 221.

There are M antennas at each BS and, in most cases, an equal number K of UEs in each cell. The values of M and K will be changed and specified every time we consider the running example.

The τ_p pilot sequences can be distributed among the UEs and reused across cells in different ways, as described in detail in Section 7.2.1 on p. 468. Unless stated otherwise, we consider $\tau_p = fK$ pilots, with the integer f being called the *pilot reuse factor*. This means that there are f times more pilots than UEs per cell and the same subset of pilots is reused in a fraction $1/f$ of the cells. We consider $f \in \{1, 2, 4\}$ in the running example and the corresponding reuse patterns are illustrated in Figure 4.4b. The cells that use the same pilots are said to belong to the same *pilot group*. The pilots are randomly assigned to the UEs in every cell in the sense that the k th UE in two cells, that belong to the same pilot group, uses the same pilot. All parameter values are summarized in Table 4.2.

Remark 4.1 (LTE comparison). To make comparisons with contemporary cellular networks, we consider a typical LTE system where each cell is equipped with four antennas and has a coverage area of $\frac{3\sqrt{3}}{2}0.25^2$ km². We refer the interested reader to [110] for more details on the cell configuration in LTE. The total DL transmit power of the cell is 46 dBm and two single-antenna UEs are served by multiuser MIMO. For a TDD system [112], the UL and DL SEs are 2.8 and 3.2 bit/s/Hz/cell,

respectively. If there is 100% of UL or DL traffic over a 20 MHz bandwidth, this corresponds to a UL throughput of 56 Mbit/s/cell or to a DL throughput of 64 Mbit/s/cell, which result into UL and DL area throughputs of 344 Mbit/s/km² and 394 Mbit/s/km², respectively.

4.1.4 SE Comparison of Different Combining Schemes

We will now compare the different receive combining schemes using the setup defined in the running example above. The following Monte Carlo methodology is used to generate simulation results:

1. Macroscopic propagation effects
 - (a) Randomly drop UEs in each cell
 - (b) Compute distances d_{lk}^j and nominal angles φ_{lk}^j
 - (c) Generate random shadow fading coefficients F_{lk}^j
 - (d) Compute average channel gains β_{lk}^j , spatial correlation matrices \mathbf{R}_{lk}^j , and estimation error correlation matrices \mathbf{C}_{lk}^j
2. Microscopic propagation effects
 - (a) Generate random estimated channel vectors $\hat{\mathbf{h}}_{lk}^j$
3. SE computation
 - (a) Compute receive combining vectors \mathbf{v}_{jk} and resulting $\text{SINR}_{jk}^{\text{UL}}$
 - (b) Compute “instantaneous” SE:

$$\text{SE}_{jk}^{\text{UL,inst.}} = \frac{\tau_u}{\tau_c} \log_2 \left(1 + \text{SINR}_{jk}^{\text{UL}} \right)$$
 - (c) Average $\text{SE}_{jk}^{\text{UL,inst.}}$ over estimated channels to obtain $\text{SE}_{jk}^{\text{UL}}$
 - (d) Obtain simulation results by considering the SEs of all UEs for different shadow fading realizations and UE locations

In this simulation, we consider $K = 10$ UEs per cell and a varying number of BS antennas. There are fK pilots in each coherence block and the remaining $\tau_c - fK$ samples are used for UL data transmission. We use Gaussian local scattering with ASD $\sigma_\varphi = 10^\circ$ as channel model.

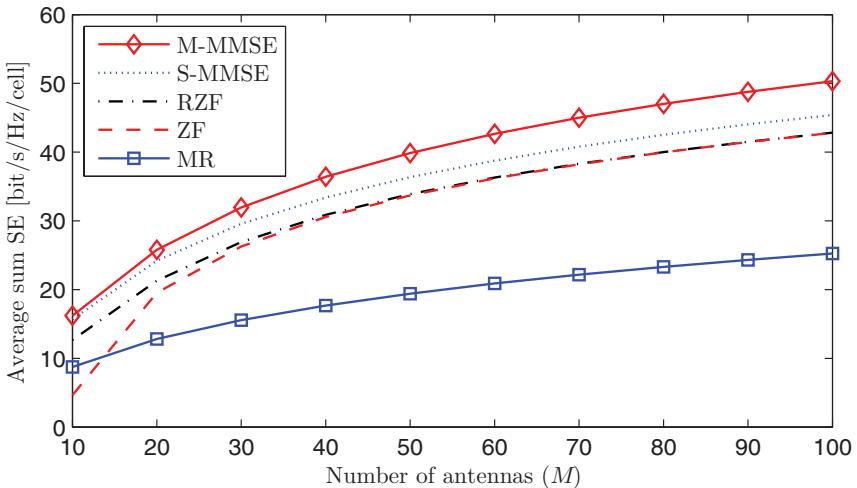


Figure 4.5: Average UL sum SE as a function of the number of BS antennas for different combining schemes. There are $K = 10$ UEs per cell and the same K pilots are reused in every cell.

Figure 4.5 shows the average UL sum SE as a function of the number of BS antennas for universal pilot reuse with $f = 1$. M-MMSE gives the largest SE in Figure 4.5. The SE reduces a little with every approximation that is made to obtain a scheme with lower complexity than M-MMSE. The S-MMSE scheme provides lower SE than M-MMSE, but 5%–10% higher SE than RZF and ZF. Note that RZF and ZF give essentially the same SE in the range $M \geq 20$ that is of main interest in Massive MIMO, but the SE with ZF deteriorates quickly for $M < 20$ since the BS does not have enough degrees of freedom to cancel the interference without also canceling a large part of the desired signal. Hence, ZF should be avoided to achieve a robust implementation. Interestingly, MR provides only half the SE of the other schemes, but looking at Figure 4.3 we know that it also reduces complexity by 10% as compared to RZF and requires no matrix inversions.

Figure 4.6 shows the average sum SE with a non-universal pilot reuse f . In particular, we consider cases where each pilot is reused in every second or fourth cells, according to the pattern shown in Figure 4.4b. This is referred to as having a pilot reuse factor of $f = 2$ and $f = 4$, respectively. The increased number of pilots reduces the

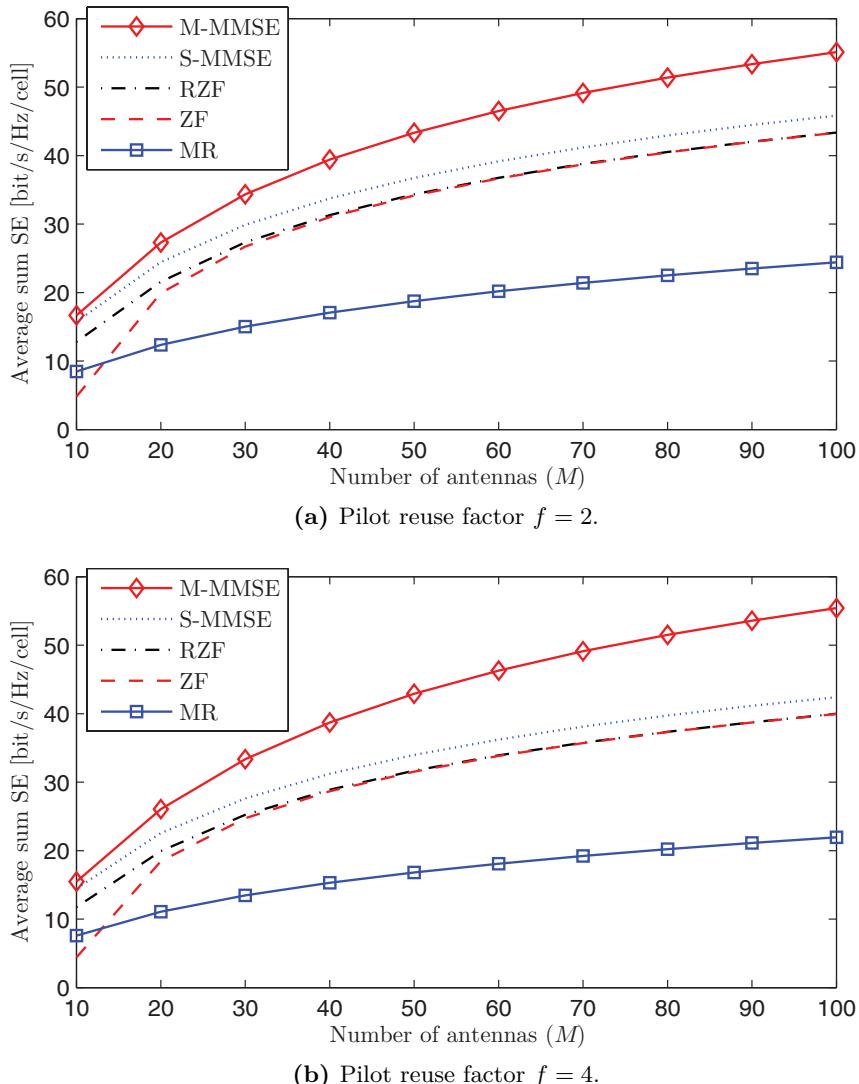


Figure 4.6: Average UL sum SE as a function of the number of BS antennas for different combining schemes. There are $K = 10$ UEs per cell and either $2K$ or $4K$ pilots that are reused across cells according to the pattern in Figure 4.4b.

Scheme	$f = 1$	$f = 2$	$f = 4$
M-MMSE	50.32	55.10	55.41
S-MMSE	45.39	45.83	42.41
RZF	42.83	43.37	39.99
ZF	42.80	43.34	39.97
MR	25.25	24.41	21.95

Table 4.3: Average UL sum SE [bit/s/Hz/cell] for $M = 100$ and $K = 10$ for different pilot reuse factors f . The largest value for each scheme is in bold face. The results are summarized from Figures 4.5 and 4.6.

pre-log factor in (4.2) since $\tau_u = \tau_c - fK$, but it also increases the instantaneous SINR in (4.3) since better channel estimates with less pilot contamination are obtained. M-MMSE benefits particularly much from having $f > 1$, because it can better suppress the interference from UEs in the surrounding cells when these UEs use other pilots. A reuse factor of 4 gives the highest SE with M-MMSE. S-MMSE, RZF, and ZF give comparable SE to each other for all f , and achieve the highest SE with $f = 2$. The SE of MR reduces when f is increased since the improved estimation quality does not outweigh the reduced pre-log factor when the estimate is only used to coherently combine the desired signal and not to cancel interference. These properties are quantified in Table 4.3, which summarizes the sum SE of all schemes with $M = 100$ and different f . The numbers can be compared with the SE 2.8 bit/s/Hz/cell achieved by a contemporary LTE system (see Remark 4.1). With all pilot reuse factors, M-MMSE and RZF provide more than an order-of-magnitude higher SE per cell. With MR, the gain is a factor 7–9.

In summary, there are basically three combining schemes to choose from, if the running example is implemented in practice. M-MMSE provides the highest SE using the highest complexity, and should be implemented using non-universal pilot reuse. MR has the lowest complexity, but also delivers the lowest SE. Finally, RZF strikes a good balance between SE and complexity; it can double the SE as compared to MR while the computational complexity is only some tens of percentages higher. In practice, RZF is always a better choice than ZF,

since it achieves similar or better SE and lacks ZF's robustness issues when $M \approx K$. However, ZF is a fairly common scheme in the literature since it allows to compute closed-form SE expressions in the special case of spatially uncorrelated channels [244, 357, 210]. Approximate closed-form expressions can be computed with M-MMSE, S-MMSE, and RZF [148, 193]. The closed-form expressions predict the SE that is practically achievable with different schemes and are particularly useful for resource allocation and optimization.

Remark 4.2 (Polynomial expansion). The ultimate receive combining scheme would provide the SE of M-MMSE or RZF, but would have a computational complexity similar to that of MR. Since the matrix inversions in M-MMSE and RZF are particularly computationally heavy, one way to reduce the complexity is to approximate the inversion by a matrix polynomial [229]. Note that for a real scalar a we can make the Taylor series expansion $(1 + a)^{-1} = \sum_{\ell=0}^{\infty} (-a)^\ell$ if $|a| < 1$. Similarly, we have $(\mathbf{I}_N + \mathbf{A})^{-1} = \sum_{\ell=0}^{\infty} (-\mathbf{A})^\ell$ if \mathbf{A} is an $N \times N$ Hermitian matrix with eigenvalues $\lambda_1, \dots, \lambda_N$ that all satisfy $|\lambda_n| < 1$. The intuition is that $(\mathbf{I}_N + \mathbf{A})^{-1}$ keeps the eigenvectors of $\mathbf{I}_N + \mathbf{A}$ (which coincide with those of \mathbf{A}) but inverts all the eigenvalues as $(1 + \lambda_n)^{-1}$, thus we can apply the Taylor expansion separately to the inversion of each eigenvalue. By truncating the polynomial expansion to only the first L_p terms, which have the dominant impact in a Taylor series, we can obtain an efficient approximation that does not involve any matrix inversion. This technique has been considered for various multiuser detection scenarios in the last decades [229, 190, 145, 233, 292, 149]. Weighted truncations of the form $\sum_{\ell=0}^{L_p} v_\ell \mathbf{A}^\ell$, for scalar weights v_ℓ , are often considered to fine-tune the approximation. The weights can be computed using scaling properties [190, 292, 170] or asymptotic random matrix analysis [233, 149, 302]. A key benefit of the polynomial expansion technique is that it allows for efficient multistage/pipelined hardware implementation [229]. The computational complexity is proportional to $L_p N^2$, where $N = M_j$ with M-MMSE and $N = K_j$ with RZF. Note that L_p does not need to scale with N since each of the N eigenvalues is approximated separately. Instead, L_p can be selected to balance between computational complexity and communication performance. Polynomial expansion in

UL Massive MIMO was studied in [149], where $L_p = 1$ coincides with MR and every additional term gives an improvement towards the SE with RZF. There is also a related concept of Neumann series expansions that can be used to approximate matrix inverses [352].

4.1.5 Impact of Spatial Channel Correlation

We know that spatial channel correlation has a major impact on channel hardening, favorable propagation, and channel estimation quality (see Figures 2.7, 2.8, and 3.3). On the positive side, we have observed that the estimation quality improves under spatial correlation and that UEs with different spatial characteristics exhibit more favorable propagation. On the negative side, we have observed a slower convergence to asymptotic channel hardening under spatial correlation and also less favorable propagation for UEs with similar spatial characteristics. We will now quantify the impact of spatial channel correlation on the SE by continuing the running example that was defined in Section 4.1.3. We use the Gaussian local scattering channel model with varying ASD σ_φ and will compare the results with uncorrelated fading. Based on the conclusion from the SE-complexity tradeoff analysis above, we only consider M-MMSE, RZF, and MR combining, which represent three distinctively different tradeoffs. We consider $M = 100$ and $K = 10$, and for each scheme and σ_φ , we use the pilot reuse factor that maximizes the SE. Apart from the pilots, the remaining $\tau_c - fK = 200 - 10f$ samples per coherence block are used for UL data transmission. The SEs are computed using Theorem 4.1.

Figure 4.7 shows the average sum SE as a function of the ASD. As expected, M-MMSE provides the highest SE, followed by RZF, and then MR. We notice that the SE is a decreasing function of the ASD, for all three combining schemes. This indicates that the dominant effect of having high spatial channel correlation (i.e., small ASD) is the reduced interference caused between UEs that have sufficiently different spatial correlation matrices. For very small ASDs, the channel resembles a LoS scenario and we then know from Section 1.3.3 on p. 193 that the interference is low, except when two UEs have very similar angles to a BS.

The figure also shows dotted lines that represent the SE achieved

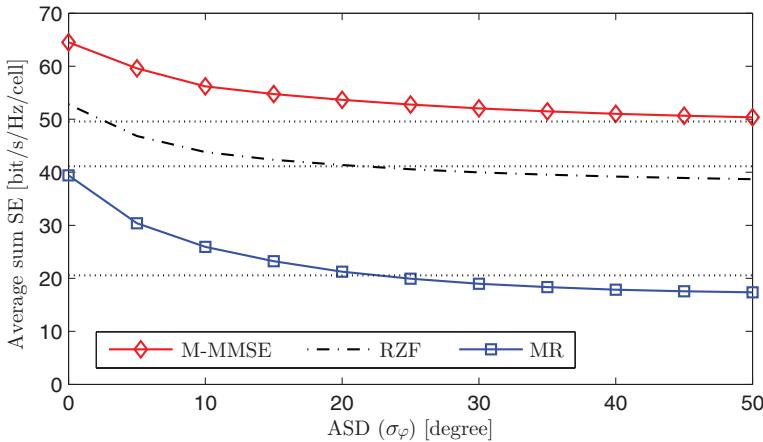


Figure 4.7: Average UL sum SE for the running example using the Gaussian local scattering channel model with varying ASD. We consider $M = 100$ and $K = 10$. The dotted lines indicate the SE achieved with uncorrelated Rayleigh fading channels. The largest SE is achieved by M-MMSE, followed by RZF, and then MR.

with uncorrelated Rayleigh fading channels, as defined in the running example. The combining schemes provide SEs in the same performance order as with spatial channel correlation. Spatially correlated channels provide higher average SEs for most ASDs; for example, M-MMSE benefits from spatial correlation if the ASD is below 50° , while MR and RZF perform better for ASDs smaller than 20° . However, for large ASDs, the SE is slightly lower than with uncorrelated fading. This is due to the geometry of a ULA that makes it better at resolving UE angles near to the boresight of the array than UE angles that are nearly parallel to the array.⁵

While the curves in Figure 4.7 represent the average SEs, Figure 4.8 shows CDF curves of the variations in SE per UE, for an arbitrary UE in the network. The randomness is due to random UE locations and shadow fading realizations. Simulation results are given for uncorrelated Rayleigh fading channels and for the Gaussian local scattering channel

⁵If the random angle $\bar{\varphi}$ of the scatterers is distributed such that $\sin(\bar{\varphi})$ is uniformly distributed between -1 and $+1$, then the ULA will behave as in uncorrelated Rayleigh fading. With the local scattering model, with essentially any angular distribution, it is instead $\bar{\varphi}$ that is uniformly distributed between $-\pi$ and $+\pi$ when σ_φ is large.

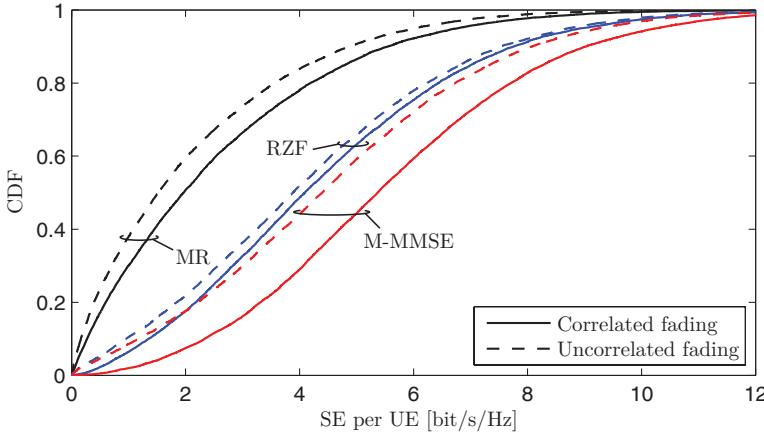


Figure 4.8: CDF of the UL SE per UE in the running example with $M = 100$, $K = 10$, and $f = 2$. Uncorrelated Rayleigh fading channels are compared with the Gaussian local scattering channel model with ASD $\sigma_\varphi = 10^\circ$.

model with ASD $\sigma_\varphi = 10^\circ$, which represents strong spatial channel correlation. The main observation is that in situations where the spatial correlation improves the sum SE, all UEs will statistically benefit from a higher SE since the CDF curves with spatial correlation are to the right of the corresponding curves with uncorrelated fading. This does not mean that spatial correlation is always beneficial. A UE at a given location might achieve higher SE with uncorrelated fading than with spatial correlation, but this cannot be seen from CDF curves. By investigating the simulation results further, we notice that this happens with 17%–35% probability. However, as a UE moves around in the network the probability of achieving a particular SE is consistently higher under spatial correlation.

4.1.6 Channel Hardening under Spatial Channel Correlation

The effective channel after receive combining is $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j$. Similar to Definition 2.4 on p. 231, we can say that the effective channel hardens if $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j / \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} \approx 1$ for any channel realization. To quantify how close to asymptotic channel hardening we are with a certain channel model, receive combining scheme, and a finite number of antennas, we

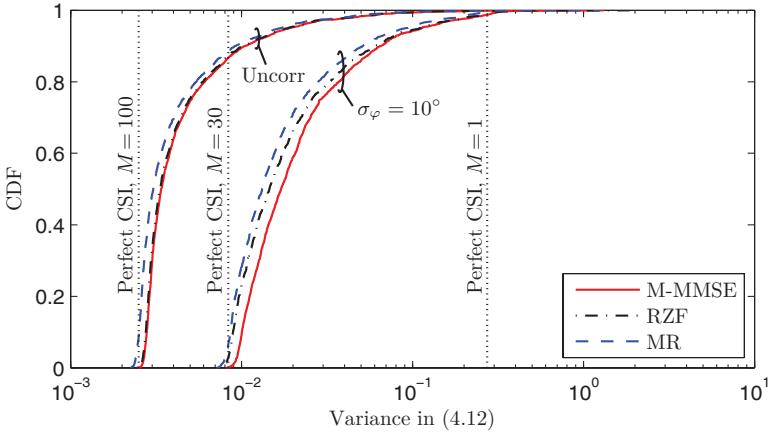


Figure 4.9: CDF curves of the channel variations after receive combining, $\mathbb{V}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}/(\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\})^2$, with different receive combining schemes and channel models. We consider $M = 100$, $f = 2$, and $K = 10$. The vertical reference curves that show the channel gain variations with $M \in \{1, 30, 100\}$ using MR with uncorrelated fading and perfect CSI. Note that the horizontal axis has a logarithmic scale.

can measure the variance of $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j / \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}$:

$$\frac{\mathbb{V}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}}{(\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\})^2}. \quad (4.12)$$

This approach is similar to the analysis in Section 2.5.1 on p. 231 and we note that the variance should ideally be almost zero.

Figure 4.9 shows CDFs of (4.12) in the running example, where the randomness is induced by random UE locations. At each location, the variance is computed numerically from many channel realizations. We consider $M = 100$, $K = 10$, and $f = 2$. Results are shown for uncorrelated fading and the Gaussian local scattering model with 10° ASD. There are also three vertical reference curves in Figure 4.9 that show the variance of the channel hardening with $M \in \{1, 30, 100\}$ using MR, uncorrelated fading, and perfect CSI.

First, we notice that the choice of receive combining scheme (MR, RZF, or M-MMSE) has little impact on the results, although the interference rejection in RZF and M-MMSE leads to a small reduction in hardening. Second, the imperfect CSI greatly impacts the hardening.

With uncorrelated Rayleigh fading, some UEs achieve the same amount of channel hardening as the reference case with perfect CSI and the same number of antennas. This happens for cell-center UEs that have high SNR and estimation quality, while the cell-edge UEs will observe a substantial loss in hardening due to the channel estimation errors. The local scattering model gives a similar trend, but all UEs observe substantially less hardening; as shown in the figure, it is basically equivalent to $M = 30$ with uncorrelated fading. This is not a coincidence, but follows from the fact that only around 40% of the eigenvalues of the spatial correlation matrix are non-negligible (see Figure 2.6 and Section 2.5.1 on p. 231). Nevertheless, there is much more hardening under strong spatial correlation than in a single-antenna system.

In summary, in the considered scenario, spatial channel correlation (and also estimation errors) leads to a substantial loss in channel hardening, due to the larger variations in the effective channel after receive combining. In contrast, the choice of receive combining scheme has an almost negligible impact on the hardening.

4.2 Alternative UL SE Expressions and Key Properties

The SE expression in Theorem 4.1 can be computed by Monte Carlo simulations for any combining scheme, by generating many realizations of the instantaneous SINR in (4.3). There is an alternative SE expression that may lead to closed-form expressions. The key idea behind this approach is to utilize the channel estimates only for computing the receive combining vectors, while this side-information is not exploited in the signal detection. This simplification makes sense when there is substantial channel hardening, such that $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j / M_j \approx \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} / M_j$.

More precisely, the receive combined signal in (4.1) is rewritten as

$$\begin{aligned} \mathbf{v}_{jk}^H \mathbf{y}_j = & \underbrace{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} s_{jk}}_{\text{Desired signal over average channel}} + \underbrace{(\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}) s_{jk}}_{\text{Desired signal over "unknown" channel}} \\ & + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} \mathbf{v}_{jk}^H \mathbf{h}_{ji}^j s_{ji}}_{\text{Intra-cell interference}} + \underbrace{\sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j s_{li}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{v}_{jk}^H \mathbf{n}_j}_{\text{Noise}} \end{aligned} \quad (4.13)$$

by adding and subtracting $\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} s_{jk}$. Only the part of the desired signal received over the average precoded channel $\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}$ is treated as the true desired signal. The part of s_{jk} received over the deviation from the mean value, $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}$, has zero mean and can thus be treated as an uncorrelated noise signal in the detection. The following theorem provides an alternative capacity bound, which is referred to as the use-and-then-forget (UatF) bound since the channel estimates are used for combining and then effectively “forgotten” before the signal detection [210].

Theorem 4.4. The UL ergodic channel capacity of UE k in cell j is lower bounded by $\underline{\text{SE}}_{jk}^{\text{UL}} = \frac{\tau_u}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{UL}})$ [bit/s/Hz] with

$$\underline{\text{SINR}}_{jk}^{\text{UL}} = \frac{p_{jk} |\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} - p_{jk} |\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2 + \sigma_{\text{UL}}^2 \mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} \quad (4.14)$$

where the expectations are with respect to the channel realizations.

Proof. The proof is available in Appendix C.3.4 on p. 596. \square

The lower bound on the capacity provided by Theorem 4.4 is intuitively less tight than the bound provided in Theorem 4.1, since the channel estimates are not utilized in the signal detection. However, it does not require the use of MMSE channel estimation, but can be applied along with any channel estimator and any combining scheme. In fact, it can be applied with any channel distributions or even measured channels. Since the SE takes the form $\frac{\tau_u}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{UL}})$, it

is convenient to refer to $\underline{\text{SINR}}_{jk}^{\text{UL}}$ as the effective SINR of the fading channel from UE k in cell j . Note that $\underline{\text{SINR}}_{jk}^{\text{UL}}$ is deterministic and the expression contains several expectations over the random channel realizations. Each of the expectations in (4.14) can be computed separately by means of Monte Carlo simulation. For MR combining, they can be obtained in closed form.

Corollary 4.5. If MR combining with $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j$ is used, based on the MMSE estimator, then

$$\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} = p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j) \quad (4.15)$$

$$\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\} = p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j) \quad (4.16)$$

$$\begin{aligned} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} &= p_{jk} \tau_p \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j) \\ &+ \begin{cases} p_{li} p_{jk} (\tau_p)^2 |\text{tr}(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)|^2 & (l, i) \in \mathcal{P}_{jk} \\ 0 & (l, i) \notin \mathcal{P}_{jk} \end{cases} \end{aligned} \quad (4.17)$$

where Ψ_{jk}^j was defined in (3.10). The SE expression in Theorem 4.4 becomes $\underline{\text{SE}}_{jk}^{\text{UL}} = \frac{\tau_u}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{UL}})$ with

$$\begin{aligned} \underline{\text{SINR}}_{jk}^{\text{UL}} &= \frac{p_{jk}^2 \tau_p \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\underbrace{\sum_{l=1}^L \sum_{i=1}^{K_l} \frac{p_{li} \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}}_{\text{Non-coherent interference}} + \underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{p_{li}^2 \tau_p |\text{tr}(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)|^2}{\text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}}_{\text{Coherent interference}} + \sigma_{\text{UL}}^2}. \end{aligned} \quad (4.18)$$

In the special case of spatially uncorrelated fading (i.e., $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{I}_{M_j}$ for $l = 1, \dots, L$ and $i = 1, \dots, K_l$), (4.18) simplifies to

$$\begin{aligned} \underline{\text{SINR}}_{jk}^{\text{UL}} &= \frac{(p_{jk} \beta_{jk}^j)^2 \tau_p \psi_{jk} M_j}{\underbrace{\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \beta_{li}^j}_{\text{Non-coherent interference}} + \underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} (p_{li} \beta_{li}^j)^2 \tau_p \psi_{jk} M_j}_{\text{Coherent interference}} + \sigma_{\text{UL}}^2} \end{aligned} \quad (4.19)$$

where

$$\psi_{jk} = \left(\sum_{(l',i') \in \mathcal{P}_{jk}} p_{l'i'} \tau_p \beta_{l'i'}^j + \sigma_{\text{UL}}^2 \right)^{-1}. \quad (4.20)$$

Proof. The proof is available in Appendix C.3.5 on p. 597. \square

The closed-form SE expression in Corollary 4.5 provides important insights into the basic behaviors of Massive MIMO. The signal term in the numerator of (4.18) is

$$p_{jk}^2 \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j) = p_{jk} \text{tr}(\mathbf{R}_{jk}^j - \mathbf{C}_{jk}^j) \quad (4.21)$$

where the equality follows from (3.11). This is the transmit power multiplied with the trace of the correlation matrix of the channel estimate (see Corollary 3.2 on p. 250). Hence, the estimation quality determines the signal strength and it is reduced by pilot contamination. Since the trace is the sum of the M_j diagonal elements, the signal term increases linearly with M_j , which proves that the signal is coherently combined over the M_j antennas. This array gain is explicit in the special case of uncorrelated fading:

$$p_{jk}^2 \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j) = (p_{jk} \beta_{jk}^j)^2 \tau_p \psi_{jk} M_j. \quad (4.22)$$

The denominator of (4.18) contains three terms. The first one is a summation over all UEs in all cells, where UE i in cell l contributes with the interference $p_{li} \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j) / \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j)$. This term is referred to as *non-coherent interference*, because it does not increase linearly with M_j ; this is easily seen in the special case of uncorrelated fading when the interference term becomes $p_{li} \beta_{li}^j$ and thus is the product between the transmit power and the average channel gain. In general, the relation between the spatial correlation matrices \mathbf{R}_{li}^j and \mathbf{R}_{jk}^j determine how large the interference terms are; the strength of the interference is basically determined by whether $\text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j) / \text{tr}(\mathbf{R}_{jk}^j)$ is large or small. It is small when the interfering UE is far from the receiving BS and/or the spatial channel correlation properties are very different. The latter can be illustrated similarly to Figure 3.4, where UE pairs with similar angles cause more interference to each other and UEs with very different angles cause less interference. In the extreme case of $\mathbf{R}_{li}^j \mathbf{R}_{jk}^j = \mathbf{0}_{M_j \times M_j}$,

there is no interference between the two UEs, since their channels “live” in separate eigenspaces.

The second term in the denominator of (4.18) only involves the UEs in $\mathcal{P}_{jk} \setminus (j, k)$, which are those using the same pilots as the desired UE. This interference term contains the square of a trace term divided by a single trace term. As explained above, each trace term increases linearly with M_j and thus the entire interference term scales as M_j . This is seen explicitly in the special case of uncorrelated fading when the term becomes $(p_{li}\beta_{li}^j)^2\tau_p\psi_{jk}M_j$ and thus resembles the signal term, but involves the power and average channel gain of the interfering UE. This term is referred to as *coherent interference* and is a consequence of the pilot contamination. Recall that the channel estimates are statistically correlated for UEs that use the same pilot, as proved in Corollary 3.3 on p. 251. When the BS uses such an estimate to coherently combine the signal from its own UEs, it will also partially coherently combine the interfering signals. The strength of the coherent interference depends on the spatial correlation matrices of the desired and interfering UEs, similar to the case of non-coherent interference.

Figure 4.10 illustrates the coherent interference power divided by the desired signal power in the numerator of (4.18) for the same scenario as in Figure 3.4; that is, there is one desired UE at a fixed angle of 30° (as seen from the receiving BS) and an interfering UE at a varying angle between -180° and 180° . The local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$ is used, while the effective SNR from the desired UE is 10 dB and the interfering signal is 10 dB weaker than that. Figure 4.10 shows that when the desired and interfering UEs have similar angles, the relative interference power is independent of M , since both the desired signal power and the interference power grow proportionally to the number of antennas. However, when the BS can separate the UEs spatially, the coherent interference is reduced and the interference situation is vastly better than in a single-antenna scenario. Hence, even with simple MR combining, the coherent interference is not worse than in a single-antenna system and it can potentially be much lower. This explains the simulation results in Section 4.1.5, which indicated that spatial correlation improves the SE.

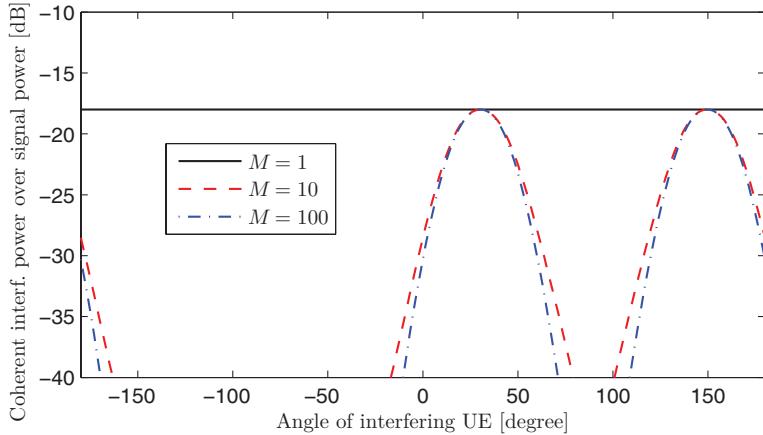


Figure 4.10: Coherent interference power (normalized by signal power) caused to a desired UE from an interfering UE that uses the same pilot, when MR combining is used. The local scattering model with Gaussian angular distribution is used and the desired UE has a nominal angle of 30° , while the angle of the interfering UE is varied between -180° and 180° . The effective SNR from the desired UE is 10 dB and the interfering signal is 10 dB weaker than that.

The third term in the denominator of (4.18) is the noise variance.

4.2.1 Tightness of the UatF Bound

We will compare the UatF bound in Theorem 4.4 with the original UL capacity bound in Theorem 4.1 by continuing the running example. We assume $M = 100$, $K = 10$, and $f = 2$. This simulation only considers MR combining, since one of the primary reasons to use the UatF bound is that one can obtain the closed-form SE expression for MR in Corollary 4.5. That corollary considers MR combining as it was defined in (4.11): $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|$. We will also evaluate MR with the following alternative vector normalizations:

$$\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\| \quad (4.23)$$

and

$$\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|^2. \quad (4.24)$$

Figure 4.11 shows the average sum SE as a function of the ASD. The top curve is obtained from Theorem 4.1 and the bottom curve is

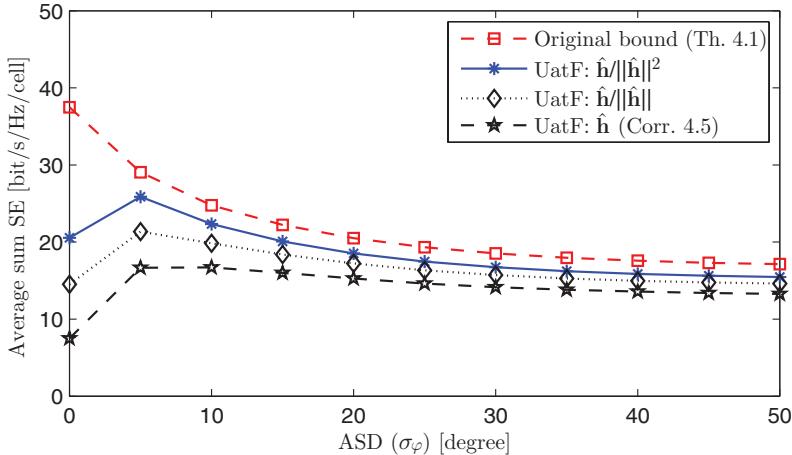


Figure 4.11: Average UL sum SE as a function of the ASD, using the local scattering model with Gaussian angular distribution. We consider $M = 100$, $K = 10$, and $f = 2$. The original SE expression for MR combining is compared with the UatF expression when the MR combining vector is normalized in three different ways.

obtained from Corollary 4.5. There is a substantial gap between these curves, particularly for small ASDs. The reason is that the UatF bound relies on channel hardening and less hardening occurs when the spatial channel correlation is strong (see Section 4.1.6).⁶ For large ASDs, the gap between the top and bottom curve is around 30%.

The SE should not be affected by the normalization of the combining vector because all parts of the received signal are scaled by the same known variable. The SE expression in Theorem 4.1 satisfies this basic property, but the normalization can actually affect the tightness of the UatF bound. For example, this bound is only tight when the combined channel $\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j$ has nearly hardened and a random normalization factor can either improve or degrade the channel hardening effect. Figure 4.11 compares the SE from Theorem 4.1, using MR, and the UatF bound with MR normalized as $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j$ or as in (4.23) or (4.24). The former option was used to compute the closed-form expression in Corollary 4.5, but it is the normalization in (4.24) that gives the SE closest to the top

⁶In the extreme case when the spatial correlation matrix has rank one (e.g., $\sigma_\varphi = 0$), the squared channel norm $\|\mathbf{h}\|^2$ has an exponential distribution irrespective of the number of antennas. No channel hardening occurs in this special case.

curve (representing Theorem 4.1). The difference between the normalizations is particularly large for small ASDs, since there is less channel hardening. The normalization-dependence is an artifact of the UatF bounding technique, which “forgets” the combining vector before the signal detection. It is important to keep this artifact in mind when using UatF-type of bounds (as often done in the Massive MIMO literature). The reason that the normalization $\hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|^2$ gives the highest SE is quite intuitive: the gain of the estimated channel is equalized as $\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j = (\hat{\mathbf{h}}_{jk}^j)^H \hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|^2 = 1$, which ideally leads to a deterministic channel. Note that the non-MR combining schemes considered in this section are created by taking the channel estimate $\hat{\mathbf{h}}_{jk}^j$ and multiplying it with the inverse of a matrix that contains the outer product $\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{jk}^j)^H$. This leads to a normalization of the combining vector that resembles $\hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|^2$ and thus we can use these combining schemes in the UatF bound without having to change the normalization. In summary, UatF bounds on the capacity can be convenient to obtain analytical insights (particularly when using MR), but they may have some unexpected behaviors, such as systematically underestimating the achievable SE.

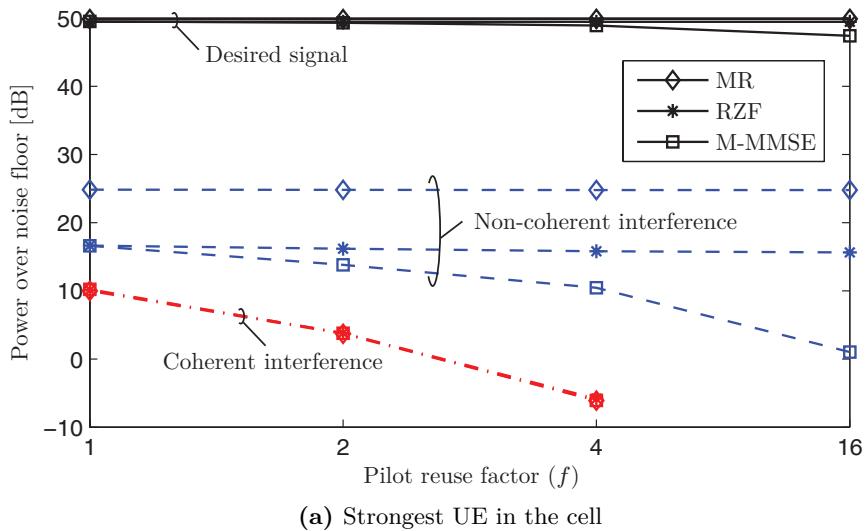
4.2.2 Pilot Contamination and Coherent Interference

Pilot contamination has two effects on the UL. First, it increases the MSE of the channel estimation (see Section 3.3.2 on p. 256), which impairs the ability to select combining vectors that provide strong array gains and that reject non-coherent interference. Second, it gives rise to coherent interference that is amplified by the array gain, similar to the desired signal. We will now investigate the impact and relative importance of these effects by continuing the running example that was defined in Section 4.1.3. We consider $M = 100$, $K = 10$, and will compare uncorrelated fading and the Gaussian local scattering channel model with $\sigma_\varphi = 10^\circ$. The average power of the desired signal, the non-coherent interference, and the coherent interference are estimated by Monte Carlo simulations, by averaging over fading realizations. Since UEs at different locations exhibit very different power levels, we focus on the average power of the strongest and the weakest UEs in an arbitrary cell, defined as the ones that achieve the largest and smallest

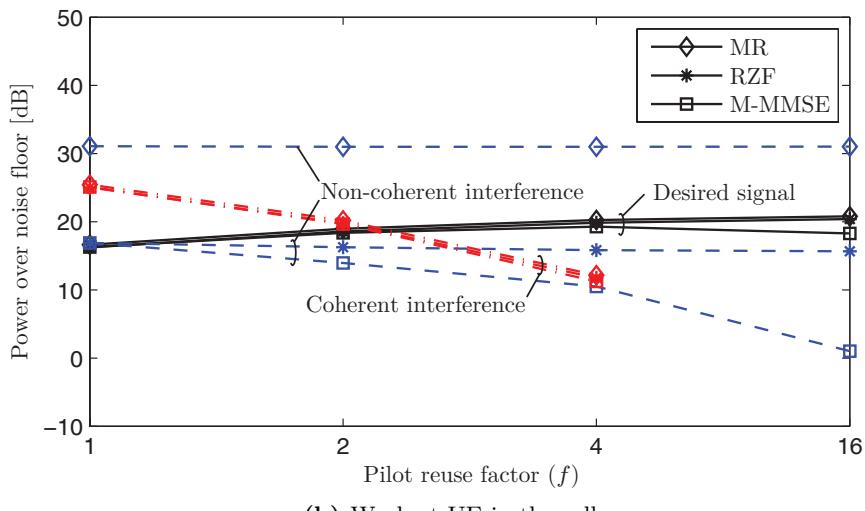
desired signal power in a given UE drop, respectively. The non-coherent interference comes from all UEs, while the coherent interference is estimated as the additional interference caused by pilot-contaminating UEs. The powers are normalized with respect to the receiver noise power, meaning that 0 dB represents a signal that is equally strong as the noise.

Figure 4.12 shows the signal power and interference power with uncorrelated fading. The horizontal axes show different pilot reuse factors $f \in \{1, 2, 4, 16\}$, where 16 represents the extreme case of having different orthogonal pilots in every cell (and thus no pilot contamination). We compare M-MMSE, RZF, and MR combining. Figure 4.12a considers the strongest UE and Figure 4.12b considers the weakest UE in an arbitrary cell. For the strongest UE, the desired signal power is almost the same for every f , which shows that pilot contamination has a minor impact on the MSE of channel estimates. The desired signal power is 20–30 dB higher than the non-coherent interference. MR gives the highest signal power—this is the main purpose of MR—while RZF and M-MMSE sacrifice a few dB of signal power to find combining vectors that suppress the interference by 10 dB or more. This explains why RZF and M-MMSE delivered substantially higher SE than MR in the previous simulations. M-MMSE benefits the most from increasing the pilot reuse factor because it can then obtain useful estimates of the channels to UEs in other cells and use them to suppress the corresponding non-coherent interference. In all the studied cases, the coherent interference that affects the strongest UE is negligible, as compared to the non-coherent interference, since the interfering UEs are much further away from the receiving BS than the desired UE.

The situation is very different for the weakest UE in the cell, which is typically at the cell edge. The additional pathloss makes the desired signal power many tens of dB lower than for the strongest UE. The channel estimation quality is also lower, thus the desired signal power after receive combining can be substantially increased by having a larger f . With MR, the non-coherent interference power is around 10 dB stronger than the desired signal power since the intra-cell interference is not suppressed. RZF and M-MMSE can still sacrifice a few dB of



(a) Strongest UE in the cell



(b) Weakest UE in the cell

Figure 4.12: Average UL power of the desired signal, non-coherent interference, and coherent interference. We consider $M = 100$, $K = 10$, and uncorrelated Rayleigh fading. Different combining schemes and pilot reuse factors are considered.

signal power to find combining vectors that suppress the non-coherent interference by 10 dB or more. It is then the coherent interference that is the dominant interference source when using these schemes. The coherent interference is basically the same for all schemes, when having uncorrelated fading, and it can be reduced by increasing f . As for the strongest UE, M-MMSE benefits the most from increasing f , since it becomes substantially better at suppressing inter-cell interference.

Figure 4.13 shows the power levels in the same setup as in Figure 4.12, but for the Gaussian local scattering model. Many of the observations made for uncorrelated fading are still applicable, but there are some important differences. First, the coherent interference is substantially lower since only pilot-sharing UEs with matching spatial correlation cause strong interference to each other. In fact, the coherent interference is now negligible, as compared to the non-coherent interference, even for the weakest UE in a cell. Interestingly, M-MMSE can reject coherent interference when there is spatial channel correlation, as seen from the substantially lower coherent interference as compared to MR and RZF. This property will play a key role in the asymptotic analysis in Section 4.4. Another consequence of the spatial correlation is that the desired signal power of the weakest UE increases rather slowly with f , which shows that it is the pathloss and not the pilot contamination that has the dominant impact on the channel estimation quality, even for cell-edge UEs.

In summary, in the running example, pilot contamination has little impact on the channel estimation quality, except for cell-edge UEs that exhibit uncorrelated fading. Pilot contamination gives rise to coherent interference that is stronger than conventional non-coherent interference in some cases (e.g., for cell-edge UEs with uncorrelated fading), but is often substantially lower. When coherent interference is an issue, it can be alleviated by increasing the pilot reuse factor. In these situations, the remaining impact of pilot contamination is on the pre-log factor of the SE expression, which decreases with the number of pilots. Recall from Section 4.1.4 that the highest SE is achieved when using M-MMSE combining and $f = 4$ as pilot reuse factor. This scheme uses $f = 4$ to obtain estimates of inter-cell channels that are then used to suppress

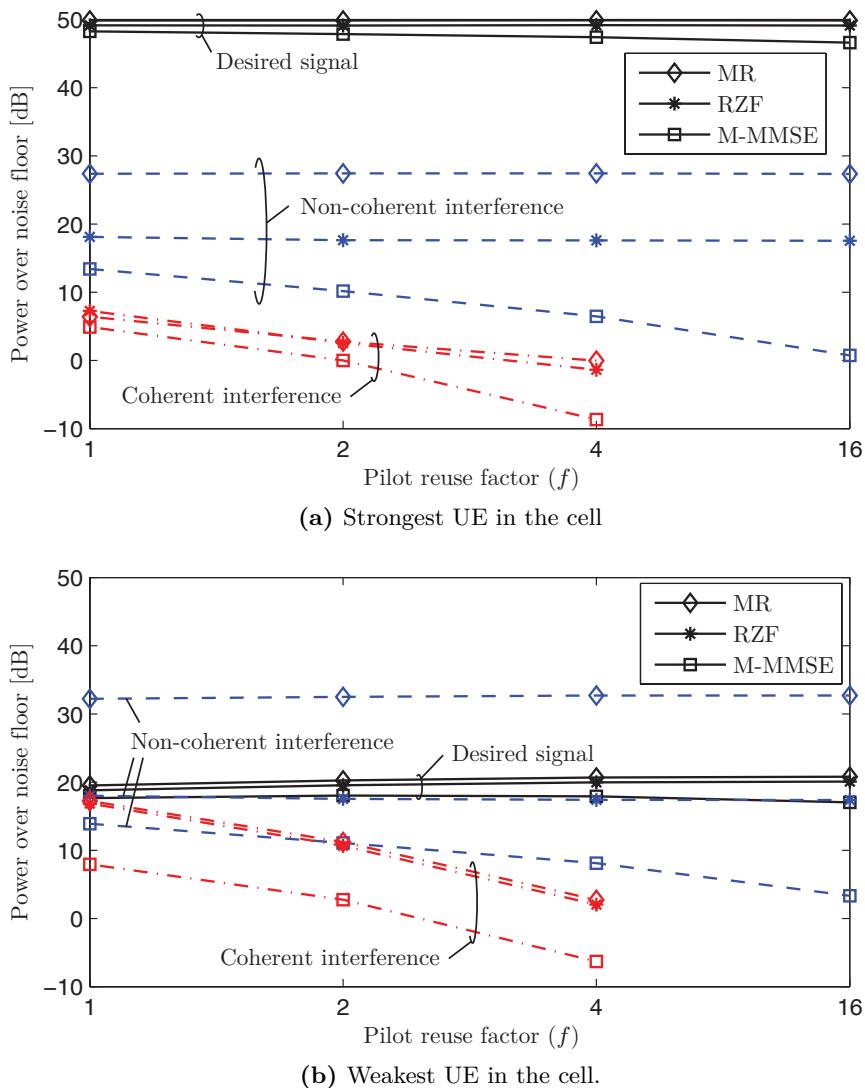


Figure 4.13: Average UL power of the desired signal, non-coherent interference, and coherent interference. We consider $M = 100$, $K = 10$, and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. Different combining schemes and pilot reuse factors are compared.

both coherent and non-coherent inter-cell interference. The simulation was performed with $M = 100$ and $K = 10$. Adding more UEs will increase the non-coherent interference level, while adding more antennas will increase the desired signal power and coherent interference, but also the ability to suppress the latter. Note that equal power allocation was used in this simulation, while the results can be different under power control. In particular, the intra-cell interference that affects cell-edge UEs can be reduced by lowering the transmit power of cell-center UEs.

4.2.3 SE with Other Channel Estimation Schemes than MMSE

The SE simulations in this section have so far been based on MMSE channel estimation. Recall that the alternative EW-MMSE and LS channel estimators were defined in Section 3.4.1 on p. 265 to reduce the computational complexity, at the cost of reduced estimation quality. We will now compare the SEs that are achieved when using these different channel estimators, to figure out by how much the loss in estimation quality translates into an SE loss. We continue the example from Figures 4.5 and 4.6, but focus on $K = 10$ UEs and $M = 100$ BS antennas. Each combining scheme uses the pilot reuse factor that maximizes the SE. We will utilize the UatF bound in Theorem 4.4, which can be applied along with any channel estimator.

Figure 4.14 shows a bar diagram of the average sum SE with M-MMSE, RZF, and MR combining. The highest SEs are obtained when using the MMSE estimator, as expected. If the EW-MMSE is used instead, then there is an 8%–12% loss in SE, depending on the combining scheme. The difference in SE between the EW-MMSE and LS estimators is very small when using RZF or MR combining, but M-MMSE combining performs poorly with LS. This is because the LS estimator does not give the right scaling of the channel estimates in the presence of pilot contamination, but rather acts as an estimator of the sum of the interfering channels. The consequence is that the norm of the channels from UEs in other cells are greatly overestimated and M-MMSE will therefore overemphasize the need for suppressing inter-cell interference.⁷

⁷By overemphasizing on interference suppression, M-MMSE behaves like a ZF-type of scheme that attempts to cancel all interference from all UEs in the network.

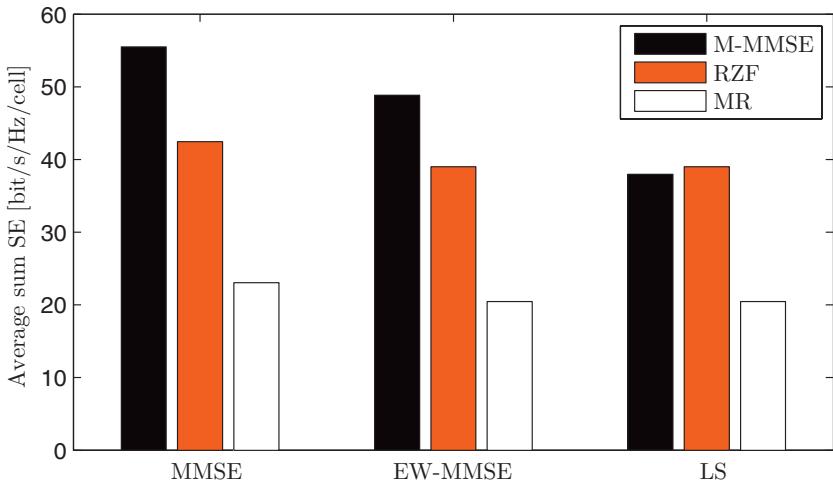


Figure 4.14: Average UL sum SE when using MMSE, EW-MMSE, or LS channel estimators, for a setup with $M = 100$ BS antennas and $K = 10$ UEs per cell. Three different combining schemes are considered.

The general observation is that the substantial SE gains of using RZF or M-MMSE combining compared with low-complexity MR combining remain, irrespective of which channel estimator is used. Note that the SEs with S-MMSE and ZF are not shown in the figure since they are very similar to that of RZF.

The example above considered a spatially correlated scenario with ASD $\sigma_\varphi = 10^\circ$. This is a case where the complexity of the MMSE estimator is relatively high as compared to EW-MMSE, while these estimators coincide in the special case of spatially uncorrelated fading. In summary, the loss in SE incurred by using a suboptimal channel estimator under strong spatial channel correlation is only 10%, for most combining schemes, which implies that high-complexity estimation schemes are not needed in Massive MIMO. The EW-MMSE channel estimator is suitable for all combining schemes, while the combination of LS estimation and M-MMSE combining is discouraged.

This type of scheme is called full-pilot ZF in [49] and multi-cell ZF in [193]. It has the benefit of only exploiting the direction of the channel estimates and therefore LS estimation can be used.

4.2.4 Synchronous versus Asynchronous Pilot Transmission

The SE analysis in this monograph is based on the assumption of synchronous pilot transmission, which means that all UEs in all cells simultaneously send pilot sequences from the same pilot book. This has been a common assumption in the Massive MIMO literature since the seminal paper [208], but there are good reasons to question this basic assumption. Even if all transmitters are time-synchronized, the signals will arrive asynchronously at every receiver due to the different propagation delays. The timing mismatches between neighboring cells are usually negligible; for example, in OFDM systems, the cyclic prefix can compensate for propagation path differences of several kilometers. The cyclic prefix compensates for path differences up to 74 km in digital audio broadcast (DAB) [143], while the normal and extended cyclic prefixes in LTE compensates for path differences up to 1.6 km and 5 km, respectively [165]. Since the strongest interference originates from the own and neighboring cells, having a more detailed model of the timing mismatches from distant cells will probably have little impact on the communication performance, but this remains to be proved rigorously. An accurate model needs to take inter-symbol interference into account [375, 263].

The main benefit of pilot synchronism is to control the inter-cell interference during pilot transmission, for example, by using a pilot reuse factor $f > 1$. However, if we do not exploit this possibility but set $f = 1$ and $\tau_p = K$, the interference will be roughly the same irrespective of whether the UEs in adjacent cells are sending pilots or data when the considered UEs are sending their pilots. This scenario was studied in [244, 48] for a case with K UEs per interfering cell. Looking at an arbitrary UE, in the case of synchronous pilot transmission, one UE per interfering cell reuses its pilot and causes coherent interference with a power proportional to M . In case of non-synchronous pilot signaling, the K UEs in an interfering cell send random data sequences. On average only a fraction $1/\tau_p = 1/K$ of their transmit power is sent in parallel with the pilot sequence of the considered UE, thus each interfering UE causes coherent interference with a power proportional to M/K . Since there are K interfering UEs per cell, the total coherent interference

is proportional to M . Hence, the total coherent interference power is the same with both synchronous and asynchronous pilot signaling, which means that the same SE is achieved in both cases. In summary, synchronous pilot transmission is helpful to mitigate interference, but nothing fundamentally different happens if we relax the synchronism and let other cells send random interfering signals.

4.3 Downlink Spectral Efficiency and Transmit Precoding

Each BS transmits payload data to its UEs in the DL, using linear precoding as defined in Section 2.3.2 on p. 227. Recall that $\varsigma_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \rho_{jk})$ denotes the random data signal intended for UE k in cell j , for $j = 1, \dots, L$ and $k = 1, \dots, K_j$. This UE is associated with the precoding vector $\mathbf{w}_{jk} \in \mathbb{C}^{M_j}$ that determines the spatial directivity of the transmission. The precoding vector satisfies $\mathbb{E}\{\|\mathbf{w}_{jk}\|^2\} = 1$, so that the signal power ρ_{jk} is also the transmit power allocated to this UE. One way to implement the precoding normalization is to make $\|\mathbf{w}_{jk}\|^2 = 1$ in every coherence block, but it is sometimes more analytically tractable to have an average normalization over many coherence blocks. The difference between these normalizations is small when there is substantial channel hardening.

The UL and DL channels are reciprocal within a coherence block, which enables the BS to use the UL channel estimates also for the computation/selection of precoding vectors. The desired signal to UE k in cell j propagates over the precoded channel $g_{jk} = (\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}$. The UE does not know g_{jk} a priori, but can either approximate it with the mean value $\mathbb{E}\{g_{jk}\} = \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}$ or estimate it from the received DL signals. We begin with the former case, which has been the dominating approach in the Massive MIMO literature since the early works [169, 148], while we consider the latter case in Section 4.3.3. The received DL

signal y_{jk} in (2.8) can then be expressed as

$$\begin{aligned}
 y_{jk} = & \underbrace{\mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \varsigma_{jk}}_{\text{Desired signal over average channel}} + \underbrace{\left((\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \right) \varsigma_{jk}}_{\text{Desired signal over unknown channel}} \\
 & + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} (\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji} \varsigma_{ji}}_{\text{Intra-cell interference}} + \underbrace{\sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li}}_{\text{Inter-cell interference}} + n_{jk} . \quad (4.25)
 \end{aligned}$$

Noise

The first term in (4.25) is the desired signal received over the deterministic average precoded channel $\mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}$, while the remaining terms are random variables with realizations that are unknown to the UE. An achievable SE can be computed by treating these terms as noise in the signal detection, by utilizing Corollary 1.3 on p. 171. We then obtain the following capacity bound, which we call the *hardening bound*. It holds for any choice of precoding vectors and channel estimation schemes.

Theorem 4.6. The DL ergodic channel capacity of UE k in cell j is lower bounded by $\underline{\text{SE}}_{jk}^{\text{DL}} = \frac{\tau_d}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{DL}})$ [bit/s/Hz] with

$$\underline{\text{SINR}}_{jk}^{\text{DL}} = \frac{\rho_{jk} |\mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \mathbb{E}\{|\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2\} - \rho_{jk} |\mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\}|^2 + \sigma_{\text{DL}}^2} \quad (4.26)$$

where the expectations are with respect to the channel realizations.

Proof. The proof is available in Appendix C.3.6 on p. 599. \square

The SE in Theorem 4.6 has the form $\frac{\tau_d}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{DL}})$ which makes it convenient to refer to $\underline{\text{SINR}}_{jk}^{\text{DL}}$ as the effective SINR of the fading DL channel to UE k in cell j . It is a deterministic scalar and the expression contains several expectations over the random channel realizations. The numerator of $\underline{\text{SINR}}_{jk}^{\text{DL}}$ contains the gain of the desired signal received over the average precoded channel. The first term in the denominator is the received power of all signals, while the second term subtracts the part of the desired signal that appeared in the numerator (i.e., the part useful for signal detection). The third term

is the noise variance. The SE expression can be computed numerically for any channel model and precoding scheme. The pre-log factor $\frac{\tau_d}{\tau_c}$ is the fraction of samples per coherence block that are used for DL data. Since $\tau_d = \tau_c - \tau_p - \tau_u$, the pre-log factor increases if we shorten the length τ_p of the pilot sequences (i.e., reduce the pilot overhead) and/or reduce the number of samples τ_d used for DL data.

The DL SE of UE k in cell j depends on the precoding vectors of all UEs in the entire network, in contrast to the UL SE in Theorem 4.1 that only depends on its own combining vector \mathbf{v}_{jk} . Hence, the precoding vectors should ideally be selected jointly across the cells, which makes precoding optimization difficult in practice [40]. A heuristic approach to precoding selection is later described in Section 4.3.2.

One simple and popular choice is MR precoding, which for UE k in cell j is based on the channel estimate $\hat{\mathbf{h}}_{jk}^j$. One variant of MR precoding is

$$\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}} \quad (4.27)$$

where the scaling is selected to satisfy the precoding normalization constraint $\mathbb{E}\{\|\mathbf{w}_{jk}\|^2\} = 1$. The SE expression in Theorem 4.6 can be computed in closed form for this average-normalized MR precoding.

Corollary 4.7. If average-normalized MR precoding is used, with $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}}$ based on the MMSE channel estimate, then the SE expression in Theorem 4.6 becomes $\underline{\text{SE}}_{jk}^{\text{DL}} = \frac{\tau_d}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{DL}})$ with

$$\begin{aligned} \underline{\text{SINR}}_{jk}^{\text{DL}} = & \frac{\rho_{jk} p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\underbrace{\sum_{l=1}^L \sum_{i=1}^{K_l} \frac{\rho_{li} \text{tr}(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)}{\text{tr}(\mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)}}_{\text{Non-coherent interference}} + \underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{\rho_{lip} p_{jk} \tau_p |\text{tr}(\mathbf{R}_{jk}^l \Psi_{li}^l \mathbf{R}_{li}^l)|^2}{\text{tr}(\mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)}}_{\text{Coherent interference}} + \sigma_{\text{DL}}^2} \end{aligned} \quad (4.28)$$

where Ψ_{jk}^j and Ψ_{li}^l were defined in (3.10).

In the special case of spatially uncorrelated fading (i.e., $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{I}_{M_j}$

for $l = 1, \dots, L$ and $i = 1, \dots, K_l$, (4.28) simplifies to

$$\text{SINR}_{jk}^{\text{DL}} = \frac{\rho_{jk} p_{jk} (\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j}{\underbrace{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \beta_{jk}^l}_{\text{Non-coherent interference}} + \underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \rho_{li} p_{jk} (\beta_{jk}^l)^2 \tau_p \psi_{li} M_l}_{\text{Coherent interference}} + \sigma_{\text{DL}}^2} \quad (4.29)$$

where ψ_{li} was defined in (4.20).

Proof. The proof is available in Appendix C.3.7 on p. 600. \square

4.3.1 Pilot Contamination in the Downlink

The closed-form DL SE expression in Corollary 4.7 has a structure that resembles that of the UL SE in Corollary 4.5. The signal term in the numerator is the same as in the UL, except that $p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j)$ is multiplied with the DL power ρ_{jk} instead of the UL power. This term increases linearly with M_j , since the trace adds up the M_j signal components coherently. This is the array gain from the precoding and the scaling with M_j is explicit in the special case of uncorrelated fading. Note that $p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j) = \text{tr}(\mathbf{R}_{jk}^j - \mathbf{C}_{jk}^j)$, thus the array gain is directly related to the estimation quality. The first term in the denominator is the non-coherent interference from all UEs. The strength of the interference is determined by how similar the spatial correlation matrices \mathbf{R}_{jk}^l and \mathbf{R}_{li}^l are, which can be measured in terms of how large $\text{tr}(\mathbf{R}_{li}^l \mathbf{R}_{jk}^l) / \text{tr}(\mathbf{R}_{jk}^l)$ is. The two correlation matrices describe the channels from the interfering BS l to the receiving UE and to one of the interfering UEs in cell l . The interference is small when the interfering BS is far away and/or the spatial channel correlation properties are very different for the two UEs. In the extreme case of $\mathbf{R}_{li}^l \mathbf{R}_{jk}^l = \mathbf{0}_{M_l \times M_l}$, there is no interference between the two UEs, since their channels “live” in separate eigenspaces. Note that the non-coherent interference term simplifies to $\rho_{li} \beta_{jk}^l$ in uncorrelated fading, which is independent of the location of the interfering UE (except if the power allocation is based on it).

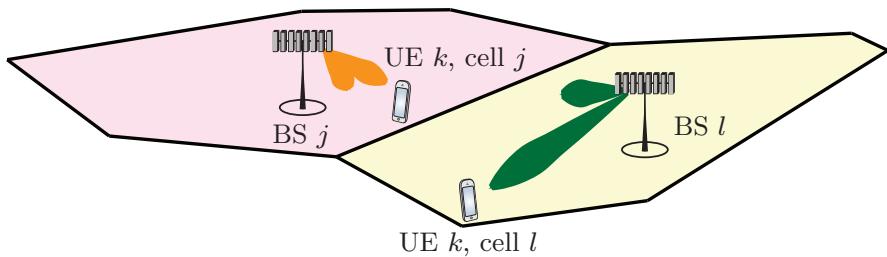


Figure 4.15: When two UEs transmit the same pilot sequence, as illustrated Figure 3.1, the pilot contamination affects the DL signals. When a BS attempts to direct a signal towards its own UE using MR precoding, it will also partially direct it towards the pilot-interfering UE in the other cell. Each color represents one precoded DL signal.

The second term in the denominator is the additional coherent interference that scales with M_l and originates from the signals to UEs that share the same pilot; that is, pilot contamination also affects the DL. In this case, the BS uses precoding to direct the signals towards the intended receivers, but partially also direct them towards the UEs that interfered with the pilot transmission. This phenomenon is illustrated in Figure 4.15. Note that the scaling factor M_l is determined by the number of antennas at the interfering BS l in the DL. Hence, a BS with many antennas must be careful not to interfere too much with other cells that have fewer BS antennas. The third term in the denominator is the noise power, which might be different between the DL and UL since different receiver hardware is used.

4.3.2 Principle for Precoding Design: Uplink-Downlink Duality

It is nontrivial to select precoding vectors. This is because each UE is affected by all precoding vectors in the network and network-wide precoding optimization is highly impractical. Clearly, the precoding must balance between selfishly directing the signal towards the desired UE and altruistically avoiding to cause interference to other UEs [167]. The difficulty is to find the right balance between these two goals, particularly when many UEs are involved. It is therefore desirable to have a judicious, yet tractable, design principle for precoding. Many heuristic precoding design principles can be found in the literature, but

the well-designed ones are usually rather similar and strongly connected to a fundamental property called *UL-DL duality* [46, 40]. We describe this duality below and explain how it can guide us in the precoding design.

There is a strong connection between the SE expression for the UL in Theorem 4.4 and the DL in Theorem 4.6. Except for the different notation for the transmit powers, the signal terms are the same and the interference terms are similar but the indices (j, k) and (l, i) are interchanged for every UE: $p_{li}\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}$ in the UL is replaced by $p_{li}\mathbb{E}\{|\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2\}$ in the DL. This represents the fact that the UL interference from cell l is received over K_l different UE channels (and processed using a single combining vector), while all the DL interference from cell l is received over the channel from BS l (and depends on K_l precoding vectors). Figure 4.16 illustrates this property from the perspective of two UEs in the network. In this example, BS l can separate the UEs well spatially, while BS j cannot (illustrated here as having a small angular difference between the UEs). The consequence is that the UE in cell j is affected by high interference from the other-cell UE in the UL, while the UE in cell l receives high interference from BS j in the DL. It can, therefore, happen that a UE exhibits very different interference levels in UL and DL.

Despite the differences in how the interference is generated, there is a symmetry that creates a fundamental connection between the achievable SEs in UL and DL, which is called the UL-DL duality.

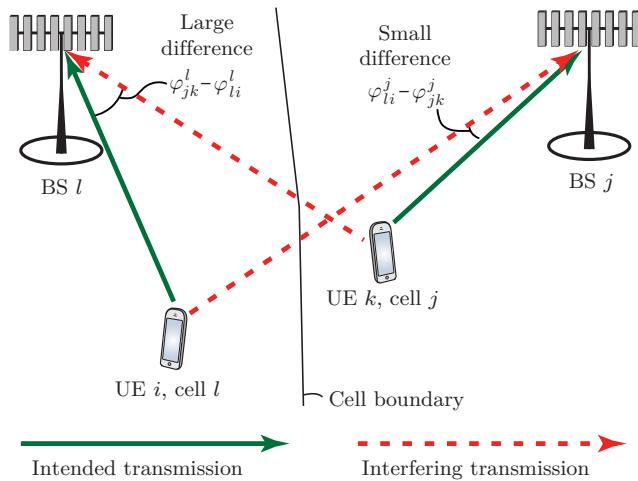
Theorem 4.8. Let $\mathbf{p} = [\mathbf{p}_1^T \dots \mathbf{p}_L^T]^T$, with $\mathbf{p}_j = [p_{j1} \dots p_{jK_j}]^T$, be the $K_{\text{tot}} \times 1$ vector with all UL transmit powers, where $K_{\text{tot}} = \sum_{l=1}^L K_l$ denotes the total number of UEs in the network.

Consider the UL $\underline{\text{SINR}}_{jk}^{\text{UL}}$ in (4.14) and the DL $\underline{\text{SINR}}_{jk}^{\text{DL}}$ in (4.26). For any given set of receive combining vectors $\{\mathbf{v}_{li}\}$ and given \mathbf{p} , we can achieve

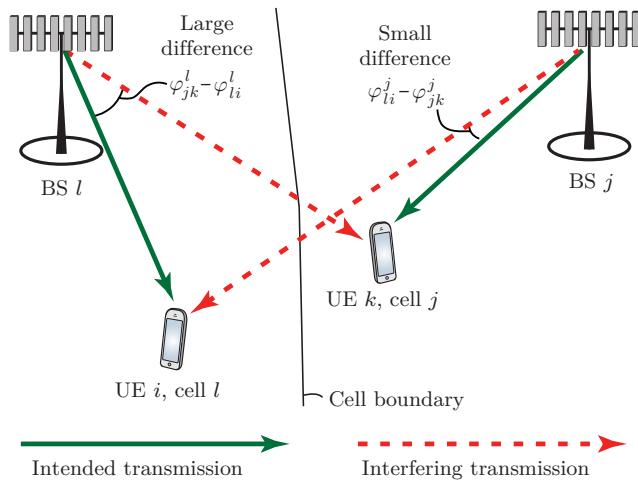
$$\underline{\text{SINR}}_{jk}^{\text{DL}} = \underline{\text{SINR}}_{jk}^{\text{UL}} \quad j = 1, \dots, L, \quad k = 1, \dots, K_j \quad (4.30)$$

if the precoding vectors are selected as

$$\mathbf{w}_{jk} = \mathbf{v}_{jk} / \sqrt{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} \quad (4.31)$$



(a) UL: UE k in cell j is affected by high interference from UE i in cell l .



(b) DL: UE i in cell l receives high interference from BS j .

Figure 4.16: Illustration of how the interference situation can change between the UL and DL. The UL interference comes from the interfering UE, while the DL interference is caused by the BS that serves the interfering UE. In this setup, BS j cannot separate the two UEs due to the similar spatial channel correlation (illustrated as a small angular difference), thus its own UE is affected by high UL interference and the other-cell UE will get high DL interference. In contrast, BS l can separate the UEs well, which leads to little interference in both UL and DL.

for all j and k . Moreover, the vector $\boldsymbol{\rho} = [\boldsymbol{\rho}_1^T \dots \boldsymbol{\rho}_L^T]^T$ with DL transmit powers, with $\boldsymbol{\rho}_j = [\rho_{j1} \dots \rho_{jK_j}]^T$, are selected based on the UL transmit powers as

$$\boldsymbol{\rho} = \frac{\sigma_{\text{DL}}^2}{\sigma_{\text{UL}}^2} \left(\mathbf{D}^{-1} - \mathbf{B} \right)^{-1} \left(\mathbf{D}^{-1} - \mathbf{B}^T \right) \mathbf{p}. \quad (4.32)$$

The sum transmit powers in the DL and UL are related as

$$\frac{\mathbf{1}_{K_{\text{tot}}}^T \boldsymbol{\rho}}{\sigma_{\text{DL}}^2} = \frac{\mathbf{1}_{K_{\text{tot}}}^T \mathbf{p}}{\sigma_{\text{UL}}^2}. \quad (4.33)$$

In (4.32), $\mathbf{B} \in \mathbb{R}^{K_{\text{tot}} \times K_{\text{tot}}}$ is a block matrix with $L \times L$ blocks. The (j, l) th block is denoted by \mathbf{B}_{jl} , has dimension $K_j \times K_l$, and its (k, i) th element is

$$[\mathbf{B}_{jl}]_{ki} = \begin{cases} \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{jk}^l|^2\} - |\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^l\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} & \text{if } k = i \text{ and } j = l \\ \frac{\mathbb{E}\{|\mathbf{v}_{li}^H \mathbf{h}_{jk}^l|^2\}}{\mathbb{E}\{\|\mathbf{v}_{li}\|^2\}} & \text{otherwise.} \end{cases} \quad (4.34)$$

Finally, $\mathbf{D} = \text{diag}(\mathbf{D}_1, \dots, \mathbf{D}_L) \in \mathbb{R}^{K_{\text{tot}} \times K_{\text{tot}}}$ is a (block) diagonal matrix. The j th diagonal block is denoted by \mathbf{D}_j and is a diagonal matrix with the k th element being

$$[\mathbf{D}_j]_{kk} = \underline{\text{SINR}}_{jk}^{\text{UL}} \frac{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}. \quad (4.35)$$

Proof. The proof is available in Appendix C.3.8 on p. 602. \square

This UL-DL duality theorem shows that the SE achieved in the UL can be achieved also in the DL, if the UL combining vectors are used as DL precoding vectors and the DL transmit power is allocated according to (4.32). Simply speaking, each BS should “listen” to the signal from a UE by directing its “hearing” towards a particular spatial direction (selected to balance between high signal power and low interference). The BS then transmits back to the UE in the same spatial direction. If $\sigma_{\text{DL}}^2 = \sigma_{\text{UL}}^2$, the sum transmit power is the same in the UL and DL, but the total power is generally allocated differently among the UEs. For example, consider a setup with one UE at the cell edge and one at the cell center. The cell-center UE has a stronger channel and should use a low UL power not to interfere too much with the cell-edge UE,

which has a weaker channel. It can, however, get a higher DL power since the desired and interfering signals are subject to the same average channel gain in the DL, meaning that the cell-edge UE is less sensitive to intra-cell interference in the DL than in the UL.

The UL-DL duality in cellular networks has been analyzed for decades and some notable early works are [370, 63, 335, 163]. Various generalizations of the duality concept have been established to take multicell properties, power constraints, and transceiver hardware impairments into account; see, for example, [345, 367, 39, 71]. The duality property in Massive MIMO, described in Theorem 4.8, is different in the sense that it applies to the ergodic SE of fading channels, in contrast to the deterministic channels considered in earlier works. This result was first established in [49]. Note that exact duality holds between the UatF capacity bound of the UL and hardening bound for the DL.

The UL-DL duality motivates a simple precoding design principle: select the DL precoding vectors based on the UL receive combining vectors as

$$\mathbf{w}_{jk} = \frac{\mathbf{v}_{jk}}{\|\mathbf{v}_{jk}\|} \quad (4.36)$$

where

$$\begin{bmatrix} \mathbf{v}_{j1} & \dots & \mathbf{v}_{jK_j} \end{bmatrix} = \begin{cases} \mathbf{V}_j^{\text{M-MMSE}} & \text{with M-MMSE precoding} \\ \mathbf{V}_j^{\text{S-MMSE}} & \text{with S-MMSE precoding} \\ \mathbf{V}_j^{\text{RZF}} & \text{with RZF precoding} \\ \mathbf{V}_j^{\text{ZF}} & \text{with ZF precoding} \\ \mathbf{V}_j^{\text{MR}} & \text{with MR precoding.} \end{cases} \quad (4.37)$$

The corresponding combining matrices were defined in (4.7)–(4.11). Note that this design principle is not unique to Massive MIMO, but different flavors of it have been proposed over the past two decades; the early works build indirectly on UL-DL duality [373, 164, 282], while later works refer directly to the duality [98, 368, 40]. The precoding design principle in (4.36) gives $\|\mathbf{w}_{jk}\|^2 = 1$ in every coherence block, which satisfies the required precoding normalization.⁸

⁸The UL-DL duality suggests an average-normalized precoding where $\mathbf{w}_{jk} =$

Scheme	Transmission Multiplications	Computing precoding vectors Multiplications
Any	$\tau_d M_j K_j$	$M_j K_j$

Table 4.4: Computational complexity per coherence block of any transmit precoding scheme, when the precoding is selected based on the combining scheme as in (4.36). Only complex multiplications are considered, while additions and subtractions are neglected; see Appendix B.1.1 on p. 558 for details.

The five precoding schemes in (4.37) depend on the UL transmit power used during pilot signaling, while none of them depends on the DL transmit power (which however appears in the DL effective SINR expression). One important benefit of using the receive combining vectors for transmit precoding is that the computational complexity of computing the precoding vectors reduces to $M_j K_j$ complex multiplications, which corresponds to computing $\|\mathbf{v}_{jk}\|$ in (4.36) for every UE.⁹ The complexity of computing the τ_d transmit signals $\sum_{k=1}^{K_j} \mathbf{w}_{jk} \varsigma_{jk}$ at BS j is $\tau_d M_j K_j$ complex multiplications per coherence block; see Appendix B.1.1 on p. 558 for details. These numbers are summarized in Table 4.4 and we stress that the complexity of precoding is the same irrespective of the choice of precoding scheme, since the combining vectors are used.

There are other precoding schemes in the literature than those listed in (4.37). For example, the polynomial expansion method described in Remark 4.2 can be utilized to reduce the computational complexity of RZF precoding. This has been studied in [173, 231, 372].

4.3.3 Spectral Efficiency with Downlink Channel Estimation

The SE in Theorem 4.8 was derived under the simplifying assumption that the receiving UE has only access to the mean of its precoded

$\mathbf{v}_{jk} / \sqrt{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}$, but in practice one should use the stricter precoding normalization $\mathbf{w}_{jk} = \mathbf{v}_{jk} / \|\mathbf{v}_{jk}\|$ instead to reduce the random variations in the precoded channel $(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}$. This gives better SE since the desired signal ς_{jk} encodes information as phase and amplitude variations; see Figure 4.11 for an illustration of the importance of using instantaneous normalization.

⁹The normalization can be absorbed into the scalar signal ς_{jk} and the corresponding complexity is thus negligible.

channel. Reception without instantaneous CSI only makes sense when the precoded channel is nearly deterministic so that the variations are small. This is approximately the case under channel hardening, but there is generally a performance loss. The loss grows with the channel variations and is particularly large for special types of channels that exhibit little or no hardening [243]. In this section, we consider the alternative approach of estimating the realizations of the precoded channels at the UEs. Since the precoded channel g_{jk} is constant within a coherence block, UE k in cell j can estimate it blindly from the received DL signals, without sending any DL pilots. An explicit algorithm for such estimation can be found in [243], but here we will instead derive a lower bound on the DL capacity that implicitly takes the acquisition of the precoded channels into account. We generalize the bounding technique from [72] to the multicell scenario considered herein and obtain the following result, which we call the *estimation bound*.

Theorem 4.9. The DL ergodic channel capacity of UE k in cell j is lower bounded by $\text{SE}_{jk}^{\text{DL}}$ [bit/s/Hz] given by

$$\text{SE}_{jk}^{\text{DL}} = \frac{\tau_d}{\tau_c} \mathbb{E} \left\{ \log_2 \left(1 + \text{SINR}_{jk}^{\text{DL}} \right) \right\} - \sum_{i=1}^{K_j} \frac{1}{\tau_c} \log_2 \left(1 + \frac{\rho_{ji} \tau_d \mathbb{V}\{\mathbf{w}_{ji}^H \mathbf{h}_{jk}^j\}}{\sigma_{\text{DL}}^2} \right) \quad (4.38)$$

if each BS computes its precoding vectors using only its own channel estimates: $\hat{\mathbf{h}}_{li}^j$ for all l and i . The expectation/variances in (4.38) are computed with respect to the channels \mathbf{h}_{li}^j , for all l and i , and

$$\text{SINR}_{jk}^{\text{DL}} = \frac{\rho_{jk} |\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j|^2}{\sum_{\substack{i=1 \\ i \neq k}}^{K_j} \rho_{ji} |\mathbf{w}_{ji}^H \mathbf{h}_{jk}^j|^2 + \sum_{l=1}^{K_l} \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \rho_{li} \mathbb{E} \left\{ |\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2 \right\} + \sigma_{\text{DL}}^2} \quad (4.39)$$

where the expectations are computed with respect to all other channels.

Proof. The proof is available in Appendix C.3.9 on p. 604. \square

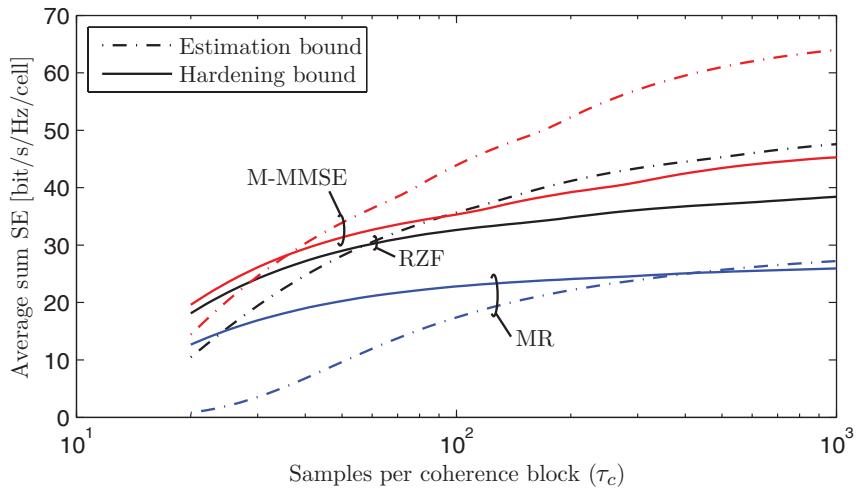
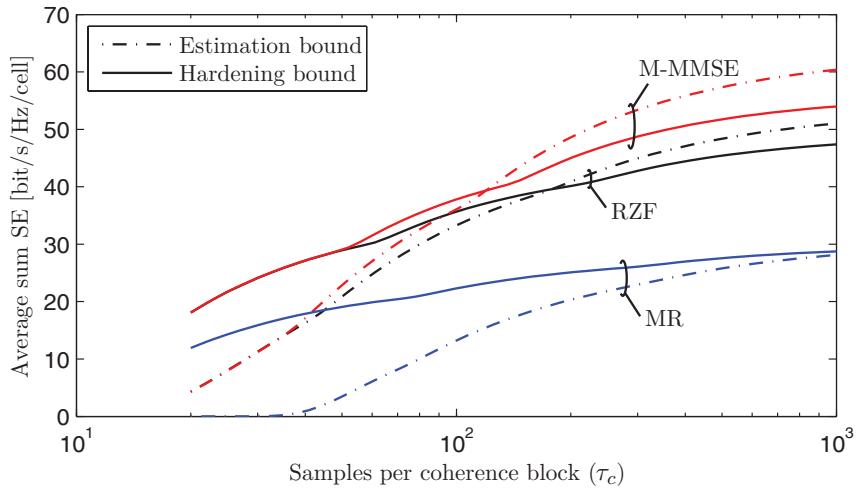
The SE provided in (4.38) is obtained as the difference between two terms, which have intuitive interpretations. The first term can

be called “*SE with perfect intra-cell CSI*”, because it represents the SE that is achieved if the receiving UE knows the precoded intra-cell channels $\mathbf{w}_{ji}^H \mathbf{h}_{jk}^j$ for $i = 1, \dots, K_j$.¹⁰ The second term can be called “*CSI uncertainty loss*”, because it compensates for the imperfect intra-cell CSI at the UE. This term depends on τ_d and τ_c and goes to zero as $\tau_c \rightarrow \infty$, even if also τ_d increases (recall that $\tau_d \leq \tau_c$). This proves that if a large coherence block is used for DL transmission, the UE can estimate the precoded channel perfectly [261]. Since the capacity is unknown, the lower bound that gives the largest value is the best performance indicator that we have. Clearly, the estimation bound in Theorem 4.9 will be larger than the hardening bound in Theorem 4.6 when τ_d and τ_c are sufficiently large, but it is hard to identify the crossing point analytically. We will show numerically that the estimation bound gives larger values when using precoding schemes that cause less interference. On the other hand, it can happen for small channel coherence blocks and/or large intra-cell interference that the second term in (4.38) is so large that the theorem provides a negative SE value. This is an artifact from the bounding technique that neglects a positive term that would have made the SE positive in these special cases (see [72] for a solution).

We compare the two DL SE bounds by continuing the running example that was defined in Section 4.1.3. We consider $M = 100$ antennas, $K = 10$ UEs per cell, and equal DL power allocation of 20 dBm per UE. We consider both the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$ and uncorrelated Rayleigh fading. The length of the coherence block τ_c is varied and $\tau_d = \tau_c - \tau_p$ samples are used for DL data transmission per block.

Figure 4.17 shows the sum SE with the hardening bound from Theorem 4.6 and the estimation bound from Theorem 4.9 when using M-MMSE, RZF, or MR precoding. The horizontal axis shows the coherence block length, using a logarithmic scale that emphasizes the behaviors for small τ_c . The SE is maximized with respect to the pilot reuse factor $f \in \{1, 2, 4\}$ for each value of τ_c , which results into the “bumps” in

¹⁰One can also derive an SE bound where the first term represents that the UE knows the precoded channels from all BSs. That bound gives larger SE as $\tau_c \rightarrow \infty$, but we have noticed that for practical value of τ_c , the bound in Theorem 4.9 gives larger values.

(a) Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$.

(b) Uncorrelated fading.

Figure 4.17: Average DL sum SE based on either the hardening bound in Theorem 4.6 or the estimation bound in Theorem 4.9, as a function of the length τ_c of the coherence block. There are $M = 100$ antennas, $K = 10$ UEs, and the SE is maximized with respect to the pilot reuse factor.

the curves. All curves increase monotonically with τ_c , since the pre-log factor $\frac{\tau_u}{\tau_c} = 1 - \frac{\tau_p}{\tau_c}$ increases. The estimation bound also benefits from an increasing τ_c in terms of improved DL channel estimation, thus this bound is the better choice when the coherence block is large. The crossing point depends on the spatial channel correlation and the precoding scheme because spatial correlation increases the variations in the precoded channel and the precoding determines the interference level under which the DL channels are estimated. M-MMSE benefits the most from using the estimation bound, which is desirable for $\tau_c > 36$ with the local scattering model and for $\tau_c > 120$ with uncorrelated fading. RZF has a similar behavior, but slightly larger τ_c is required before the estimation bound becomes advantageous. The case is different for MR, where the hardening bound gives the highest values, except for very large coherence blocks. This is because MR leads to strong interference that makes the DL channel estimation challenging, leading to a large subtractive term in (4.38).

In summary, it is harder to characterize the DL capacity than the UL capacity, because there are multiple bounds and none of them is always the preferable one; that is, the one providing the largest value. The UE should estimate the realization of the precoded channel from the received signals, but it is challenging to find the best estimator since it is hard to quantify the exact SE for a given estimation scheme.

4.3.4 Comparison of Precoding Schemes

We will now compare the SE achieved with different precoding schemes by continuing the running example that was defined in Section 4.1.3. We consider the same scenario as in the UL example in Section 4.1.4, which means $K = 10$ UEs per cell and a varying number of BS antennas. Equal DL power allocation of 20 dBm per UE and the Gaussian local scattering channel model with ASD $\sigma_\varphi = 10^\circ$ are assumed. Each coherence block consists of $\tau_c = 200$ samples, whereof $\tau_d = \tau_c - fK$ samples are used for DL data transmission and there are $\tau_p = fK$ pilot sequences. For each scheme and number of antennas, we use the DL capacity bound and pilot reuse factor $f \in \{1, 2, 4\}$ that gives the largest SE.

Figure 4.18 shows the average DL sum SE with $f = 1$. We consider

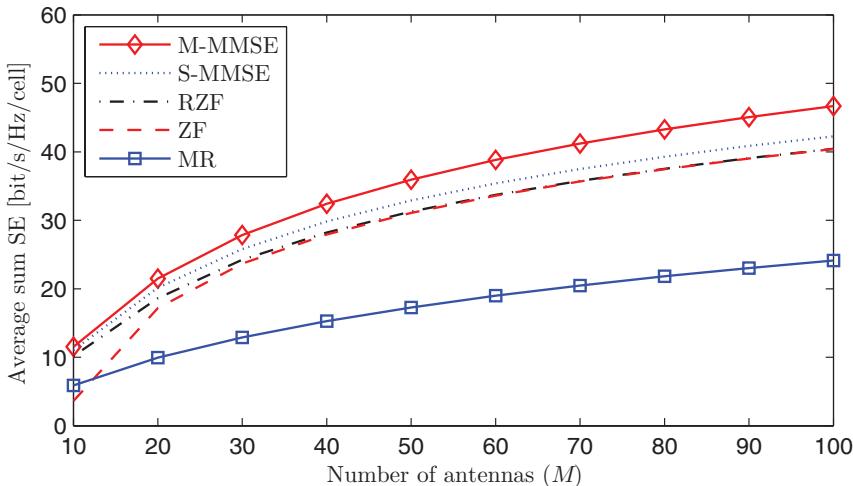


Figure 4.18: Average DL sum SE as a function of the number of BS antennas for different precoding schemes. There are $K = 10$ UEs per cell and the same K pilots are reused in every cell.

M-MMSE, S-MMSE, RZF, ZF, and MR precoding. These precoding schemes behave in a similar way as their UL counterparts. M-MMSE provides the highest SE for any number of antennas. S-MMSE, RZF, and ZF provide almost the same SE, except that ZF has robustness issues for $M < 20$ antennas. Finally, MR provides the lowest SE among all schemes and it is also the only scheme that prefers the hardening bound over the estimation bound. MR achieves only 40%–50% of the SE provided by M-MMSE and 50%–60% of the SE provided by RZF.

Figure 4.19 shows the corresponding sum SE with $f = 2$ and $f = 4$ as pilot reuse factors. The results are once again similar to the UL, both in terms of the SE values and the fact that M-MMSE gives its highest performance with $f = 4$, S-MMSE, RZF, and ZF prefer $f = 2$, and MR gives its highest SE with $f = 1$. This observation is emphasized in Table 4.5, which lists the sum SEs with $M = 100$ for the different precoding schemes. As in the UL, the computational complexity is higher for the precoding/combining schemes that provide higher SEs, and we can appoint M-MMSE, RZF, and MR as three distinct tradeoffs between high SE and low complexity. These are the schemes to choose between in a practical implementation.

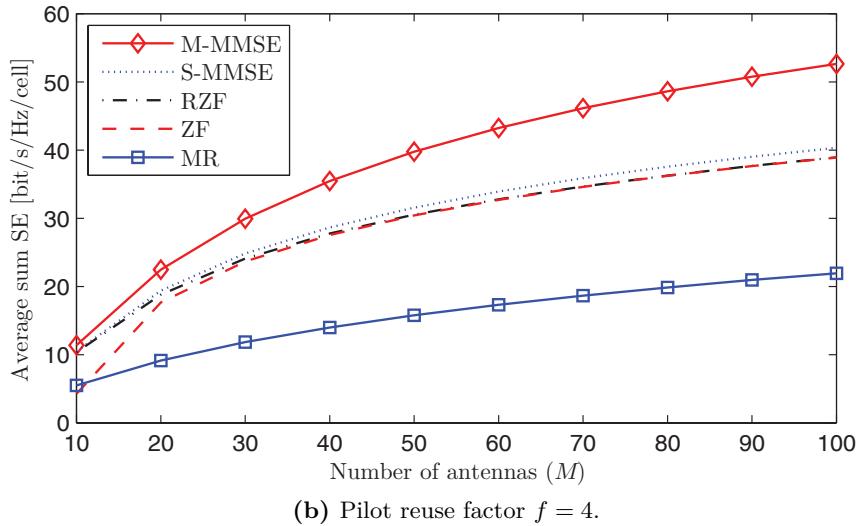
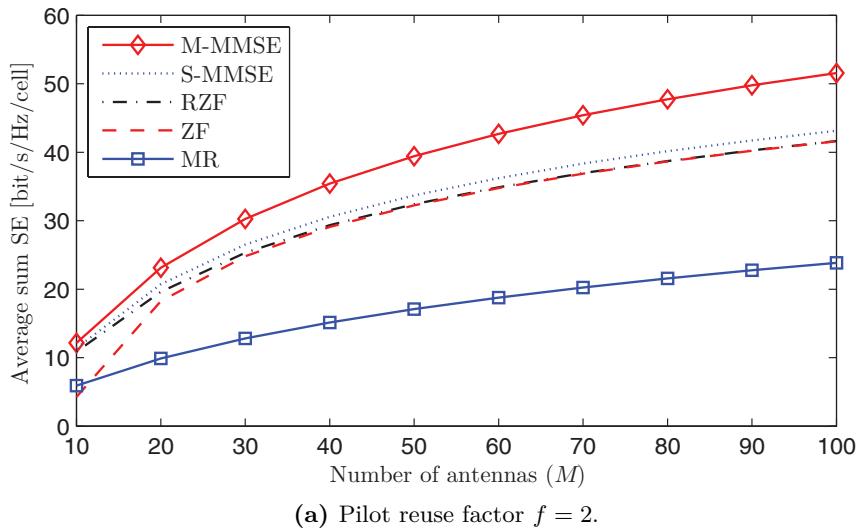


Figure 4.19: Average DL sum SE as a function of the number of BS antennas for different precoding schemes. There are $K = 10$ UEs per cell and either $2K$ or $4K$ pilots that are reused across cells according to the pattern in Figure 4.4b.

Scheme	$f = 1$	$f = 2$	$f = 4$
M-MMSE	46.67	51.57	52.63
S-MMSE	42.24	43.13	40.32
RZF	40.44	41.63	38.92
ZF	40.40	41.60	38.89
MR	24.12	23.83	21.93

Table 4.5: Average DL sum SE [bit/s/Hz/cell] for $M = 100$ and $K = 10$ for different pilot reuse factors f . The largest value for each scheme is in bold face. The results are based on Figures 4.18 and 4.19.

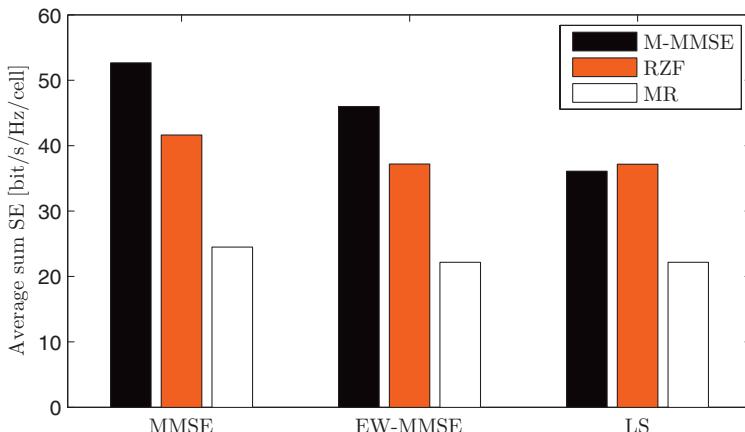


Figure 4.20: Average DL sum SE when using the MMSE, EW-MMSE, or LS channel estimators, for a setup with $M = 100$ BS antennas and $K = 10$ UEs per cell. Three different precoding schemes are considered. The UL counterpart was provided in Figure 4.14 and shows similar results.

4.3.5 SE with Other Channel Estimation Schemes than MMSE

The previous DL SE simulation was based on MMSE channel estimation, but we will now investigate how the SE is affected by using the low-complexity EW-MMSE and LS channel estimators. We continue the example from Figures 4.18 and 4.19, but focus on $K = 10$ UEs, $M = 100$ BS antennas, and for each precoding scheme we use the pilot reuse factor that maximizes the SE.

Figure 4.20 shows a bar diagram over the average sum SE with M-MMSE, RZF, and MR precoding. As expected, the highest SEs are

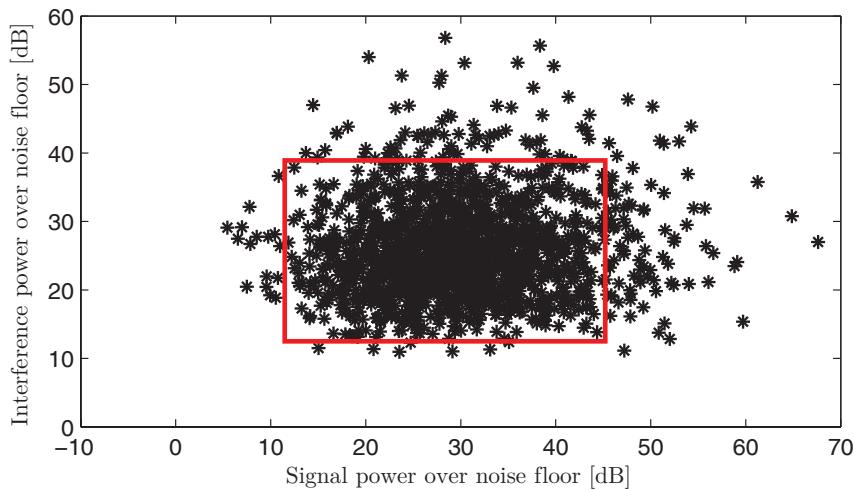
obtained with the MMSE estimator. There is an SE loss of 10%–12% if the EW-MMSE estimator is used instead. The difference in SE between the EW-MMSE and LS estimators is very small when using RZF or MR precoding, while M-MMSE precoding performs poorly with LS. As previously discussed in Section 4.2.3, this is because the LS estimator does not give the right scaling when estimating the channels to UEs in other cells, which leads to an overemphasis on mitigating inter-cell interference. This issue can be solved in practice by reducing the norm of the inter-cell channel estimates.

In summary, the SE loss incurred in the DL by using a suboptimal channel estimator is only 10% in the considered scenario (similar results were obtained for the UL in Section 4.2.3). Hence, the substantial SE gains of using RZF or M-MMSE precoding compared with low-complexity MR precoding remain. It is only the suppression of inter-cell interference in M-MMSE precoding that is particularly sensitive to the choice of channel estimator.

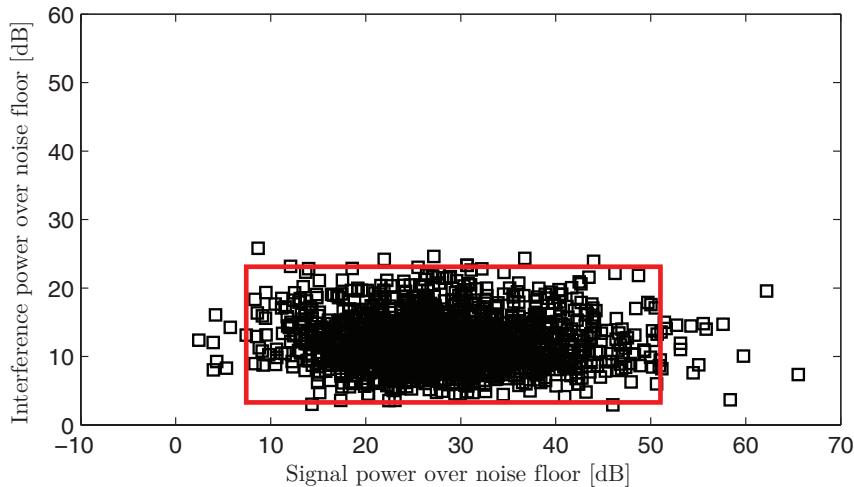
4.3.6 Differences in Interference Between UL and DL

Despite the UL-DL duality, there are important differences between UL and DL. We will exemplify one of the key differences, namely how the interference that affects a UE is formed. To this end, we continue the running example that was defined in Section 4.1.3 and measure the average desired signal power and interference power, normalized by the noise power, at 1600 random UE locations with corresponding shadow fading realizations. We compare MR and M-MMSE combining/precoding using $M = 100$ antennas, $K = 10$ UEs per cell, and $f = 1$. Each setup with 16 cells gives 160 UE locations. The simulation shows 1600 random locations from ten setups.

The desired signal power decays with the propagation distance in both UL and DL so that large values are achieved in the cell center and small values at the cell edge. The signal power and interference power of a UE are essentially independent in the UL, because all signals are received at the same BS. This is illustrated by the scatter plots in Figure 4.21, where each point represents the signal and interference power of one UE. The Gaussian local scattering model with ASD



(a) MR: Signal and interference values are basically independent.



(b) M-MMSE: Signal and interference values are basically independent.

Figure 4.21: Scatter plot of the average UL signal power and UL interference power received at different UE locations. M-MMSE and MR combining with universal pilot reuse are considered. The Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$ is used. The boxes indicate the shape of the point clouds.

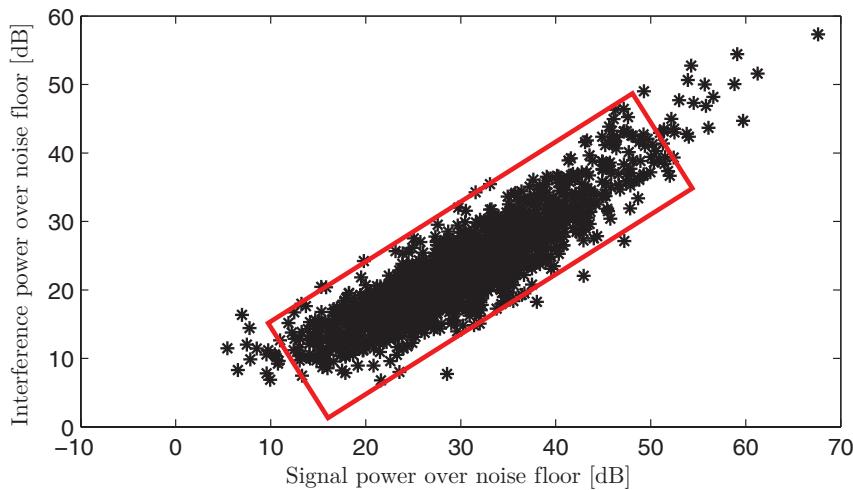
$\sigma_\varphi = 10^\circ$ is considered, but uncorrelated fading would result in the same behavior (but smaller variations). The interference powers have the same spread irrespective of the signal power, thus making the point cloud resemble a horizontal rectangle, as illustrated in the figure. The interference power and signal power are statistically higher with MR than with M-MMSE, which sacrifices some of the signal power to reduce the interference powers with tens of dB.

In contrast, the desired signal power and interference power are coupled in the DL, because all desired and interfering signals from a particular cell are received through the same channel from the cell's BS. UEs with strong channels are more likely to receive strong intra-cell interference and vice versa. This is illustrated by the scatter plots in Figure 4.22, using the same fading model as in the previous figure. MR precoding shows the expected behavior where the point cloud resembles a rectangle that has been rotated 45° , as illustrated in the figure. Interestingly, M-MMSE precoding suppresses the coupling between the desired signal power and interference power, leading to a situation that resembles the UL. This is because M-MMSE identifies and mitigates the strongest sources of interference, leading to more interference suppression between UEs in the cell-center than UEs at the cell-edge.

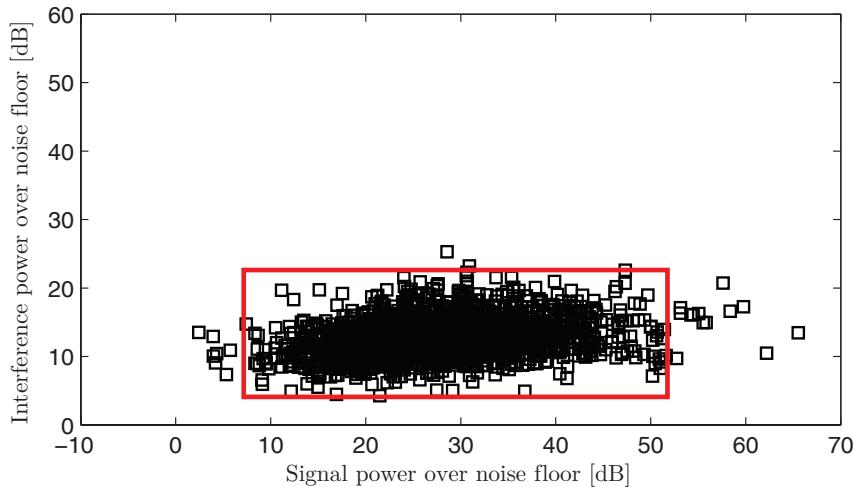
In summary, there are distinct differences between the interference sources that affect a UE in UL and DL. Although the use of M-MMSE (and similar schemes that suppress interference) can reduce these differences, it is important to take them into account when designing the power allocation (see Section 7.1 on p. 452) and other resource allocation tasks.

4.4 Asymptotic Analysis

In this section, we analyze how the SE behaves when the number of BS antennas is very large. A convenient way to analyze this is to investigate the asymptotic regime where $M_j \rightarrow \infty$, as was done in Marzetta's seminal work [208] on Massive MIMO. Similar asymptotic studies were carried out in [169, 244, 281, 43] and numerous other papers. There is also a branch of literature that studies the alternative asymptotic



(a) MR: Signal and interference values are strongly coupled.



(b) M-MMSE: Signal and interference values are basically independent.

Figure 4.22: Scatter plot of the average DL signal power and DL interference power received at different UE locations. M-MMSE and MR precoding with universal pilot reuse are considered. The Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$ is used. The boxes indicate the shape of the point clouds.

regime where both M_j and K_j go to infinity, with a finite non-zero ratio; for example, [148, 173, 336]. The purpose of asymptotic analysis is not that any of these parameters will be nearly infinite in practice (that is physically impossible as explained in Remark 1.3 on p. 192), but is to understand the scaling behaviors and whether there are any fundamental SE limits.

We will let $M_j \rightarrow \infty$ for a fixed number of UEs per cell, as in [208]. While [208] focused on uncorrelated Rayleigh fading and MR, we will analyze the impact of spatial channel correlation and compare different combining/precoding schemes. This will uncover that spatially correlated channels behave fundamentally different than spatially uncorrelated channels in the asymptotic regime, which is a relatively new discovery [44, 43, 239]. We have considered arbitrary spatial correlation matrices so far, but two technical assumptions are needed to enable asymptotic analysis.

Assumption 1. The spatial correlation matrix \mathbf{R}_{li}^j satisfies

1. $\liminf_{M_j} \frac{1}{M_j} \text{tr}(\mathbf{R}_{li}^j) > 0$
2. $\limsup_{M_j} \|\mathbf{R}_{li}^j\|_2 < \infty$

for $l = 1, \dots, L$ and $i = 1, \dots, K_l$.

The operators “ \liminf ” and “ \limsup ” are the sequence-counterparts of the conventional limit operator for functions. They give the smallest and largest values in the asymptotic tail of the sequence (the smallest and largest value are different if the sequence oscillates). These operators are applied to the sequence of correlation matrices $\mathbf{R}_{li}^j \in \mathbb{C}^{M_j \times M_j}$ with different dimensions that is generated when M_j grows. The first condition in Assumption 1 implies that the array gathers an amount of signal energy that is proportional to the number of antennas, which is natural if the array aperture grows with M_j . The second condition implies that the increasing signal energy is spread over many spatial dimensions and does not concentrate on only a few very strong directions; that is, all eigenvalues of \mathbf{R}_{li}^j remain bounded as M_j grows. A consequence of these conditions is that the rank of \mathbf{R}_{li}^j must be proportional to M_j , but full

rank is not required; for example, a fraction of the array can be totally blocked from the UE. Assumption 1 is actually a sufficient condition for asymptotic channel hardening and favorable propagation (see Section 2.5 on p. 231). In case of uncorrelated fading with $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{I}_{M_j}$, we have $\frac{1}{M_j} \text{tr}(\mathbf{R}_{li}^j) = \beta_{li}^j$ and $\|\mathbf{R}_{li}^j\|_2 = \beta_{li}^j$, thus Assumption 1 requires β_{li}^j to be strictly positive and finite in this case.

4.4.1 Linearly Independent and Orthogonal Correlation Matrices

The asymptotic results will depend on how different the spatial correlation matrices of the UEs are. Two measures of such difference play a key role, whereof the first one is linear independence.

Linear independence

A set of vectors is linearly independent if no vector in the set can be written as a linear combination of the others. In this section, we will apply this concept to matrices.

Definition 4.1 (Linearly independent correlation matrices). Consider the correlation matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$. This matrix is *linearly independent* of the correlation matrices $\mathbf{R}_1, \dots, \mathbf{R}_N \in \mathbb{C}^{M \times M}$ if

$$\left\| \mathbf{R} - \sum_{i=1}^N c_i \mathbf{R}_i \right\|_F^2 > 0 \quad (4.40)$$

for all scalars $c_1, \dots, c_N \in \mathbb{R}$. We further say that \mathbf{R} is *asymptotically linearly independent* of $\mathbf{R}_1, \dots, \mathbf{R}_N$ if

$$\liminf_M \frac{1}{M} \left\| \mathbf{R} - \sum_{i=1}^N c_i \mathbf{R}_i \right\|_F^2 > 0 \quad (4.41)$$

for all scalars $c_1, \dots, c_N \in \mathbb{R}$.

Note that linear independence means that the correlation matrix \mathbf{R} cannot be written as a linear combination of the matrices $\mathbf{R}_1, \dots, \mathbf{R}_N$. These matrices can all have full rank, but different eigenvalues (and also different eigenvectors). The asymptotically linear independence

condition is a more restrictive condition, since it does not only require linear independence, but also that the subspace in which the matrices differ has a norm (e.g., a sum of eigenvalues) that grows at least linearly with M . We will give two examples to describe the implications of this definition.

Example of Asymptotic Linear Independence

First, we consider the correlation matrices

$$\mathbf{R} = \begin{bmatrix} 2\mathbf{I}_{M'} & 0 \\ 0 & \mathbf{I}_{M-M'} \end{bmatrix} \quad \text{and} \quad \mathbf{R}_1 = \begin{bmatrix} \mathbf{I}_{M'} & 0 \\ 0 & \mathbf{I}_{M-M'} \end{bmatrix} \quad (4.42)$$

where the two matrices are different in the first M' diagonal elements, for some integer $M' \geq 1$. These matrices are linearly independent since none of them can be written as a scalar times the other matrix. Moreover, we have that

$$\begin{aligned} \frac{1}{M} \|\mathbf{R} - c_1 \mathbf{R}_1\|_F^2 &= \frac{M'(2 - c_1)^2 + (M - M')(1 - c_1)^2}{M} \\ &\geq \frac{(M - M')M'}{M^2} \end{aligned} \quad (4.43)$$

where the lower bound follows from minimizing the expression with respect to c_1 (the minimum is achieved by $c_1 = (M + M')/M$). If $M' = aM$, for some a satisfying $0 < a < 1$, then (4.43) becomes

$$\frac{1}{M} \|\mathbf{R} - c_1 \mathbf{R}_1\|_F^2 \geq \frac{(M - M')M'}{M^2} = (1 - a)a. \quad (4.44)$$

The lower bound is then non-zero for any M , thus (4.41) is satisfied and we conclude that \mathbf{R} and \mathbf{R}_1 are also asymptotically linearly independent. If either M' or $M - M'$ are instead constant, then $\frac{(M - M')M'}{M^2} \rightarrow 0$ as $M \rightarrow \infty$ and there is no asymptotic linear independence. In other words, the asymptotic definition in (4.41) requires the subspace where the matrices are linearly independent to have a dimension that is proportional to M .

Linearly Dependent Matrices are Sensitive to Perturbations

Channels with uncorrelated fading, where the correlation matrices are scaled identity matrices, is a notable example of matrices that are not

linearly independent. However, any such example is non-robust to minor variations in the matrix elements. For example, consider

$$\mathbf{R} = \begin{bmatrix} \epsilon_1 & 0 & \dots \\ 0 & \ddots & 0 \\ \dots & 0 & \epsilon_M \end{bmatrix} \quad \text{and} \quad \mathbf{R}_1 = \mathbf{I}_M \quad (4.45)$$

where $\epsilon_1, \dots, \epsilon_M$ are i.i.d. random variables. The two matrices are linearly independent, except in the special case of $\epsilon_1 = \epsilon_2 = \dots = \epsilon_M$, which has zero probability if the random variables have continuous distributions. Moreover, we have that

$$\begin{aligned} \frac{1}{M} \|\mathbf{R} - c_1 \mathbf{R}_1\|_F^2 &= \frac{1}{M} \sum_{m=1}^M (\epsilon_m - c_1)^2 \\ &\geq \frac{1}{M} \sum_{m=1}^M \left(\epsilon_m - \frac{1}{M} \sum_{n=1}^M \epsilon_n \right)^2 \rightarrow \mathbb{E}\{(\epsilon_m - \mathbb{E}\{\epsilon_m\})^2\} \end{aligned} \quad (4.46)$$

almost surely as $M \rightarrow \infty$, due to the law of large numbers (see Lemma B.12 on p. 564). The inequality in (4.46) follows from setting $c_1 = \frac{1}{M} \sum_{n=1}^M \epsilon_n$, which minimizes the expression with respect to c_1 . The last expression in (4.46) is identified as the variance of an arbitrary element ϵ_m . The variance is non-zero for any random variable, thus \mathbf{R} and \mathbf{R}_1 are also asymptotically linearly independent, almost surely.

This example shows that small random perturbations are sufficient to satisfy the asymptotic definition in (4.41). In practice, the correlation matrices of an arbitrary UE can be viewed as realizations from an underlying continuous random distribution. For example, in our simulations, it is the random UE location along with the channel model that randomly generates the correlation matrices. Under such cases, the correlation matrices of the UEs will be almost surely linearly independent, while the probability of getting linearly dependent matrices is zero.¹¹ In spatially correlated fading, two correlation matrices are linearly independent unless the received signals from the corresponding UEs have

¹¹The principle is the same as if we would generate L i.i.d. random vectors $\mathbf{x}_1, \dots, \mathbf{x}_L \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I}_N)$. One can then show that the vectors will be linearly independent almost surely when $L < N$.

an angular distribution that is identical and a power distribution over these angles that is also identical. In summary, we conclude that all practical collections of correlation matrices are linearly independent.

Spatial Orthogonality

Another measure of the difference between spatial correlation matrices is spatial orthogonality.

Definition 4.2 (Orthogonal correlation matrices). Two correlation matrices $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{C}^{M \times M}$ are *spatially orthogonal* if

$$\text{tr}(\mathbf{R}_1 \mathbf{R}_2) = 0 \quad (4.47)$$

which also implies that $\mathbf{R}_1 \mathbf{R}_2 = \mathbf{0}_{M \times M}$. We further say that \mathbf{R}_1 and \mathbf{R}_2 *asymptotically spatially orthogonal* if

$$\frac{1}{M} \text{tr}(\mathbf{R}_1 \mathbf{R}_2) \rightarrow 0 \quad \text{as } M \rightarrow \infty. \quad (4.48)$$

The definition of asymptotically spatially orthogonal matrices in (4.48) implies that the common subspace of the matrices has a dimension and eigenvalues that are constant, or grow sublinearly with M . This is a less restrictive condition than in the definition of spatially orthogonal matrices in (4.47), where there can be no common subspace. However, both spatial orthogonality conditions are much stronger than the linear independence conditions, since they imply that both correlation matrices are strongly rank-deficient. For example, \mathbf{R} and \mathbf{R}_1 in (4.45) are only orthogonal if $\epsilon_1 = \dots = \epsilon_M = 0$, which makes $\mathbf{R} = \mathbf{0}_{M \times M}$. In spatially correlated fading, it was shown in [7, 363] that two correlation matrices become asymptotically spatially orthogonal if the BS is equipped with a ULA and the channels from the two UEs have non-overlapping supports of their angular distributions. However, the measurements in [121] indicate that such angular separation is unlikely to occur in practice, at least at the frequencies used in the coverage tier of cellular networks. Angular separability is, however, more likely to arise in hotspots operating at mmWave frequencies [275, 8].

4.4.2 Asymptotic Insights

We begin the asymptotic analysis by considering MR combining and pre-coding, for which closed-form expressions were presented in Corollary 4.5 and Corollary 4.7, respectively.

Theorem 4.10 (MR combining). Under Assumption 1, if MR combining with $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j$ is used, it follows that $\underline{\text{SINR}}_{jk}^{\text{UL}} \rightarrow \infty$ as $M_j \rightarrow \infty$ if \mathbf{R}_{jk}^j is asymptotically spatially orthogonal to \mathbf{R}_{li}^j for all $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$. If this is not the case, then, as $M_j \rightarrow \infty$, it follows that

$$\underline{\text{SINR}}_{jk}^{\text{UL}} - \frac{p_{jk}^2 \text{tr} (\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} p_{li}^2 \frac{\left| \text{tr} (\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j) \right|^2}{\text{tr} (\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}}}_{\text{Coherent interference}} \rightarrow 0. \quad (4.49)$$

Proof. The proof is available in Appendix C.3.10 on p. 607. \square

Theorem 4.11 (MR precoding). Under Assumption 1, if MR precoding with $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}}$ is used for all UEs, it follows that $\underline{\text{SINR}}_{jk}^{\text{DL}} \rightarrow \infty$ as $M_1 = \dots = M_L \rightarrow \infty$ if \mathbf{R}_{jk}^l and \mathbf{R}_{li}^l are asymptotically spatially orthogonal for all $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$. If this is not the case, then as $M_1 = \dots = M_L \rightarrow \infty$ it follows that

$$\underline{\text{SINR}}_{jk}^{\text{DL}} - \frac{\rho_{jk} \text{tr} (\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \rho_{li} \frac{\left| \text{tr} (\mathbf{R}_{jk}^l \Psi_{li}^l \mathbf{R}_{li}^l) \right|^2}{\text{tr} (\mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)}}}_{\text{Coherent interference}} \rightarrow 0. \quad (4.50)$$

Proof. The proof is available in Appendix C.3.10 on p. 607. \square

These theorems show that the UL and DL SINRs achieved by MR asymptotically approach the simplified expressions in (4.49) and (4.50), respectively, where the noise and non-coherent interference terms are not present. This does not mean that the “missing” terms go asymptotically to zero, but that they are negligibly small compared to the signal term

and the coherent interference from the UEs that utilized the same pilot sequence, both of which grow proportionally to M . This is a consequence of pilot contamination that makes the estimates of these UEs' channels correlated with the MR vector.

In the special case when the desired UE has a correlation matrix that is asymptotically spatially orthogonal to all the pilot-contaminating UEs' correlation matrices, the SINR will instead grow without bound. As discussed above, this is a very strong condition that is unlikely to hold at the cellular frequencies used in the coverage tier (recall Figure 1.2) [121], but there are theoretical channel models that can give rise to such low-rank effects [7, 363], thus one should always be careful when exploring the asymptotic behaviors. Even with full-rank correlation matrices, by allocating the pilot sequences in such a way that the pilot-sharing UEs have rather different support, the asymptotic SE limit with MR can be increased and this should be taken into account when assigning pilot sequences to the UEs [363, 7, 192].

Apart from the removal of some of the interference/noise terms, the asymptotic formulas have the same characteristics as before. The expressions are particularly clean in the case of uncorrelated fading, where (4.49) and (4.50) become

$$\text{SINR}_{jk}^{\text{UL}} \rightarrow \frac{(p_{jk}\beta_{jk}^j)^2\tau_p\psi_{jk}}{\sum_{(l,i)\in\mathcal{P}_{jk}\setminus(j,k)}(p_{li}\beta_{li}^j)^2\tau_p\psi_{jk}} = \frac{(p_{jk}\beta_{jk}^j)^2}{\sum_{(l,i)\in\mathcal{P}_{jk}\setminus(j,k)}(p_{li}\beta_{li}^j)^2} \quad (4.51)$$

$$\text{SINR}_{jk}^{\text{DL}} \rightarrow \frac{\rho_{jk}p_{jk}(\beta_{jk}^j)^2\tau_p\psi_{jk}}{\sum_{(l,i)\in\mathcal{P}_{jk}\setminus(j,k)}\rho_{li}p_{jk}(\beta_{jk}^l)^2\tau_p\psi_{li}} = \frac{\rho_{jk}(\beta_{jk}^j)^2\psi_{jk}}{\sum_{(l,i)\in\mathcal{P}_{jk}\setminus(j,k)}\rho_{li}(\beta_{jk}^l)^2\psi_{li}}. \quad (4.52)$$

The values given by these asymptotic limits depend on the ratio between the signal power and interference power, while the exact values of these terms are unimportant since the noise term vanishes asymptotically. It is desirable to make $\beta_{li}^j/\beta_{jk}^j$ small in the UL, which corresponds to the interfering UE having a relatively weak channel to BS j . Similarly, it is desirable to make $\beta_{jk}^l/\beta_{jk}^j$ small in the DL, which corresponds to the interfering BS having a weak channel to UE k in cell j . Notice that it

is possible that one of these ratios is small, while the other one is large. These asymptotic insights can be utilized as a heuristic to assign pilots to the UEs.

Based on the asymptotic results with MR and the fact that only coherent interference caused by pilot contamination remains, one might suspect that the SE has a finite limit when using any combining/precoding scheme in a scenario with pilot contamination. To investigate if this is the case, let us consider the “optimal” combining scheme, namely M-MMSE.

Theorem 4.12 (M-MMSE combining). If BS j uses M-MMSE combining with MMSE channel estimation, then the UL SE of UE k in cell j grows without bound as $M_j \rightarrow \infty$, if Assumption 1 holds and the correlation matrix \mathbf{R}_{jk}^j is asymptotically linearly independent of the set of correlation matrices \mathbf{R}_{li}^j with $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$.

Proof. The rigorous proof of this result is quite involved and therefore we will only validate it numerically. The interested reader can find the proof in [43]. \square

This theorem proves that the UL SE of a UE grows without bound as $M_j \rightarrow \infty$ when M-MMSE combining is used, which is in sharp contrast to the case of MR combining. The fact that the noise and non-coherent interference vanish asymptotically is expected, as this was also the case with MR. The impact of coherent interference vanishes if the spatial correlation matrices are asymptotically linearly independent. This is a mild condition that is generally satisfied in practice (as previously explained), but not in the special case of uncorrelated Rayleigh fading that was studied in [208] and many following papers. Asymptotic analysis that is carried out with uncorrelated fading is bound to give overly pessimistic results, and should thus be handled with care.

The reason that the BS can reject coherent interference from UEs that caused pilot contamination is that the MMSE estimated channel vectors, despite being correlated, are linearly independent when the correlation matrices are linearly independent. More precisely, Theorem 3.1 on p. 249 gives the channel estimates $\hat{\mathbf{h}}_{jk}^j = \sqrt{p_{jk}} \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{y}_{jk}^p$

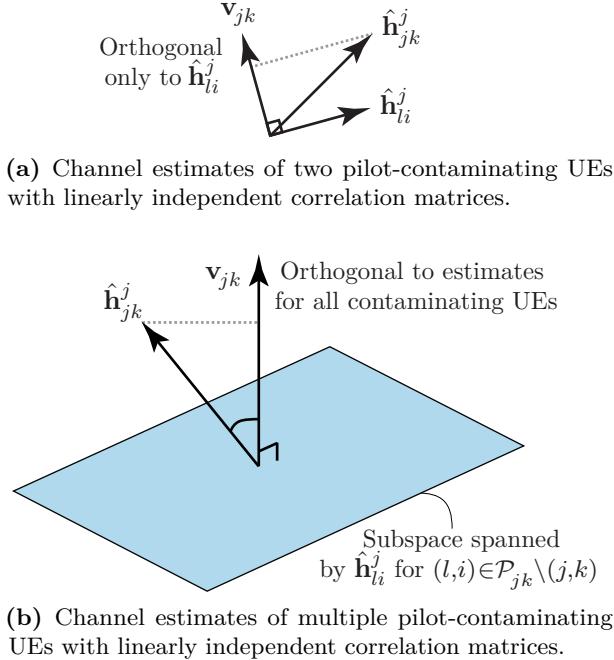


Figure 4.23: Geometric illustration of the linearly independent channel estimates, for UEs that reuse the same pilots but have linearly independent spatial correlation matrices. The indicated combining vector \mathbf{v}_{jk} rejects the coherent interference that is received along the interfering channel estimates, while a non-zero part of the desired signal remains.

and $\hat{\mathbf{h}}_{li}^j = \sqrt{p_{li}} \mathbf{R}_{li}^j \boldsymbol{\Psi}_{jk}^j \mathbf{y}_{jjk}^p$ for any $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$. This implies that the difference

$$\hat{\mathbf{h}}_{jk}^j - c\hat{\mathbf{h}}_{li}^j = \left(\sqrt{p_{jk}} \mathbf{R}_{jk}^j - c\sqrt{p_{li}} \mathbf{R}_{li}^j \right) \boldsymbol{\Psi}_{jk}^j \mathbf{y}_{jjk}^p \quad (4.53)$$

is only zero for some $c \in \mathbb{R}$ if \mathbf{R}_{jk}^j and \mathbf{R}_{li}^j are linearly dependent. This principle is illustrated geometrically in Figure 4.23. The key insight is that, for linearly independent channel estimates, we can find a direction of the combining vector \mathbf{v}_{jk} that is orthogonal to the channel estimates of the contaminating UEs (i.e., $\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{li}^j = 0$) and have a non-zero inner product $\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j$ with the channel estimate of the desired UE. By using this (suboptimal) combining vector, or the “optimal” M-MMSE combining, we can always reject the coherent interference. If the correlation

matrices are also asymptotically linearly independent, the combining vector will also provide an array gain that makes the SINR grow unboundedly. This is why Theorem 4.12 requires asymptotically linearly independent correlation matrices.

A very similar asymptotic behavior applies to the DL SE when using M-MMSE precoding, which is logical due to the UL-DL duality.

Theorem 4.13 (M-MMSE precoding). If all BSs use M-MMSE precoding with MMSE channel estimation, then the DL SE of UE k in cell j grows without bound as $M_1 = \dots = M_L \rightarrow \infty$, if Assumption 1 holds and the correlation matrices \mathbf{R}_{li}^j , for $(l, i) \in \mathcal{P}_{jk}$, are all asymptotically linearly independent.

Proof. The rigorous proof of this result is quite involved and therefore we will only validate it numerically. The interested reader can find the proof in [43]. \square

The asymptotic results presented above rely on the use of MMSE channel estimation, which requires knowledge of the spatial correlation matrices at the BSs. Methods to estimate the correlation matrices were discussed earlier in Section 3.3.3 on p. 260.

Asymptotic SE with EW-MMSE and LS Estimation

As discussed in Section 3.4.1 on p. 265, an alternative to the MMSE estimator is the EW-MMSE estimator in Corollary 3.4, which does not require full knowledge of the spatial correlation matrices. It utilizes only the main diagonals of \mathbf{R}_{li}^j , which can be estimated efficiently, as done in (3.26), by using a small number of samples that does not need to grow with M_j [57, 299]. In [43], it is proved that the SE can grow unboundedly with the number of antennas also when using the EW-MMSE estimator. To reach this result, it is required that the diagonals of the correlation matrices $\mathbf{R}_{l'i'}^j$, with $(l', i') \in \mathcal{P}_{li}$ (between pilot-sharing UEs) are known and asymptotic linearly independent. This condition is likely to hold in practice, as indicated by the measurements in [122].

If also the diagonals of the spatial correlation matrices \mathbf{R}_{li}^j are unknown or unreliable (e.g., due to rapid changes in the UE scheduling

in other cells), it is necessary to consider the LS estimator in (3.35) that does not require any prior statistical information. In this case, the channel estimates of the pilot contaminating UEs are parallel vectors that only differ in the scaling. It appears to be challenging to reject coherent interference in this case, but one can always let the pilot-sharing UEs take turns in being active. This removes the pilot contamination and thus the coherent interference disappears, but it also multiplies the SE with a pre-log factor proportional to $1/L$. Hence, in a large network with a practical number of antennas, it is not an attractive solution.

Remark 4.3 (Pilot contamination precoding). Suppose the BSs are allowed to cooperate, in a coherent joint transmission mode, where the signal to each UE is sent from all BSs. There is then a method called pilot contamination precoding (or large-scale fading precoding/decoding) that can reject the coherent interference in the asymptotic regime [23, 191, 6] and achieve unbounded SE under pilot contamination. Since each UE is served by multiple spatially separated BSs in this setup, the correlation matrix of the joint channel from all BSs is strongly spatially correlated. Moreover, the correlation matrices of the pilot-sharing UEs are likely to be linearly independent, which is also a requirement for the method to work [43]. Hence, pilot contamination precoding relies on the same basic properties as the asymptotic analysis provided above. The drawback with pilot contamination precoding, as compared to M-MMSE combining/precoding, is that all BSs need to process the data signals of all UEs, which might not be practically feasible.

4.4.3 Asymptotic Behavior with Strong Interference

To illustrate the asymptotic behavior, this example considers the specific UL scenario in Figure 4.24, where the interference is particularly strong. There are $L = 2$ BSs and $K = 2$ UEs per cell. Both UEs are located at a distance of 140 m from their serving BS and there is only a 3.6° difference in the angles seen from a BS. The setup is symmetric as illustrated in the figure. There are two pilot sequences, each being reused in every cell by the UE with the same index. The transmit powers and channel propagation model are the same as in the running example, defined in Section 4.1.3, except that we neglect the shadow

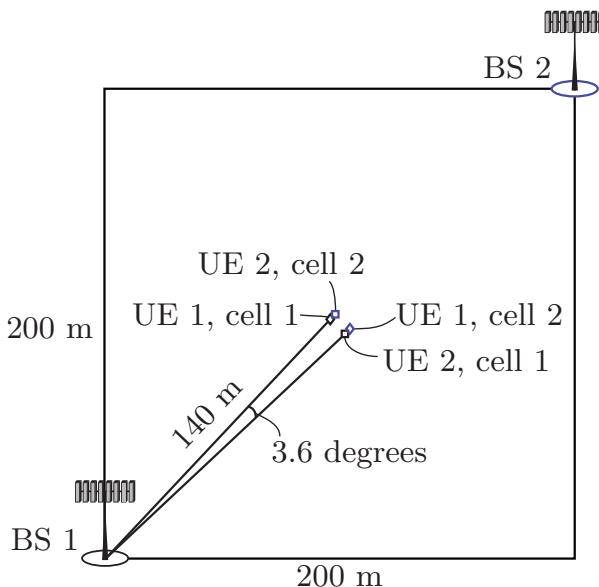


Figure 4.24: The two-cell setup with two UEs per cell that is used to illustrate the asymptotic behaviors of Massive MIMO.

fading. In particular, the average SNR $p_{jk}\text{tr}(\mathbf{R}_{jk}^l)/(M_l\sigma_{\text{UL}}^2)$ is -2 dB for the serving BS ($l = j$) and -2.3 dB for the other BS ($l \neq j$).

We begin by considering the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$, which is substantially larger than the angular difference between the UEs. The UL sum SE per cell is given in Figure 4.25, where the horizontal axis shows the number of BS antennas $M = M_1 = M_2$, using logarithmic scale. We consider M-MMSE, S-MMSE, RZF, ZF, and MR combining. These schemes provide approximately the same SE when $M = 10$, but there are substantial differences for larger values of M . The SE with M-MMSE grows without bound, which is in line with Theorem 4.12. The slope of the curve increases with M and approaches a $\frac{\tau_u}{\tau_c} \log_2(M)$ scaling per UE. The SEs with all other combining schemes approach finite asymptotic limits, due to their inability to reject the coherent interference from the pilot-contaminating UE in the other cell. RZF, ZF, and MR seem to have the same limit, while S-MMSE has a slightly higher limit since the

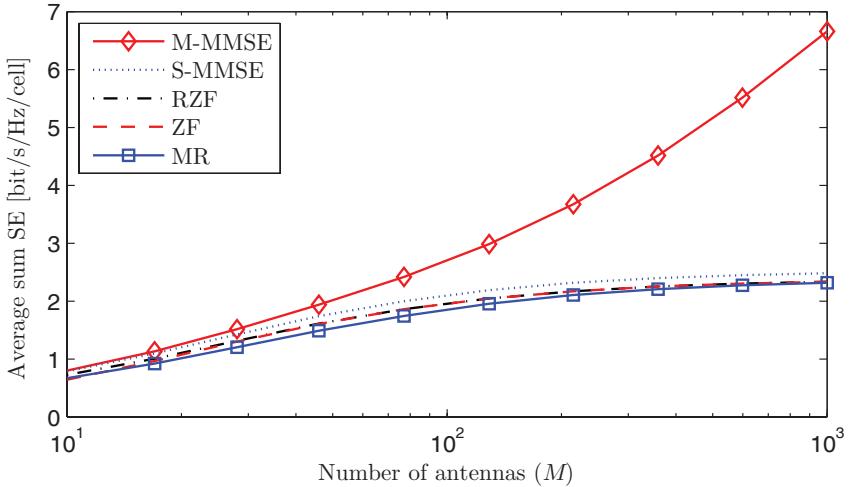


Figure 4.25: The UL sum SE per cell for the setup in Figure 4.24, as a function of the number of BS antennas (notice the logarithmic scale). The local scattering model is considered with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$.

estimation error correlation matrices are taken into account. In this example, the difference between M-MMSE and the other schemes is large already at $M = 100$, while the divergence in SE can appear at substantially higher antenna numbers when the coherent interference is weaker (as it would be in the running example).

The behavioral difference is not limited to scenarios with strong spatial channel correlation. To illustrate this fact, we change channel model to uncorrelated Rayleigh fading that is perturbed by some mild large-scale fading variations over the antenna arrays; that is, $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{D}_{li}^j$, where \mathbf{D}_{li}^j is a diagonal matrix with the m th diagonal element $[\mathbf{D}_{li}^j]_{mm}$ being independently distributed as $10 \log_{10}([\mathbf{D}_{li}^j]_{mm}) \sim \mathcal{N}(0, \sigma_{\text{variation}}^2)$. This large-scale fading variations are motivated by the NLoS measurements reported in [122], which show that there are 4 dB differences in received signal power over the antennas in a ULA. Figure 4.26 shows the UL sum SE per cell with $M = 200$ antennas and varying standard deviation $\sigma_{\text{variation}} \in [0, 4]$. M-MMSE gives an SE similar to the other schemes when there are no large-scale fading variations, but the difference increases rapidly with the standard deviation. The reason is that the

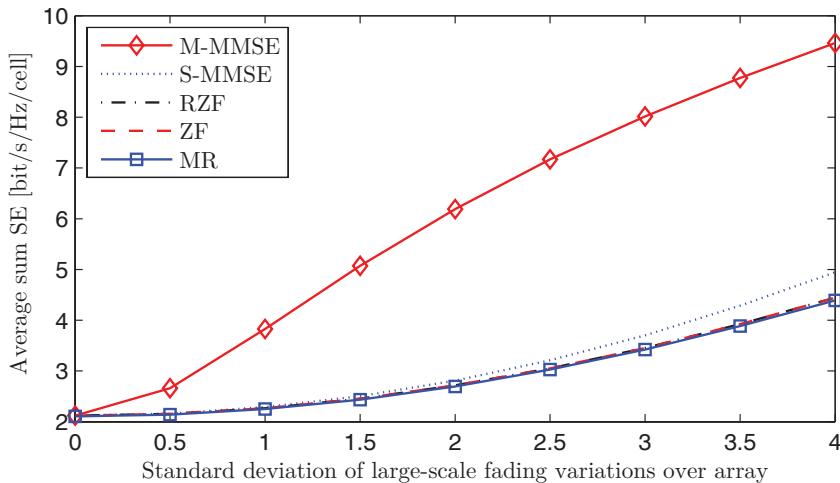


Figure 4.26: The UL sum SE per cell for the setup in Figure 4.24 with $M = 200$ antennas and uncorrelated fading with large-scale fading variations over the array. The standard deviation $\sigma_{\text{variation}}$ is varied along the horizontal axis.

correlation matrices become asymptotically linearly independent when the large-scale fading variations are introduced. The more different the matrices are, the smaller is the loss in signal power from the interference rejection carried out by M-MMSE, and the larger the SE becomes. The SE also increases when using other schemes than M-MMSE, since the random variations in the correlation matrices make the channels more likely to be spatially separated. Such spatial correlation is beneficial irrespective of which scheme that is used, but it is only M-MMSE that utilizes it to reject coherent interference. Hence, the performance gap between M-MMSE and the other schemes is substantial for $M = 200$ and will continue to grow as $M \rightarrow \infty$, since the other schemes have finite asymptotic limits.

4.5 Summary of Key Points in Section 4

- SE expressions for the UL and DL were derived in this section and can be computed numerically for any channel model. Spatial channel correlation can have a positive impact on the SE since most UEs cause less interference to each other. However, there are also larger variations in SE since UEs that happen to have similar spatial correlation matrices interfere more with each other.
- The BSs should use the same vectors for UL receive combining and DL transmit precoding, motivated by the UL-DL duality. The M-MMSE scheme provides the highest SE and requires the highest computational complexity, while the MR scheme has the lowest complexity and SE. The RZF scheme provides a good SE-complexity tradeoff. The channel estimates provided by the low-complexity EW-MMSE estimator are sufficient for these schemes to work well, thus high-complexity channel estimators are not needed.
- Different power allocations are needed in UL and DL since the signal and interference levels of a UE can be very different. This is further studied in Section 7.1 on p. 452.
- The received signal power increases linearly with the number of BS antennas M , thanks to coherent signal processing. This happens even with pilot contamination. However, pilot contamination gives rise to coherent interference that grows with M , unless this interference is suppressed by using M-MMSE combining/precoding. The coherence interference is in addition to the conventional non-coherent interference that is unaffected by M . The impact of pilot contamination can be made negligible if the pilots are not reused in every cell, which leaves an increased pilot overhead as the main practical impact of pilot contamination.

- The SINRs increase with the pilot length τ_p and the pre-log factors decrease with τ_p . Thus, it is non-trivial to find the pilot length that maximizes the SE.
- The DL transmission can be performed without DL channel estimation, by relying only on channel hardening. However, all precoding schemes, except MR, can benefit substantially from estimating the precoded channel from the received DL data transmission. The accuracy increases with the length of the coherence block.
- The SE always grows with the number of BS antennas. In the cases of practical interest, there is no upper SE limit when using M-MMSE, despite the common belief that a fundamental upper limit exists. This holds with either MMSE or EW-MMSE channel estimation since the noise and all types of interference are rejected and their impact vanishes asymptotically. Spatial correlation enables the BS to reject the coherent interference. Other schemes (e.g., RZF and MR) have asymptotic upper SE limits determined by the coherent interference from pilot contamination, since only the impact of noise and non-coherent interference vanish.

5

Energy Efficiency

In this section, we analyze the energy efficiency (EE) of Massive MIMO based on a realistic *circuit power (CP)* consumption model. Before looking into this, we explain in Section 5.1 why power consumption (PC) is a major concern for future cellular networks. In Section 5.2, we show that Massive MIMO can potentially improve the area throughput while providing substantial power savings. The asymptotic behavior of the transmit power when the number of BS antennas grows towards infinity is also studied and a power-scaling law is established, which proves how quickly the transmit power can be reduced with the number of antennas while achieving a non-zero asymptotic SE. Section 5.3 formally introduces the EE metric and provides basic insights into the EE-SE tradeoff, as a function of the key system parameters, such as the number of BS antennas and UEs. A tractable and realistic CP model for Massive MIMO networks is developed in Section 5.4. This model is used in Section 5.5 to examine the EE-throughput tradeoff of Massive MIMO, and also in Section 5.6, to design a cellular network that achieves maximal EE. Finally, the key points are summarized in Section 5.7.

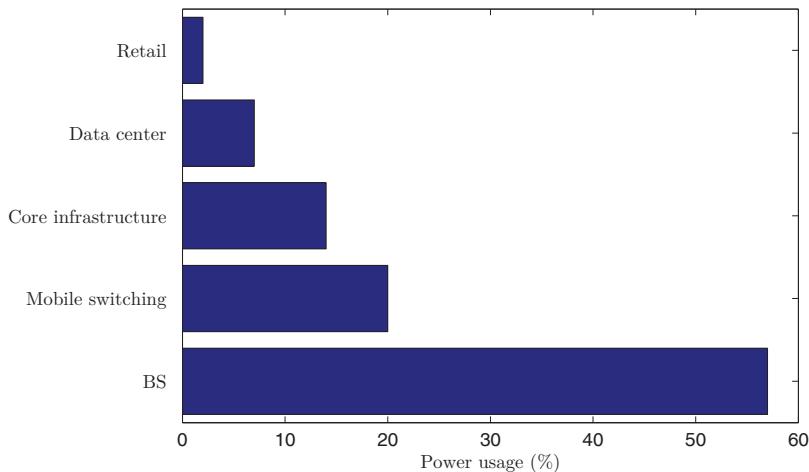


Figure 5.1: Breakdown of power consumed by cellular networks [138].

5.1 Motivation

As mentioned in Section 1.1.1 on p. 163, if the annual traffic growth rate of cellular networks continues to be in the range of 41%–59%, the area throughput will have to increase by a factor of 1000 over the next 15–20 years [271]. If no active countermeasures are taken, the solution to the “1000× data challenge” will increase the PC prohibitively. This is because current networks are based on a rigid central infrastructure, that is powered by the electric grid and designed to maximize the throughput and the traffic load that each cell can handle. The PC is mainly determined by the peak throughput and varies very little with the actual throughput of the cell. This is problematic since the number of active UEs in a cell can change rapidly due to changes in user behaviors and the bursty nature of packet transmission (see Section 7.2.3 on p. 479 for further discussion). The measurements reported in [27] show that the daily maximum network load is 2–10 times higher than the daily minimum load. Hence, a lot of energy is wasted at the BSs in non-peak hours.

A quite remarkable effort has been devoted to reducing the PC of UEs, in order to enhance their battery lifetime. Academia and industry alike have recently shifted their attention towards the BSs. According

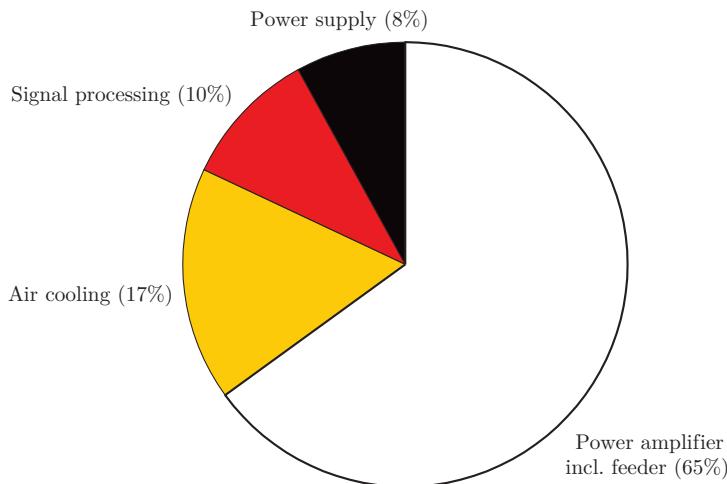


Figure 5.2: Percentage of power consumed by different components of a coverage tier BS [140].

to figures from Vodafone [138] shown in Figure 5.1, BSs account for almost 60% of the total power consumed by a cellular network, while 20% is consumed by mobile switching equipment, and around 15% by the core infrastructure. The rest is consumed by data centers and retail/offices. The total power consumed by a BS is composed of fixed (traffic-independent) and variable (traffic-dependent) parts. Figure 5.2 breaks down how different parts of a BS in the coverage tier contribute to the total PC [140]. The fixed part, including control signaling and power supply, accounts for around one quarter of the total consumed power. This amount is not efficiently used during non-peak traffic hours or, even worse, it is completely wasted when no UE is active within the coverage area of a BS (as frequently happens in rural areas). The most significant portion of power is consumed in the power amplification process. Shockingly, 80%–95% of this power is dissipated as heat in the power amplifiers (PAs), since the total efficiency of currently deployed PAs is generally in the range of 5%–20% (depending on the communication standard and the equipment’s condition). This is due to the fact that the modulation schemes used in contemporary communication standards, such as LTE, are characterized by strongly varying signal

envelopes with peak-to-average power ratios that exceed 10 dB. To avoid distortions of the transmitted signals, the PAs have to operate well below saturation.

Massive MIMO aims at evolving the coverage tier BSs by using arrays with a hundred or more antennas, each transmitting with a relatively low power. This allows for coherent multiuser MIMO transmission with tens of UEs being spatially multiplexed in both UL and DL of each cell. The area throughput is improved by the multiplexing gain. However, the throughput gains provided by Massive MIMO come from deploying more hardware (i.e., multiple RF chains per BS) and digital signal processing (i.e., SDMA combining/precoding) which, in turn, increase the CP per BS. Hence, the overall EE of the network, defined later as “how much energy it takes to achieve a certain amount of work”, can be optimized only if these benefits and costs are properly balanced. The aim of Section 5 is to explore the potential of Massive MIMO to improve the network-wide EE. Before looking into this, we show in Section 5.2 that the array gain can be utilized to reduce the transmit power.

Remark 5.1 (A brief look at the hotspot tier). Hotspot BSs—providing additional capacity to small areas within the coverage of the coverage tier BSs—have a big role to play in the years to come (see Section 1.1.1 on p. 163), not only for increasing the area throughput but also for reducing the transmit power, by shortening the distances between UEs and their serving BSs. However, this comes at the price of deploying a larger amount of hardware and network infrastructure, which could substantially increase the network’s PC. A possible solution is to equip hotspot BSs with mechanisms that monitor the traffic load and save power by deciding whether to turn on or off certain components [24, 351]. These techniques are promising for reducing the consumed power of the hotspot tier, without sacrificing the area throughput, but they are not suitable for the coverage tier (which is the main focus of this monograph) since they would inevitably degrade the coverage and mobility support. This is why most of the research activities for the coverage tier aim at making the consumed power fully proportional to the network load, to avoid the need for dynamically turning anything on or off.

5.2 Transmit Power Consumption

A metric that is used to measure the transmit power consumed by a wireless network is the area transmit power (ATP), which is defined as the network-average power usage for data transmission per unit area. This metric is measured in W/km²:

$$\text{ATP} = \text{transmit power [W/cell]} \cdot D \text{ [cells/km}^2\text{]} \quad (5.1)$$

where D is the average cell density, as defined in (1.1). Consider the DL of a Massive MIMO network with L cells. BS j communicates with K_j UEs. As described in Section 4.3 on p. 316, BS j uses the precoding vector $\mathbf{w}_{jk} \in \mathbb{C}^{M_j}$ to transmit the data signal $\varsigma_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \rho_{jk})$ intended for UE k in cell j . Since the precoding vector is normalized as $\mathbb{E}\{\|\mathbf{w}_{jk}\|^2\} = 1$, the transmit power allocated to this UE is equal to the signal variance ρ_{jk} . The ATP of BS j is thus given by

$$\text{ATP}_j^{\text{DL}} = D \sum_{k=1}^{K_j} \rho_{jk}. \quad (5.2)$$

The corresponding UL expression is obtained if ρ_{jk} is replaced with p_{jk} .

To quantitatively evaluate ATP_j^{DL} , we consider the running example defined in Section 4.1.3 on p. 288 with pilot reuse $f = 1$, $K = 10$ UEs in each cell and a DL transmit power of 20 dBm per UE, which corresponds to $\rho_{jk} = 100 \text{ mW } \forall j, k$. Then, the total DL transmit power per BS is 30 dBm. Each BS covers a square area of 0.25 km × 0.25 km and is equipped with the same number of antennas M . The ATP of BS j is $\text{ATP}_j^{\text{DL}} = 16 \text{ W/km}^2$, which is smaller than in current LTE networks of a factor 15 (cf. Remark 4.1 on p. 291). However, in order to be meaningful, the ATP needs to be complemented by a quality metric; for example, the area throughput. Table 5.1 summarizes the average DL sum throughput per cell over a 20 MHz channel. The results are obtained using the SE values of Figure 4.18. In the case of $M = 100$, we see that the DL throughput can be as large as 482 Mbit/s/cell with MR and 1053 Mbit/s/cell with M-MMSE, which is 8–16 times larger than in LTE (cf. Remark 4.1 on p. 291). These per-cell throughputs correspond to area throughputs of 7.72 Gbit/s/km² and 16.8 Gbit/s/km², respectively.

Scheme	$M = 10$	$M = 50$	$M = 100$
M-MMSE	243 Mbit/s	795 Mbit/s	1053 Mbit/s
RZF	217 Mbit/s	648 Mbit/s	832 Mbit/s
MR	118 Mbit/s	345 Mbit/s	482 Mbit/s

Table 5.1: Average DL throughput per cell over a 20 MHz channel for $K = 10$ with M-MMSE, RZF, and MR precoding. The results follow from Figure 4.18 and are obtained for a DL ATP of 16 W/km².

In summary, the above analysis shows that, for the considered scenario and a sufficiently large number of BS antennas, Massive MIMO can achieve more than an order-of-magnitude higher area throughput than current networks, while also providing more than an order-of-magnitude ATP savings. Notice that the division of the total transmit power among M antennas results into a low transmit power per antenna. With $M = 100$ and a total DL transmit power of 1 W, we have only 10 mW per antenna in the considered scenario. This allows replacing the expensive high-power PAs used in current cellular networks (that consume most of the power in a BS) by hundreds of low-cost low-power PAs with output power in the mW range. With a sufficiently low power per antenna, we might not even need to amplify the signal by a dedicated PA, but feed each antenna directly from a circuit. This can have very positive effects on the consumed power. It is important to note that these savings are obtained at the cost of deploying multiple RF chains per BS and using combining/precoding schemes, whose computational complexities depend on the number of BS antennas and UEs (cf. Table 4.1 on p. 287). This, in turn, increases the CP of the network, as will be quantified in Section 5.4. Therefore, the ATP metric does not provide the right insights into the net reduction in consumed power provided by Massive MIMO. This is why we advocate the use of the EE metric, that will be defined in Section 5.3 and studied in the remainder of this section, which accounts not only for the transmit power and throughput, but also for the CP.

5.2.1 Asymptotic Analysis of Transmit Power

Before delving into the EE and CP analysis, in what follows we briefly describe an interesting power-scaling result, which establishes how the SE and transmit power interact as the number of antennas grows. As in Section 4.4 on p. 335, the analysis is performed in the asymptotic regime where $M_j \rightarrow \infty$, while the number of UEs per cell is kept fixed and the spatial correlation matrices satisfy Assumption 1 on p. 337. The purpose is to show that, as the number of antennas grows, we can trade away parts of the array gain for reducing the transmit powers; in particular, the transmit power can asymptotically go to zero while approaching a non-zero SE limit. This result provides evidence that Massive MIMO can operate at very low transmit power levels.

For simplicity, we focus on the DL and consider MR precoding with $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}}$, such that $\mathbb{E}\{\|\mathbf{w}_{jk}\|^2\} = 1$. Since other precoding schemes generally provide larger SE than MR, if we can establish that the SE of MR approaches a non-zero asymptotic limit, we expect that the same result holds for other precoding schemes. As shown in Corollary 4.7 on p. 318, the DL channel capacity of UE k in cell j with MR precoding is lower bounded by $\underline{\text{SE}}_{jk}^{\text{DL}}$ [bit/s/Hz], given by

$$\underline{\text{SE}}_{jk}^{\text{DL}} = \frac{\tau_d}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{DL}}) \quad (5.3)$$

where

$$\begin{aligned} \underline{\text{SINR}}_{jk}^{\text{DL}} = & \frac{\rho_{jk} p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \frac{\text{tr}(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)}{\text{tr}(\mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)} + \sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \rho_{li} \frac{p_{jk} \tau_p |\text{tr}(\mathbf{R}_{jk}^l \Psi_{li}^l \mathbf{R}_{li}^l)|^2}{\text{tr}(\mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l)} + \sigma_{\text{DL}}^2} \end{aligned} \quad (5.4)$$

and Ψ_{li}^j was defined in (3.10), and reported below for convenience:

$$\Psi_{li}^j = \left(\sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'} \tau_p \mathbf{R}_{l'i'}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1}. \quad (5.5)$$

Recall that p_{jk} denotes the UL power used for transmitting the pilot

sequence of length τ_p whereas ρ_{jk} denotes the DL signal power. The above expressions can be used to obtain the following result.

Lemma 5.1. Consider $M = M_1 = \dots = M_L$, $p_{jk} = \overline{P}/M^{\varepsilon_1}$, and $\rho_{jk} = \underline{P}/M^{\varepsilon_2}$, where $\overline{P}, \underline{P}, \varepsilon_1, \varepsilon_2 > 0$ are constants. If MR precoding with $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}}$ is used and Assumption 1 holds, then

$$\frac{\text{SINR}_{jk}^{\text{DL}} - \frac{\frac{1}{M} \text{tr}(\mathbf{R}_{jk}^j \mathbf{R}_{jk}^j)}{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{\left(\frac{1}{M} \text{tr}(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l)\right)^2}{\frac{1}{M} \text{tr}(\mathbf{R}_{li}^l \mathbf{R}_{li}^l)}}}{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{\left(\frac{1}{M} \text{tr}(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l)\right)^2}{\frac{1}{M} \text{tr}(\mathbf{R}_{li}^l \mathbf{R}_{li}^l)}} \rightarrow 0 \quad (5.6)$$

as $M \rightarrow \infty$ if $\varepsilon_1 + \varepsilon_2 < 1$, while $\text{SINR}_{jk}^{\text{DL}} \rightarrow 0$ if $\varepsilon_1 + \varepsilon_2 > 1$.

Proof. The proof is given in Appendix C.4.1 on p. 609. \square

Lemma 5.1 provides a transmit *power-scaling law* for Massive MIMO networks. The condition $\varepsilon_1 + \varepsilon_2 < 1$ implies that we can either decrease both p_{jk} and ρ_{jk} roughly as $1/\sqrt{M}$ or decrease one faster than the other, as long as the product $p_{jk}\rho_{jk}$ does not decay faster than $1/M$. Under these conditions, the DL SE has a non-zero asymptotic limit and behaves asymptotically as

$$\frac{\tau_d}{\tau_c} \log_2 \left(1 + \frac{\frac{1}{M} \text{tr}(\mathbf{R}_{jk}^j \mathbf{R}_{jk}^j)}{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{\left(\frac{1}{M} \text{tr}(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l)\right)^2}{\frac{1}{M} \text{tr}(\mathbf{R}_{li}^l \mathbf{R}_{li}^l)}} \right). \quad (5.7)$$

The reason why p_{jk} and ρ_{jk} play a similar role in the DL is that the product $p_{jk}\rho_{jk}$ appears in the numerator of (5.4). Since $p_{jk}\rho_{jk}$ is multiplied with $\text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)$, which grows proportionally to M , the numerator will grow without bound as $M \rightarrow \infty$ as long as $p_{jk}\rho_{jk}M$ diverges. This gives rise to a sort of “squaring effect” that limits to $1/\sqrt{M}$ the fastest rate at which the two transmit powers can be jointly decreased. In the case of fixed UL pilot powers, the “squaring effect” is absent and thus the fastest rate at which ρ_{jk} can be decreased is $1/M$ and not $1/\sqrt{M}$. If the transmit powers are reduced faster than allowed by the power-scaling law, the numerator goes asymptotically to zero and this leads to zero asymptotic SE.

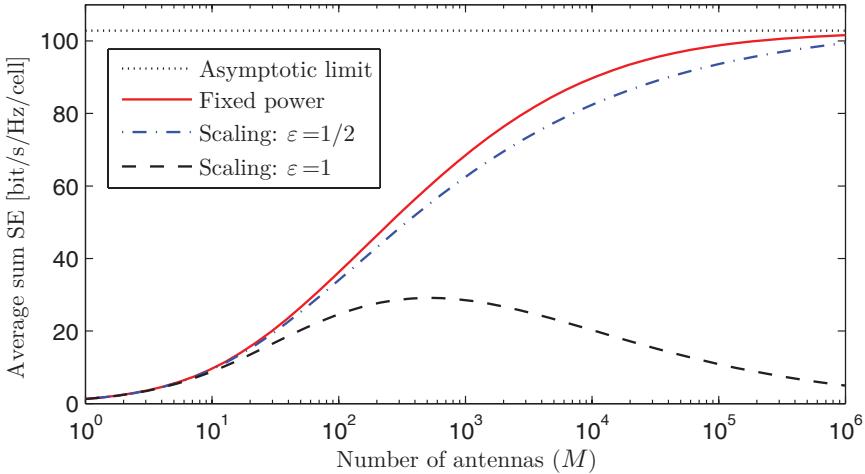


Figure 5.3: Average DL sum SE per cell as a function of the number of BS antennas with averaged-normalized MR precoding, $\underline{P} = \bar{P} = 20$ dBm, for $\varepsilon = 1/2$, $\varepsilon = 1$, and fixed power (i.e., $\varepsilon = 0$). Uncorrelated Rayleigh fading is considered. The data and pilot signal powers can be reduced both as $1/\sqrt{M}$ (i.e., $\varepsilon = 1/2$) while achieving almost the same asymptotic DL SE as for fixed power.

Figure 5.3 exemplifies the asymptotic result of Lemma 5.1 for the running example described in Section 4.1.3 on p. 288 with a UL transmit power per UE of $\bar{P} = 20$ dBm and a total DL transmit power of $K\underline{P} = 30$ dBm, for $M = 1$. Uncorrelated Rayleigh fading is considered. We assume $\varepsilon = \varepsilon_1 = \varepsilon_2$ and consider two different power-scalings for the transmit powers, namely, $\varepsilon = 1/2$ and $\varepsilon = 1$. The fixed power case (with \underline{P} and \bar{P} being the powers) and the asymptotic limit for $\varepsilon = 0$ are also indicated. As stated in Lemma 5.1, if p_{jk} and ρ_{jk} are decreased as $1/\sqrt{M}$ (i.e., $\varepsilon = 1/2$), we approach a non-zero asymptotic limit. This limit is almost identical to the fixed power case, though the convergence is slower. In particular, we obtain 55% of the asymptotic value with $M = 10^3$ and 95% with $M = 10^6$. In agreement with Lemma 5.1, the average DL sum SE vanishes asymptotically when having $\varepsilon = 1$.

In summary, the above analysis gives theoretical evidence that Massive MIMO may operate at unconventionally low transmit power levels. In fact, Figure 5.3 shows that with $M = 100$ the total transmit power per BS can be reduced from $K\underline{P} = 1$ W to $K\underline{P}/\sqrt{M} = 0.1$ W,

while achieving almost the same SE. The division of 0.1 W among the 100 antennas implies that the power per antenna is only $K\underline{P}/M^{3/2} = 1\text{ mW}$. We stress that the above reduction in transmit power comes at the cost of using a large number of BS antennas, which, in turn, increases the CP.

Remark 5.2 (A look at the UL). The DL power-scaling in Lemma 5.1 can be easily rederived for the UL (starting from Corollary 4.5 on p. 303) with potential benefits for UEs. In fact, although UEs are rapidly developing with new advanced functionalities, the battery capacity only increases at a modest 10% every two years [113, 186]. Since the wireless data traffic per device grows faster than that [109], this leads to an increasing gap between the demand for power and the battery capacity offered. Therefore, although UEs are only marginally responsible for the PC in cellular networks, Massive MIMO provides potential benefits in terms of power saving to both operators and UEs. The power-savings are particularly important in the deployment of sensors and other devices whose batteries preferably should last for a very long time.

5.3 Definition of Energy Efficiency

In a broad sense, EE refers to how much energy it takes to achieve a certain amount of work. This general definition applies to all fields of science, from physics to economics, and wireless communication is no exception [371]. Unlike many fields wherein the definition of “work” is straightforward, in a cellular network it is not easy to define what exactly one unit of “work” is. The network provides connectivity over a certain area and it transports bits to and from UEs. Users pay not only for the delivered number of bits but also for the possibility to use the network anywhere at any time. Moreover, grading the performance of a cellular network is more challenging than it first appears, because the performance can be measured in a variety of different ways and each such performance measure affects the EE metric differently (see Section 5.4 for further details and also [371]). Among the different ways to define the EE of a cellular network, one of the most popular definitions takes inspiration from the definition of SE, that is, “the SE

of a wireless communication system is the number of bits that can be reliably transmitted per complex-valued sample” (a formal definition of the SE was given in Definition 1.2 on p. 167). By replacing “SE” with “EE” and “complex-valued sample” with “unit of energy”, the following definition is obtained [371, 157]:

Definition 5.1 (Energy efficiency). The EE of a cellular network is the number of bits that can be reliably transmitted per unit of energy.

According to the definition above, we define the EE as

$$\text{EE} = \frac{\text{Throughput [bit/s/cell]}}{\text{Power consumption [W/cell]}} \quad (5.8)$$

which is measured in bit/Joule and can be seen as a benefit-cost ratio, where the service quality (throughput) is compared with the associated costs (power consumption). Hence, it is an indicator of the network’s bit-delivery efficiency.¹ The throughput can be computed using any of the UL and DL SE expressions provided in Section 4, which characterize the performance of Massive MIMO networks operating over large communication bandwidths (see also Remark 2.3 on p. 225).

Unlike the ATP, the EE metric is affected by changes in the numerator and denominator since both are variable. This means that some caution is required to avoid incomplete and potentially misleading conclusions from EE analysis. Particular attention should be paid to accurately model the PC of the network. Assume for example that the PC only comprises the transmit power. Lemma 5.1 showed that the transmit power can be reduced towards zero as $1/\sqrt{M}$ when $M \rightarrow \infty$ while approaching a non-zero asymptotic DL SE limit. This implies that the EE would grow without bound as $M \rightarrow \infty$. Clearly, this is misleading and comes from the fact that the transmit power only captures a part of the overall PC, as illustrated in Figure 5.2. Moreover, we notice that the transmit power does not represent the effective transmit power (ETP) needed for transmission since it does not account for the efficiency of the PA. The efficiency of a PA is defined as the ratio of input power to output power. When the efficiency is low, a large portion

¹The reciprocal of the bit/Joule metric, namely energy consumption per delivered information bit in Joule/bit, is referred as the power consumption ratio.

of the power is dissipated as heat (cf. Section 5.1). To correctly evaluate the EE, the PC must be computed on the basis of the ETP (not of the radiated transmit power) and of the CP required for running the cellular network:

$$\underbrace{\text{PC}}_{\text{Power consumption}} = \underbrace{\text{ETP}}_{\text{Effective transmit power}} + \underbrace{\text{CP}}_{\text{Circuit power}}. \quad (5.9)$$

A common model for CP is $\text{CP} = P_{\text{FIX}}$, where the term P_{FIX} is a constant quantity, which may account for the fixed power required for control signaling and load-independent power of baseband processors and backhaul infrastructure. However, this is not sufficiently accurate for comparing systems with different hardware setups (e.g., with a different number of antennas) and varying network loads² because it does not account for the power dissipation in the analog hardware and in the digital signal processing. Therefore, there are many ways in which an overly simplistic CP model may lead to wrong conclusions. Detailed CP models are needed to evaluate the power consumed by a practical network and to identify the non-negligible components. Clearly, the complexity of this task makes a certain level of idealization unavoidable. As we will show in Section 5.4, already a fairly simple polynomial CP model allows for a quite realistic assessment of the CP of Massive MIMO.

Remark 5.3 (Bandwidth shall not be normalized). A considerable number of papers on EE analysis have considered misleading EE metrics measured in bit/Joule/Hz, instead of bit/Joule. Such metrics are obtained by normalizing the bandwidth, but this is pointless since one cannot make the EE bandwidth-independent: the transmit power is divided over the bandwidth while the noise power is proportional to the bandwidth. An “EE” number measured in bit/Joule/Hz only applies to a system with exactly the bandwidth that was used to compute the noise power. In other words, the EE should be computed as the throughput divided by the consumed power (as defined in (5.8)), not as the SE divided

²Serving a larger number of UEs requires more CP due to the increased computational complexity of precoding/combining schemes, encoding and decoding as well as channel estimation.

by the consumed power. Some published papers have even normalized the bandwidth but forgotten to change the unit, leading to erroneous “EE” values that could be one million times smaller than in practice. As a rule-of-thumb, one should anticipate EE values at the order of kbit/Joule or Mbit/Joule.

Remark 5.4 (Alternative EE expressions). It is only reasonable to use the SE as performance metric when transmitting large data packets, for which the channel capacity can be approached. There exist a multitude of alternative figures of merit for the performance of cellular networks. Each of them accounts for specific targets and affects the EE differently [371]. For example, an alternative definition of the EE makes use of the goodput [270]; that is, the rate of successful delivery of finite-length data packets over a communication channel. The computation of the goodput, however, requires knowledge of the BER, which varies substantially between UEs and depends on many factors, such as modulation, encoding, and packet size. One way to overcome this issue is to approximate the BER as $1 - e^{-\text{SINR}}$ [216]. In slow-fading scenarios, outage events become the major channel impairment and the outage capacity becomes a suitable metric for measuring the service quality (cf. Remark 2.3 on p. 225).

5.3.1 Energy-Spectral Efficiency Tradeoff

Section 1.3 on p. 173 showed that the SE of a cell can be increased by using more transmit power, deploying multiple BS antennas, or serving multiple UEs per cell. All these approaches inevitably increase the PC of the network, either directly (by increasing the transmit power) or indirectly (by using more hardware), and therefore may potentially reduce the EE. However, this is not necessarily the case. In fact, there exist operating conditions under which it is possible to use these techniques to jointly increase SE and EE. To explore this in more detail, the EE-SE tradeoff is studied next and the impact of different network parameters and operating conditions are investigated. For simplicity, we focus on the UL of the two-cell Wyner model (i.e., $L = 2$) illustrated in Figure 1.8 (similar results can be obtained for the DL) and consider only uncorrelated Rayleigh fading channels over a

bandwidth B , under the assumption that the BSs are equipped with M antennas, have perfect channel knowledge, and use MR combining.

Impact of Multiple BS Antennas

Assume that there is only one active UE (i.e., $K = 1$) in cell 0 and that no interfering signals come from cell 1. Then, from Lemma 1.7 on p. 196, an achievable SE of the UE in cell 0 is

$$\text{SE}_0 = \log_2 (1 + (M - 1)\text{SNR}_0) = \log_2 \left(1 + (M - 1) \frac{p}{\sigma^2} \beta_0^0 \right) \quad (5.10)$$

where p is the transmit power, σ^2 is the noise power, and β_0^0 denotes the average channel gain of the active UE. We have omitted the superscript “NLoS”, since we do not consider the LoS case here. To evaluate the impact of M on the EE, we distinguish between two different cases in the computation of the PC: *i*) the CP increase due to multiple BS antennas is neglected; *ii*) the CP increase is accounted for.

Assume, for the moment, that the CP of cell 0 consists only of the fixed power P_{FIX} ; that is, $\text{CP}_0 = P_{\text{FIX}}$. Hence, the corresponding EE of cell 0 is

$$\text{EE}_0 = \frac{B \log_2 \left(1 + (M - 1) \frac{p}{\sigma^2} \beta_0^0 \right)}{\frac{1}{\mu} p + P_{\text{FIX}}} \quad (5.11)$$

where B is the bandwidth and $\frac{1}{\mu} p$ accounts for the ETP with $0 < \mu \leq 1$ being the PA efficiency. For a given SE, denoted as SE_0 , from (5.10) we obtain the required transmit power as³

$$p = \frac{(2^{\text{SE}_0} - 1) \sigma^2}{(M - 1) \beta_0^0} \quad (5.12)$$

which inserted into (5.11) yields

$$\text{EE}_0 = \frac{B \text{SE}_0}{(2^{\text{SE}_0} - 1) \frac{\nu_0}{M-1} + P_{\text{FIX}}} \quad (5.13)$$

³In such an interference-free scenario, p is an exponentially increasing function of SE_0 , meaning that increasing SE_0 is equivalent to increasing the transmit power p .

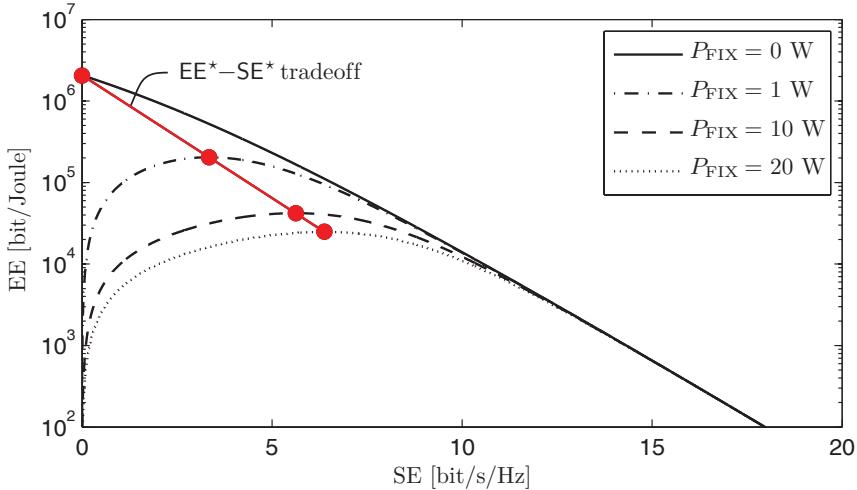


Figure 5.4: SE and EE relation in (5.16) for different values of CP = P_{FIX} when $M = 10$, $B = 100$ kHz, $\sigma^2/\beta_0^0 = -6$ dBm, and $\mu = 0.4$. The red dots represent the points at each curve in which EE_0 achieves its maximum.

with

$$\nu_0 = \frac{\sigma^2}{\mu\beta_0^0}. \quad (5.14)$$

The above expression provides the relation between EE and SE for the UE in cell 0.

Figure 5.4 illustrates the EE versus SE for $M = 10$, $B = 100$ kHz, $\sigma^2/\beta_0^0 = -6$ dBm, $\mu = 0.4$, and $P_{\text{FIX}} \in \{0, 1, 10, 20\}$ W. As we can see, if $P_{\text{FIX}} = 0$, there is a monotonic decreasing tradeoff between EE and SE (as predicted by Shannon theory [326]) because (5.13) reduces to

$$\text{EE}_0 = \frac{B \text{SE}_0}{(2^{\text{SE}_0} - 1) \frac{\nu_0}{M-1}}. \quad (5.15)$$

In other words, if the CP is not accounted for, an increased SE always comes at the price of a decreased EE. If, however, $P_{\text{FIX}} > 0$ (as it is in practice), then EE_0 is a unimodal⁴ function that increases for values of SE_0 such that $(2^{\text{SE}_0} - 1) \frac{\nu_0}{M-1} < P_{\text{FIX}}$ and decreases to zero as $\frac{\text{SE}_0}{2^{\text{SE}_0} - 1} \rightarrow P_{\text{FIX}}$.

⁴A function $f(x)$ is unimodal if, for some value m , it is monotonically increasing for $x \leq m$ and monotonically decreasing for $x > m$.

when SE_0 takes larger values. We can also see from Figure 5.4 that the EE-SE curve becomes flatter with increasing values of P_{FIX} , such that the range of SE values for which almost the same EE is achieved gets larger.

To get some analytical insights into the EE-optimal point, we take the derivative of EE_0 in (5.13) with respect to SE_0 and equate it to zero. We observe that the maximum EE (called EE^*) and its corresponding SE (called SE^*) satisfy the following identity:

$$\log_2(\text{EE}^*) + \text{SE}^* = \log_2\left((M-1)\frac{B}{\nu_0 \log_e(2)}\right) \quad (5.16)$$

where SE^* is such that

$$\text{SE}^* \left(2^{\text{SE}^*} \log_e(2)\right) = \left(2^{\text{SE}^*} - 1\right) + \frac{M-1}{\nu_0} P_{\text{FIX}}. \quad (5.17)$$

The identity (5.16) shows a linear dependence between $\log_2(\text{EE}^*)$ and SE^* . This dependence is illustrated by the red tradeoff line in Figure 5.4. This means that an exponential EE gain may be obtained at the cost of a linear SE loss. Observe that (5.17) has a unique solution, which takes the form (see Appendix C.4.2 on p. 610)

$$\text{SE}^* = \frac{W\left((M-1)\frac{P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e}\right) + 1}{\log_e(2)} \quad (5.18)$$

where $W(\cdot)$ is the Lambert function (defined in Appendix B.3 on p. 567) and e is Euler's number. Inserting (5.18) into (5.16) yields

$$\text{EE}^* = \frac{(M-1)Be^{-W\left((M-1)\frac{P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e}\right)-1}}{\nu_0 \log_e(2)} \quad (5.19)$$

where we have used the fact that $2^{-1/\log_e(2)} = e^{-1}$. Equations (5.18) and (5.19) provide SE^* and EE^* in closed form and thus allow us to get insights into how both are affected by the system parameters. From (5.18), taking into account that $W(x)$ is an increasing function for $x \geq e$, it turns out that SE^* increases with M (as intuitively expected), but also with P_{FIX} , as shown in Figure 5.4. This can be explained as follows: the higher P_{FIX} , the higher SE can be afforded before the transmit power

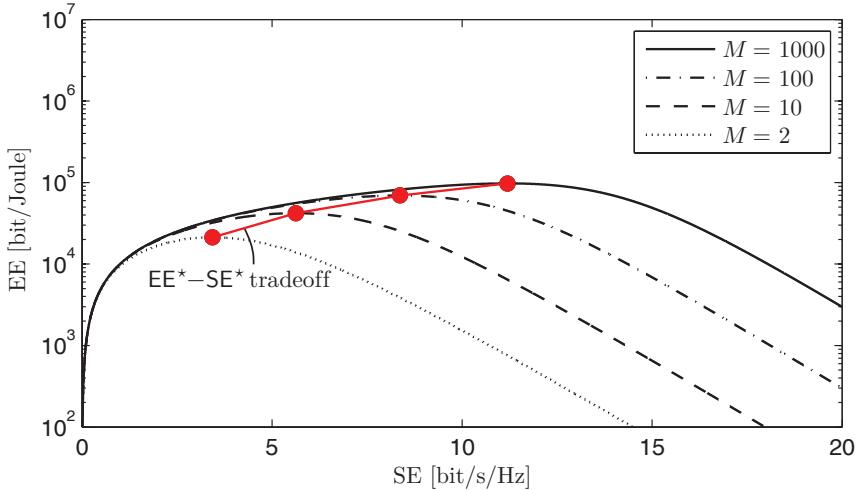


Figure 5.5: EE and SE relation in (5.16) for different values of M when $P_{\text{FIX}} = 10 \text{ W}$, $B = 100 \text{ kHz}$, $\sigma^2/\beta_0^0 = -6 \text{ dBm}$, and $\mu = 0.4$. Both SE^* and EE^* increase as M grows large. This erroneously happens if the CP does not account for the additional power consumed by having multiple antennas. The red dots represent the points at which the EE achieves its maximum.

$(2^{\text{SE}_0} - 1) \frac{\nu_0}{M-1}$ in (5.13) becomes a limiting factor for the EE. On the other hand, EE^* in (5.19) decreases with P_{FIX} (as shown in Figure 5.4) and increases without bound with the number of antennas M . The impact of M is illustrated in Figure 5.5 for $P_{\text{FIX}} = 10 \text{ W}$, $B = 100 \text{ kHz}$, $\sigma^2/\beta_0^0 = -6 \text{ dBm}$, and $\mu = 0.4$. As anticipated by the analysis, both EE and SE increase with M .

The following corollary provides further insights into the scaling behavior of SE^* and EE^* with respect to M and P_{FIX} .

Corollary 5.2 (Scaling law with M and/or P_{FIX}). If M or P_{FIX} grow large, then

$$\text{SE}^* \approx \log_2(MP_{\text{FIX}}) \quad (5.20)$$

and

$$\text{EE}^* \approx \frac{eB}{(1+e)} \frac{\log_2(MP_{\text{FIX}})}{P_{\text{FIX}}}. \quad (5.21)$$

Proof. The proof is given in Appendix C.4.3 on p. 610. \square

Corollary 5.2 shows that SE^* increases logarithmically with M and P_{FIX} . On the other hand, EE^* grows logarithmically with M and is an almost linearly decreasing function of P_{FIX} .⁵ Therefore, it seems that an unbounded EE^* can be achieved by adding more and more antennas. This result is due to the simplified model $\text{CP}_0 = P_{\text{FIX}}$, which ignores the fact that the CP increases with M in practice. In other words, there is a *cost-performance tradeoff* in practical systems. This tradeoff is particularly important when implementing a multiantenna system because a BS equipped with M antennas needs M RF chains, each containing many components; for example, PAs, analog-to-digital converters (ADCs), digital-to-analog converters (DACs), local oscillators (LOs), filters, in-phase/quadrature (I/Q) mixers, and OFDM modulation/demodulation. The CP of such an implementation will be, roughly, M times higher than the CP of a single-antenna transceiver. In what follows, we thus consider the CP model

$$\text{CP}_0 = P_{\text{FIX}} + M P_{\text{BS}} \quad (5.22)$$

where P_{BS} is the power consumed by the circuit components (e.g., ADCs, DACs, I/Q mixers, LOs, filters, and OFDM modulation/demodulation) needed for the operation of each BS antenna. Then, (5.13) becomes

$$\text{EE}_0 = B \frac{\text{SE}_0}{(2^{\text{SE}_0} - 1) \frac{\nu_0}{M-1} + P_{\text{FIX}} + M P_{\text{BS}}}. \quad (5.23)$$

Figure 5.6 shows the EE versus SE in the same operating conditions as in Figure 5.5, except that now $\text{CP}_0 = P_{\text{FIX}} + M P_{\text{BS}}$ with $P_{\text{BS}} = 1$ W. The EE^* - SE^* tradeoff curve is now a unimodal function of M : monotonically increasing for $M \leq 10$ and monotonically decreasing for $M > 10$. The maximum value is obtained for $M = 10$. This is in sharp contrast to the results of Figure 5.5, where the EE^* - SE^* tradeoff curve is always increasing with M . This demonstrates that an accurate modeling of the CP is of paramount importance when dealing with the design of energy-efficient multiantenna systems.

⁵Note that $\log_e(x)/x \approx \log_e(A)/x$ for $x \geq A$ with A being some large constant.

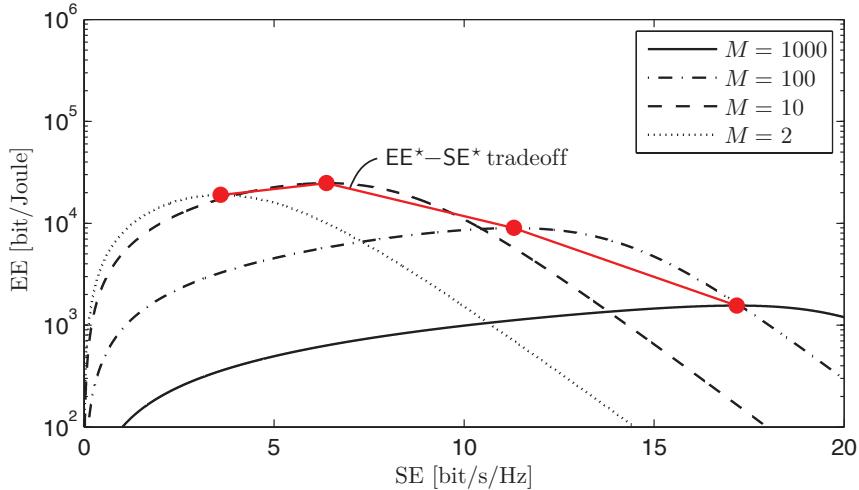


Figure 5.6: SE and EE relation in (5.23) for different values of M when $P_{\text{FIX}} = 10 \text{ W}$, $P_{\text{BS}} = 1 \text{ W}$, $B = 100 \text{ kHz}$, $\sigma^2/\beta_0^0 = -6 \text{ dBm}$, and $\mu = 0.4$. In contrast to Figure 5.5, the EE^*-SE^* tradeoff (red curve) does not grow unboundedly with the number of antennas. This is because each additional antenna increases the CP by P_{BS} , as it does in practice. The red dots on each curve represent the points in which EE achieves its maximum.

Corollary 5.3 (Scaling law with M , P_{FIX} and/or P_{BS}). If M , P_{FIX} , and/or P_{BS} grow large, then

$$\text{SE}^* \approx \log_2(M(P_{\text{FIX}} + MP_{\text{BS}})) \quad (5.24)$$

and

$$\text{EE}^* \approx \frac{eB}{(1+e)} \frac{\log_2(M(P_{\text{FIX}} + MP_{\text{BS}}))}{(P_{\text{FIX}} + MP_{\text{BS}})}. \quad (5.25)$$

Proof. The proof is given in Appendix C.4.4 on p. 611. \square

From (5.24), we can see that SE^* scales logarithmically with M^2 (rather than with M as in (5.20)) since higher SE can be afforded before the transmit power has a detrimental effect on the EE given by (5.23). In contrast to (5.21), EE^* in (5.25) is an almost linearly decreasing function of MP_{BS} . In summary, increasing the number of antennas M monotonically improves SE^* , which even grows without bound as $M \rightarrow \infty$, but the positive effect on EE^* vanishes quickly as increasing M requires more hardware and, thus, higher CP.

Impact of Multiple UEs

As shown in Sections 1.3.3 and 1.3.4 on p. 193 and p. 204, respectively, increasing the number of simultaneously active UEs by SDMA transmission is the most efficient way to improve the per-cell SE. Next, we investigate the potential benefits that SDMA can bring to the EE by considering the two-cell Wyner model in Figure 1.8 with K single-antenna UEs in each cell and the relative strength $\bar{\beta} = \beta_1^0/\beta_0^0 = \beta_0^1/\beta_1^1$ of the inter-cell interference. Then, if MR combining is used with perfect channel knowledge at the BS, a UL SE of each UE is

$$\text{SE}_0 = \log_2 \left(1 + \frac{M - 1}{(K - 1) + K\bar{\beta} + \frac{\sigma^2}{p\beta_0^0}} \right) \quad (5.26)$$

by using Lemma 1.7 on p. 196. A given SE_0 value is thus achieved by

$$p = \left(\frac{M - 1}{2^{\text{SE}_0} - 1} - K\bar{\beta} + 1 - K \right)^{-1} \frac{\sigma^2}{\beta_0^0}. \quad (5.27)$$

The corresponding EE of cell 0 is

$$\text{EE}_0 = \frac{BK\text{SE}_0}{K \left(\frac{M-1}{2^{\text{SE}_0}-1} - K\bar{\beta} + 1 - K \right)^{-1} \nu_0 + \text{CP}_0} \quad (5.28)$$

where ν_0 was defined in (5.14) and we have taken into account that the sum SE in cell 0 is $K\text{SE}_0$ and that the total transmit power is $\frac{1}{\mu}Kp$. To account for the additional CP consumed by all the active UEs, we assume that

$$\text{CP}_0 = P_{\text{FIX}} + MP_{\text{BS}} + KP_{\text{UE}} \quad (5.29)$$

where P_{UE} accounts for the power required by all circuit components (e.g., DAC, I/Q mixer, filter, and so forth) of each single-antenna UE.

Taking the derivative of EE_0 in (5.28) with respect to SE_0 and equating to zero yields the expression

$$\begin{aligned} & K \left(\frac{M - 1}{2^{\text{SE}^*} - 1} - K\bar{\beta} + 1 - K \right)^{-1} \nu_0 + P_{\text{FIX}} + MP_{\text{BS}} + KP_{\text{UE}} = \\ & = K\text{SE}^* \left(1 - \left(\frac{2^{\text{SE}^*} - 1}{M - 1} \right) (K\bar{\beta} - 1 + K) \right)^{-2} \frac{\nu_0 \log_e(2)}{M - 1} 2^{\text{SE}^*} \end{aligned} \quad (5.30)$$

from which the SE^* that maximizes the EE is obtained. Inserting this expression into (5.28) yields

$$\text{EE}^* = \frac{B}{\left(1 - \left(\frac{2^{\text{SE}^*}-1}{M-1}\right)(K\bar{\beta} - 1 + K)\right)^{-2} \frac{\nu_0 \log_e(2)}{M-1} 2^{\text{SE}^*}} \quad (5.31)$$

or, equivalently,

$$\begin{aligned} \log_2(\text{EE}^*) + \text{SE}^* - 2 \log_2 \left(1 - \left(\frac{2^{\text{SE}^*}-1}{M-1}\right)(K\bar{\beta} - 1 + K)\right) \\ = \log_2 \left((M-1) \frac{B}{\nu_0 \log_e(2)}\right). \end{aligned} \quad (5.32)$$

The expression in (5.32) has a similar form as (5.16), except for the extra terms due to intra-cell and inter-cell interference. Unlike (5.17), the solution to (5.30) cannot be provided in closed form, due to the presence of interference. In what follows, we evaluate numerically how the relative strength $\bar{\beta}$ of the inter-cell interference and the number of UEs K impact the EE-SE tradeoff. Figure 5.7 shows the EE of cell 0 as a function of the sum SE with $K \in \{5, 10, 30\}$ and $\bar{\beta} = -15 \text{ dB}$ or -3 dB . Additionally, we assume that $M = 10$, $B = 100 \text{ kHz}$, $\sigma^2/\beta_0^0 = -6 \text{ dBm}$, $\mu = 0.4$, $P_{\text{FIX}} = 10 \text{ W}$, $P_{\text{BS}} = 1 \text{ W}$, and $P_{\text{UE}} = 0.5 \text{ W}$. Increasing $\bar{\beta}$ has a detrimental effect on both EE and SE, since the inter-cell interference term $K\bar{\beta}$ in (5.26) increases linearly with $\bar{\beta}$. On the other hand, the EE^* - SE^* tradeoff curve is a unimodal function of K (as it is for M , cf. Figure 5.6). For the considered setup, the maximum value is obtained for $K = 10$. This is because the sum SE is a slowly increasing function of K in the case of $M = 10$ (cf. Figure 1.16) while each additional UEs increases the PC by $P_{\text{UE}} = 0.5 \text{ W}$. Therefore, the degradation in EE for a given sum SE increases as K or $\bar{\beta}$ grow large.

The previous figure seems to indicate that SDMA cannot improve the EE due to the increased interference and additional hardware. However, when studying the impact of K on the sum SE only (cf. Figure 1.17), we observed that multiple UEs can be served simultaneously without decreasing the SE per UE if a proportional number of antennas is added to counteract the increased interference. This leads to an operating regime with the antenna-UE ratio $M/K \geq c$, for some preferably

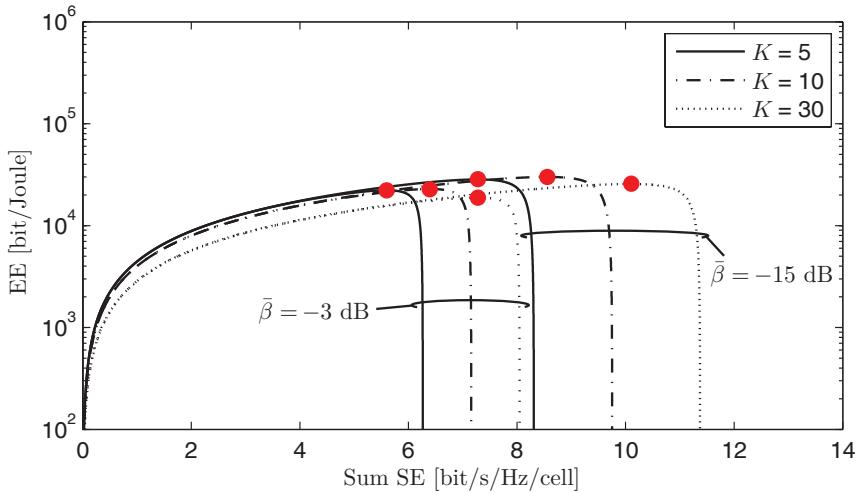


Figure 5.7: Sum SE and EE relation in (5.32) for different values of the inter-cell interference $\bar{\beta}$ and the number of UEs K when $M = 10$, $P_{\text{FIX}} = 10$ W, $P_{\text{BS}} = 1$ W and $P_{\text{UE}} = 0.5$ W, $B = 100$ kHz, $\sigma^2/\beta_0^0 = -6$ dBm, and $\mu = 0.4$. Increasing the strength $\bar{\beta}$ of the inter-cell interference has a detrimental effect on both EE and SE. As observed for M , the EE*-SE* tradeoff does not grow unboundedly with K since each additional UE increases the CP by P_{UE} .

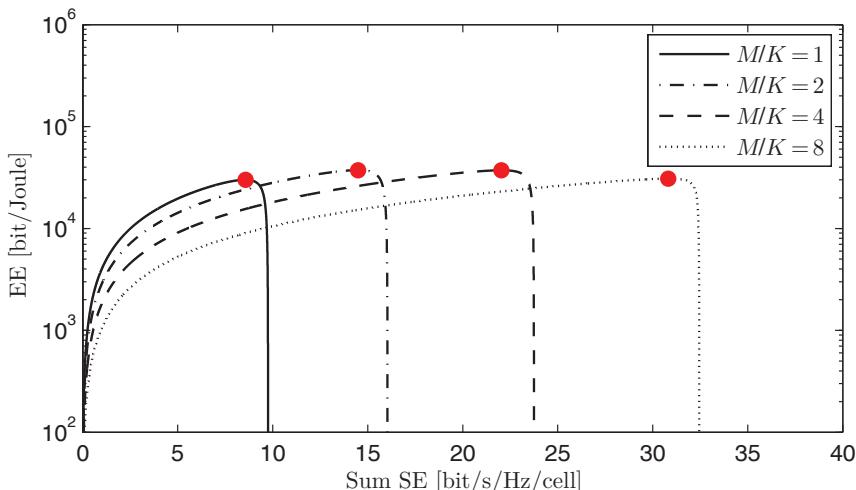


Figure 5.8: Sum SE and EE relation in (5.32) for different values of the antenna-UE ratio M/K when $K = 10$, $\bar{\beta} = -10$ dB, $P_{\text{FIX}} = 10$ W, $P_{\text{BS}} = 1$ W, $P_{\text{UE}} = 0.1$ W, $B = 100$ kHz, $\sigma^2/\beta_0^0 = -6$ dBm, and $\mu = 0.4$.

large constant c , where we can provide K -fold gains in sum SE by SDMA. An identical result cannot be achieved for the EE since adding more antennas not only increases the SE but also the PC through $M P_{\text{BS}}$. Intuitively, this means that there will exist an optimal pair of values (M, K) such that the EE attains its maximum. To exemplify this conclusion, Figure 5.8 shows the EE of cell 0 for different values of the antenna-UE ratio M/K when $K = 10$. Unlike the sum SE in Figure 1.17, which grows monotonically with the antenna-UE ratio M/K , EE^* is a unimodal function of M/K . For the considered setup, it increases up to $M/K = 2$ and then slowly decreases as M/K grows larger. In summary, serving multiple UEs while simultaneously increasing the number of BS antennas (to compensate for the higher interference) may improve the EE of the network only when the benefits and costs of deploying more RF hardware are properly balanced. The EE-optimum configuration of BS antennas and number of UEs will be evaluated in Section 5.6.

5.4 Circuit Power Consumption Model

In the previous section, we have used the simple two-cell Wyner network model to show that a PC model accounting for the transmit power as well as for the CP consumed by the transceiver hardware at the BS and UEs is necessary to avoid misleading conclusions about the EE. These are not the only contributions that must be taken into account to appropriately evaluate the CP of the UL and DL of Massive MIMO. We will show that we must also consider the power consumed by digital signal processing, backhaul signaling, encoding, and decoding [26]. Building on [312, 26, 358, 95, 219, 185], a CP model for a generic BS j in a Massive MIMO network is:

$$\begin{aligned} \text{CP}_j = & \underbrace{P_{\text{FIX},j}}_{\text{Fixed power}} + \underbrace{P_{\text{TC},j}}_{\text{Transceiver chains}} + \underbrace{P_{\text{CE},j}}_{\text{Channel estimation}} + \underbrace{P_{\text{C/D},j}}_{\text{Coding/Decoding}} \\ & + \underbrace{P_{\text{BH},j}}_{\text{Load-dependent backhaul}} + \underbrace{P_{\text{SP},j}}_{\text{Signal processing}} \end{aligned} \quad (5.33)$$

where $P_{\text{FIX},j}$ was defined before as a constant quantity accounting for the fixed power required for control signaling and load-independent

power of backhaul infrastructure and baseband processors. Furthermore, $P_{\text{TC},j}$ accounts for the power consumed by the transceiver chains, $P_{\text{CE},j}$ for the channel estimation process (performed once per coherence block), $P_{\text{C/D},j}$ for the channel encoding and decoding units, $P_{\text{BH},j}$ for the load-dependent backhaul signaling, and $P_{\text{SP},j}$ for the signal processing at the BS. Note that neglecting the power consumed by transceiver chains, channel estimation, precoding, and combining was previously the norm in multiuser MIMO. More precisely, the small numbers of antennas and UEs, before Massive MIMO was introduced, were such that the CP for all those operations was negligible compared to the fixed power. The CP associated with those operations was modeled for single-cell systems in [358, 58, 59], while multicell systems were considered in [60]. Inspired by these works, we provide in what follows a tractable and realistic model for each term in (5.33), as a function of the main system parameters M_j and K_j . This is achieved by characterizing the hardware setup using a variety of fixed hardware coefficients, which are kept generic in the analysis; typical values will be given later and strongly depend on the actual hardware equipment and the state-of-the-art in circuit implementation.

Remark 5.5 (Economical efficiency). In this monograph, we are mainly concerned with the PC and not with the economical expenses such as deployment cost, site renting, and so forth. However, we stress that economical expenses can be added into the CP model developed below; for example, by dividing the cost rate of the network (measured in $\$/s$) with the energy price (measured in Joule/ $\$$) to get a number in Watt that is an equivalent PC. The main economical expenses are likely to be proportional to the number of BSs and would thus increase the load-independent term $P_{\text{FIX},j}$ in (5.33).

5.4.1 Transceiver Chains

As described in [95] and [185], $P_{\text{TC},j}$ of cell j can be quantified as

$$P_{\text{TC},j} = \underbrace{M_j P_{\text{BS},j} + P_{\text{LO},j}}_{\text{BS circuit components}} + \underbrace{K_j P_{\text{UE},j}}_{\text{UE circuit components}} \quad (5.34)$$

where $P_{\text{BS},j}$ is the power required to run the circuit components (such as ADCs, DACs, I/Q mixers, filters, and OFDM modulation/demodulation) attached to each antenna at BS j (which has to be multiplied by the number of antennas M_j) and $P_{\text{LO},j}$ is the power consumed by the LO.⁶ The term $P_{\text{UE},j}$ accounts for the power required by all circuit components (such as ADCs, DACs, I/Q mixer, LO, filters, and OFDM modulation/demodulation) of each single-antenna UE.

5.4.2 Coding and Decoding

In the DL, BS j applies channel coding and modulation to K_j sequences of information symbols and each UE applies some practical fixed-complexity algorithm for decoding its own received data sequence. The opposite is done in the UL. The term $P_{\text{C/D},j}$ accounting for these processes in cell j is thus proportional to the number of information bits that is transferred [219] and can be quantified as

$$P_{\text{C/D},j} = (P_{\text{COD}} + P_{\text{DEC}})\text{TR}_j \quad (5.35)$$

where TR_j stands for the throughput (in bit/s) of cell j , while P_{COD} and P_{DEC} are the encoding and decoding powers (in W per bit/s), respectively. For simplicity, we assume that P_{COD} and P_{DEC} are the same in the UL and DL for all UEs in the network, but it is straightforward to assign them different values. Note also that P_{COD} and P_{DEC} highly depend on the employed channel coding technique; for example, in [219, 181], the authors consider a low-density parity-check code and express P_{COD} and P_{DEC} as functions of code parameters. The throughput TR_j accounts for the UL and DL of all UEs in cell j and can be obtained using the SE expressions provided in Section 4 (see (5.43) for an example).

5.4.3 Backhaul

The backhaul is used to transfer UL and DL data between the BS and the core network, and can be either wired or wireless depending on the

⁶In general, a single LO is used for frequency synthesis for all BS antennas. This is why this term is independent of M_j . If multiple LOs are used (e.g., for BSs with distributed antenna arrays), we can set $P_{\text{LO},j} = 0$ and include the power consumed by LOs in $P_{\text{BS},j}$ instead.

network deployment. The power consumed by the backhaul is commonly modeled as the sum of two parts [312]: one load-independent and one load-dependent. The first part was included in $P_{\text{FIX},j}$ and it is typically the most significant part of the backhaul consumed power (around 80%), while the load-dependent part of each BS j is proportional to the sum throughput of its served UEs. Looking jointly at the UL and DL, the load-dependent backhaul term $P_{\text{BH},j}$ in cell j is computed as

$$P_{\text{BH},j} = P_{\text{BT}} \cdot \text{TR}_j \quad (5.36)$$

where P_{BT} is the backhaul traffic power (in W per bit/s), which is, for simplicity, assumed to be the same for all cells in the network.

5.4.4 Channel Estimation

As discussed in Section 3.1 on p. 244, UL channel estimation plays a major role in Massive MIMO to make efficient use of a large number of antennas. All processing for UL channel estimation is carried out locally at the BS in each coherence block and has a computational cost, which translates into a consumed power. The UL channel estimation is carried out using the MMSE estimator developed in Section 3.2 on p. 248 or alternative techniques, namely the EW-MMSE and LS estimators, defined in Section 3.4.1 on p. 265. The computational complexities of the above estimators are summarized in Table 3.1 on p. 270, in terms of number of complex multiplications per UE. To transform these figures into consumed power, let L_{BS} be the computational efficiency of the BS measured in [flops/W]⁷ and recall that a complex multiplication requires three real floating-point multiplications.⁸ Since there are B/τ_c coherence blocks per second (see Definition 2.2 on p. 219), from Table 3.1 the power consumed by the considered estimators is

$$P_{\text{CE},j} = \frac{3B}{\tau_c L_{\text{BS}}} K_j \cdot \begin{cases} M_j \tau_p + M_j^2 & \text{with MMSE} \\ M_j \tau_p + M_j & \text{with EW-MMSE} \\ M_j \tau_p & \text{with LS} \end{cases} \quad (5.37)$$

⁷ Literally, it measures the number of operations per second that can be delivered per Watt of power consumed.

⁸ Let $x = a + jb$ and $y = c + jd$, then $xy = (ac - bd) + j((a + b)(c + d) - ac - bd)$ whose computation requires three real multiplications: ac , bd and $(a + b)(c + d)$.

where K_j is the number of UEs in cell j and τ_p is the pilot sequence length, typically chosen such that $\tau_p \geq \max_l K_l$. In Massive MIMO, we have τ_p in the order of tens, thus EW-MMSE and LS estimation have approximately the same power consumption. Observe that we have neglected the complexity of precomputing statistical matrices since these computations must only be redone when the channel statistics change. Notice also that (5.37) quantifies only the power consumed for estimating the intra-cell channels, which is sufficient except when using M-MMSE. The extra cost of estimating inter-cell channels will be quantified in Section 5.4.5.

From (5.37), we notice that a power model that is linear in M_j (as often assumed in other literature) is only valid for the EW-MMSE and LS estimators. The MMSE estimator consumes power proportionally to M_j^2 . This is the price to pay for the improved channel estimation accuracy (cf. Figure 3.7) and it cannot be neglected for a fair comparison of different estimation schemes in terms of EE.⁹ The functional dependence on the number of UEs is not only due to K_j but also to τ_p , which scales linearly with $\max_l K_l$ (or, in other words, with the maximum UE load). It follows that the power required by channel acquisition at BS j increases proportionally to $K_j \max_l K_l$. Hence, also a consumed power model that it is linear in K_j cannot be considered accurate enough, especially in Massive MIMO where K_j is large.

Note that we have neglected the complexity of DL channel estimation since its complexity is substantially lower than that of UL channel estimation; each UE only needs to estimate the precoded scalar channel from the received data signals (cf. Section 4.3.3 on p. 325).

5.4.5 Receive Combining and Transmit Precoding

We can use the computational complexity analysis from Sections 4.1.2 and 4.3.2 on p. 284 and p. 320, respectively, to compute the power $P_{SP,j}$ consumed by BS j for receive combining and transmit precoding. This

⁹Note that, in the special case when all spatial correlation matrices are diagonal (so that it is optimal to estimate each channel element separately), the computational complexity of MMSE estimation becomes the same as that of EW-MMSE (see Section 3.4.1 on p. 265 for further details).

can be quantified as

$$P_{\text{SP},j} = \underbrace{P_{\text{SP-R/T},j}}_{\text{Reception/transmission}} + \underbrace{P_{\text{SP-C},j}^{\text{UL}}}_{\text{Computing combining}} + \underbrace{P_{\text{SP-C},j}^{\text{DL}}}_{\text{Computing precoding}} \quad (5.38)$$

where $P_{\text{SP-R/T},j}$ accounts for the total power consumed by UL reception and DL transmission of data signals (for given combining and precoding vectors) whereas $P_{\text{SP-C},j}^{\text{UL}}$ and $P_{\text{SP-C},j}^{\text{DL}}$ are the powers required for the computation of the combining and precoding vectors at BS j , respectively.

UL Reception and DL Transmission

In the UL with \mathbf{v}_{jk} given, the complexity for computing $\mathbf{v}_{jk}^H \mathbf{y}_j$, for the τ_u received UL signals \mathbf{y}_j and every UE in the cell, is $\tau_u M_j K_j$ complex multiplications per coherence block. In the DL with \mathbf{w}_{jk} given, the computation of $\mathbf{x}_j = \sum_{k=1}^{K_j} \mathbf{w}_{jk} s_{jk}$ requires a total of $\tau_d M_j K_j$ complex multiplications per coherence block. Therefore, we obtain

$$P_{\text{SP-R/T},j} = \frac{3B}{\tau_c L_{\text{BS}}} M_j K_j (\tau_u + \tau_d). \quad (5.39)$$

Note that the CP for reception and transmission is the same irrespective of the choice of combining and precoding schemes.

Computation of the Combining/Precoding Vectors

Thanks to the UL-DL duality (see Section 4.3.2 on p. 320), a natural choice of the precoding vectors is $\mathbf{w}_{jk} = \mathbf{v}_{jk} / \|\mathbf{v}_{jk}\|$ (except when the UL and DL are designed very differently). If \mathbf{v}_{jk} is given, then the complexity for computing \mathbf{w}_{jk} reduces to first evaluate $\|\mathbf{v}_{jk}\|$ and then $\mathbf{v}_{jk} / \|\mathbf{v}_{jk}\|$. This costs

$$P_{\text{SP-C},j}^{\text{DL}} = \frac{4B}{\tau_c L_{\text{BS}}} M_j K_j. \quad (5.40)$$

As shown in Table 4.1 on p. 287, the complexity of computing \mathbf{v}_{jk} largely depends on the receive combining scheme. If MR is used, \mathbf{v}_{jk} is obtained directly from the channel estimates and there is no extra

complexity, except for the normalization required by the decoding unit. This cost is K_j divisions per BS, which in total gives

$$P_{\text{SP-C},j}^{\text{UL}} = \frac{7B}{\tau_c L_{\text{BS}}} K_j \quad (5.41)$$

where we have taken into account that a complex division requires seven real multiplication/division operations.¹⁰ Similarly, the consumed power with RZF is

$$P_{\text{SP-C},j}^{\text{UL}} = \frac{3B}{\tau_c L_{\text{BS}}} \left(\frac{3K_j^2 M_j}{2} + \frac{3K_j M_j}{2} + \frac{K_j^3 - K_j}{3} + \frac{7}{3} K_j \right). \quad (5.42)$$

where the last term accounts for divisions and the other terms for multiplications. Table 5.2 provides the power consumed by all the combining and precoding schemes considered in this monograph. Note that in case of M-MMSE combining we have also included the cost for estimating the inter-cell channels and for correlating the received pilot signal with the $\tau_p - K_j$ pilot sequences that are only used in other cells.

5.4.6 Comparison of CP with Different Processing Schemes

We will compare the CP consumed with different combining/precoding schemes by continuing the running example that was defined in Section 4.1.3 on p. 288. There are M antennas at each BS and K UEs in each cell. The values of M and K will be changed and specified in each figure. The pilot reuse factor is $f = 1$, such that each pilot sequence consists of $\tau_p = K$ samples. The number of samples per coherence block that is used for data is $\tau_c - \tau_p = 190 - K$, whereof 1/3 is used for UL and 2/3 for DL. This yields $\tau_u = \frac{1}{3}(\tau_c - \tau_p)$ and $\tau_d = \frac{2}{3}(\tau_c - \tau_p)$. We consider UL and DL transmit powers of 20 dBm per UE (i.e., $p_{jk} = \rho_{jk} = 100$ mW). The Gaussian local scattering with ASD $\sigma_\varphi = 10^\circ$ is used as channel model. The throughput of cell j for computing the consumed power for backhaul, encoding, and decoding is obtained using the UL and DL SE expressions of Section 4. For each scheme and number of antennas, we use the capacity bound of Theorem 4.1 on p. 276 for the UL and the

¹⁰Let $x = a + jb$ and $y = c + jd$, then $x/y = xy^*/|y|^2$. The computation of xy^* requires three real operations whereas two real multiplications are needed for $|y|^2 = c^2 + d^2$ and two real divisions are required for computing the ratio.

Scheme	$P_{\text{SP-R}/\Gamma,j}$	$P_{\text{SP-C},j}^{\text{UL}}$	$P_{\text{SP-C},j}^{\text{DL}}$
M-MMSE (MMSE estimation)	$\frac{3B}{\tau_c L_{\text{BS}}} K_j M_j (\tau_u + \tau_d)$	$\frac{3B}{\tau_c L_{\text{BS}}} \left(\sum_{l=1}^L \frac{(3M_j^2 + M_j)K_l}{2} + \frac{M_j^3}{3} \right. \\ \left. + 2M_j + M_j \tau_p (\tau_p - K_j) \right)$	$\frac{3B}{\tau_c L_{\text{BS}}} M_j K_j$
M-MMSE (EW-MMSE estimation)	$\frac{3B}{\tau_c L_{\text{BS}}} K_j M_j (\tau_u + \tau_d)$	$\frac{3B}{\tau_c L_{\text{BS}}} \left(\sum_{l=1}^L \frac{(M_j^2 + 3M_j)K_l}{2} + (M_j^2 - M_j)K_j + \frac{M_j^3}{3} \right. \\ \left. + 2M_j + M_j \tau_p (\tau_p - K_j) \right)$	$\frac{3B}{\tau_c L_{\text{BS}}} M_j K_j$
S-MMSE	$\frac{3B}{\tau_c L_{\text{BS}}} K_j M_j (\tau_u + \tau_d)$	$\frac{3B}{\tau_c L_{\text{BS}}} \left(\frac{3M_j^2 K_j}{2} + \frac{M_j K_j}{2} + \frac{M_j^3 - M_j}{3} + \frac{7}{3} M_j \right)$	$\frac{3B}{\tau_c L_{\text{BS}}} M_j K_j$
RZF	$\frac{3B}{\tau_c L_{\text{BS}}} K_j M_j (\tau_u + \tau_d)$	$\frac{3B}{\tau_c L_{\text{BS}}} \left(\frac{3K_j^2 M_j}{2} + \frac{3K_j M_j}{2} + \frac{K_j^3 - K_j}{3} + \frac{7}{3} K_j \right)$	$\frac{3B}{\tau_c L_{\text{BS}}} M_j K_j$
ZF	$\frac{3B}{\tau_c L_{\text{BS}}} K_j M_j (\tau_u + \tau_d)$	$\frac{3B}{\tau_c L_{\text{BS}}} \left(\frac{3K_j^2 M_j}{2} + \frac{K_j M_j}{2} + \frac{K_j^3 - K_j}{3} + \frac{7}{3} K_j \right)$	$\frac{3B}{\tau_c L_{\text{BS}}} M_j K_j$
MR	$\frac{3B}{\tau_c L_{\text{BS}}} K_j M_j (\tau_u + \tau_d)$	$\frac{7B}{\tau_c L_{\text{BS}}} K_j$	$\frac{3B}{\tau_c L_{\text{BS}}} M_j K_j$

Table 5.2: Power required by each BS for the signal processing with different combining/precoding schemes over a bandwidth B , under the assumption that the precoding vectors are chosen as normalized versions of the receive combining vectors. The results are obtained on the basis of the computational complexity analysis performed in Section 4.1.1 on p. 281 with complex multiplications and divisions requiring three and seven real operations, respectively.

Parameter	Value set 1	Value set 2
Fixed power: P_{FIX}	10 W	5 W
Power for BS LO: P_{LO}	0.2 W	0.1 W
Power per BS antennas: P_{BS}	0.4 W	0.2 W
Power per UE: P_{UE}	0.2 W	0.1 W
Power for data encoding: P_{COD}	0.1 W/(Gbit/s)	0.01 W/(Gbit/s)
Power for data decoding: P_{DEC}	0.8 W/(Gbit/s)	0.08 W/(Gbit/s)
BS computational efficiency: L_{BS}	75 Gflops/W	750 Gflops/W
Power for backhaul traffic: P_{BT}	0.25 W/(Gbit/s)	0.025 W/(Gbit/s)

Table 5.3: Parameters in the CP model. Two different set of values are exemplified.

one that gives the largest SE between those of Theorem 4.6 on p. 317 and Theorem 4.9 on p. 326 for the DL:

$$\text{TR}_j = B \sum_{k=1}^{K_j} (\text{SE}_{jk}^{\text{UL}} + \max(\text{SE}_{jk}^{\text{DL}}, \text{SE}_{jk}^{\text{DL}})). \quad (5.43)$$

Two sets of CP parameters are given in Table 5.3. The first set is inspired by a variety of recent works: baseband power modeling from [174, 185], backhaul power according to [311], and the computational efficiencies from [358]. In the future, these parameters will take very different values that we cannot predict at the time of writing of this monograph. For the sake of the analysis, in what follows we also consider a setup in which the transceiver hardware's PC is reduced by a factor two whereas the computational efficiencies (which benefit from Moore's law) are increased by a factor ten. This leads to the second set of values in Table 5.3. We stress that these parameters tend to be extremely hardware-specific and thus may take substantially different values.¹¹ The Matlab code that is available online enables testing of other values.

Figure 5.9 illustrates the total CP per cell for the combined UL and DL scenario with different combining/precoding schemes. The MMSE estimator is used for channel estimation to fully exploit the spatial

¹¹IMEC has developed a power model to predict the power consumption of contemporary and future cellular BSs, which supports a broad range of operating conditions and BS types, and incorporates hardware technology forecasts. The model is available at <https://www.imec-int.com/en/powermodel> as an online web tool.

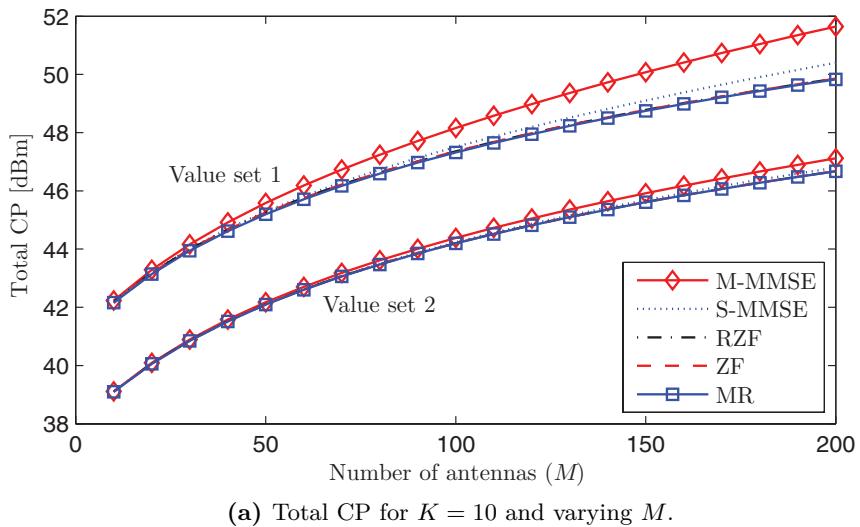
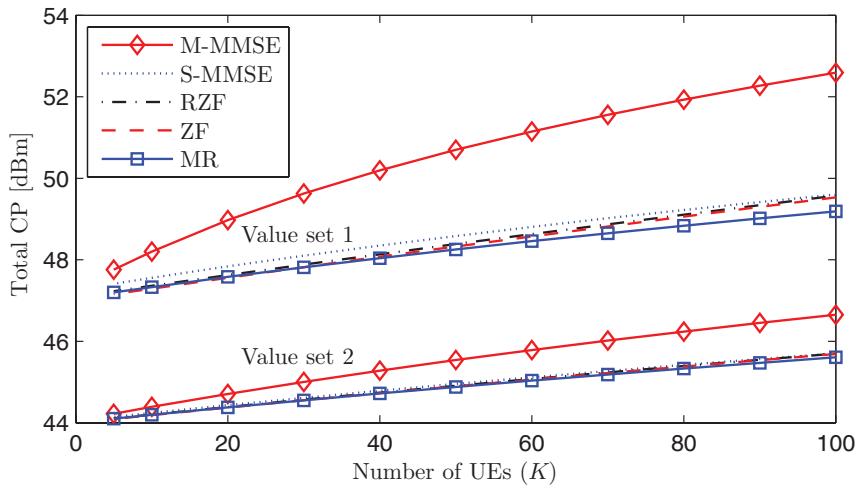
(a) Total CP for $K = 10$ and varying M .(b) Total CP for $M = 100$ and varying K .

Figure 5.9: Total CP per cell of both UL and DL for the running example. The two sets of CP parameter values reported in Table 5.3 are considered.

channel correlation. Note that the vertical axis is reported in dBm. In Figure 5.9a, we consider $K = 10$ and let M vary from 10 to 200. The CP increases with M for all schemes and for both value sets. The highest CP is required by M-MMSE, followed by S-MMSE. For Value set 1, S-MMSE reduces the CP by 0.5%–25% since inter-cell channel estimates are not computed. Note, however, that M-MMSE provides higher SE than S-MMSE. Quantitatively speaking, the CP required by M-MMSE for $M = 100$ and $K = 10$ is 48.16 dBm (65.48 W) whereas S-MMSE needs 47.5 dBm (56.35 W), which is roughly a 14% reduction. From Section 4, we know that this CP increase with M-MMSE is compensated by a 10% higher SE than with S-MMSE in both UL and DL (cf. Tables 4.3 and 4.5 on p. 295 and p. 332, respectively). For Value set 2, the CP required by M-MMSE is only 0.1%–7% higher than with S-MMSE. This is mainly due to the increased computational efficiency. RZF and ZF consume less CP, since both invert matrices of dimensions $K \times K$, rather than $M \times M$. Compared to M-MMSE, when $M = 100$, this reduces the CP by 17% for Value set 1 and by 4% for Value set 2. MR is characterized by the lowest CP since no matrix inversions are required. However, the CP reduction compared to RZF and ZF is marginal for both value sets; MR only provides a substantial complexity reduction compared to RZF and ZF when the number of UEs is very large; see Figure 4.3a. In Figure 5.9b, we consider $M = 100$ and let K vary from 10 to 100. The CP increases with the number of UEs, but with a smaller slope than when M is changed (especially for Value set 2). Although the general trends for the two set of values are the same (e.g., M-MMSE requires the highest CP and MR the lowest), we see that for Value set 1 the CP required by M-MMSE is 8%–100% higher than with S-MMSE. This CP increase reduces to 2%–25% CP for Value set 2.

The CP of all schemes, for the two different sets of parameter values, $M = 100$, and $K = 10$ are summarized in Table 5.4. As we can see, in this considered setup the CP required by the different schemes is marginally different. This happens because the CP of the transceiver hardware dominates over that of signal processing. Moreover, comparisons in this section are made for given configurations of (M, K) , which do not

Scheme	Value set 1	Value set 2
M-MMSE	65.48 W	27.42 W
S-MMSE	56.35 W	26.51 W
RZF	54.43 W	26.32 W
ZF	54.43 W	26.32 W
MR	53.96 W	26.27 W

Table 5.4: CP per cell with $M = 100$ and $K = 10$ for different schemes and the two sets of values reported in Table 5.3. The results are summarized from Figure 5.9. There are only marginal differences in the considered scenario, although the different schemes are characterized by different computational complexities. This happens because the CP of the hardware dominates over that of signal processing.

necessarily represent the optimal ones for maximizing the EE of the network, as we will see later in Section 5.6.

Figure 5.10 shows a bar diagram that shows how different parts contribute to the CP for $M = 100$ and $K = 10$ with the Value set 1 in Table 5.3. Only M-MMSE, RZF, and MR are considered, since the CPs of S-MMSE and ZF are similar to that of RZF. Note that the vertical axis is reported in dBm. The CP contributed by the fixed power, transceiver chains, signal processing for UL reception, DL transmission, and precoding computation are the same for all schemes. These four terms contribute as illustrated in Figure 5.10a and require a total of 47.23 dBm, which is the majority of the total CP. The largest CP is required by transceiver chains, followed by the fixed power. The signal processing required for UL reception and DL transmission of data consumes around 28.8 dBm, while the smallest part is the computation of precoding vectors, roughly 7 dBm. The breakdown of the CP required by the different processing schemes for channel estimation, computation of receive combining vectors, backhaul, and encoding/decoding is reported in Figure 5.10b. The CP consumed by intra-cell channel estimation is approximately 26 dBm (440 mW) and independent of the processing scheme. The CP for computing the receive combining vectors depends on the scheme and the highest CP is required by M-MMSE, for which it is approximately 40 dBm (10.96 W). Together with the consumed power by channel estimation, they account for 91% of the CP required

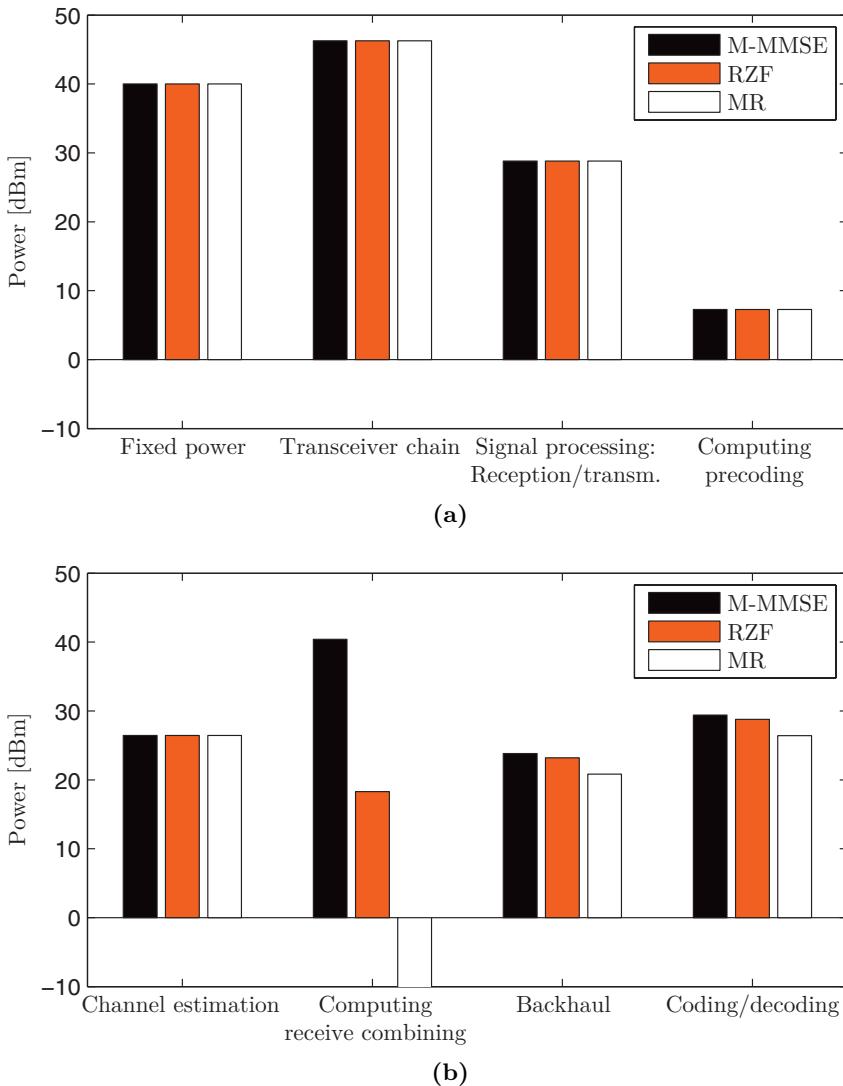


Figure 5.10: Breakdown of the CP per cell when using the first set of values in Table 5.3 with M-MMSE, RZF or MR. A setup with $K = 10$ UEs and $M = 100$ BS antennas per cell is considered. Note that the vertical axis is in dBm.

by M-MMSE for performing the operations considered in Figure 5.10b. A substantially lower CP would be required by M-MMSE with the EW-MMSE estimator, which reduces the computational complexity by 45%–90% (cf. Figure 3.8) by not exploiting the correlation between antenna elements. This, however, impacts the estimation accuracy (cf. Figure 3.7) for the considered channel model. If RZF is used, then the CP required for computing combining vectors is around 18.28 dBm (67 mW), which corresponds to a 99% reduction compared to M-MMSE. With MR, it is further reduced to 0.09 mW, which is negligible compared to all other contributions.

Figure 5.11 shows the CP terms with Value set 2 in Table 5.3. Compared to Figure 5.10a, the CP common to all schemes (accounting for the fixed power, transceiver chains, and signal processing) is reduced by 50%. Computing the receive combining vectors with M-MMSE still represents the most power-consuming operation in Figure 5.11b, though in this case it requires only 30 dBm rather than 40 dBm, which corresponds to a 90% reduction. Quantitatively speaking, the CP required for all the operations of Figure 5.11b is roughly 31 dBm with M-MMSE, 21.6 dBm with RZF, and 20 dBm with MR.

In summary, the above analysis shows that the CP model developed in this section highly depends on the hardware setup (i.e., number of BS antennas and UEs) and on the choice of the model parameters in Table 5.3. However, some general observations can be made irrespective of the specific parameter values. The CP increases faster with M than with K for all processing schemes. The highest CP is required by M-MMSE due to the extra cost for estimating the inter-cell channels, and is followed by S-MMSE. RZF and ZF require lower CP than S-MMSE, since both invert matrices of dimensions $K \times K$, rather than $M \times M$. MR has the lowest CP since no matrix inversion is required. However, since the computational efficiency of the signal processing is very high in modern systems, these differences have a marginal impact on the total CP, which is roughly the same for all the schemes (cf. Table 5.4 on p. 386). The transceiver chains contribute to the largest part of the total CP, followed by the fixed power and then by the channel estimation (except when using M-MMSE for which the computation of combining

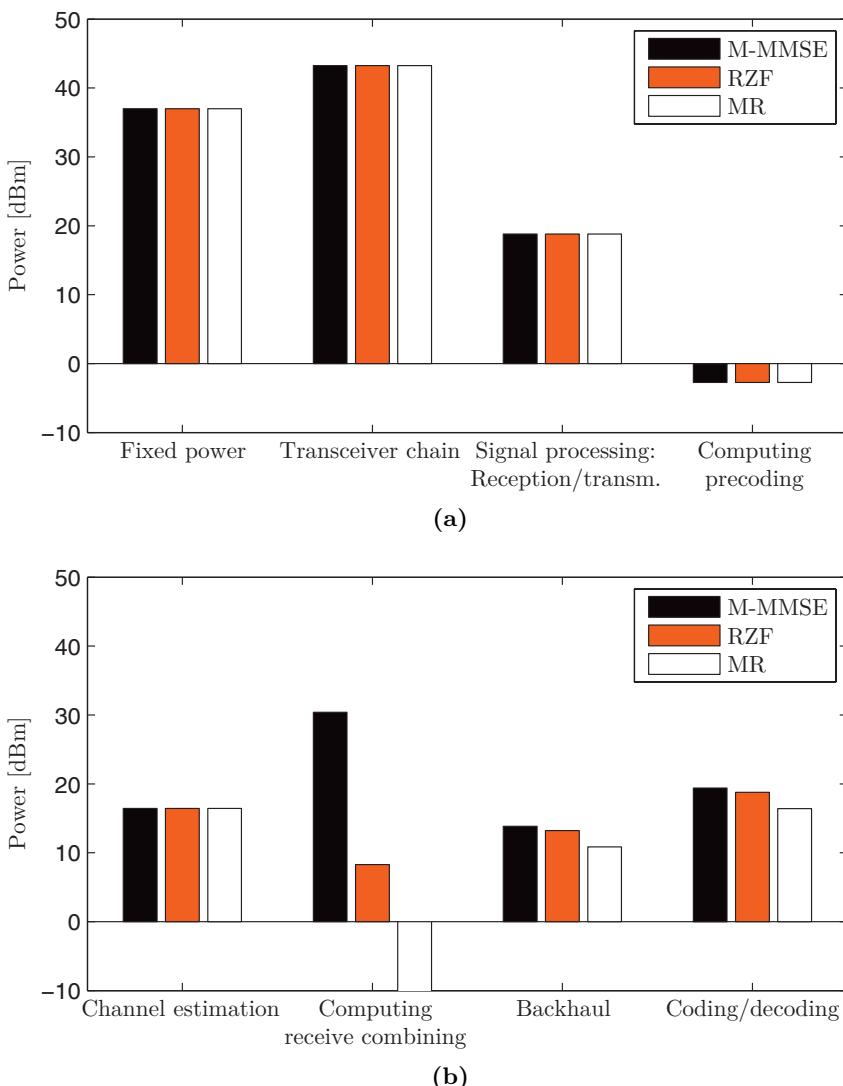


Figure 5.11: Breakdown of the CP per cell when using the second set of values in Table 5.3 with M-MMSE, RZF or MR. A setup with $K = 10$ UEs and $M = 100$ BS antennas per cell is considered. Note that the vertical axis is in dBm.

vectors is higher)—these parts are the same for all processing schemes. Moreover, the power consumed by backhaul, encoding, and decoding accounts only for a small fraction of the total CP in Massive MIMO, where the number of BS antennas is relatively large.

5.5 Tradeoff Between Energy Efficiency and Throughput

We will now examine the tradeoff between EE and throughput, using the CP model introduced in the previous section and the two sets of CP values reported in Table 5.3. Unlike the case-study analysis of Section 5.3.1 wherein the EE-SE tradeoff was considered, we concentrate on the throughput of the Massive MIMO network to emphasize that one cannot carry out EE analysis without specifying the bandwidth (cf. Remark 5.3). We use the running example defined in Section 4.1.3 on p. 288. There are M antennas at each BS and K UEs in each cell. The values of M and K will be changed and specified in each figure. The pilot reuse factor is $f = 1$, such that each pilot sequence consists of $\tau_p = K$ samples. The number of samples per coherence block used for UL and DL are $\tau_u = \frac{1}{3}(\tau_c - \tau_p)$ and $\tau_d = \frac{2}{3}(\tau_c - \tau_p)$, respectively. We consider UL and DL transmit powers of 20 dBm per UE (i.e., $p_{jk} = \rho_{jk} = 100$ mW). The Gaussian local scattering with ASD $\sigma_\varphi = 10^\circ$ is used as channel model. The throughput is obtained as in (5.43) by using the UL and DL SE expressions of Section 4.

The EE of cell j is computed as

$$\text{EE}_j = \frac{\text{TR}_j}{\text{ETP}_j + \text{CP}_j} \quad (5.44)$$

where ETP_j denotes the ETP of cell j . This term accounts for the power consumed by the transmission of the pilot sequences as well as of UL and DL signals:

$$\text{ETP}_j = \underbrace{\frac{\tau_p}{\tau_c} \sum_{k=1}^{K_j} \frac{1}{\mu_{\text{UE},jk}} p_{jk}}_{\text{ETP for pilots}} + \underbrace{\frac{\tau_u}{\tau_c} \sum_{k=1}^{K_j} \frac{1}{\mu_{\text{UE},jk}} p_{jk}}_{\text{ETP in the UL}} + \underbrace{\frac{1}{\mu_{\text{BS},j}} \frac{\tau_d}{\tau_c} \sum_{k=1}^{K_j} \rho_{jk}}_{\text{ETP in the DL}} \quad (5.45)$$

where $\mu_{\text{UE},jk}$ ($0 < \mu_{\text{UE},jk} \leq 1$) is the PA efficiency at UE k in cell j and $\mu_{\text{BS},j}$ ($0 < \mu_{\text{BS},j} \leq 1$) is that of BS j . The EE and throughput tradeoff of

different schemes will be compared by continuing the previous example of Figure 5.9 with the additional assumption of $\mu_{\text{UE},jk} = 0.4$ and $\mu_{\text{BS},j} = 0.5$. Notice that we have deliberately chosen a PA efficiency higher than 25% (i.e., higher than in contemporary PAs). This is motivated by the fact that in Massive MIMO, the low power levels per antenna (in the mW range, cf. Section 5.2) allow using more efficient PAs.

Figure 5.12 illustrates the EE as a function of the average throughput per cell with all processing schemes. The different throughput values are achieved with $K = 10$ UEs and by letting the number of BS antennas vary from $M = 10$ to $M = 200$, in steps of 10. The two sets in Table 5.3 are considered. We notice that the EE is a unimodal function of the throughput for all schemes and both sets of CP values. This implies that we can jointly increase the throughput and EE up to the maximum EE point, but further increases in throughput can only come at a loss in EE. The curves are quite smooth around the maximum EE point; thus, there is a variety of throughput values or, equivalently, numbers of BS antennas that provide nearly maximum EE. M-MMSE provides the highest EE for any value of the throughput, followed by S-MMSE. MR has the lowest performance. This shows that, in the considered setup, the additional computational complexity of M-MMSE processing pays off both in terms of SE and EE.

From Figure 5.12a, we can see that the maximal EE value with M-MMSE is 21.26 Mbit/Joule and is achieved at $M = 30$ for a throughput of 600 Mbit/s/cell, which corresponds to an area throughput of 9.6 Gbit/s/km². For $M = 40$, the EE is nearly the same and equal to 20.73 Mbit/Joule while the area throughput increases to 11 Gbit/s/km². With S-MMSE, the maximal EE is also obtained with $M = 30$ but is reduced by 3.2% and achieved at a 6% lower throughput than with M-MMSE. RZF and ZF provide similar performance and achieve the maximal EE of roughly 19 Mbit/Joule with $M = 30$ and a corresponding area throughput of 8.38 Gbit/s/km². Interestingly, RZF and ZF tend to perform as M-MMSE and S-MMSE when the throughput increases. This happens since the higher the throughput, the higher is also the CP, due to the larger number of antennas. Since the CP grows faster with M-MMSE and S-MMSE than with RZF and ZF, it counteracts the SE gain

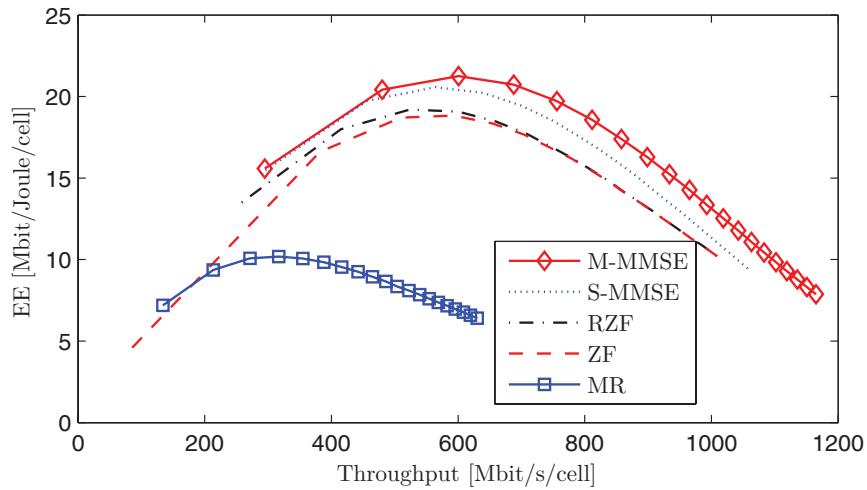
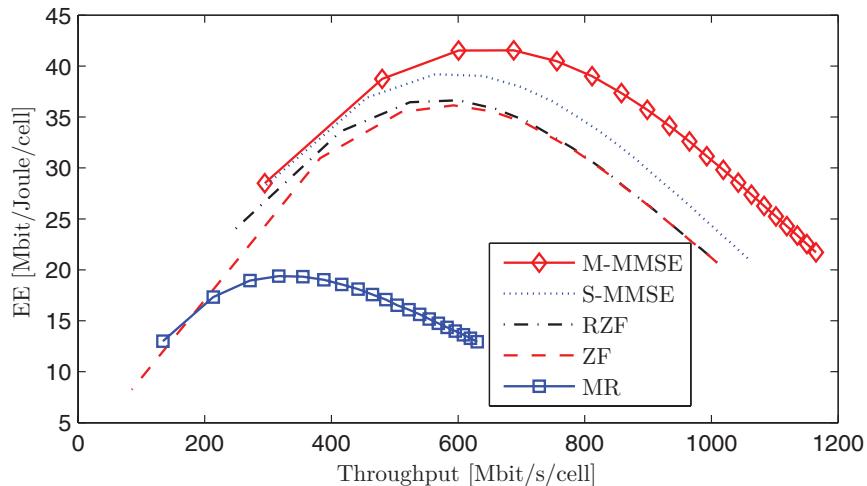
(a) $K = 10$ with the first set of values in Table 5.3.(b) $K = 10$ with the second set of values in Table 5.3.

Figure 5.12: EE versus throughput for the running example defined in Section 4.1.3 on p. 288. The hardware parameters are modeled as in Table 5.3. The different values of throughput are achieved varying M from $M = 10$ to $M = 200$, in steps of 10. Notice that all schemes allow to jointly increase the EE and throughput. M-MMSE provides the highest EE for any value of throughput.

of using M-MMSE and S-MMSE (cf. Figure 4.5 on p. 293 for UL). Note that the EE of ZF deteriorates quickly for throughput values smaller than 380 Mbit/s/cell, which are achieved by $M \approx K$ where ZF is known to perform badly. MR provides a maximum EE of 10.18 Mbit/Joule with $M = 40$ and an area throughput of 5.07 Gbit/s/km².

Figure 5.12b is obtained with the second set of values in Table 5.3. Compared to Figure 5.12a, the EE of all schemes is roughly doubled (since most of the CP coefficients are reduced by a factor two), but the general trends are the same. With M-MMSE and S-MMSE, an EE of 41.52 Mbit/Joule and 39.2 Mbit/Joule, respectively, is achieved using $M = 30$. Table 5.5 reports the EE and area throughout of M-MMSE, RZF, and MR with $K = 10$ and $M = 40$. In summary, the above analysis shows that different schemes provide slightly different EE. However, with all of them the EE is a unimodal function of the throughput and, for the considered scenario, achieves its maximum at roughly $M = 30$ or 40 irrespective of the CP parameter values. Note that these values for M are far from what it is envisioned for Massive MIMO, but the antenna-UE ratio achieving the maximum EE is on the order of $M/K = 3$ or 4 as expected. In what follows, we show that a larger value for M is obtained if K increases, especially with the more power-efficient hardware setup.

Figure 5.13 considers the same setup but with $K = 20$. The throughput values are obtained for all schemes, except for ZF, by varying M from 10 to 200, in steps of 10. With ZF, M varies from 20 to 200 since M must be larger than K . Compared to Figure 5.12, we can see that increasing the number of UEs per cell may have or not a positive effect on the EE. Unlike Figure 5.12a, Figure 5.13a shows that, with $K = 20$ and the first set of CP values, the highest EE is not achieved with M-MMSE but with S-MMSE. Quantitatively speaking, S-MMSE provides a maximal EE of 22.86 Mbit/Joule at $M = 50$ for an area throughput of 15.05 Gbit/s/km². On the other hand, with M-MMSE the maximal EE is obtained with $M = 40$ and is 1.75% lower than with S-MMSE. Interestingly, with the first set of CP values, M-MMSE performs even worse than RZF and RZF when the throughput increases. This happens because the CP with M-MMSE is much higher (cf. Figure 5.9b) than

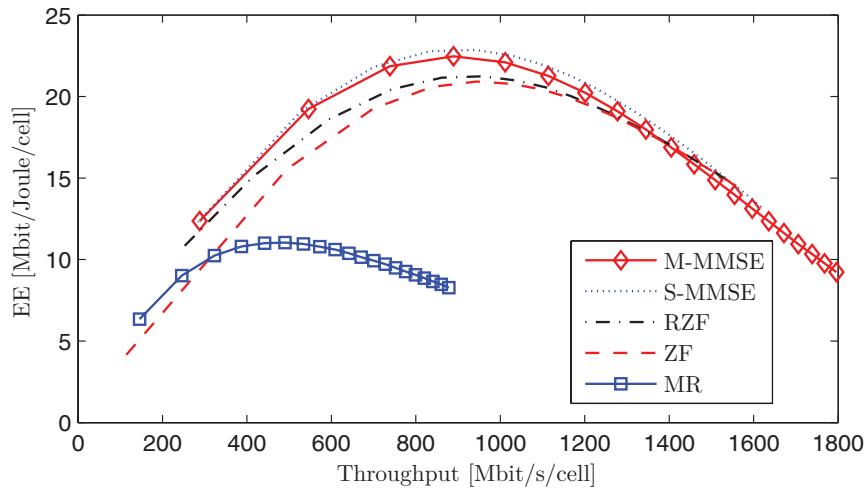
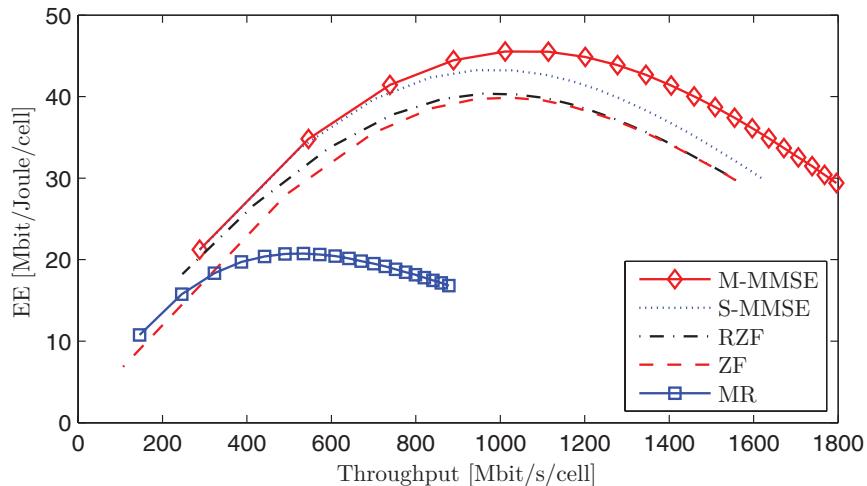
(a) $K = 20$ with the first set of CP values in Table 5.3.(b) $K = 20$ with the second set of CP values in Table 5.3.

Figure 5.13: EE versus throughput for the running example defined in Section 4.1.3 on p. 288. The hardware parameters are modeled as in Table 5.3. The different values of throughput are achieved by varying M from $M = 20$ to $M = 200$, in steps of 10. Compared to results of Figure 5.12, we see that increasing K improves the EE of all schemes.

Scheme	EE, set 1	EE, set 2	Area throughput
M-MMSE	20.73 Mbit/Joule	41.53 Mbit/Joule	11 Gbit/s/km ²
RZF	19.07 Mbit/Joule	36.63 Mbit/Joule	9.6 Gbit/s/km ²
MR	10.18 Mbit/Joule	19.38 Mbit/Joule	5.07 Gbit/s/km ²

(a) With $K = 10$ and $M = 40$ (the results are summarized from Figure 5.12).

Scheme	EE, set 1	EE, set 2	Area throughput
M-MMSE	21.27 Mbit/Joule	45.5 Mbit/Joule	17.82 Gbit/s/km ²
RZF	21.24 Mbit/Joule	40.35 Mbit/Joule	15.33 Gbit/s/km ²
MR	11.04 Mbit/Joule	20.7 Mbit/Joule	7.84 Gbit/s/km ²

(b) With $K = 20$ and $M = 60$ (the results are summarized from Figure 5.13).

Table 5.5: Maximal EE per cell with the two sets of CP values in Table 5.3 for M-MMSE, RZF and MR. The corresponding area throughputs are also reported.

with S-MMSE, RZF, and ZF, and thus counteracts the SE gain of using M-MMSE. Different observations can be made for the second set of CP values as we can see from Figure 5.13b. In this case, the general trends are the same of Figure 5.12, where M-MMSE provides the highest EE and throughput. Moreover, increasing the number of UEs per cell has a positive effect on the EE of all schemes, which is larger for any throughput value. Unlike with $K = 10$ in Figure 5.12b, wherein the maximal EE was achieved at $M = 30$ or 40 , with $K = 20$ we see that $M = 50$ or 60 provides the highest EE. Table 5.5b summarizes the results of Figure 5.13 for M-MMSE, RZF, and MR with $M = 60$.

5.6 Network Design for Maximal Energy Efficiency

In the previous section, we examined the EE of Massive MIMO networks for a given number of UEs and varying number of BS antennas (or, equivalently, for a given throughput value per cell). In the following, we look at the EE from a different perspective: we design the network from scratch to achieve maximal EE, without a-priori assumptions on the number of antennas or UEs. We ask the following questions:

1. What is the optimal number of BS antennas?

2. How many UEs should be served?
3. When should different processing schemes be used?

To answer these questions, we consider the running example, defined in Section 4.1.3 on p. 288, for the same scenario as in Section 5.5, with M antennas at each BS and K UEs per cell.

Figure 5.14 illustrates the set of achievable EE values with M-MMSE, RZF, and MR for different combinations of M and K . We consider $K \in \{10, 20, \dots, 100\}$ and $M \in \{20, 30, \dots, 200\}$. The first set of CP values in Table 5.3 is considered. We notice that the EE-surfaces are concave and a global EE-optimum exists for each scheme. With M-MMSE, a maximal EE of 20.73 Mbit/Joule is achieved by $(M, K) = (40, 20)$, which provides an area throughput of 13.71 Gbit/s/km² and a total PC of 41.35 W per cell. With RZF, the EE surface is smoother than with M-MMSE; thus, there is a variety of pairs (M, K) that provides nearly optimal EE. The global EE maximum is 20.25 Mbit/Joule, which is only 2.3% higher than with M-MMSE. It is achieved by $(M, K) = (90, 30)$ resulting in an area throughput of 20.97 Gbit/s/km², which is 53% higher than with M-MMSE. This is achieved using 64.76 W CP per cell, which is also 56% higher. The intuition behind this result is that, although M-MMSE is the best from a throughput perspective for any given (M, K) , the CP increases faster with M and K with M-MMSE compared to RZF due to the higher computational complexity. This discourages the use of M-MMSE with larger M and K values, where the slightly higher throughput than with RZF comes with a disproportionately larger CP that degrades the EE. Hence, M-MMSE achieves its EE-optimum at a lower throughput value. The price to pay with RZF is the higher PC per cell, thus having an energy-efficient network does not imply that the power is low. The EE-optimum with MR provides an EE of 10.63 Mbit/Joule, which is roughly 47% smaller than with M-MMSE and RZF, and is achieved at $(M, K) = (60, 20)$ for an area throughput of 7.64 Gbit/s/km² and a PC of 44.9 W per cell. The above results are summarized in Table 5.6a.

Figure 5.15 considers the second set of CP values in Table 5.3. The EE-surfaces with RZF and MR show similar behaviors as in Figure 5.14.

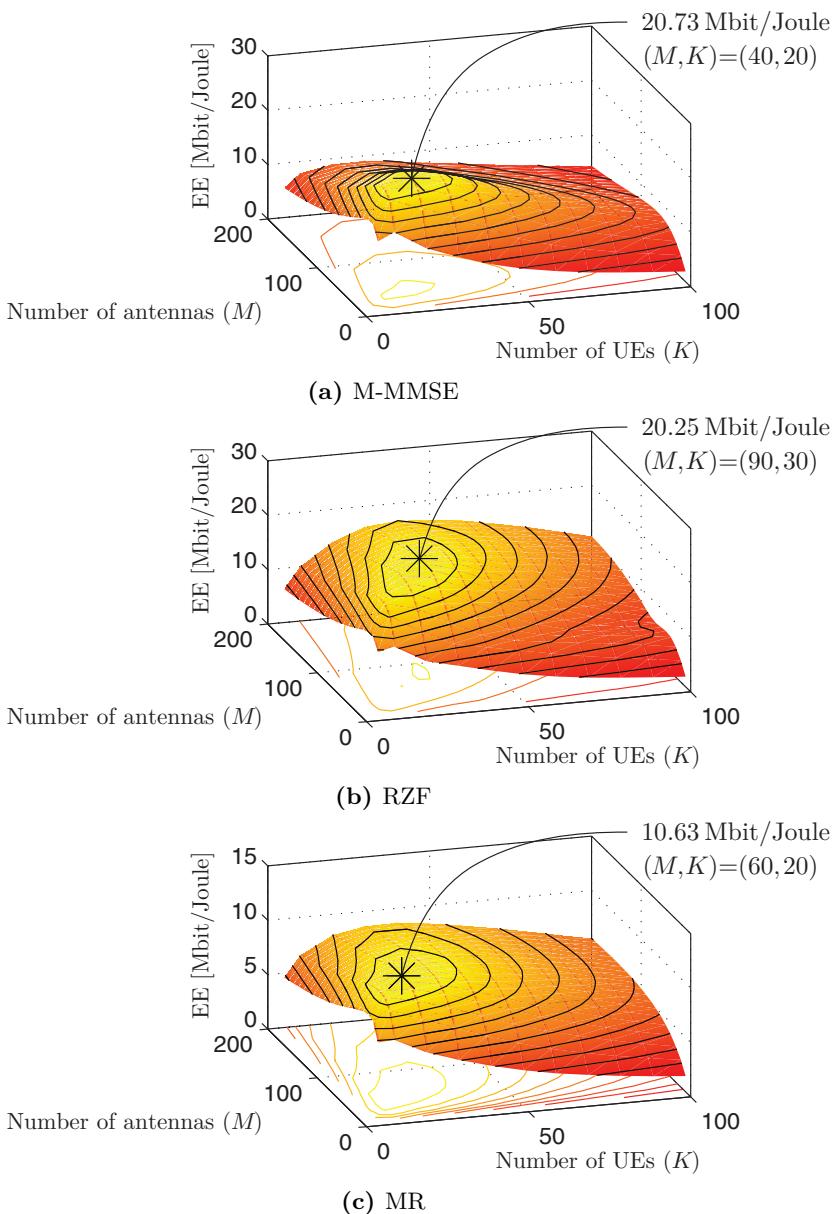


Figure 5.14: EE per cell as a function of M and K with M-MMSE, RZF, and MR. The first set of values in Table 5.3 is used.

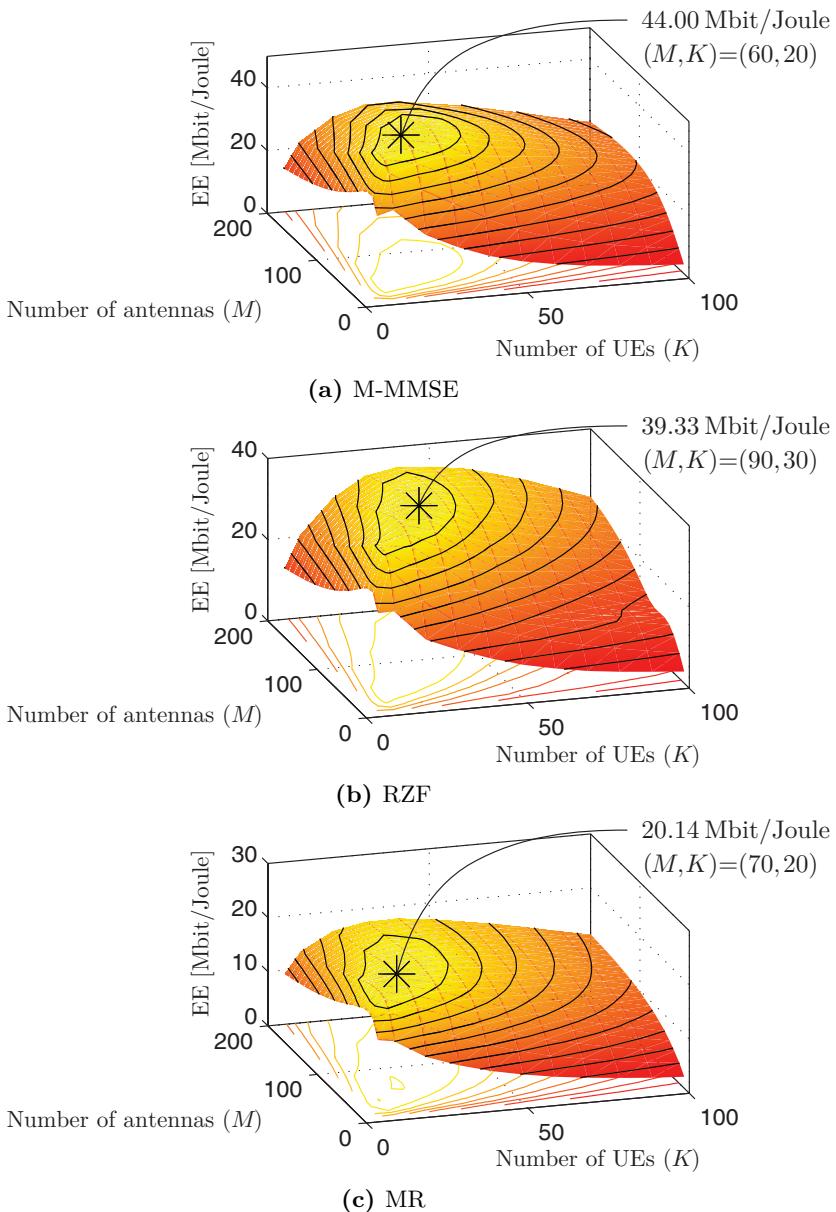


Figure 5.15: EE per cell as a function of M and K with M-MMSE, RZF, and MR. The second set of values in Table 5.3 is used.

Scheme	(M, K)	Maximal EE	Area throughput	PC
M-MMSE	(40, 20)	20.73 Mbit/Joule	13.71 Gbit/s/km ²	41.35 W
RZF	(90, 30)	20.25 Mbit/Joule	20.97 Gbit/s/km ²	64.76 W
MR	(60, 20)	10.63 Mbit/Joule	7.64 Gbit/s/km ²	44.9 W

(a) With the first set of values in Table 5.3, the results are obtained from Figure 5.14.

Scheme	(M, K)	Maximal EE	Area throughput	PC
M-MMSE	(60, 20)	44.00 Mbit/Joule	17.33 Gbit/s/km ²	24.62 W
RZF	(90, 30)	39.33 Mbit/Joule	20.97 Gbit/s/km ²	33.34 W
MR	(70, 20)	20.14 Mbit/Joule	8.3 Gbit/s/km ²	25.75 W

(b) With the second set of values in Table 5.3, the results are obtained from Figure 5.15.

Table 5.6: Maximal EE per cell with the two sets of CP values in Table 5.3 for M-MMSE, RZF and MR. The corresponding values of area throughput and PC per cell are also reported. The results are summarized from Figures 5.14 and 5.15.

However, the surface with M-MMSE is substantially smoother than in Figure 5.14. The EE, area throughput, and PC values of the EE-optimal operating points for all schemes are summarized in Table 5.6b. The values of M and K at the EE-optimum points are both increased due to the reduction in the CP parameters. In particular, the higher computational efficiency encourages the use of more network infrastructure and spatial multiplexing of more UEs. We notice that M-MMSE still provides a 17% lower throughput than RZF, while achieving a 12% higher EE and a 26% PC saving per cell. Compared to the results of Figure 5.14, it thus follows that M-MMSE becomes a potential solution for high EE and low PC when more energy-efficient hardware is used.

Interestingly, we observe that all the EE-optimum configurations of (M, K) fall within the class of Massive MIMO networks, with a number of antennas in the range of tens and an antenna-UE ratio M/K that varies between 2 and 3.5 at the different EE-optima. This is noteworthy since it is the output of a numerical search in which we did not restrict the system dimensions whatsoever. Notice also that the PCs are within a realistic range when using any of the schemes.

Remark 5.6 (From numerical to analytical analysis). Numerical results were used in this section to find the numbers of BS antennas and UEs

that jointly achieve high EE in the running example. However, we stress that in some simplified scenarios, the EE maximization problem can potentially be solved analytically with respect to M and K . Other system parameters (e.g., the transmit power) can also be optimized. This approach was taken in [58, 59, 224] for a single-cell Massive MIMO system using ZF in UL and DL over spatially uncorrelated channels. Closed-form expressions for the optimal (M, K) and transmit power were derived, from which valuable insights into the interplay between the optimization variables, hardware characteristics, and propagation environment were provided. The UL multicell case was considered in [60, 266], which determine the optimal M , K , transmit power, cell density, and pilot reuse factor. The analysis showed that reducing the cell size is undoubtably the way towards high EE, but the positive effect of increasing the cell density saturates when the CP dominates over the transmit power. A further leap in EE can then be achieved by adding more BS antennas to spatially multiplex UEs in every cell. The corresponding EE gains come from suppressing intra-cell interference by the many antennas and by sharing the per-cell CP costs among multiple UEs. Moreover, the analysis showed that a large pilot reuse factor can protect against inter-cell interference and can be tailored to guarantee a certain SE.

5.7 Summary of Key Points in Section 5

- Higher area throughput on the one hand and less consumed power on the other are two seemingly contradictory requirements for future networks.
- Massive MIMO can potentially achieve a higher area throughput than current networks while providing substantial ATP savings. The transmit power can be gradually reduced with the number of antennas while approaching a non-zero asymptotic SE. Therefore, Massive MIMO networks reduce the transmit power required to achieve a given SE.
- The EE of a cellular network, defined as the number of bits that can be reliably transmitted per unit of energy (measured in bit/Joule), is a good performance metric to balance the throughput and consumed power.
- While increasing the number of antennas has always a positive effect on the SE, the EE first increases with M , due to the improved SE, and then decreases with M , due to the additional hardware that increases the CP.
- Substantial SE gains are achieved by multiplexing K UEs per cell, if a proportional number of antennas M is used to counteract the increased interference. A similar result cannot be achieved for the EE since adding more antennas increases the SE but also the CP of the network. This means that the EE attains its maximum at a finite value of the antenna-UE ratio M/K .
- Realistic CP models are needed to evaluate the PC for different numbers of antennas and UEs. The modeling complexity makes a certain level of idealization unavoidable, but a fairly accurate polynomial CP model was developed in this section to account for the dissipation in analog hardware, digital

signal processing, backhaul signaling, and channel estimation. The model depends on a variety of fixed parameters that were kept generic in the analysis. Typical values are given in Table 5.3. The MR scheme has the lowest CP, while the interference suppressing schemes, such as RZF and M-MMSE, require higher CP.

- Massive MIMO allows to jointly increase the EE and throughput, as compared to a system with few antennas. In the running example, M-MMSE provides the highest EE for any throughput value only when more energy-efficient hardware is used. MR achieves the lowest EE. RZF provides a good tradeoff between EE and throughput.
- A numerical example was used to demonstrate how a cellular network should be designed for maximal EE. The results show that a Massive MIMO setup, wherein a large number of antennas (in the order of a hundred) is used to serve many tens of UEs, is the EE-optimal solution, even using contemporary circuit technology.

6

Hardware Efficiency

In this section, we analyze how the use of non-ideal transceiver hardware affects the SE. The goal is to show that Massive MIMO improves the *hardware efficiency (HE)*, in the sense that the SE loss from using hardware components of lesser quality than in conventional systems is negligible. For example, we will show that one can compensate for increasing hardware distortion by adding additional BS antennas. Section 6.1 provides a brief overview of the non-idealities that exist in practical transceivers and develops a model that captures the hardware impairment characteristics that affect the SE, without limiting the analysis to a particular hardware setup. This model is then used to generalize the Massive MIMO system model that was used in previous sections. UL channel estimation with hardware impairments is considered in Section 6.2, while the achievable SEs are analyzed in Section 6.3. The improved HE is then established by a hardware-quality scaling law in Section 6.4, which proves how quickly the hardware quality can be reduced when increasing the number of antennas if the SE loss should remain small. The key points are summarized in Section 6.5.

6.1 Transceiver Hardware Impairments

Wireless communication channels are commonly modeled as linear filters that take an analog input signal from the transmitter and produce a distorted output signal, which is measured at the receiver in the presence of additive noise. This classic setup is illustrated in Figure 6.1a. The discrete-time complex baseband model described in Section 2.3 on p. 226 is an equivalent representation of such an analog channel, under certain conditions. For the equivalence to hold, the transmitter needs to generate the correct modulated passband signal from the complex baseband samples, the receiver needs to demodulate and sample the received signal correctly, and the transmitter and receiver need to be synchronized in time and frequency. None of these conditions are fully satisfied in practice, because the closer to ideal a hardware component is, the more challenging it is to implement—it becomes bulkier, more expensive, and consumes more power. In other words, there is a *cost-quality tradeoff* in practical systems. As already shown in Section 5, this tradeoff is particularly important when implementing Massive MIMO systems because if BS j is equipped with M_j antennas it needs M_j copies of many components; for example, PAs, ADCs, filters, I/Q mixers, and DACs. The cost of such an implementation will be, roughly, M_j times higher than the cost of a single-antenna transceiver, unless we can compensate by reducing the quality of the individual components.

To investigate how the quality of the hardware components affects the communication performance, we need a model of how the hardware affects the transmitted and received signals and particularly the SE. Different from the linear channel filter, many hardware components act as non-linear filters. For example, the PA in the transmitter does generally not have a linear amplification, but provides less amplification gain to stronger input signals; in other words, the output power saturates. The finite-resolution quantization in the receiver is another non-linear operation, which additionally is destructive and cannot be undone. These non-idealities are referred to as *hardware impairments* in this monograph. There is plenty of literature on the modeling of different types of hardware impairments, including PA non-linearities, amplitude/phase imbalance in I/Q mixers, phase noise in LOs, sampling jitter, and

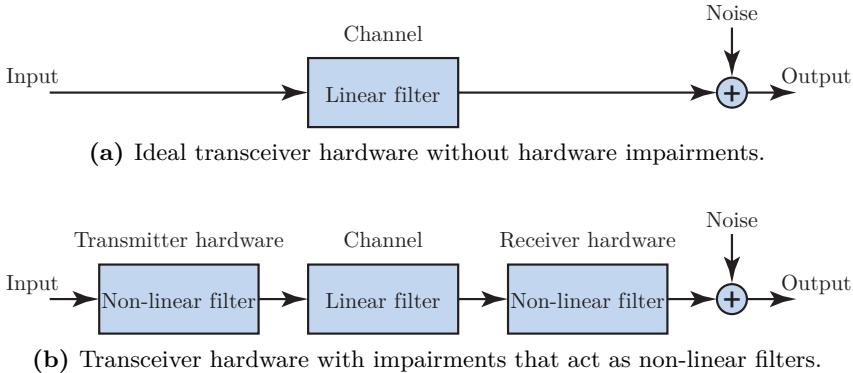


Figure 6.1: A communication system is modeled differently depending on whether the impact of the transceiver hardware can be neglected or not. All filters in this figure are assumed to be memoryless.

finite-resolution quantization in ADCs. These models are typically used to devise analog or digital compensation algorithms that mitigate the impairments, which can substantially reduce their impact on the communication. Residual impairments will still exist due to modeling inaccuracies and the destructive nature of some impairments. We will not cover the detailed modeling and compensation of hardware impairments here, but we refer to [147, 290, 322, 344] and references therein. Instead, we will focus on the impact that (residual) hardware impairments have on the SE, by modeling the non-ideal hardware as non-linear memoryless filters at the transmitter and the receiver [42]. This setup is illustrated in Figure 6.1b. Note that additive noise exists in many components of the receiver and is being filtered and amplified, similarly to the desired signals. Thermal noise is a main cause of additive noise in communications, but we refer to it as receiver noise since it accounts also for the noise amplification due to hardware impairments. For simplicity, we represent all the noise created in the receiver by a single equivalent receiver noise term that is added to the output of the receiver hardware.

6.1.1 Basic Modeling of Residual Hardware Impairments

To understand the basic impact of hardware impairments, we consider an information signal $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$ that is fed into a non-linear memoryless

function $g(\cdot)$. This function produces the output $y = g(x)$ that is also a random variable, but its distribution is generally non-Gaussian. The output y is correlated with the input x [69]. In particular, we can express the output signal as

$$y = \frac{\mathbb{E}\{yx^*\}}{p}x + \eta \quad (6.1)$$

where we have defined the distortion term

$$\eta = y - \frac{\mathbb{E}\{yx^*\}}{p}x \quad (6.2)$$

and utilized the assumption $\mathbb{E}\{|x|^2\} = p$. Note that the expression in (6.1) does not account for the receiver noise, which can later be added to the output. The expectation $\mathbb{E}\{yx^*\} = \mathbb{E}\{g(x)x^*\}$ can be computed for any given $g(\cdot)$, either analytically or numerically. The distortion term η is uncorrelated with x , since

$$\mathbb{E}\{\eta x^*\} = \mathbb{E}\{yx^*\} - \frac{\mathbb{E}\{yx^*\}}{p} \underbrace{\mathbb{E}\{|x|^2\}}_{=p} = 0. \quad (6.3)$$

However, the input and the distortion term are generally not independent. For example, suppose $g(x) = a|x|^2x + bx$ for some scalars $a, b \neq 0$. Then, from the definition of the distortion term, we have that $\eta = a|x|^2x - 2apx$, which is clearly dependent on x .

We can exploit the fact that the distortion term is uncorrelated with the input to bound the capacity C of the non-linear memoryless channel between x and y [377]. In particular, we can utilize the lower bound in Corollary 1.3 on p. 171 for deterministic channels by setting $h = \mathbb{E}\{yx^*\}/p$, $p_v = \mathbb{E}\{|\eta|^2\} = \mathbb{E}\{|y|^2\} - |\mathbb{E}\{yx^*\}|^2/p$, and $\sigma^2 = 0$. This results in

$$C \geq \log_2 \left(1 + \frac{p|h|^2}{p_v} \right) = \log_2 \left(1 + \frac{|\mathbb{E}\{yx^*\}|^2/p}{\mathbb{E}\{|y|^2\} - |\mathbb{E}\{yx^*\}|^2/p} \right). \quad (6.4)$$

This lower bound on the capacity is an achievable SE and reveals that from the total power $\mathbb{E}\{|y|^2\}$ of the output, the part that is correlated with the input (which has power $|\mathbb{E}\{yx^*\}|^2/p$) is always useful for communication. The distortion term η can also carry useful information, but in the worst case it is an independent complex Gaussian variable

with zero mean and power $\mathbb{E}\{|y|^2\} - |\mathbb{E}\{yx^*\}|^2/p$. This is the worst-case condition under which the lower bound in (6.4) is attained. We stress that the SE in (6.4) is generally finite, although there is no noise included in the non-linear channel relation between x and y .

In what follows, we will utilize the observations above to define an analytically tractable model of the impact of (residual) hardware impairments on the communication. As noted above, the key modeling characteristics are that the desired signal is scaled by a deterministic factor and that an uncorrelated memoryless distortion term is added, which is Gaussian distributed in the worst case. Assume that the compensation algorithms that are used to mitigate hardware impairments are calibrated to make the average power of the input and output equal: $\mathbb{E}\{|y|^2\} = \mathbb{E}\{|x|^2\} = p$. If we define the parameter

$$\kappa = \frac{|\mathbb{E}\{yx^*\}|^2}{\mathbb{E}\{|x|^2\}} = \frac{|\mathbb{E}\{yx^*\}|^2}{p} \quad (6.5)$$

then the output of the non-linear hardware is modeled as

$$y = \sqrt{\kappa}x + \eta \quad (6.6)$$

where $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$ is the input and the distortion term η is worst-case modeled as $\eta \sim \mathcal{N}_{\mathbb{C}}(0, (1 - \kappa)p)$ and independent of x . The distortion power is therefore proportional to the input power p , with the proportionality constant $(1 - \kappa)$. This makes the additive distortion term different from conventional receiver noise, which is independent of the input power. When operating at high SNR, where the noise is negligible, the distortion becomes a main limiting factor for the SE [56]. We refer to $\kappa \in (0, 1]$ as the *hardware quality factor*, where $\kappa = 1$ represents ideal hardware with $y = x$. In contrast, $\kappa = 0$ is the pathological case where the output signal is uncorrelated with the input. Note that, by definition, $\mathbb{E}\{|y|^2\} = \kappa p + (1 - \kappa)p = p$ for any κ .

The input-output model in (6.6) is by no means a full characterization of the transceiver hardware characteristics, but it captures the essential detrimental impact that hardware impairments have on the SE. For example, the capacity lower bound in (6.4) becomes $\log_2(1 + \kappa/(1 - \kappa))$ when using the notation defined in (6.5). This expression is characterized only by κ and is an increasing function for $\kappa \in (0, 1]$. The true

capacity is possibly higher, but the lower bound represents what can be achieved without any complicated hardware-adapted signal processing that attempts to extract information from η .

6.1.2 A Practical Measure of Hardware Quality

The error vector magnitude (EVM) is a common metric for measuring the distortion level in practical transceiver hardware. It is defined as the ratio between the average distortion magnitude and the signal magnitude, after basic equalization. For the model in (6.6), the input signal is x , the output signal is y , and the equalized output that minimizes¹ the MSE is $y\sqrt{\kappa}$. The EVM definition then gives

$$\text{EVM} = \sqrt{\frac{\mathbb{E}\{|y\sqrt{\kappa} - x|^2\}}{\mathbb{E}\{|x|^2\}}} = \sqrt{\frac{(1-\kappa)p}{p}} = \sqrt{1-\kappa}. \quad (6.7)$$

The EVM is one of the key metrics that are specified on the data sheets of RF transceivers. The LTE standard requires $\text{EVM} \leq 0.08$ in the transmitter hardware if 64-QAM should be supported [144, Sec. 14.3.4]. This corresponds to $\kappa = 1 - \text{EVM}^2 \geq 0.994$. If the transmitter should only support 4-PSK, then LTE only requires $\text{EVM} \leq 0.175$ and this corresponds to $\kappa \geq 0.97$. While practical LTE transceivers typically support 64-QAM, larger EVMs than 0.08 are of interest in Massive MIMO to relax the hardware design constraints. The analysis in this section applies to any κ between zero and one.

6.1.3 Extending the Canonical Massive MIMO Model

We will now incorporate hardware impairments into the Massive MIMO system model with ideal hardware that was used for analysis in the previous sections. We consider impairments in both the transmitter and the receiver hardware, as illustrated in Figure 6.1b. These impairments affect the analog signals that are transmitted over the channel, but we model the impairments on the complex-baseband representation. The impairment model from Section 6.1.1 will be applied to model both

¹The MSE $\mathbb{E}\{|ya - x|^2\}$ between the input x and the equalized output ya is minimized by selecting $a = \sqrt{\kappa}$, which leads to the MSE value $(1-\kappa)p$.

the transmitter distortion and the receiver distortion, in both UL and DL. We assume that the distortion is independent between samples, which is a worst-case assumption since a correlation could be utilized for distortion mitigation. A consequence of this assumption is that the coherence block dimensionality of the end-to-end channel, including the hardware, is the same as for the propagation channel. The model described below resembles the one in [42], but the notation is different.

Uplink Transmission

In the UL, the data signal of UE k in cell j is $s_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, p_{jk})$, as defined in Section 2.3.1 on p. 226. This complex Gaussian signal is distorted by the transmitter hardware and we apply the model in (6.6) to obtain that $\sqrt{\kappa_t^{\text{UE}}} s_{jk} + \eta_{jk}^{\text{UE}}$ is sent over the channel instead of s_{jk} , where $\eta_{jk}^{\text{UE}} \sim \mathcal{N}_{\mathbb{C}}(0, (1 - \kappa_t^{\text{UE}}) p_{jk})$ is the hardware distortion term. The factor $\kappa_t^{\text{UE}} \in (0, 1]$ determines the quality of the UE's transmitter hardware and is assumed to be the same for all UEs, for notational convenience. Following (2.5), the signal $\check{\mathbf{y}}_j \in \mathbb{C}^{M_j}$ that reaches the M_j receive antennas of BS j (before noise is added) is

$$\check{\mathbf{y}}_j = \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbf{h}_{li}^j \left(\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}} \right). \quad (6.8)$$

In a given coherence block with a set $\{\mathbf{h}_{li}^j\}$ of channel realizations, the signal $\check{\mathbf{y}}_j$ is conditionally complex Gaussian distributed with zero mean and correlation matrix

$$\mathbb{E} \left\{ \check{\mathbf{y}}_j \check{\mathbf{y}}_j^H \mid \{\mathbf{h}_{li}^j\} \right\} = \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbf{h}_{li}^j (\mathbf{h}_{li}^j)^H \quad (6.9)$$

since $\mathbb{E}\{|\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}}|^2\} = p_{li}$. Hence, we can apply (6.6) once more to model the impairments in the receiver hardware. We assume that the transceiver hardware attached to the different BS antennas are decoupled so that their respective distortion terms are independent.²

²Even if the transceiver hardware is decoupled, the received signals are correlated since the same transmitted signals are observed on all antennas. This fact can make

The hardware distortion in the receiver replaces $\check{\mathbf{y}}_j$ with $\sqrt{\kappa_r^{\text{BS}}} \check{\mathbf{y}}_j + \boldsymbol{\eta}_j^{\text{BS}}$, where the distortion term $\boldsymbol{\eta}_j^{\text{BS}} \in \mathbb{C}^{M_j}$ has the conditional distribution

$$\boldsymbol{\eta}_j^{\text{BS}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \mathbf{D}_{j,\{\mathbf{h}\}}) \quad (6.10)$$

for the given set of channel realizations $\{\mathbf{h}_{li}^j\}$ and the conditional correlation matrix $\mathbf{D}_{j,\{\mathbf{h}\}} \in \mathbb{C}^{M_j \times M_j}$ is given by

$$\mathbf{D}_{j,\{\mathbf{h}\}} = (1 - \kappa_r^{\text{BS}}) \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \text{diag} \left(|\mathbf{h}_{li}^j|_1^2, \dots, |\mathbf{h}_{li}^j|_{M_j}^2 \right) \quad (6.11)$$

where $[\mathbf{h}_{li}^j]_m$ denotes the m th element of \mathbf{h}_{li}^j . The factor $\kappa_r^{\text{BS}} \in (0, 1]$ determines the quality of the BS's receiver hardware and is the same for all BSs, for notational convenience. Note that the diagonal of $\mathbf{D}_{j,\{\mathbf{h}\}}$ is the same as that of the matrix in (6.9), except for the scaling factor $(1 - \kappa_r^{\text{BS}})$, which means that the receiver distortion term at each receive antenna is proportional to the power received at that antenna (which is generally different between antennas). The off-diagonal elements in $\mathbf{D}_{j,\{\mathbf{h}\}}$ are zero due to the assumed independence of the distortion terms over the array. The marginal (unconditional) distribution of the receiver distortion can be expressed as

$$\boldsymbol{\eta}_j^{\text{BS}} = \sqrt{1 - \kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{i=1}^{K_l} \sqrt{p_{li}} \mathbf{h}_{li}^j \odot \bar{\boldsymbol{\eta}}_{jli}^{\text{BS}} \quad (6.12)$$

where $\bar{\boldsymbol{\eta}}_{jli}^{\text{BS}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \mathbf{I}_{M_j})$ is an independent random variable and \odot denotes the Hadamard product. Note that the receiver distortion term in (6.12) is the element-wise product of two independent complex Gaussian vectors and is, thus, not Gaussian distributed.

In summary, by adding the receiver noise \mathbf{n}_j to $\sqrt{\kappa_r^{\text{BS}}} \check{\mathbf{y}}_j + \boldsymbol{\eta}_j^{\text{BS}}$, the

the distortion terms correlated as well. Simply speaking, the hardware reacts similarly to correlated input signals. The measurements reported in [223] confirm that such correlation exists but also that it is rather small. Thus the correlation matrix of the distortion has a different eigenstructure than the correlation matrix of the signal in (6.9). This is the key property that we capture in our model, while the assumption of independence between the distortion terms is only made for analytical tractability.

received UL signal $\mathbf{y}_j \in \mathbb{C}^{M_j}$ at BS j is modeled as

$$\mathbf{y}_j = \sqrt{\kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbf{h}_{li}^j \left(\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}} \right) + \boldsymbol{\eta}_j^{\text{BS}} + \mathbf{n}_j \quad (6.13)$$

and represents the signal that is available in the digital baseband, for signal processing and data detection.

Downlink Transmission

In the DL, BS l is supposed to transmit the signal $\mathbf{x}_l = \sum_{i=1}^{K_l} \mathbf{w}_{li} s_{li}$, where $s_{li} \sim \mathcal{N}_{\mathbb{C}}(0, \rho_{li})$ is the data signal intended for UE i in cell l , as defined in Section 2.3.2 on p. 227. The precoding vectors depend on the current channel realizations (or rather on the estimates of the channels). Hence, in a given coherence block with the set $\{\mathbf{h}_{lk}^j\}$ of channel realizations, the precoding vectors are fixed. The signal \mathbf{x}_l is then conditionally complex Gaussian distributed with zero mean and correlation matrix

$$\mathbb{E} \left\{ \mathbf{x}_l \mathbf{x}_l^H \mid \{\mathbf{h}_{lk}^j\} \right\} = \sum_{i=1}^{K_l} \rho_{li} \mathbf{w}_{li} \mathbf{w}_{li}^H. \quad (6.14)$$

We can now apply the impairment model in (6.6) to describe the impairments in the hardware attached to each transmit antenna of BS l . We once again assume that the transceiver hardware of the different BS antennas are decoupled so that the respective distortion terms are independent. Hence, the signal $\check{\mathbf{x}}_l = \sqrt{\kappa_t^{\text{BS}}} \mathbf{x}_l + \boldsymbol{\mu}_l^{\text{BS}}$ is transmitted over the channel instead of \mathbf{x}_l , where the hardware distortion term $\boldsymbol{\mu}_l^{\text{BS}} \in \mathbb{C}^{M_l}$ has the conditional distribution

$$\boldsymbol{\mu}_l^{\text{BS}} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_{M_l}, \mathbf{D}_{l,\{\mathbf{w}\}} \right) \quad (6.15)$$

for the given set $\{\mathbf{w}_{li}\}$ of precoding vectors. The conditional correlation matrix $\mathbf{D}_{l,\{\mathbf{w}\}} \in \mathbb{C}^{M_l \times M_l}$ is given by

$$\mathbf{D}_{l,\{\mathbf{w}\}} = (1 - \kappa_t^{\text{BS}}) \sum_{i=1}^{K_l} \rho_{li} \text{diag} \left(|[\mathbf{w}_{li}]_1|^2, \dots, |[\mathbf{w}_{li}]_{M_l}|^2 \right) \quad (6.16)$$

where $[\mathbf{w}_{li}]_m$ denotes the m th element of \mathbf{w}_{li} . The factor $\kappa_t^{\text{BS}} \in (0, 1]$ determines the quality of the BS's transmitter hardware and is the same for all BSs, for notational convenience. The marginal distribution of the transmitter distortion can then be expressed as

$$\boldsymbol{\mu}_l^{\text{BS}} = \sqrt{1 - \kappa_t^{\text{BS}}} \sum_{i=1}^{K_l} \sqrt{\rho_{li}} \mathbf{w}_{li} \odot \bar{\boldsymbol{\mu}}_{li}^{\text{BS}} \quad (6.17)$$

where $\bar{\boldsymbol{\mu}}_{li}^{\text{BS}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_l}, \mathbf{I}_{M_l})$ is an independent random variable. Note that this transmitter distortion term is the element-wise product between independent complex Gaussian vectors, thus it is not Gaussian distributed.

The signal that reaches UE k in cell j (before receiver noise is added) is $\sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \check{\mathbf{x}}_l$ and, in the given coherence block, it is complex Gaussian distributed with zero mean and conditional variance

$$\begin{aligned} \zeta_{jk,\{\mathbf{w}\}} &= \mathbb{E} \left\{ \left| \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \check{\mathbf{x}}_l \right|^2 \mid \{\mathbf{h}_{lk}^j\} \right\} \\ &= \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \left(\sum_{i=1}^{K_l} \rho_{li} \kappa_t^{\text{BS}} \mathbf{w}_{li} \mathbf{w}_{li}^H + \mathbf{D}_{l,\{\mathbf{w}\}} \right) \mathbf{h}_{jk}^l \\ &= \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \left(\kappa_t^{\text{BS}} |(\mathbf{h}_{jk}^l)^H \mathbf{w}_{li}|^2 + (1 - \kappa_t^{\text{BS}}) \|\mathbf{w}_{li} \odot \mathbf{h}_{jk}^l\|^2 \right) \end{aligned} \quad (6.18)$$

where the second equality follows from (6.16). We can once again apply (6.6) to model impairments in the receiver hardware of UE k in cell j . This changes the received signal to $\sqrt{\kappa_r^{\text{UE}}} \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \check{\mathbf{x}}_l + \mu_{jk}^{\text{UE}}$, where μ_{jk}^{UE} is the hardware distortion term with conditional distribution $\mu_{jk}^{\text{UE}} \sim \mathcal{N}_{\mathbb{C}}(0, (1 - \kappa_r^{\text{UE}}) \zeta_{jk,\{\mathbf{w}\}})$. The marginal distribution is

$$\mu_{jk}^{\text{UE}} = \sqrt{(1 - \kappa_r^{\text{UE}}) \zeta_{jk,\{\mathbf{w}\}}} \bar{\mu}_{jk}^{\text{UE}} \quad (6.19)$$

where $\bar{\mu}_{jk}^{\text{UE}} \sim \mathcal{N}_{\mathbb{C}}(0, 1)$ is an independent random variable. The factor $\kappa_r^{\text{UE}} \in (0, 1]$ determines the quality of the UE's receiver hardware and is the same for all UEs, for notational convenience. Finally, the independent receiver noise $n_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{DL}}^2)$ is added to the signal.

In summary, by taking transceiver hardware impairments into account, the received DL sample $y_{jk} \in \mathbb{C}$ at UE k in cell j is modeled as

$$y_{jk} = \sqrt{\kappa_r^{\text{UE}}} \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \left(\sqrt{\kappa_t^{\text{BS}}} \sum_{i=1}^{K_l} \mathbf{w}_{li} \varsigma_{li} + \boldsymbol{\mu}_l^{\text{BS}} \right) + \mu_{jk}^{\text{UE}} + n_{jk} \quad (6.20)$$

and represents the signal that is available in the digital baseband, for signal processing and data detection.

6.2 Channel Estimation with Hardware Impairments

The UL channel estimation is carried out as described in Section 3.1 on p. 244, with the only difference that the received signal $\mathbf{Y}_j^p \in \mathbb{C}^{M_j \times \tau_p}$ at BS j now contains distortion from hardware impairments. Recall that $\phi_{li} \in \mathbb{C}^{\tau_p}$ is the pilot sequence used by UE i in cell l . When it is transmitted over τ_p instances of the UL model in (6.13), we have³

$$\mathbf{Y}_j^p = \sqrt{\kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbf{h}_{li}^j \left(\sqrt{p_{li} \kappa_t^{\text{UE}}} \phi_{li}^T + (\boldsymbol{\eta}_{li}^{\text{UE}})^T \right) + \mathbf{G}_j^{\text{BS}} + \mathbf{N}_j^p \quad (6.21)$$

where $\mathbf{N}_j^p \in \mathbb{C}^{M_j \times \tau_p}$ is receiver noise whose elements are i.i.d. as $\mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{UL}}^2)$. The transmitter distortion $\boldsymbol{\eta}_{li}^{\text{UE}} \in \mathbb{C}^{\tau_p}$ contains τ_p independent realizations of η_{li}^{UE} in (6.13), thus $\boldsymbol{\eta}_{li}^{\text{UE}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{\tau_p}, (1 - \kappa_t^{\text{UE}}) p_{li} \mathbf{I}_{\tau_p})$. Each column of the receiver distortion matrix $\mathbf{G}_j^{\text{BS}} \in \mathbb{C}^{M_j \times \tau_p}$ has the same distribution as $\boldsymbol{\eta}_j^{\text{BS}}$ in (6.13). More precisely, for a given set of channel realizations $\{\mathbf{h}\}$, the columns are independently distributed as $\mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \mathbf{D}_{j,\{\mathbf{h}\}})$.

When BS j estimates the channel \mathbf{h}_{li}^j from UE i in cell l , it first correlates \mathbf{Y}_j^p with the pilot sequence ϕ_{li} used by this UE, which results

³The system model in (6.13) assumes that the transmitted signals are Gaussian distributed, while we use it here for transmission of deterministic pilot sequences. This issue can be resolved by rotating the pilot book by a Haar distributed matrix, which turns each pilot sequence into a scaled i.i.d. Gaussian vector. Since the estimation performance, in that case, is the same as with the simplified model considered in this section, we have for simplicity omitted the rotation matrices.

in $\mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^*$. For example, for the k th UE in cell j we get

$$\begin{aligned} \mathbf{y}_{jjk}^p = \mathbf{Y}_j^p \phi_{jk}^* &= \underbrace{\sqrt{p_{jk}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}}\tau_p \mathbf{h}_{jk}^j}_{\text{Desired pilot}} + \underbrace{\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \sqrt{p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}}\tau_p \mathbf{h}_{li}^j}_{\text{Interfering pilots}} \\ &\quad + \underbrace{\sum_{l=1}^L \sum_{i=1}^{K_l} \sqrt{\kappa_r^{\text{BS}}} \mathbf{h}_{li}^j (\boldsymbol{\eta}_{li}^{\text{UE}})^T \phi_{jk}^*}_{\text{Transmitter distortion}} + \underbrace{\mathbf{G}_j^{\text{BS}} \phi_{jk}^*}_{\text{Receiver distortion}} + \underbrace{\mathbf{N}_j^p \phi_{jk}^*}_{\text{Noise}}. \end{aligned} \quad (6.22)$$

Since the pilot sequences are deterministic and $\|\phi_{jk}\|^2 = \tau_p$, we have $(\boldsymbol{\eta}_{li}^{\text{UE}})^T \phi_{jk}^* \sim \mathcal{N}_{\mathbb{C}}(0, \tau_p(1 - \kappa_t^{\text{UE}})p_{li})$, $\mathbf{N}_j^p \phi_{jk}^* \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \sigma_{\text{UL}}^2 \tau_p \mathbf{I}_{M_j})$, and the conditional distribution $\mathbf{G}_j^{\text{BS}} \phi_{jk}^* \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \tau_p \mathbf{D}_{j,\{\mathbf{h}\}})$ given $\{\mathbf{h}\}$.

The processed signal \mathbf{y}_{jjk}^p in (6.22) will be used to estimate \mathbf{h}_{jk}^j . As expected, \mathbf{y}_{jjk}^p depends on the channel from UE i in cell l to BS j , for all $(l, i) \in \mathcal{P}_{jk}$, which are the UEs that use the same pilot. In fact, $\mathbf{y}_{jli}^p = \mathbf{y}_{jjk}^p$ for $(l, i) \in \mathcal{P}_{jk}$, thus the same processed signal can be used for estimating \mathbf{h}_{li}^j . In addition, \mathbf{y}_{jjk}^p is affected by the transmitter distortion from all UEs in the entire network, since the random distortion terms are (almost surely) non-orthogonal to the pilot sequences. The receiver distortion also depends on the transmissions from all UEs. This implies that, with hardware impairments, there is pilot contamination between every UE.

Channel estimation is theoretically more challenging with hardware impairments because it is not only the desired pilot term in (6.22) that depends on the channel realization but also the distortion terms. In particular, the received signal \mathbf{y}_{jli}^p is not Gaussian distributed, since some of the terms are formed as products between two Gaussian variables. Hence, the standard MMSE estimation results for Gaussian distributed channels that are observed in independent Gaussian noise cannot be applied here. In principle, we could compute another MMSE estimation expression for the scenario at hand, but since we observed in Section 4.2.3 on p. 313 and Section 4.3.5 on p. 332 that suboptimal estimators can be used with little performance loss, we take the more practical approach of computing the LMMSE estimator. The linear estimator $\hat{\mathbf{h}}_{li}^j$ that minimizes the MSE $\mathbb{E}\{\|\mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j\|^2\}$ takes the following form.

Theorem 6.1. With hardware impairments, the LMMSE estimate of \mathbf{h}_{li}^j , based on $\mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^*$, is

$$\hat{\mathbf{h}}_{li}^j = \sqrt{p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}} \mathbf{R}_{li}^j \Psi_{li}^j \mathbf{y}_{jli}^p \quad (6.23)$$

where

$$\begin{aligned} \Psi_{li}^j &= \left(\sum_{(l', i') \in \mathcal{P}_{li}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \tau_p \mathbf{R}_{l'i'}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right. \\ &\quad \left. + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} (1 - \kappa_t^{\text{UE}}) \kappa_r^{\text{BS}} \mathbf{R}_{l'i'}^j + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} (1 - \kappa_r^{\text{BS}}) \mathbf{D}_{\mathbf{R}_{l'i'}^j} \right)^{-1} \end{aligned} \quad (6.24)$$

and

$$\mathbf{D}_{\mathbf{R}_{l'i'}^j} = \text{diag} \left([\mathbf{R}_{l'i'}^j]_{11}, \dots, [\mathbf{R}_{l'i'}^j]_{M_j M_j} \right) \quad (6.25)$$

is an $M_j \times M_j$ diagonal matrix with the diagonal elements from $\mathbf{R}_{l'i'}^j$.

The estimation error $\tilde{\mathbf{h}}_{li}^j = \mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j$ has the correlation matrix $\mathbf{C}_{li}^j = \mathbb{E}\{\tilde{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H\}$ given by

$$\mathbf{C}_{li}^j = \mathbf{R}_{li}^j - p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}\tau_p \mathbf{R}_{li}^j \Psi_{li}^j \mathbf{R}_{li}^j \quad (6.26)$$

while the LMMSE estimate satisfies

$$\mathbb{E}\{\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H\} = \mathbf{R}_{li}^j - \mathbf{C}_{li}^j = p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}\tau_p \mathbf{R}_{li}^j \Psi_{li}^j \mathbf{R}_{li}^j. \quad (6.27)$$

Proof. The proof is available in Appendix C.5.1 on p. 612. \square

The LMMSE estimator with hardware impairment in (6.23) has a structure similar to that of the MMSE estimator for ideal hardware in Theorem 3.1 on p. 249. An important difference is that the transmit power p_{li} of the UE is reduced to $p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}$, since only a fraction $\kappa_t^{\text{UE}}\kappa_r^{\text{BS}} \leq 1$ of the power is received without distortion. Moreover, the matrix Ψ_{li}^j contains not only receiver noise and interference from UEs that reuse the same pilot sequence, but also hardware distortion caused by the signaling from all UEs in the entire network.

The LMMSE estimate $\hat{\mathbf{h}}_{li}^j$ in (6.23) is computed by multiplying the observation \mathbf{y}_{jli}^p with the deterministic matrix $\sqrt{p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}} \mathbf{R}_{li}^j \Psi_{li}^j$. Since

\mathbf{y}_{jli}^p is not Gaussian distributed, the estimate is not Gaussian either. The estimate $\hat{\mathbf{h}}_{li}^j$ and the estimation error $\tilde{\mathbf{h}}_{li}^j$ are uncorrelated by the definition of LMMSE estimation, meaning that $\mathbb{E}\{\hat{\mathbf{h}}_{li}^j(\tilde{\mathbf{h}}_{li}^j)^H\} = \mathbf{0}_{M_j \times M_j}$, but they are not independent. This is different from the MMSE estimator computed with ideal hardware, which provides Gaussian distributed estimates and then the fact that the estimate and the estimation error are uncorrelated implies that they are also independent. The different behavior with hardware impairments has no major practical implications, but it is analytically important when we compute SEs in Section 6.3.

The special case of ideal hardware is represented by $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 1$. In this case, the LMMSE estimator in (6.23) coincides with the MMSE estimator for ideal hardware in Theorem 3.1.

6.2.1 Impact of Hardware Impairments on Channel Estimation

We will now illustrate the basic impact that hardware impairments have on the channel estimation by considering a single-cell single-UE setup where the channel is denoted by $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R})$ and has the correlation matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$. The estimation error correlation matrix in (6.26) then becomes

$$\mathbf{C} = \mathbf{R} - p\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}\tau_p \mathbf{R} \boldsymbol{\Psi} \mathbf{R} \quad (6.28)$$

where

$$\boldsymbol{\Psi} = \left(p(1 + \kappa_t^{\text{UE}}(\tau_p - 1))\kappa_r^{\text{BS}} \mathbf{R} + p(1 - \kappa_r^{\text{BS}}) \mathbf{D}_{\mathbf{R}} + \sigma_{\text{UL}}^2 \mathbf{I}_M \right)^{-1}. \quad (6.29)$$

The expression in (6.28) is different from $\mathbf{R} - \mathbf{R}(\mathbf{R} + \frac{\sigma_{\text{UL}}^2}{p\tau_p} \mathbf{I}_M)^{-1} \mathbf{R}$, which is obtained with ideal hardware (i.e., $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 1$). The difference is particularly visible at high SNR, when $p \rightarrow \infty$, since the error correlation matrix in (6.28) approaches

$$\mathbf{C} = \mathbf{R} - \mathbf{R} \left(\frac{1 + \kappa_t^{\text{UE}}(\tau_p - 1)}{\kappa_t^{\text{UE}}\tau_p} \mathbf{R} + \frac{(1 - \kappa_r^{\text{BS}})}{\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}\tau_p} \mathbf{D}_{\mathbf{R}} \right)^{-1} \mathbf{R} \quad (6.30)$$

which is equal to $\mathbf{R} - \mathbf{R}\mathbf{R}^{-1}\mathbf{R} = \mathbf{0}_{M \times M}$ for $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 1$, but otherwise is non-zero. In other words, asymptotically error-free channel

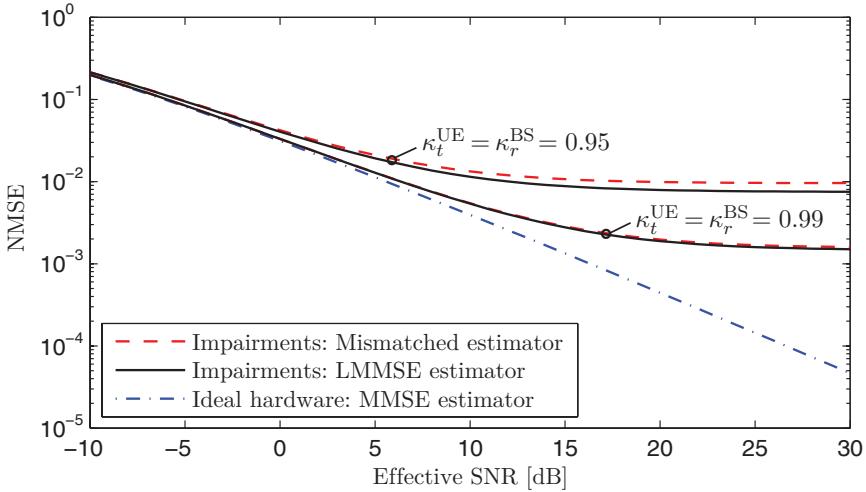


Figure 6.2: NMSE of a spatially correlated channel with hardware impairments, based on the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$. The result is averaged over different nominal angles.

estimates are achieved at high SNR with ideal hardware, while there is a non-zero estimation error floor with hardware impairments.

If we further assume $\mathbf{R} = \beta \mathbf{I}_M$, the error correlation matrix in (6.30) simplifies to

$$\begin{aligned} \mathbf{C} &= \beta \mathbf{I}_M - \frac{\beta^2}{\frac{1+\kappa_t^{\text{UE}}(\tau_p-1)}{\kappa_t^{\text{UE}}\tau_p}\beta + \frac{(1-\kappa_r^{\text{BS}})}{\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}\tau_p}\beta} \mathbf{I}_M \\ &= \frac{\beta(1-\kappa_t^{\text{UE}}\kappa_r^{\text{BS}})}{1+\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}(\tau_p-1)} \mathbf{I}_M. \end{aligned} \quad (6.31)$$

The expression in front of the identity matrix is the error floor and it is determined by the hardware quality (κ_t^{UE} and κ_r^{BS}), the length of the pilot sequence τ_p , and the average channel gain β . One straightforward way to lower the error floor is to increase τ_p . Note that the transmitter and receiver distortion have an identical impact on \mathbf{C} in this case. This is not the case when there is spatial channel correlation.

The average NMSE, $\text{tr}(\mathbf{C})/\text{tr}(\mathbf{R})$, is shown in Figure 6.2 for correlation matrices generated by the local scattering model, defined in (2.23), with Gaussian angular distribution, ASD $\sigma_\varphi = 10^\circ$, and uniformly distributed nominal angle. The figure shows the NMSE as a function

of the effective SNR $p\tau_p/\sigma_{\text{UL}}^2$ for $\tau_p = 10$ and with different hardware qualities. The upper curves have $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.95$, the middle curves have $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.99$, and the bottom curve has ideal hardware (i.e., $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 1$). The NMSE is shown for the LMMSE estimator from Theorem 6.1 and for a mismatched estimator that ignores the hardware impairments (i.e., we use the estimator from Theorem 3.1, which is not the MMSE estimator with hardware impairments).

Figure 6.2 confirms that there are error floors in the channel estimation at high SNR and that the error floor increases as the hardware quality decreases. The impact of hardware impairments is small at low SNR, but substantial at high SNR. The NMSE with hardware impairments is close to the error floor already at an effective SNR of 20 dB, thus a further increase in SNR only brings minor improvements. An effective SNR of 20–30 dB is sufficient to achieve nearly the maximal estimation quality with hardware impairments. Another key observation is that the mismatched estimator provides almost the same NMSE as the LMMSE estimator. This indicates that although hardware impairments greatly affect the estimation quality, the loss of using a simple estimator that ignores the impairments is minor.

6.2.2 Estimation with Interference and Hardware Impairments

To showcase the joint impact of hardware distortion and inter-cell interference, we continue the running example defined in Section 4.1.3 on p. 288 with $M = 100$, $K = 10$, and UL power 20 dBm per UE. We consider the NMSE, $\text{tr}(\mathbf{C}_{jk}^j)/\text{tr}(\mathbf{R}_{jk}^j)$, from an arbitrary BS j to an arbitrary UE k in its cell. The results are presented as CDF curves, where the randomness is the UE locations and shadow fading realizations. We consider the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$.

Figure 6.3 shows the CDFs with either ideal hardware or hardware impairments with $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.95$. In both cases, the NMSE reduces substantially as the pilot reuse factors f is increased. This is natural since the pilot processing gain τ_p is increased and the inter-cell interference is reduced. The distortion from hardware impairments has the opposite effect of increasing the NMSE, since the transmitter distortions break the orthogonality of the pilot sequences, leading to interference from all

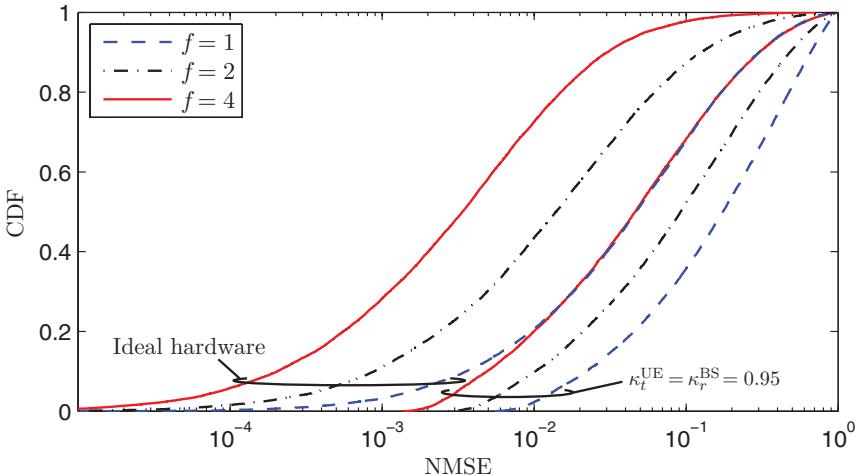


Figure 6.3: CDF of the NMSE in the channel estimation for an arbitrary UE in the running example. The results are shown for different UL hardware qualities ($\kappa_t^{\text{UE}}, \kappa_r^{\text{BS}}$) and pilot reuse factors f .

UEs, including itself. The conclusion is that hardware impairments can greatly degrade the channel estimation in cellular networks and might be a more critical performance-reducing factor than conventional pilot contamination.

6.3 Spectral Efficiency with Hardware Impairments

The achievable SEs in systems with hardware impairments will now be analyzed. SE expressions and numerical results will be provided to show how hardware distortion affects the performance. This will reveal that the quality of the UE and the BS hardware have a very different impact on the SE.

6.3.1 Uplink SE Expressions

In the UL, BS j selects a receive combining vector $\mathbf{v}_{jk} \in \mathbb{C}^{M_j}$ for its k th UE, similar to Section 4.1 on p. 275. By applying this vector to the

received signal $\mathbf{y}_j \in \mathbb{C}^{M_j}$ in (6.13), the BS obtains

$$\begin{aligned}
\mathbf{v}_{jk}^H \mathbf{y}_j &= \sqrt{\kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j (\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}}) + \mathbf{v}_{jk}^H \boldsymbol{\eta}_j^{\text{BS}} + \mathbf{v}_{jk}^H \mathbf{n}_j \\
&= \underbrace{\sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} s_{jk}}_{\text{Desired signal over average channel}} \\
&\quad + \underbrace{\sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} (\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}) s_{jk}}_{\text{Self-interference}} + \underbrace{\sqrt{\kappa_r^{\text{BS}}} \mathbf{v}_{jk}^H \mathbf{h}_{jk}^j \eta_{jk}^{\text{UE}}}_{\text{Self-distortion}} \\
&\quad + \underbrace{\sqrt{\kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j \left(\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}} \right)}_{\text{Inter-user interference and transmitter distortion}} + \underbrace{\mathbf{v}_{jk}^H \boldsymbol{\eta}_j^{\text{BS}}}_{\text{Receiver distortion}} + \underbrace{\mathbf{v}_{jk}^H \mathbf{n}_j}_{\text{Noise}}. \tag{6.32}
\end{aligned}$$

The signal consists of the desired signal, self-interference, self-distortion, interference, transmitter/receiver distortion, and noise, as indicated in (6.32). The second equality follows from adding and then subtracting the desired signal that is received over the average effective channel $\sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}$. With this formulation, we can compute a UatF bound on the channel capacity of the UE, as on p. 302, by treating everything that is not received over the average effective channel as worst-case noise. The UatF bound is considered since there is no simple counterpart to the tighter UL bound in Theorem 4.1 on p. 276. That bound requires the estimation error to have zero conditional mean, for every given channel realization, which might not be the case when the estimate and estimation error are statistically dependent.

Theorem 6.2. With hardware impairments, the UL ergodic channel capacity of UE k in cell j is lower bounded by

$$\overline{\mathsf{SE}}_{jk}^{\text{UL-imp}} = \frac{\tau_u}{\tau_c} \log_2 \left(1 + \overline{\mathsf{SINR}}_{jk}^{\text{UL-imp}} \right) \tag{6.33}$$

with

$$\overline{\text{SINR}}_{jk}^{\text{UL-imp}} = \frac{p_{jk} \frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}}{\sum_{l,i} p_{li} \left(\frac{\kappa_r^{\text{BS}} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} + (1 - \kappa_r^{\text{BS}}) \mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\}}{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} \right) - p_{jk} \frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} + \frac{\sigma_{\text{UL}}^2}{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}}} \quad (6.34)$$

where the expectations are with respect to the channel realizations.

Proof. The proof is available in Appendix C.5.2 on p. 614. \square

This theorem provides an achievable UL SE for systems with hardware impairments. The SE is measured in bit/s/Hz, as in previous sections, and this implicitly assumes that the signal's bandwidth is kept constant under the presence of hardware impairments. Since hardware impairments generally lead to spectral regrowth, in practice, one might need to reduce the bandwidth to comply with out-of-band radiation regulations, depending on how these are defined; see Section 6.4.3 for a further discussion. The expression can be computed numerically for any receive combining scheme and any spatial correlation matrices. The same schemes as presented for ideal hardware impairments in Section 4.1.1 on p. 281 (e.g., M-MMSE, RZF, and MR) can be applied also with hardware impairments, as will be demonstrated later.

There are many similarities between (6.34) and the corresponding effective SINR expression with ideal hardware in (4.14). The numerator contains the signal power that is useful for detection, while the denominator contains the total received power minus the expression from the numerator. One key difference is that the desired signal power is reduced by a factor $\kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \in (0, 1]$, which in (6.34) is represented by multiplying the interference and noise terms by $1/(\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}) \geq 1$. This factor describes the SNR loss caused by desired signals being turned into distortion. Another key difference is that a fraction $(1 - \kappa_r^{\text{BS}})$ of the conventional interference term $\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}$ is replaced by the alternative term $\mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\}$, where the inner product is changed to an element-wise multiplication. The latter term can be substantially smaller than the former; for example, suppose $\mathbf{v}_{jk} = \mathbf{h}_{li}^j = \mathbf{1}_{M_j}$, then

$|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2 = |M_j|^2 = M_j^2$ while $\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2 = \|\mathbf{1}_{M_j}\|^2 = M_j$. The intuition is that the former term adds the signals coherently in the amplitude domain, while the latter only sums up the average power per antenna. This indicates that having hardware impairments in the receiving BS can also reduce the interference level.

In addition to changing the structure of the SINR, the combining vectors are also affected since they typically depend on channel estimates that are already degraded by hardware distortion. The expectations in (6.34) can sometimes be computed in closed form with MR combining; the case with diagonal spatial correlation matrices was considered in [53, 376], while arbitrary correlation matrices were treated in [42]. The general impact of spatial channel correlation was analyzed in Section 4. For this reason, we focus on spatially uncorrelated channels in this section to show concisely how hardware impairments affect the SE.

Corollary 6.3. If MR combining with $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j$ is used, based on the LMMSE estimator in Theorem 6.1, and the channels are spatially uncorrelated (i.e., $\mathbf{R}_{li}^j = \beta_{li}^j \mathbf{I}_{M_j}$ for $l = 1, \dots, L$ and $i = 1, \dots, K_l$), then

$$\frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} = p_{jk} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j \quad (6.35)$$

$$\begin{aligned} \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} &= \beta_{li}^j + p_{li} (\beta_{li}^j)^2 \psi_{jk} \left(1 - \kappa_r^{\text{BS}} + (1 - \kappa_t^{\text{UE}}) \kappa_r^{\text{BS}} M_j\right) \\ &+ \begin{cases} p_{li} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\beta_{li}^j)^2 \tau_p \psi_{jk} M_j & (l, i) \in \mathcal{P}_{jk} \\ 0 & (l, i) \notin \mathcal{P}_{jk} \end{cases} \end{aligned} \quad (6.36)$$

$$\begin{aligned} \frac{\mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} &= \beta_{li}^j + p_{li} (\beta_{li}^j)^2 \psi_{jk} \left(1 - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}\right) \\ &+ \begin{cases} p_{li} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\beta_{li}^j)^2 \tau_p \psi_{jk} & (l, i) \in \mathcal{P}_{jk} \\ 0 & (l, i) \notin \mathcal{P}_{jk} \end{cases} \end{aligned} \quad (6.37)$$

where

$$\psi_{jk} = \frac{1}{\sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \tau_p \beta_{l'i'}^j + \sum_{l', i'} p_{l'i'} (1 - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}) \beta_{l'i'}^j + \sigma_{\text{UL}}^2}. \quad (6.38)$$

The SE in Theorem 6.2 becomes $\overline{\text{SE}}_{jk}^{\text{UL-imp}} = \frac{\tau_u}{\tau_c} \log_2(1 + \overline{\text{SINR}}_{jk}^{\text{UL-imp}})$ with

$$\begin{aligned} \overline{\text{SINR}}_{jk}^{\text{UL-imp}} = & \frac{(p_{jk}\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j}{\sum_{l,i} p_{li}\beta_{li}^j \overline{F}_{li}^{jk} + \sum_{(l,i) \in \mathcal{P}_{jk}} (p_{li}\beta_{li}^j)^2 \tau_p \psi_{jk} M_j \overline{G}_j - (p_{jk}\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j + \frac{\sigma_{\text{UL}}^2}{(\kappa_t^{\text{UE}} \kappa_r^{\text{BS}})^2}} \end{aligned} \quad (6.39)$$

where

$$\overline{F}_{li}^{jk} = \frac{1 + p_{li}\beta_{li}^j \psi_{jk} \left(1 - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} + (1 - \kappa_t^{\text{UE}})(\kappa_r^{\text{BS}})^2(M_j - 1)\right)}{(\kappa_t^{\text{UE}} \kappa_r^{\text{BS}})^2} \quad (6.40)$$

$$\overline{G}_j = \frac{1 + \kappa_r^{\text{BS}}(M_j - 1)}{M_j \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}}. \quad (6.41)$$

Proof. The proof is available in Appendix C.5.3 on p. 615. \square

The SE expression in Corollary 6.3 generalizes the previous expression in Corollary 4.5 on p. 303 to include hardware impairments. The general structure is otherwise the same; the signal term in the numerator of the SINR is identical, while the denominator contains noise as well as non-coherent and coherent interference terms, where the latter refers to the terms that grow with M_j . The noise term has effectively increased by a factor $1/(\kappa_t^{\text{UE}} \kappa_r^{\text{BS}})^2 \geq 1$, which does not mean that the noise itself has grown, but that the signal power has been reduced by $(\kappa_t^{\text{UE}} \kappa_r^{\text{BS}})^2$. This type of “squaring effect” also appeared in the transmit power-scaling law in Section 5.2.1 on p. 359. In this section, it is the combined effect of losing a factor $1 - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}$ of the signal power during both channel estimation and data transmission.

The conventional non-coherent interference term $\sum_{l,i} p_{li}\beta_{li}^j$ is replaced by $\sum_{l,i} p_{li}\beta_{li}^j \overline{F}_{li}^{jk}$, where the factor \overline{F}_{li}^{jk} is an affine increasing function of M_j . This implies that all UEs, irrespective of their pilot sequences, cause some coherent interference to each other. The reason is that the distortion breaks the orthogonality between the pilot sequences. The term $\sum_{(l,i) \in \mathcal{P}_{jk}} (p_{li}\beta_{li}^j)^2 \tau_p \psi_{jk} M_j \overline{G}_j$ represents the additional interference from UEs that use the same pilot as UE k in cell j .

This term contains coherent interference caused by conventional pilot contamination, but it also contains a small portion of non-coherent interference (the first term in \bar{G}_j does not grow with M_j) since the pilot contamination reduces when a distorted version of the pilot is transmitted.

The desired UE causes coherent interference to itself when there is hardware impairments, which we refer to as *self-distortion*. For example, there are two terms in the denominator of (6.39) that grow with M_j and originate from the UE itself: $(p_{li}\beta_{jk}^j)^2\tau_p\psi_{jk}M_j\bar{G}_j$ in the second sum and $-(p_{jk}\beta_{jk}^j)^2\tau_p\psi_{jk}M_j$. These terms cancel out for $\bar{G}_j = 1$, but with hardware impairments we have $\bar{G}_j > 1$. The coherent self-distortion originates from the fact that the BS coherently combines both the desired signal and the transmitter distortion that the UE caused during data transmission.

In summary, hardware impairments reduce the effective signal power, reduce the coherent interference from UEs having the same pilot, and add coherent interference from all other UEs. To further study the coherent interference characteristics, we consider the asymptotic regime with a very large number of BS antennas.

Corollary 6.4. Under the same conditions as in Corollary 6.3, $\overline{\text{SINR}}_{jk}^{\text{UL-imp}}$ with MR combining has the asymptotic limit

$$\frac{(p_{jk}\beta_{jk}^j)^2}{\sum_{l,i}(p_{li}\beta_{li}^j)^2\frac{1-\kappa_t^{\text{UE}}}{(\kappa_t^{\text{UE}})^2\tau_p} + \sum_{(l,i)\in\mathcal{P}_{jk}\setminus(j,k)}(p_{li}\beta_{li}^j)^2\frac{1}{\kappa_t^{\text{UE}}} + (p_{jk}\beta_{jk}^j)^2\frac{1-\kappa_t^{\text{UE}}}{\kappa_t^{\text{UE}}}} \quad (6.42)$$

as $M_j \rightarrow \infty$.

Proof. This result follows from taking the limit in (6.39) and noting that $\bar{F}_{li}^j/M_j \rightarrow p_{li}\beta_{li}^j\psi_{jk}\frac{1-\kappa_t^{\text{UE}}}{(\kappa_t^{\text{UE}})^2}$ and $\bar{G}_j \rightarrow 1/\kappa_t^{\text{UE}}$. \square

This corollary shows that the noise and non-coherent interference vanish asymptotically when the BS has many antennas, just as with ideal hardware. Interestingly, also the impact of hardware impairments at the receiving BS vanishes, as seen from the fact that the limit in (6.42) is independent of κ_r^{BS} . The reason that the receiver distortion is

combined non-coherently by MR is that the distortion vector points in a direction that is asymptotically orthogonal to the channel, which is similar to the concept of asymptotically favorable propagation (see Section 2.5.2 on p. 233). The first interference term in the denominator of (6.42) is caused by the imperfect pilot orthogonality due to distortion. This term is proportional to $1/\tau_p$, which indicates that systems with long pilot sequences are less affected by such hardware distortion.

The second term in the denominator is the coherent interference due to reuse of pilots and it is a factor $1/\kappa_t^{\text{UE}}$ larger than when having ideal hardware. It is actually not the interference that has increased but the desired signal that has decreased. Note that the transmitter distortion does not change the interference power during data transmission, but only the content of the interfering signals, which is not important for a receiver that treats interference as noise. The last term is the coherent self-distortion. This term remains even if all other UEs are silent, thus the asymptotic limit in (6.42) can be upper bounded as

$$\overline{\text{SINR}}_{jk}^{\text{UL-imp}} \leq \frac{(p_{jk}\beta_{jk}^j)^2}{(p_{jk}\beta_{jk}^j)^2 \frac{1-\kappa_t^{\text{UE}}}{\kappa_t^{\text{UE}}}} = \frac{\kappa_t^{\text{UE}}}{1 - \kappa_t^{\text{UE}}}. \quad (6.43)$$

Larger effective SINRs than this cannot be achieved with hardware impairments, with the model and MR scheme considered in this section. The SINR limit is 166 for a high-quality transceiver with $\kappa_t^{\text{UE}} = 0.99$, which corresponds to 7.4 bit/s/Hz and can be achieved in practice using 256-QAM signaling with a high coding rate. For a very low-quality transceiver with $\kappa_t^{\text{UE}} = 0.9$, the SINR limit is 9 and the corresponding SE is 3.3 bit/s/Hz, which is achievable by 16-QAM and a high coding rate. These numbers give an indication of the maximum modulation sizes that are useful in practice.

6.3.2 Impact of Hardware Impairment on UL SE

To quantify the impact that hardware distortions have on the SE in systems with a practical number of antennas, we continue the running example that was defined in Section 4.1.3 on p. 288. We consider $M = 100$ and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. We use Theorem 6.2 to compute the SE with M-MMSE, RZF, and MR.

The MR scheme is normalized as $\mathbf{v}_{jk} = \hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|^2$, which was shown in Section 4.2.1 on p. 306 to provide the tightest UatF bound. Except for pilots, all samples in every coherence block are used for UL data. The pilot reuse factor that maximizes the SE is considered.

Figure 6.4 shows the average sum SE as a function of the hardware quality factors $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} \in [0.9, 1]$, which are assumed to be equal for simplicity. We consider $K = 10$ and $K = 20$ UEs. Recall from Section 6.1.2 that hardware qualities in the range from 0.97 to 1 are typical with today's hardware. A substantial loss in SE is observed in this interval, where the steepest loss is from 1 to 0.99. M-MMSE and RZF are particularly sensitive to the distortion that is added when the hardware quality reduces, while MR is less affected. The explanation is that M-MMSE and RZF suppress interference, which makes the unsuppressed distortion stronger relative to the interference. In contrast, MR is already limited by high levels of interference. The fact that systems that achieve higher SEs require better hardware is fully in line with the LTE requirements (see Section 6.1.2). The relative loss in SE is higher with $K = 10$ than with $K = 20$, because in the latter case there is more interference. In other words, a system that provides low SE to many UEs is less affected by hardware impairments than a system that provides high SE to fewer UEs. Note that M-MMSE and RZF continue to provide substantial gains in SE over MR over the entire range of hardware qualities.

6.3.3 Comparison of Interference and Distortion Sources

The terms in the UL effective SINR expression in (6.34) can be decomposed into the six components:

1. Desired signal: $\frac{p_{jk}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}}{\sigma_{\text{UL}}^2} \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}$;
2. Interference from UEs having the same pilot:

$$\sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}}{\sigma_{\text{UL}}^2} \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}$$
;
3. Interference from UEs having different pilots:

$$\sum_{(l,i) \notin \mathcal{P}_{jk}} \frac{p_{li}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}}{\sigma_{\text{UL}}^2} \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}$$
;

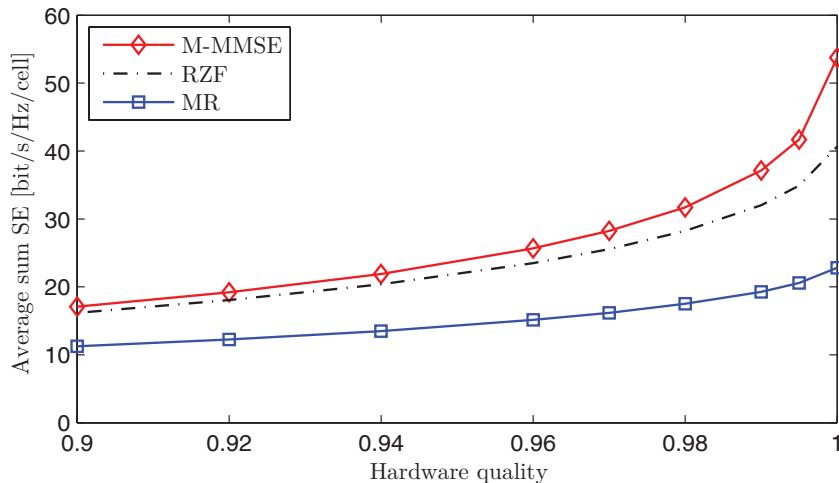
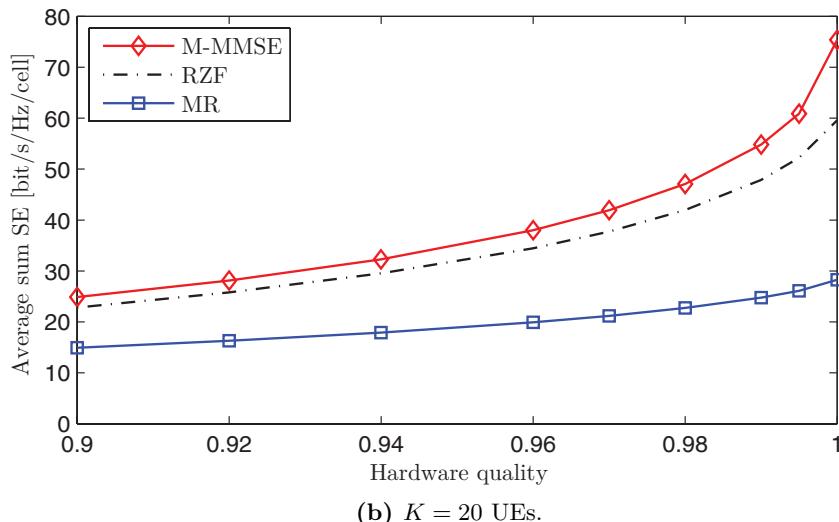
(a) $K = 10$ UEs.(b) $K = 20$ UEs.

Figure 6.4: Average UL sum SE as a function of the hardware quality $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}}$, using the local scattering model with Gaussian angular distribution. We consider $M = 100$ and $K \in \{10, 20\}$, and for each point on the curves we use the pilot reuse factor that maximizes the SE.

4. Transmitter distortion: $\sum_{(l,i) \neq (j,k)} \frac{p_{li}(1-\kappa_t^{\text{UE}})\kappa_r^{\text{BS}}}{\sigma_{\text{UL}}^2} \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}};$
5. Receiver distortion: $\sum_{l,i} \frac{p_{li}(1-\kappa_r^{\text{BS}})}{\sigma_{\text{UL}}^2} \frac{\mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}};$
6. Self-distortion/interference: $\frac{p_{jk}\kappa_r^{\text{BS}}}{\sigma_{\text{UL}}^2} \frac{(\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j|^2\} - \kappa_t^{\text{UE}} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j|\}^2)}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}.$

Note that these power terms have been normalized with respect to the noise power and will be measured in dB. The term called self-distortion/interference contains both the transmitter distortion that the UE causes to itself and the interference that originates from the fact that the BS has imperfect CSI, thus this term will be non-zero even with ideal hardware. We continue the previous example with $M = 100$ and $K = 10$ by analyzing the average power of each of these components.

The average powers (over different UE locations and channel realizations) of the six components are presented in Figure 6.5 for MR, RZF, and M-MMSE combining. For each scheme, we consider three different hardware qualities: $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} \in \{0.95, 0.99, 1\}$. The pilot reuse factor is $f = 2$, which means that each UE is affected by interference from 7 UEs that use the same pilot and 152 UEs that use different pilots. Since there are many more UEs in the second category, if the average interference caused per UE is the same in both categories, the latter category would cause $10 \log_{10}(152/7) \approx 13.4$ dB more interference.

All three combining schemes provide average signal power levels of around 40 dB, with the highest value for MR and the lowest value for M-MMSE, since the latter sacrifices some array gain to suppress interference. With ideal hardware, the (non-coherent) interference from UEs with different pilots dominates the (partially coherent) interference from UEs that reuse the same pilot. The power difference is as much as 24–27 dB for MR, which means that UEs in the own and neighboring cells cause much more interference (on average) than the UEs in more distant cells that reuse the same pilot. With RZF and M-MMSE, the UEs that reuse the pilot actually cause slightly more interference per UE, but since there are much fewer such UEs, their total interference power is still negligible. Quantitatively speaking, the total interference power is 34–36 dB with MR, 18–19 dB with RZF, and 11–16 dB with M-MMSE.

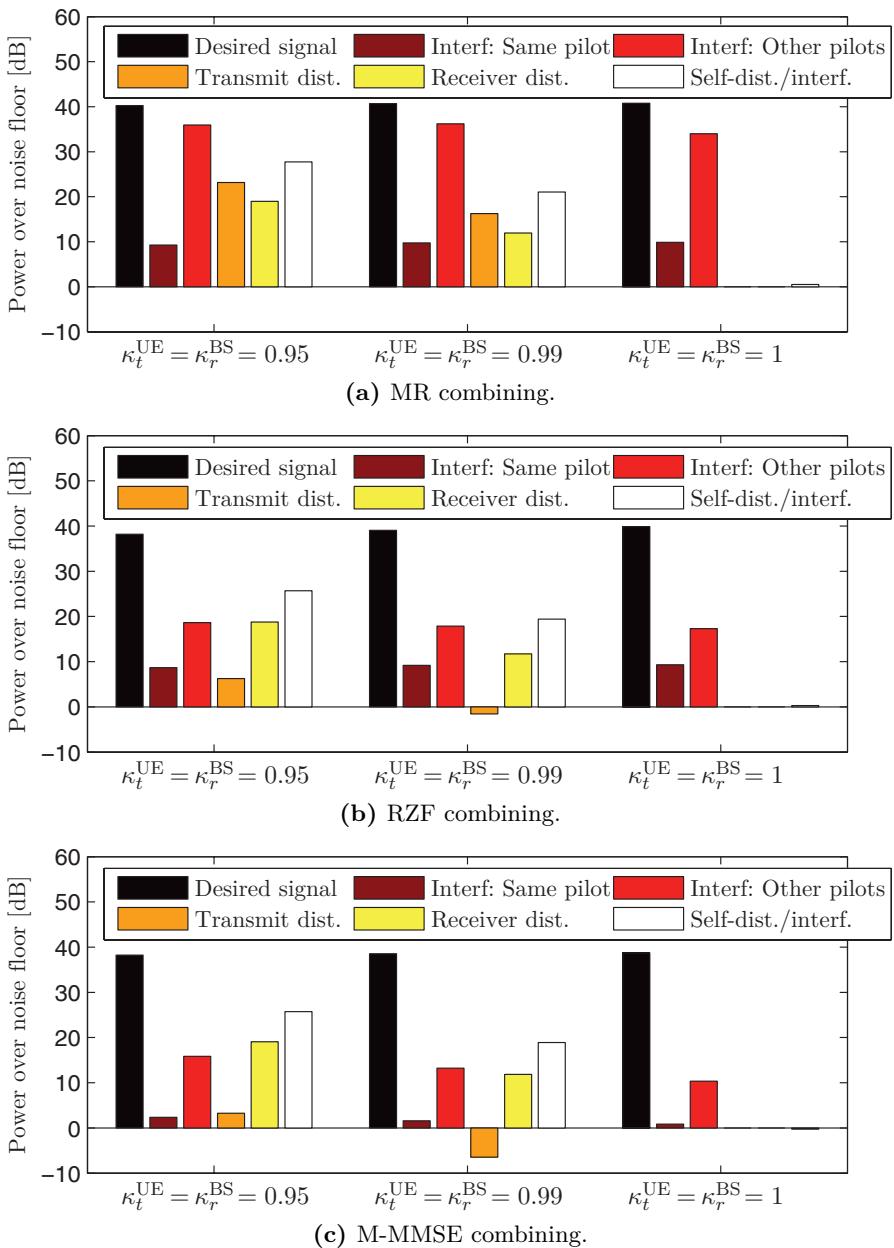


Figure 6.5: Average UL power of desired signal, interference from UEs with same or different pilots, transmitter and receiver distortion, and self-distortion/interference.

The conventional interference terms are little affected by the inclusion of hardware impairments, but three new interference components are added: transmitter distortion, receiver distortion, and self-distortion, where the latter refers to the transmitter distortion that the UE causes to itself and it is lumped together with the conventional self-interference caused by having imperfect CSI. Figure 6.5 shows that the transmitter distortion power is potentially very different between the combining schemes. The transmitter distortion has the same spatial directivity as the conventional interference, but it is $\kappa_t^{\text{UE}}/(1 - \kappa_t^{\text{UE}})$ times smaller. Hence, M-MMSE and RZF suppress the transmitter distortion from other UEs in the receive combining, while MR does not. Nonetheless, the conventional interference is $\kappa_t^{\text{UE}}/(1 - \kappa_t^{\text{UE}}) = 19$ times stronger than the transmitter distortion for $\kappa_t^{\text{UE}} = 0.95$, thus showing that one can basically neglect its impact on the SE.

The receiver distortion depends on the total received power, which is dominated by the desired signal power. The self-distortion/interference is also roughly proportional to the signal power. Hence, these terms are almost the same for all schemes, but varies with the hardware quality. In the case of $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.95$, the self-distortion/interference is 26–27 dB and the receiver distortion is around 19 dB. Although these numbers are nearly the same for all schemes, the impact they have on the SE is very different. The conventional interference power with MR is twice as strong as the total distortion and self-distortion/interference (with $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.95$), while with M-MMSE the total distortion and self-distortion/interference power is much stronger (11 times) than the interference. The latter explains the substantial reduction in SE due to hardware impairments that was observed for M-MMSE in Figure 6.4.

In summary, the self-distortion/interference is typically stronger than both the transmitter distortion from other UEs and the receiver distortion when having hardware impairments. The total power of transmitter/receiver distortion and self-distortion/interference is roughly the same for all combining schemes, but M-MMSE and RZF are more susceptible to it since these schemes can greatly suppress the conventional types of interference, but not the self-distortion or receiver distortion. Moreover, we know from the asymptotic analysis that the non-coherent

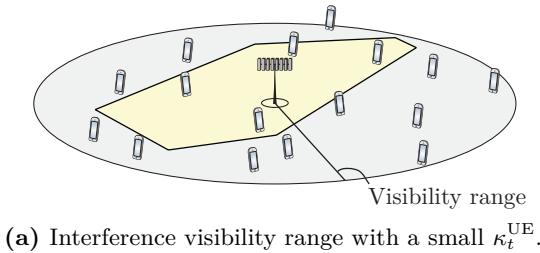
interference and receiver distortion vanish asymptotically, thus leaving the self-distortion as the main limiting factor.

6.3.4 Interference Visibility Range

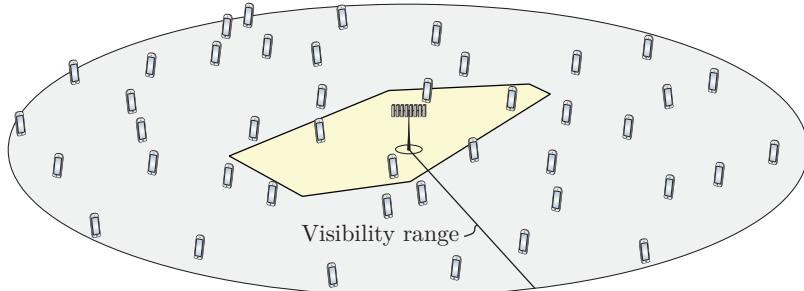
As illustrated in Figure 6.5, the self-distortion can be an important performance limiting factor in practice, since it has the same spatial characteristics as the desired signal and thus cannot be rejected by receive combining. Recall that the self-distortion creates the upper bound in (6.43) on the SINR, even in a single-UE system, because the self-distortion is consistently $(1 - \kappa_t^{\text{UE}})/\kappa_t^{\text{UE}}$ times weaker than the desired signal. Any source of interference or distortion that is substantially weaker than the self-distortion is basically negligible. Qualitatively speaking, for BS j , any interfering UE with a channel gain to BS j that is much more than $10 \log_{10}((1 - \kappa_t^{\text{UE}})/\kappa_t^{\text{UE}})$ dB weaker than the channel gain for a cell-edge UEs in cell j has no practical impact on the UL in cell j . For example, with $\kappa_t^{\text{UE}} = 0.97$ we get $(1 - \kappa_t^{\text{UE}})/\kappa_t^{\text{UE}} \approx -15$ dB. Let $\bar{\beta}_j^j$ denote the average channel gain for a UE at the edge of cell j and let $\bar{\beta}_l^j$ be the average channel gain to BS j from an arbitrary UE in cell $l \neq j$. If $\bar{\beta}_l^j/\bar{\beta}_j^j \ll -15$ dB, then any interference (coherent or non-coherent) from this UE to cell j can be neglected. This can be visualized as an *interference visibility range*; see Figure 6.6. Only the UEs within this range impact the SEs in the center cell, which is an insight to take into account in the resource allocation (e.g., the center cell's pilots can be reused freely outside the interference visibility range). The visibility range shrinks as the UE hardware quality reduces, which thus reduces the need for interference management across cells. If we instead improve the hardware quality of the UEs, motivated by the fact that the SE is limited by it, then the interference visibility range increases and each BS should coordinate its resource allocation with a larger number of adjacent cells.

6.3.5 Downlink SE Expressions

In the DL, we consider the hardening bounding technique, which will provide the baseline performance with hardware impairments. Similar



(a) Interference visibility range with a small κ_t^{UE} .



(b) Interference visibility range with a large κ_t^{UE} .

Figure 6.6: Illustration of the limited interference visibility range that is created by self-distortion. Only UL interference from UEs within this range has an impact on the center cell, while the more distant UEs can be neglected because their interference is substantially smaller than the self-distortion caused inside the center cell.

to (6.32), the received DL signal y_{jk} in (6.20) can be expressed as

$$\begin{aligned}
 y_{jk} = & \underbrace{\sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \varsigma_{jk}}_{\text{Desired signal over average channel}} + \underbrace{\mu_{jk}^{\text{UE}}}_{\text{Self-distortion}} \\
 & + \underbrace{\sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \left((\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \right) \varsigma_{jk}}_{\text{Desired signal over "unknown" channel}} + \underbrace{\sqrt{\kappa_r^{\text{UE}}} \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \boldsymbol{\mu}_l^{\text{BS}}}_{\text{Transmitter distortion}} \\
 & + \underbrace{\sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \sum_{\substack{i=1 \\ i \neq k}}^{K_j} (\mathbf{h}_{jk}^i)^H \mathbf{w}_{ji} \varsigma_{ji}}_{\text{Intra-cell interference}} + \underbrace{\sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \sum_{\substack{l=1 \\ l \neq j}}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li}}_{\text{Inter-cell interference}} + \underbrace{n_{jk}}_{\text{Noise}}
 \end{aligned} \tag{6.44}$$

The first term in (6.44) is the desired signal received over the deterministic average effective channel $\sqrt{\kappa_t^{\text{BS}}\kappa_r^{\text{UE}}}\mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}$, while the remaining terms are random variables with realizations that are unknown to the UE. An achievable SE is obtained by treating these terms as noise in the signal detection, as shown in the following theorem.

Theorem 6.5. With hardware impairments, the DL ergodic channel capacity of UE k in cell j is lower bounded by

$$\underline{\text{SE}}_{jk}^{\text{DL-imp}} = \frac{\tau_d}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{DL-imp}}) \quad (6.45)$$

with

$$\begin{aligned} \underline{\text{SINR}}_{jk}^{\text{DL-imp}} = & \frac{\rho_{jk}|\mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\sum_{l,i} \rho_{li} \frac{(\kappa_t^{\text{BS}}\mathbb{E}\{|\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2\} + (1 - \kappa_t^{\text{BS}})\mathbb{E}\{|\mathbf{w}_{li} \odot \mathbf{h}_{jk}^l|^2\})}{\kappa_t^{\text{BS}}\kappa_r^{\text{UE}}} - \rho_{jk}|\mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\}|^2 + \frac{\sigma_{\text{DL}}^2}{\kappa_t^{\text{BS}}\kappa_r^{\text{UE}}}} \end{aligned} \quad (6.46)$$

where the expectations are with respect to the channel realizations.

Proof. The proof is available in Appendix C.5.4 on p. 618. \square

The DL SE in Theorem 6.5 generalizes Theorem 4.6 on p. 317 to include hardware impairments. The new expression can be computed numerically for any precoding scheme and any spatial correlation matrices. There is a close connection to the UL expression in Theorem 6.2; if we select $\mathbf{w}_{jk} = \mathbf{v}_{jk}/\sqrt{\mathbb{E}\{|\mathbf{v}_{jk}|^2\}}$, then the same expectations appear in both the UL and DL. The main difference is that the indices (l, i) and (j, k) are swapped in the interference terms, since UL interference comes from the UEs and DL interference comes from the BSs. A UL-DL duality result, similar to Theorem 4.8 on p. 321, can be established also with hardware impairments, under the additional condition that $\kappa_t^{\text{UE}} = \kappa_r^{\text{UE}}$ and $\kappa_t^{\text{BS}} = \kappa_r^{\text{BS}}$. Since the transmitter and receiver hardware in a device are fundamentally different, there is no practical reason for these equalities to hold, thus we will exclude the exact duality details. However, the duality suggests that it is sensible to select each precoding vector based on the corresponding combining vector, also when having hardware impairments.

The key impacts of hardware impairments are clearly visible in (6.46). There is a loss in desired signal power by a factor $\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}$, represented by scaling the interference and noise powers as $1/(\kappa_t^{\text{BS}} \kappa_r^{\text{UE}})$. A fraction $(1 - \kappa_t^{\text{BS}})$ of the transmitted interference power is also turned into transmitter distortion, which is not coherently combined over the channel; that is, $\mathbb{E}\{|w_{li}^H h_{jk}^l|^2\}$ is replaced by $\mathbb{E}\{\|w_{li} \odot h_{jk}^l\|^2\}$ for this distortion. Another important factor, that is less visible in Theorem 6.5, is the impact of the distortion in the UL channel estimation, which affects the selection of the precoding vectors.

To gain further insights into the impact of hardware impairments, we consider the case of MR precoding with $w_{jk} = \hat{h}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{h}_{jk}^j\|^2\}}$, where the average normalization is used to enable the derivation of closed-form expressions. We will compute a closed-form SE expression for spatially uncorrelated channels. The more general case with diagonal spatial correlation matrices was considered in [54], while arbitrary correlation matrices were treated in [42].

Corollary 6.6. With hardware impairments, if average-normalized MR precoding with $w_{jk} = \hat{h}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{h}_{jk}^j\|^2\}}$ is used, based on the LMMSE estimator in Theorem 6.1, and the channels are spatially uncorrelated (i.e., $R_{li}^j = \beta_{li}^j I_{M_j}$ for $l = 1, \dots, L$ and $i = 1, \dots, K_l$), then the SE expression in Theorem 6.5 becomes $\underline{\text{SE}}_{jk}^{\text{DL-imp}} = \frac{\tau_d}{\tau_c} \log_2(1 + \underline{\text{SINR}}_{jk}^{\text{DL-imp}})$ with

$$\begin{aligned} \underline{\text{SINR}}_{jk}^{\text{DL-imp}} = & \frac{\rho_{jk} p_{jk} (\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j}{\sum_{l,i} \rho_{li} \beta_{jk}^l F_{jk}^{li} + \sum_{(l,i) \in \mathcal{P}_{jk}} \rho_{li} p_{jk} (\beta_{jk}^l)^2 \tau_p \psi_{li} M_l G_l - \rho_{jk} p_{jk} (\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j + \sigma_{\text{DL}}^2} \end{aligned} \quad (6.47)$$

where ψ_{jk} and ψ_{li} are given by (6.38) and

$$\check{\sigma}_{\text{DL}}^2 = \frac{\sigma_{\text{DL}}^2}{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \quad (6.48)$$

$$F_{jk}^{li} = \frac{1 + p_{jk} \beta_{jk}^l \psi_{li} \left(1 - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} + (1 - \kappa_t^{\text{UE}}) \kappa_t^{\text{BS}} \kappa_r^{\text{BS}} (M_l - 1) \right)}{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \quad (6.49)$$

$$G_l = \frac{1 + \kappa_t^{\text{BS}} (M_l - 1)}{M_l \kappa_t^{\text{BS}} \kappa_r^{\text{UE}}}. \quad (6.50)$$

Proof. The proof is available in Appendix C.5.5 on p. 620. \square

The DL effective SINR expression in (6.47) has the typical SINR structure, with the first term in the denominator being the interference from all UEs, the second term being additional interference from UEs that use the same pilot, the third term subtracts the desired signal power that appeared in the numerator, and the last term represents the noise power. The hardware impairments affect the SINR in multiple ways. First, there is a loss in signal power represented by increasing the effective noise power $\check{\sigma}_{\text{DL}}^2$ by a factor $1/(\kappa_t^{\text{BS}} \kappa_r^{\text{UE}} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}})$. This is another instance of the “squaring effect”, where the channel estimation causes $1/(\kappa_t^{\text{UE}} \kappa_r^{\text{BS}})$ and the data transmission causes $1/(\kappa_t^{\text{BS}} \kappa_r^{\text{UE}})$. Second, the distorted pilot sequences lead to less coherent interference from UEs using the same pilot, but new coherent interference from all other UEs in the network. Generally speaking, the distortion has the same impact on the SE as in the UL, but one important difference is that the UL expression is only affected by UL hardware quality, while the DL SE in (6.47) is affected by the hardware quality in both directions (i.e., $\kappa_t^{\text{BS}}, \kappa_r^{\text{UE}}, \kappa_t^{\text{UE}}, \kappa_r^{\text{BS}}$) since the channels are estimated in the UL.

To further study the coherent interference characteristics, we consider the asymptotic regime with a very large number of BS antennas.

Corollary 6.7. Under the same conditions as in Corollary 6.6, $\text{SINR}_{jk}^{\text{DL-imp}}$ with MR combining has the asymptotic limit

$$\frac{\rho_{jk} (\beta_{jk}^j)^2}{\sum_{l,i} \rho_{li} (\beta_{jk}^l)^2 \frac{\psi_{li}}{\psi_{jk}} \frac{1 - \kappa_t^{\text{UE}}}{\kappa_t^{\text{UE}} \kappa_r^{\text{UE}} \tau_p} + \sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \rho_{li} (\beta_{jk}^l)^2 \frac{\psi_{li}}{\psi_{jk}} \frac{1}{\kappa_r^{\text{UE}}} + \rho_{jk} (\beta_{jk}^j)^2 \frac{1 - \kappa_r^{\text{UE}}}{\kappa_r^{\text{UE}}}} \quad (6.51)$$

as $M_1 = \dots = M_L \rightarrow \infty$.

Proof. This result follows from taking the limit in (6.47) and noting that $\underline{F}_{jk}^l/M_j \rightarrow p_{jk}\beta_{jk}^l\psi_{li}\frac{1-\kappa_t^{\text{UE}}}{\kappa_t^{\text{UE}}\kappa_r^{\text{UE}}}$ and $\underline{G}_j \rightarrow 1/\kappa_r^{\text{UE}}$. \square

As expected, the noise and non-coherent interference vanish asymptotically when all BSs have a very large number of antennas. The asymptotic SINR is the ratio between the coherent signal gain and the coherent interference terms. The first interference term covers all UEs, due to the break of pilot orthogonality caused by distortion, while the second term only covers those UEs that reuse the pilot sequence. At first sight, it seems that (6.51) is independent of κ_t^{BS} and κ_r^{BS} , which are the hardware qualities of the BS. This is not the case since ψ_{li}/ψ_{jk} is actually a function of κ_r^{BS} , but it indicates that the distortion caused at the BS is non-coherently combined.

6.3.6 Impact of Hardware Impairment on DL SE

When quantifying the impact that hardware impairments have on the DL SE, we need to consider both the hardware quality in the UL (κ_t^{UE} and κ_r^{BS}) and the hardware quality in the DL (κ_t^{BS} and κ_r^{UE}). We will now investigate which set of parameters has the largest impact on the SE. To this end, we continue the running example that was defined in Section 4.1.3 on p. 288. We consider $M = 100$, $K = 10$, and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. We use Theorem 6.5 to compute the SE with M-MMSE, RZF, and MR, where the latter is normalized as $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \|\hat{\mathbf{h}}_{jk}^j\|$. Except for pilots, all samples in every coherence block are used for DL data. The pilot reuse factor that maximizes the SE is considered.

The average DL sum SE is shown in Figure 6.7. The UL hardware quality is fixed at $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.99$ in Figure 6.7a while the DL hardware quality is varied as $\kappa_t^{\text{BS}} = \kappa_r^{\text{UE}} \in [0.95, 1]$. The opposite case of fixed DL hardware quality with $\kappa_t^{\text{BS}} = \kappa_r^{\text{UE}} = 0.99$ and varying UL hardware quality with $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} \in [0.95, 1]$ is considered in Figure 6.7b. For all three precoding schemes, we notice that the DL hardware has a greater impact on the DL SE than the UL hardware, particularly when using RZF or M-MMSE. This means that it is the distortion caused

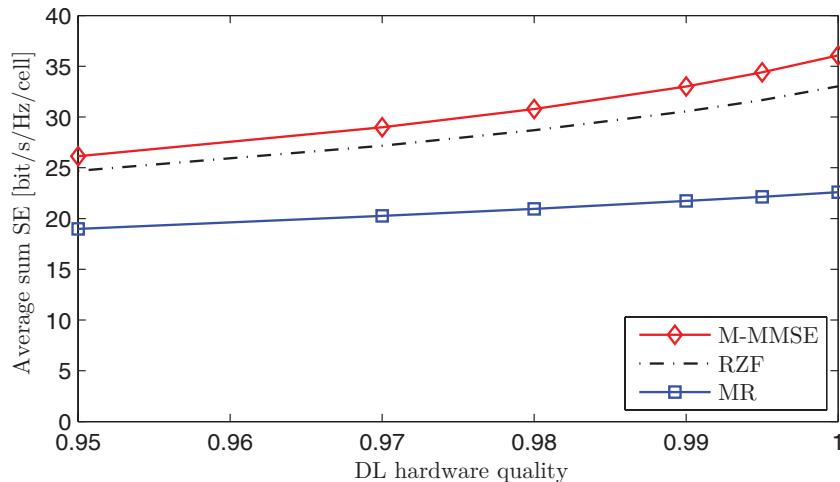
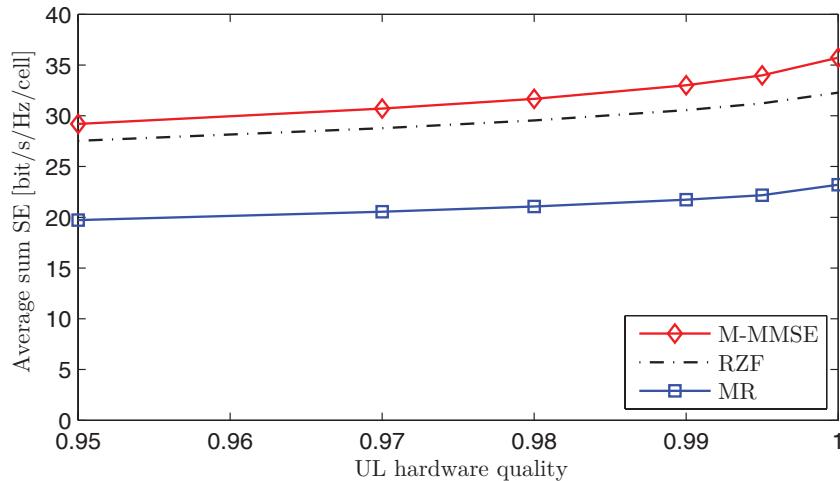
(a) Varying DL hardware quality (κ_t^{BS} and κ_r^{UE}) for $\kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = 0.99$.(b) Varying UL hardware quality (κ_t^{UE} and κ_r^{BS}) for $\kappa_t^{\text{BS}} = \kappa_r^{\text{UE}} = 0.99$.

Figure 6.7: Average DL sum SE as a function of the hardware quality, when the quality is fixed in one direction and varies in the other direction. There are $M = 100$ antennas, $K = 10$ UEs, and for each point on the curves we consider the pilot reuse factor that maximizes the SE.

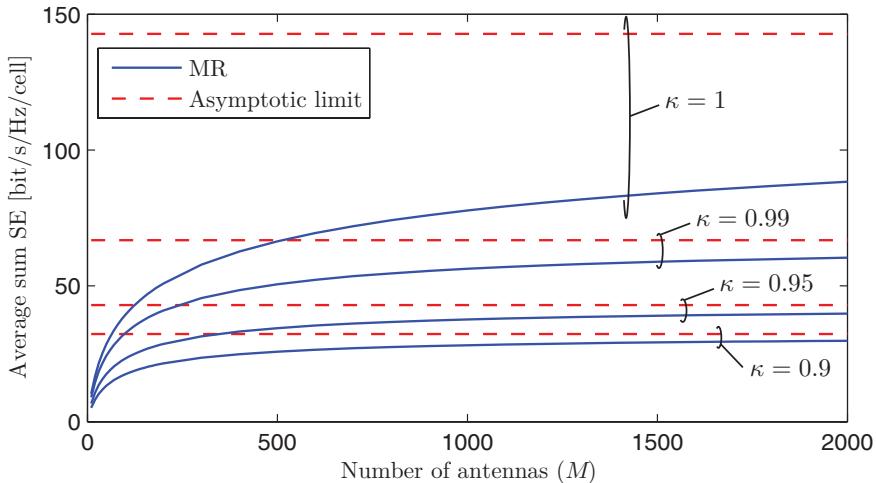


Figure 6.8: Average DL sum SE as a function of the number of BS antennas, for different hardware qualities $\kappa = \kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = \kappa_t^{\text{BS}} = \kappa_r^{\text{UE}}$. Average-normalized MR precoding is considered.

during the data transmission, and not the distortion in the channel estimation, that dominates. The precoding schemes otherwise behave as expected from the UL; that is, M-MMSE gives a substantially larger SE than MR, but is more sensitive to hardware impairments. RZF provides a performance relatively close to M-MMSE.

6.3.7 MR Precoding with Many Antennas

We will now investigate the convergence speed to the asymptotic limit in Corollary 6.7, when the number of antennas in all cells grows without bound; that is, $M = M_1 = \dots = M_L \rightarrow \infty$. To this end, we revisit the running example that was defined in Section 4.1.3 on p. 288, but this time we use the uncorrelated Rayleigh fading model. We consider $K = 10$, $f = 2$, and average-normalized MR precoding. Except for pilots, all samples in every coherence block are used for DL data.

For simplicity, we assume that all hardware quality factors are equal to κ ; that is, $\kappa = \kappa_t^{\text{UE}} = \kappa_r^{\text{BS}} = \kappa_t^{\text{BS}} = \kappa_r^{\text{UE}}$. The average DL sum SE is shown in Figure 6.8 for different hardware qualities: $\kappa \in \{0.9, 0.95, 0.99, 1\}$. The asymptotic limit is indicated in all cases.

We notice that the convergence to the asymptotic limit is slow when having ideal hardware (i.e., $\kappa = 1$); for example, with $M = 2000$ antennas, we have only reached 61% of the asymptotic SE. Roughly $M = 10^5$ antennas are required to reach the limit. As we reduce the hardware quality, the SE reduces substantially but the convergence speed to the asymptotic limit is also faster. With $\kappa = 0.9$, we obtain 79% of the asymptotic SE with $M = 500$ and 92% with $M = 2000$. To explain this phenomenon, we recall that in order to reach the asymptotic limit, we need the non-coherent interference and noise to become negligible as compared to the coherent interference sources. This requires a very large M when having ideal hardware, since the non-coherent interference from the own cell is received through a channel with much stronger gain than the coherent interference from other cells (cf. Figure 6.5a for an illustration of this in the UL). As the coherent self-distortion caused by hardware impairment grows, much fewer antennas are needed to make the non-coherent interference weaker than the self-distortion—a few thousand antennas are sufficient to be close to the asymptotic limit. This example suggests that there is a practical limit on the number of antennas that is useful to deploy.

Figure 6.9 considers the case of a fixed UE hardware quality of $\kappa_t^{\text{UE}} = \kappa_r^{\text{UE}} = 0.99$ and varying hardware qualities at the BS: $\kappa_t^{\text{BS}} = \kappa_r^{\text{BS}} \in \{0.9, 0.95, 0.99, 1\}$. A slightly higher SE is achieved when the BSs' hardware quality is high as compared to low, but the differences are marginal. The asymptotic limits are shown and these are not identical but almost the same. Hence, the major differences observed in Figure 6.8 are mainly due to variations in the UEs' hardware quality.

In summary, the practical implications of asymptotic results are larger with hardware impairments, than with ideal hardware, since there is a much faster convergence to the limits. It is primarily the hardware quality of the UEs that determines the SE. This is a positive result as it allows us to decrease the hardware quality of the BSs.

6.4 Hardware-Quality Scaling Law

The asymptotic UL and DL SE expressions in Section 6.3 reveal that the impact of hardware impairments at the BS vanishes almost completely

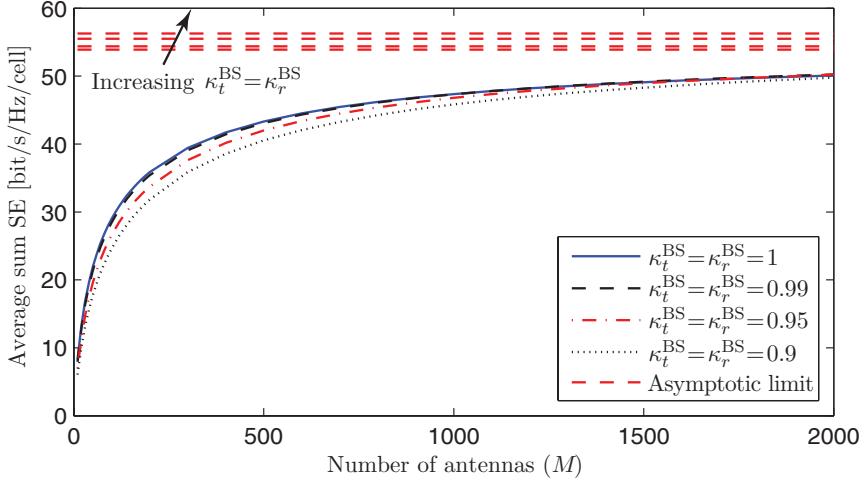


Figure 6.9: Average DL sum SE as a function of the number of BS antennas, for $\kappa_t^{\text{UE}} = \kappa_r^{\text{UE}} = 0.99$ and different hardware qualities at the BS: $\kappa_t^{\text{BS}} = \kappa_r^{\text{BS}} \in \{0.9, 0.95, 0.99, 1\}$. Average-normalized MR precoding is considered.

as the number of BS antennas grows. Hence, low-quality BS hardware can be used with only a minor SE loss, which suggests that Massive MIMO utilizes the hardware very efficiently; in other words, it achieves a high HE. The asymptotic analysis holds for any fixed values of the hardware quality factors κ_t^{BS} and κ_r^{BS} . We will now show that we can even decrease κ_t^{BS} and κ_r^{BS} with the number of antennas and still make the impact of the hardware quality vanish asymptotically. We begin by analyzing the UL.

Corollary 6.8. Consider $\kappa_r^{\text{BS}} = \bar{\kappa}/M_j^\varepsilon$, where $\bar{\kappa} \in (0, 1]$ and $\varepsilon > 0$ are constants. Under the same conditions as in Corollary 6.3, $\overline{\text{SINR}}_{jk}^{\text{UL-imp}}$ with MR combining has the asymptotic limit

$$\begin{cases} \frac{(p_{jk}\beta_{jk}^j)^2}{\sum_{l,i} (p_{li}\beta_{li}^j)^2 \frac{1-\kappa_t^{\text{UE}}}{(\kappa_t^{\text{UE}})^2 \tau_p} + \sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{(p_{li}\beta_{li}^j)^2}{\kappa_t^{\text{UE}}} + (p_{jk}\beta_{jk}^j)^2 \frac{1-\kappa_t^{\text{UE}}}{\kappa_t^{\text{UE}}}} & \varepsilon < \frac{1}{2} \\ 0 & \varepsilon > \frac{1}{2} \end{cases} \quad (6.52)$$

as $M_j \rightarrow \infty$.

Proof. This results follows from substituting $\kappa_r^{\text{BS}} = \bar{\kappa}/M_j^\varepsilon$ into (6.52)

and then taking the limit. In particular, we notice that $\bar{F}_{li}^j/M_j \rightarrow p_{li}\beta_{li}^j\psi_{jk}\frac{1-\kappa_t^{\text{UE}}}{(\kappa_t^{\text{UE}})^2}$ and $\frac{\sigma_{\text{UL}}^2}{(\kappa_t^{\text{UE}}\kappa_r^{\text{BS}})^2 M_j} \rightarrow 0$ if $\varepsilon < 1/2$, while these terms diverge for $\varepsilon > 1/2$. Moreover, $\bar{G}_j \rightarrow 1/\kappa_t^{\text{UE}}$ if $\varepsilon < 1$. \square

This corollary proves that we can gradually tolerate lower BS hardware quality as the number of antennas increases. Suppose $M = M_1 = \dots = M_L$. Corollary 6.8 then proves that we decrease κ_r^{BS} roughly as $1/\sqrt{M}$ and still achieve the same asymptotic UL SE limit as derived in Corollary 6.4 for a fixed hardware quality. However, we can expect the convergence speed to this limit to be slower if we gradually degrade the hardware quality.

We refer to Corollary 6.8 as a *hardware-quality scaling law*. The intuition behind this result can be seen in (6.39), where the desired signal power grows as M_j while the effective noise term is proportional to $1/(\kappa_r^{\text{BS}})^2$ (or $M_j^{2\varepsilon}/\bar{\kappa}^2$ using the notation in the corollary). When the scaling law is satisfied, the signal power grows faster than the effective noise term, which is sufficient to reach a non-zero limit. A similar result can be obtained in the DL.

Corollary 6.9. Consider $M = M_1 = \dots = M_L$, $\kappa_t^{\text{BS}} = \underline{\kappa}/M^{\varepsilon_1}$, and $\kappa_r^{\text{BS}} = \bar{\kappa}/M^{\varepsilon_2}$, where $\underline{\kappa}, \bar{\kappa} \in (0, 1]$ and $\varepsilon_1, \varepsilon_2 > 0$ are constants. Under the same conditions as in Corollary 6.6, $\text{SINR}_{jk}^{\text{DL-imp}}$ with average-normalized MR precoding has the asymptotic limit

$$\begin{cases} \frac{\rho_{jk}(\beta_{jk}^j)^2}{\sum_{l,i} \rho_{li}(\beta_{jk}^l)^2 \frac{\psi_l^\infty}{\psi_j^\infty} \frac{(1-\kappa_t^{\text{UE}})}{\kappa_r^{\text{UE}} \kappa_t^{\text{UE}} \tau_p} + \sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{\rho_{li}(\beta_{jk}^l)^2}{\kappa_r^{\text{UE}}} \frac{\psi_l^\infty}{\psi_j^\infty} + \rho_{jk}(\beta_{jk}^j)^2 \frac{1-\kappa_r^{\text{UE}}}{\kappa_r^{\text{UE}}}} & \varepsilon_1 + \varepsilon_2 < 1 \\ 0 & \varepsilon_1 + \varepsilon_2 \geq 1 \end{cases} \quad (6.53)$$

as $M \rightarrow \infty$, where

$$\psi_j^\infty = \frac{1}{\sum_{l',i'} p_{l'i'} \beta_{l'i'}^j + \sigma_{\text{UL}}^2}, \quad \psi_l^\infty = \frac{1}{\sum_{l',i'} p_{l'i'} \beta_{l'i'}^l + \sigma_{\text{UL}}^2}. \quad (6.54)$$

Proof. This results follows from substituting $\kappa_t^{\text{BS}} = \underline{\kappa}/M^{\varepsilon_1}$ and $\kappa_r^{\text{BS}} = \bar{\kappa}/M^{\varepsilon_2}$ into (6.53) and then taking the limit. In particular, we notice that $\check{\sigma}_{\text{DL}}^2/M \rightarrow 0$ and $\underline{F}_{jk}^{li}/M = p_{jk}\beta_{jk}^l\psi_{li}^\infty(1-\kappa_t^{\text{UE}})/(\kappa_r^{\text{UE}}\kappa_t^{\text{UE}})$ if $\varepsilon_1 + \varepsilon_2 < 1$,

while these terms diverge for $\varepsilon_1 + \varepsilon_2 > 1$. Moreover, $\overline{G}_j \rightarrow 1/\kappa_r^{\text{UE}}$ if $\varepsilon_1 < 1$, $\psi_{jk} \rightarrow \psi_j^\infty$, and $\psi_{li} \rightarrow \psi_l^\infty$. \square

This corollary provides a DL hardware-quality scaling law that is similar to the UL scaling law in Corollary 6.8, but both the transmitter and receiver hardware play a role in the DL. We can either decrease κ_t^{BS} and κ_r^{BS} jointly as $1/\sqrt{M}$ or decrease one faster than the other as long as the product $\kappa_t^{\text{BS}}\kappa_r^{\text{BS}}$ does not decay faster than $1/M$.

The UL and DL hardware-quality scaling laws give further theoretical evidence that Massive MIMO networks have a higher HE than conventional cellular networks. More precisely, the network can operate well using lower BS hardware quality than conventional networks and the quality can be gradually reduced as the number of antennas increases. This fact might be incredibly important for the practical adoption of Massive MIMO since it means that one can deploy the technology without increasing the hardware cost linearly with M . Similarly, the physical size of the hardware components and their power consumption may not grow, as compared to contemporary systems, if we reduce the hardware quality as more antennas are added. The exact consequences are hard to quantify, but it is not impossible that a well-designed Massive MIMO implementation would achieve the same, or even lower, cost, size, and/or circuit power as in conventional networks [102, 153]. Below, we briefly discuss a few specific challenges and opportunities in the hardware design.

6.4.1 Low-Resolution ADCs

An impairment source that has received particular attention from researchers is the quantization noise, caused by finite-resolution ADCs. Resolutions of 4–20 bit per I/Q component are being considered in different wireless scenarios [90]. In LTE, an ADC resolution of at least 10 bits is needed to satisfy the EVM requirements [144, Section 14.8.3], but it is common to have a substantial design margin (e.g., 15-bit ADCs) to allow for impairments in other components. Since the I/Q components are quantized separately, a Massive MIMO BS with M antennas requires $2M$ ADCs. The power consumption of these ADCs grows with

the system bandwidth, which makes it highly desirable to operate with reduced ADC resolution in wideband systems. Fortunately, the UL hardware-quality scaling law manifests that the bit resolution of each ADC can be reduced as M increases [53]. There are plenty of papers that specifically studies the impact of low-resolution ADCs on various performance metrics [103, 306, 343, 327]. The common conclusion is that, with $M = 100$ antennas, ADCs with 3–4 bits are sufficient to operate close to the performance of a system with infinite ADC resolution. These numbers are typically achieved with UL power control that makes the UEs' signals equally strong at the receiver. A larger dynamic range is needed if the BS receives a superposition of strong and weak signals, to avoid that the weaker signals drown in the quantization noise. It is also possible to operate with 1-bit ADCs at the BS, which can greatly simplify the hardware design since only the sign of the received signal must be measured and not the signal power. The channel estimation and detection are more challenging but still possible in such systems [218, 99]. In contrast to having 3–4 bit ADCs, the performance loss with 1-bit ADCs is quite substantial, particularly since it is hard to accurately estimate the channels, but the loss reduces when the number of antennas increases [220, 226]. In principle, also the bit resolution of the DACs can be reduced [136, 158, 356], but this is less desirable since it leads to increased out-of-band radiation in the DL; see Section 6.4.3.

6.4.2 Phase Noise

The hardware characteristics have been assumed to be stationary in this monograph, which resulted in the distortion terms being stationary random processes (within a coherence block). Phase noise in the LOs is an example of a non-stationary impairment source, which creates random phase drifts that accumulate over time. The modeling and impact of phase noise have been considered in [310, 101, 92, 188, 258, 213, 107], among others. These models have recently been used in the Massive MIMO literature to compute capacity bounds and analyze their behavior [53, 54, 179, 265, 262]. Since each channel is estimated once per coherence block, the phase noise can be absorbed into the channel fading if the accumulated drift within a block is small. Hence, it is

mainly in scenarios with long channel coherence time (or low hardware quality) that the phase noise effect is substantial. In those cases, it can be worthwhile to send pilots more frequently [53] or to use decoded data to track the phase offsets [215]. The analysis and modeling details depend on the modulation format (e.g., OFDM or single-carrier), but the qualitative conclusions are basically the same. If there is one LO per BS and per UE, then phase noise has a similar impact on the SE in Massive MIMO as in single-antenna systems; the self-distortion caused by phase-noise is coherently combined, just as the desired signal. However, if each BS antennas has a separate LO, the phase noise realizations are independent between the antennas. The resulting distortion is then non-coherently combined over the array, which implies that the impact reduces as the number of antennas increases. Reduced-quality LOs can then be used and a hardware-quality scaling law for phase noise has been established in [53].

6.4.3 Out-of-Band Radiation

The analysis in this section shows that Massive MIMO networks are less affected by in-band distortion, caused by hardware impairments, than conventional networks. This paves the way for using simpler hardware, but only if the higher distortion level does not impair with other systems. It is the total power of the interfering signals that matters, not the fraction of it that is distortion.

One important point that is not captured by the complex baseband model considered in this section is the out-of-band radiation. Some hardware impairments lead to spectral regrowth of the analog transmit signal and effectively increase the distortion caused to adjacent frequency bands. This is the case for PA non-linearities and low-resolution DACs. There are rigid requirements on the maximum out-of-band radiation levels in wireless networks. The requirements can be on the maximum absolute power of the out-of-band radiation or on the maximum relative power, as in the adjacent-channel leakage ratio (ACLR) metric. Initial studies on the out-of-band radiation in the Massive MIMO are provided in [136, 227, 228, 62]. There are two important messages: *i*) The average out-of-band radiation is the same as in a single-antenna system, using

the same hardware quality. To control the absolute power of the out-of-band radiation, as we reduce the hardware quality, we either need to reduce the transmit power or send a signal with a reduced bandwidth (which proportionally reduces the SE). *ii)* The DL out-of-band radiation is emitted non-isotropically. If only one UE is served in the DL by the Massive MIMO BS, the resulting out-of-band radiation has a strong spatial directivity, similar to the precoded in-band signal. This effect is smeared out when multiple UEs are spatially multiplexed.

6.4.4 Reciprocity Calibration

The DL transmission that we have described in this monograph relies on channel reciprocity; that is, if \mathbf{h}_{li}^j is the UL channel, then $(\mathbf{h}_{li}^j)^\text{T}$ is the corresponding DL channel. We have simplified the notation by letting $(\mathbf{h}_{li}^j)^\text{H}$ be the DL channel, which can be done without loss of generality. However, there is another issue with reciprocity: the RF propagation channels are reciprocal by nature, but the end-to-end channels are also affected by the transceiver hardware. Since different hardware components are used at the BS and UE for transmission and reception, there is no reason for the hardware to be reciprocal. The actual DL channel is commonly modeled as $c_{li}(\mathbf{h}_{li}^j)^\text{T} \mathbf{D}^j$ [374], where $c_{li} \in \mathbb{C}$ models the reciprocity mismatch at UE i in cell l and the diagonal matrix $\mathbf{D}^j \in \mathbb{C}^{M_j \times M_j}$ represents the mismatches at the M_j antennas of BS j . These parameters describe the scaling and phase mismatches between the UL and DL, and can be treated as deterministic parameters since it takes hours (or at least minutes) for them to change in practice [329]. This makes it feasible to estimate and calibrate the system so that channel reciprocity holds and the overhead of doing so appears to be negligible. There are plenty of reciprocity calibration algorithms, whereof some were developed for general TDD systems [248, 133, 374] and some particularly for Massive MIMO [300, 331, 279, 330]. Perfect reciprocity calibration should not be expected from any of these schemes, but if the estimation errors are independent across antennas, the distortion caused by the residual reciprocity mismatch will combine non-coherently over the array—similar to other types of hardware impairments.

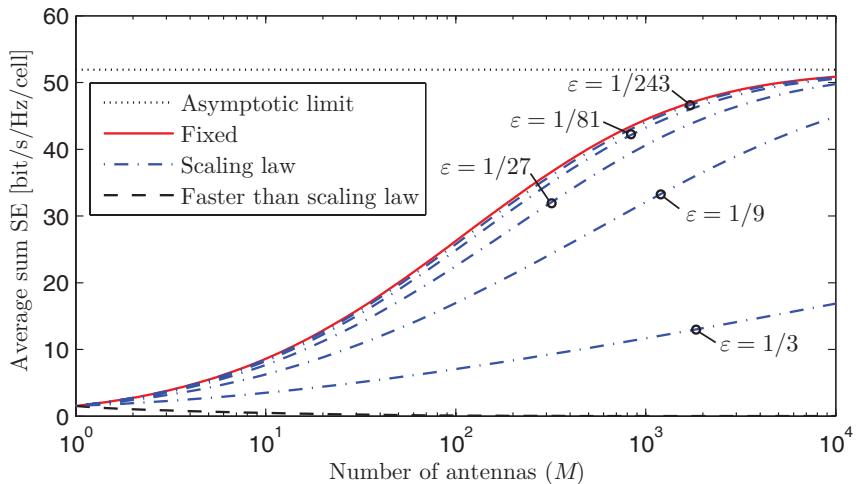
6.4.5 Example of the Hardware-Quality Scaling Law

The hardware-quality scaling law for the UL in Corollary 6.8 will now be exemplified, by continuing the running example that was defined in Section 4.1.3 on p. 288. We consider average-normalized MR combining, $K = 10$, $f = 2$, $\kappa_t^{\text{UE}} = 0.997$, and $\kappa_t^{\text{BS}} = \bar{\kappa}/M^\varepsilon$. We further assume $\bar{\kappa} = 0.997$ and consider different scaling exponents: $\varepsilon \in \{0, 1/243, 1/81, 1/27, 1/9, 1/3, 1\}$. Observe that $\varepsilon = 0$ represents a fixed BS hardware quality. Except for pilots, all samples in every coherence block are used for UL data.

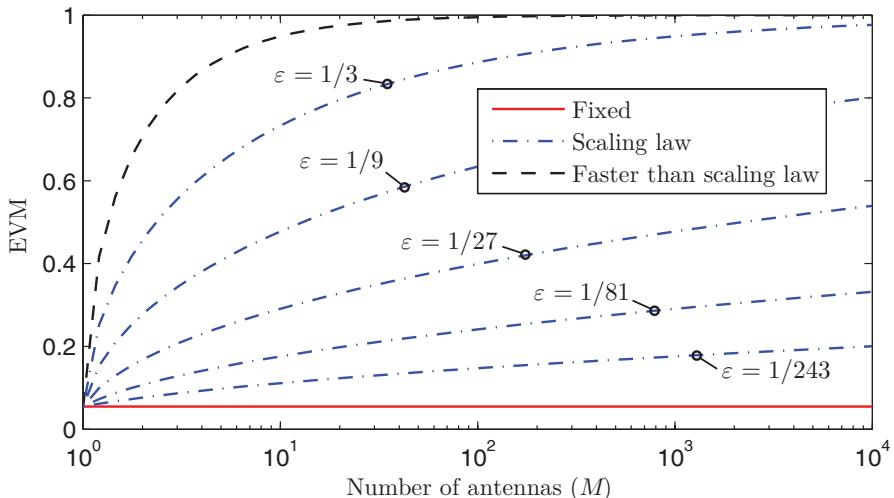
The average UL sum SE is shown in Figure 6.10a, as a function of the number of BS antennas (notice the logarithmic horizontal scale). All curves increase with M , except for $\varepsilon = 1$ which is reported as “faster than scaling law” in the figure. This is in line with the scaling law, which prescribes that the SE only approaches a non-zero limit if $\varepsilon < 1/2$. This asymptotic limit is shown in the figure. The curves $\varepsilon \in \{0, 1/243, 1/81, 1/27\}$ provide similar SE and approach the asymptotic limit at around $M = 10^4$. We can thus degrade the BS hardware quality in these ways and only get a limited SE loss. In contrast, $\varepsilon = 1/9$ and $\varepsilon = 1/3$ give more substantial performance losses, but the same asymptotic limit is approached as $M \rightarrow \infty$.

To understand the practical implications of these scaling laws, we show the corresponding EVM values in Figure 6.10b, computed based on (6.7). The EVM takes values between 0 (ideal hardware) and 1 (all signals are replaced with distortion). For any $\varepsilon > 0$, the EVM will approach 1 as $M \rightarrow \infty$. Nonetheless, we can achieve a high asymptotic SE if we use $\varepsilon < 1/2$. Practical transceivers typically have an EVM below 0.1, thus an EVM of 0.3 represents unusually low hardware quality. By looking at both parts of Figure 6.10, we notice that $\varepsilon = 1/243$, $\varepsilon = 1/81$, and $\varepsilon = 1/27$ lead to small SE losses and EVMs in the range up to 0.5 (for $M < 10^4$). These scaling exponents thus strike a good balance between high SE and the use of low-resolution hardware.

Another way to utilize Figure 6.10 is to select a target sum SE and then identify different combinations of M and ε that deliver this performance. We notice that more antennas are required when ε increases, but the hardware quality per antenna is also reduced. In other



(a) Average UL sum SE when using the scaling law with different exponents.



(b) EVM when using the scaling law with different exponents.

Figure 6.10: Average UL sum SE and EVM, as a function of the number of BS antennas, when applying the hardware-quality scaling law in Corollary 6.8.

words, one can compensate for a reduced hardware quality by adding antennas.

In summary, the hardware-quality scaling law allows for a very rapid degradation in hardware quality, while achieving the same asymptotic limit as with fixed hardware quality. A small scaling exponent (e.g., $\varepsilon = 1/27$ or $\varepsilon = 1/81$) is sufficient to allow for practical low-quality hardware at the BS and a low SE loss. Note that these results are obtained using MR combining, while schemes such as RZF and M-MMSE are likely to lose more in SE from degrading the hardware.

6.5 Summary of Key Points in Section 6

- Practical transceivers are affected by hardware impairments, which can be mitigated by compensation algorithms but not fully removed.
- To quantify the worst-case impact of hardware impairments on the SE, it is sufficient to use models with an independent distortion scalar term added at the UE and an independent distortion vector added at the BS.
- Distortions from the BS are non-coherently combined, while the self-distortion caused by the UEs is coherently combined, similar to the desired signal. This makes the hardware quality of the UEs particularly important.
- Distortion from other UEs can be suppressed by precoding and combining (e.g., RZF or M-MMSE) similar to inter-user interference.
- The distortion caused by the BS and the self-distortion are roughly the same for all precoding and combining schemes.
- Self-distortion might drown in the interference when using MR, while interference-suppressing processing schemes, such as RZF and M-MMSE, are more sensitive to it. This creates a limited interference visibility range, outside which the interference is negligible compared to the self-distortion.
- Massive MIMO networks have a high HE, since they make more efficient use of the BS hardware than conventional networks. For example, for any fixed BS hardware quality, the impact of its distortion vanishes as the number of antennas grows. The impact of distortion is also smaller the more UEs are served in the cell.

- The high HE implies that the BS hardware quality can be gradually decreased as the number of antennas increases while approaching a non-zero asymptotic SE limit. The hardware-quality scaling law describes at which rate the quality can be degraded. As a consequence, the hardware cost of Massive MIMO needs not increase linearly with the number of antennas.
- When reducing the hardware quality, the increased out-of-band radiation caused by spectral regrowth is an important practical issue. It can be dealt with by reducing the transmit power and/or the bandwidth, at the cost of reduced SE.

7

Practical Deployment Considerations

This section describes some important tradeoffs and considerations for the design, optimization, and deployment of Massive MIMO in practical networks. While the previous sections described the fundamental theory that is generally well understood, the topics covered by this section are still under development at the time of writing this monograph. Section 7.1 describes power allocation schemes for the maximization of network utility functions. Key topics in spatial resource allocation are outlined in Section 7.2, which include pilot assignment, scheduling, and load balancing. Channel modeling is considered in Section 7.3 with focus on the spatial characteristics that are resolved by large arrays. The configuration and deployment of antenna arrays are covered by Section 7.4. Massive MIMO technology for hotspots operating at mmWave frequencies is described in Section 7.5, while the role of Massive MIMO in heterogeneous networks is described in Section 7.6. We conclude with a case study in Section 7.7, taking some of the practical considerations into account, and a summary of key points in Section 7.8.

7.1 Power Allocation

Although the SE analysis in the previous sections applies to arbitrarily selected transmit powers, the numerical results were based on the assumption of equal transmit power per UE, in both UL and DL. This is generally not the optimal strategy if we want to maximize some *utility function*. Firstly, the sum SE can be increased by unequal power allocation, exploiting the different propagation conditions of the UEs. Secondly, the sum SE only measures the aggregated throughput of the network, while ignoring how fairly it is distributed among the UEs. This can lead to substantial unfairness. In addition to the sum SE utility, there are alternative utility functions that balance between aggregate throughput and fairness [46]. In this section, we describe several different network utility functions and provide power allocation schemes that maximize them. The channel hardening makes power allocation different in Massive MIMO than in single-antenna systems. There is no need to adapt the transmit powers to small-scale fading variations, but only to the large-scale fading characteristics. This unique characteristic of Massive MIMO makes advanced power allocation schemes, that previously were overly complex, practically feasible.

In a network with $\sum_{l=1}^L K_l$ UEs, there are equally many SE expressions to take into consideration—in both UL and DL. These SEs are all connected due to interference. This can, for example, be seen from the DL SE in Theorem 4.6 on p. 317, which takes the form

$$\underline{\text{SE}}_{jk}^{\text{DL}} = \frac{\tau_d}{\tau_c} \log_2 \left(1 + \frac{\rho_{jk} a_{jk}}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} b_{lijk} + \sigma_{\text{DL}}^2} \right) \quad (7.1)$$

for UE k in cell j , where

$$a_{jk} = |\mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\}|^2 \quad (7.2)$$

$$b_{lijk} = \begin{cases} \mathbb{E}\{|\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2\} & (l, i) \neq (j, k) \\ \mathbb{E}\{|\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j|^2\} - |\mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\}|^2 & (l, i) = (j, k). \end{cases} \quad (7.3)$$

Notice that $\underline{\text{SE}}_{jk}^{\text{DL}}$ in (7.1) is an increasing function of ρ_{jk} , which is the DL transmit power allocated to this UE, while it is a decreasing

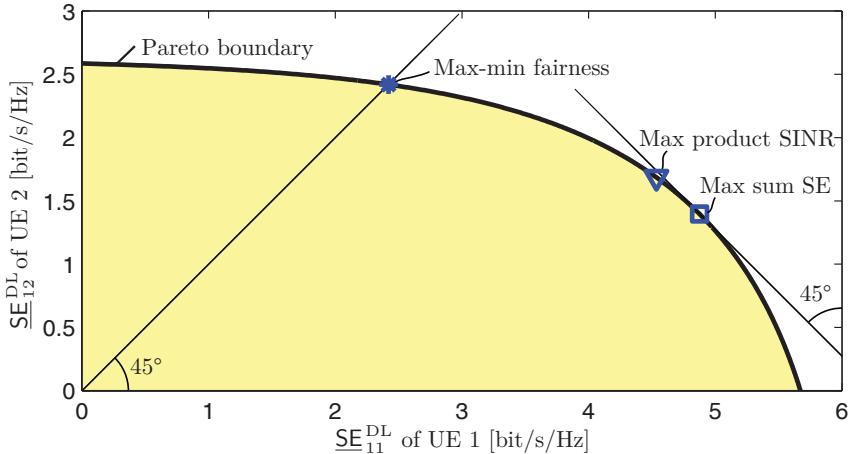


Figure 7.1: Example of an SE region (shaded) with different combinations $(\underline{SE}_{11}^{\text{DL}}, \underline{SE}_{12}^{\text{DL}})$ of SEs that can be achieved by different power allocations. The three operating points that maximize the utility functions in (7.4) are indicated.

function of the transmit power ρ_{li} of any other UE. Hence, there is a conflicting relation between the SEs of two UEs, not only due to mutual interference but also because each BS has a limited power budget to allocate among its UEs.

This conflicting relation can be illustrated by an *SE region* with $\sum_{l=1}^L K_l$ dimensions, which contains all simultaneously achievable SE combinations. An example is provided in Figure 7.1 for the DL with $L = 1$ and $K_1 = 2$ UEs. The shaded region contains all $(\underline{SE}_{11}^{\text{DL}}, \underline{SE}_{12}^{\text{DL}})$ -points that can be achieved by different allocations of power between the UEs. In this example, we assumed that UE 1 has a better channel than UE 2, which results in an SE region that contains larger SE values for UE 1. Any point in the interior of the SE region is strictly suboptimal because it is possible to jointly increase $\underline{SE}_{11}^{\text{DL}}$ and $\underline{SE}_{12}^{\text{DL}}$ by changing the power allocation. Hence, an efficient network must operate on the outer boundary of the SE region, which is referred to as the *Pareto boundary*. For points on the Pareto boundary, we cannot increase the SE of a UE without decreasing the SE of another UE. The shape of the Pareto boundary describes the conflicting relation [47].

There are infinitely many points on the Pareto boundary, so which point should we choose? No objective answer exists to this question because the typical goal of every UE is to maximize its own SE [47], which would in principle require to switch off all other UEs. As a network designer, we need to find a subjective balance between the individual goals of the UEs. A structured way to do this is by defining a network utility function, $U(\text{SE}_{11}, \dots, \text{SE}_{LK_L})$, that takes the SEs of all UEs as inputs and gives a scalar that measures the utility as output [250]; that is, the larger the better. To keep it general, SE_{jk} for UE k in cell j can either denote the UL SE $\underline{\text{SE}}_{jk}^{\text{UL}}$ or the DL SE $\underline{\text{SE}}_{jk}^{\text{DL}}$. The utility function is to be maximized and should be selected so that the input is more preferred the larger the output is. Some prominent examples of utility functions are:

$$U(\text{SE}_{11}, \dots, \text{SE}_{LK_L}) = \begin{cases} \sum_{j=1}^L \sum_{k=1}^{K_j} \text{SE}_{jk} & \text{Max sum SE} \\ \min_{j,k} \text{SE}_{jk} & \text{Max-min fairness} \\ \prod_{j=1}^L \prod_{k=1}^{K_j} \text{SINR}_{jk} & \text{Max product SINR} \end{cases} \quad (7.4)$$

where SINR_{jk} denotes the effective SINR of SE_{jk} . Notice that “max” indicates that we want to maximize these utilities, while the actual utility functions are the sum SE, minimum SE, and product of the SINRs, respectively.

By definition, the max sum SE utility leads to the highest aggregate SE, but without any fairness guarantees—some UEs with bad channel conditions might get zero SE when maximizing this utility. By normalizing the sum SE by the number of UEs, we can also interpret this utility as the arithmetic mean SE. The other extreme is max-min fairness, which provides complete fairness by only counting the SE achieved by the weakest UE in the network. One can easily convince oneself that maximizing this utility function results in the same SE for everyone, thus a UE has no benefit of having a good channel condition. There is a variety of tradeoffs between these extremes, represented either by weighing the SEs of the UEs differently in (7.4) or by using alternative functions; for example, the geometric mean or the harmonic mean of the SEs [46, 176, 221]. The book [210] advocates a heuristic approach for maximizing the minimum SE locally in each cell, while allowing for

different SE levels in different cells. This is made possible by neglecting coherent interference and assuming that every BS and at least one UE per cell transmit with full power.

The max product SINR utility is also defined in (7.4). To understand the motivation behind maximizing the product of the effective SINRs, let us consider the DL and notice that

$$\begin{aligned} \sum_{j=1}^L \sum_{k=1}^{K_j} \underline{\text{SE}}_{jk}^{\text{DL}} &= \sum_{j=1}^L \sum_{k=1}^{K_j} \frac{\tau_d}{\tau_c} \log_2 \left(1 + \underline{\text{SINR}}_{jk}^{\text{DL}} \right) \\ &\geq \sum_{j=1}^L \sum_{k=1}^{K_j} \frac{\tau_d}{\tau_c} \log_2 \left(\underline{\text{SINR}}_{jk}^{\text{DL}} \right) = \frac{\tau_d}{\tau_c} \log_2 \left(\prod_{j=1}^L \prod_{k=1}^{K_j} \underline{\text{SINR}}_{jk}^{\text{DL}} \right) \end{aligned} \quad (7.5)$$

which shows that this utility seeks to maximize a lower bound on the sum SE where the “1+” term is neglected in every logarithm. This has little effect on UEs that support high SINRs, but underestimates the SEs of the weakest UEs. Hence, a maximization of the product of the SINRs leads to higher SEs for the weakest UEs, as compared to maximizing the sum SE. The max product SINR utility also guarantees that every UE gets a non-zero SE, thus this utility function provides more fairness than the sum SE utility. Recall from Theorem 4.13 on p. 346 that, with M-MMSE precoding, the SINR of every UE increases without bound when the number of antennas grows large. The “1+” in the logarithms is then negligible and maximizing the product SINR or the sum SE is asymptotically the same thing. When there are very many antennas, equal power allocation will be nearly optimal for both utilities, since the SNR variations created by power allocation have little impact on the SEs when everyone has very high SINRs.

The points in the SE region that maximize the three utility functions in (7.4) are exemplified in Figure 7.1. The max-min fairness point is the intersection between the Pareto boundary and a line from the origin with 45° slope. This is the line that contains all points where the UEs have equal SE. The max sum SE point lies on another line with 45° slope, which is defined by the equation $\underline{\text{SE}}_{11}^{\text{DL}} + \underline{\text{SE}}_{12}^{\text{DL}} = v$, where v represents the maximum value of the sum SE. Every point on this line would give this sum SE, but the max sum SE point is the one that

intersects with the SE region. Note that this point gives substantially higher SE to UE 1, compared with max-min fairness, at the expense of decreasing the SE of UE 2. Finally, the max product SINR point gives slightly higher SE to the weaker UE 2, but in this example, the point is still very close to the line that maximizes the sum SE.

7.1.1 Downlink Power Allocation

We will now use the utility framework developed above to optimize the DL power allocation, for the case when the hardening bound in Theorem 4.6 on p. 317 is used. In practice, the transmit power of a BS is limited by regulations and hardware constraints, which we model by a maximum total transmit power $P_{\max}^{\text{DL}} \geq 0$ per BS. The resulting utility maximization problem is

$$\begin{aligned} & \underset{\rho_{11} \geq 0, \dots, \rho_{LK_L} \geq 0}{\text{maximize}} \quad U(\underline{\text{SE}}_{11}^{\text{DL}}, \dots, \underline{\text{SE}}_{L K_L}^{\text{DL}}) \\ & \text{subject to} \quad \sum_{k=1}^{K_j} \rho_{jk} \leq P_{\max}^{\text{DL}}, \quad j = 1, \dots, L \end{aligned} \quad (7.6)$$

where the DL transmit powers $\rho_{11}, \dots, \rho_{LK_L}$ are the $\sum_{l=1}^L K_l$ optimization variables. The actual power transmitted by BS j is $\sum_{k=1}^{K_j} \rho_{jk}$. We can use this problem formulation along with any precoding scheme and we keep the precoding vectors fixed while optimizing the transmit powers. The type of power allocation problems in (7.6) has been studied for decades [81, 341, 46] and the computational complexity of finding the solution strongly depends on the choice of the utility function U . The max sum SE problem is non-convex and hard to solve to global optimality [201], but there exist successive approximation algorithms that find locally optimal solutions in polynomial time [341] and global optimization algorithms that find the global optimum with a complexity that grows exponentially with the number of UEs [269]. These methods can be used for benchmarking, but are of limited practical use. See Appendix B.6 on p. 575 for details on practical optimization algorithms, as well as definitions of linear, convex, and geometric programs (the three problem categories that will be of importance in this section).

The solution to the power allocation problem in (7.6) depends on a_{jk} in (7.2) and b_{lijk} in (7.3), which are the average channel gains and average interference gains, respectively. The average is computed with respect to the small-scale fading realizations so that the optimized power allocation is only a function of the channel statistics and the choice of precoding. The same power allocation solution can be used for many coherence blocks, in both the frequency and time domains (cf. Figure 2.1). It is mainly the macroscopic mobility of UEs that determines the time interval over which the channel statistics are static; see Remark 2.3 on p. 225. This is a key feature of Massive MIMO as compared to systems with few antennas, in which there is no channel hardening and hence a need to adapt the power allocation to the substantial power variations that occur between coherence blocks in both time and frequency. The bottom line is that Massive MIMO systems can afford more complex power allocation schemes since the same solution can be used for many coherence blocks.

When the max-min fairness or max product SINR utility functions are used, the globally optimal solution to (7.6) can be obtained efficiently, as shown by the following theorems.

Theorem 7.1. The max-min fairness problem, for given values of a_{jk} and b_{lijk} , can be expressed as

$$\begin{aligned} & \underset{\rho_{11} \geq 0, \dots, \rho_{LK_L} \geq 0, \gamma \geq 0}{\text{maximize}} \quad \gamma \\ & \text{subject to} \quad \frac{\rho_{jk} a_{jk}}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} b_{lijk} + \sigma_{\text{DL}}^2} \geq \gamma, \quad j = 1, \dots, L, \quad k = 1, \dots, K_j \\ & \quad \sum_{k=1}^{K_j} \rho_{jk} \leq P_{\text{max}}^{\text{DL}}, \quad j = 1, \dots, L \end{aligned} \tag{7.7}$$

and the globally optimal solution is obtained by Algorithm 1, to a tolerance $\epsilon > 0$.

Proof. The first step is to notice that maximization of $\min_{j,k} \underline{\text{SE}}_{jk}^{\text{DL}}$ is equivalent to maximization of $\min_{j,k} \underline{\text{SINR}}_{jk}^{\text{DL}}$. Using the latter utility function, we obtain (7.7) by writing (7.6) on epigraph form [67]; that

is, to introduce the auxiliary variable γ that must satisfy the constraint $\text{SINR}_{jk}^{\text{DL}} \geq \gamma$ for all j, k , and maximize that variable instead. Moreover, we notice that the constraints in (7.7) are linear in the power parameters. The SINR constraints can also be written in the linear form $\rho_{jk}a_{jk} \geq \gamma(\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li}b_{lijk} + \sigma_{\text{DL}}^2)$ for any fixed value of γ . Hence, we can solve (7.7) as a linear feasibility problem when γ is fixed. If we solve this subproblem and the SINR constraints $\text{SINR}_{jk}^{\text{DL}} \geq \gamma$ are satisfied for a given value of γ , we need to increase the transmit powers as γ increases to keep the constraints satisfied. We can thus make a line search over γ to find the largest value for which all power constraints are satisfied when solving the feasibility problem. Algorithm 1 finds that value, to a tolerance ϵ , by bisection over the search range between $\gamma = 0$ and $\gamma = \min_{j,k} P_{\max}^{\text{DL}}a_{jk}/\sigma_{\text{DL}}^2$, where the latter is the effective SINR of the weakest UE when neglecting all interference. \square

This theorem shows that the max-min fairness problem can be solved to global optimality by Algorithm 1, which solves a sequence of linear feasibility problems. Each subproblem is similar to the classical power optimization works [64, 370, 360] that find the minimum power that satisfies given SINR constraints. Appendix B.6 on p. 575 describes how such subproblems are solved by general-purpose algorithms. The outer while-loop in the algorithm performs a bisection search for the optimal SINR value, which means that the search space for the max-min SINR value is halved in every iteration. The convergence is thus very fast.

Theorem 7.2. The max product SINR problem, for given values of a_{jk} and b_{lijk} , can be expressed as

$$\begin{aligned}
 & \underset{\rho_{11} \geq 0, \dots, \rho_{LK_L} \geq 0, c_{11} \geq 0, \dots, c_{LK_L} \geq 0}{\text{maximize}} && \prod_{j=1}^L \prod_{k=1}^{K_j} c_{jk} \\
 & \text{subject to} && \sum_{l=1}^L \sum_{i=1}^{K_l} \frac{c_{jk}\rho_{li}b_{lijk}}{\rho_{jk}a_{jk}} + \frac{c_{jk}\sigma_{\text{DL}}^2}{\rho_{jk}a_{jk}} \leq 1, \\
 & && j = 1, \dots, L, k = 1, \dots, K_j \\
 & && \sum_{k=1}^{K_j} \rho_{jk} \leq P_{\max}^{\text{DL}}, \quad j = 1, \dots, L
 \end{aligned} \tag{7.8}$$

Algorithm 1: Bisection algorithm for solving the max-min fairness problem in (7.7).

```

Input :  $\{a_{jk}\}$ ,  $\{b_{lijk}\}$ ,  $P_{\max}^{\text{DL}}$ ,  $\epsilon$ 
Output:  $\gamma^{\text{lower}}$ ,  $\{\rho_{jk}^{\text{opt}}\}$ 

/* Initialization */  

 $\gamma^{\text{lower}} \leftarrow 0$   

 $\gamma^{\text{upper}} \leftarrow \min_{j,k} P_{\max}^{\text{DL}} a_{jk} / \sigma_{\text{DL}}^2$   

 $\rho_{jk}^{\text{opt}} \leftarrow 0$  for all  $j, k$   

/* Bisection over potential SINR values */  

do  

     $\gamma^{\text{candidate}} \leftarrow \frac{\gamma^{\text{lower}} + \gamma^{\text{upper}}}{2}$   

    Solve the linear feasibility problem  

        find  $\rho_{11} \geq 0, \dots, \rho_{LK_L} \geq 0$   

        subject to  $\gamma^{\text{candidate}} \left( \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} b_{lijk} + \sigma_{\text{DL}}^2 \right) - \rho_{jk} a_{jk} \leq 0,$   

         $j = 1, \dots, L, \quad k = 1, \dots, K_j$   

         $\sum_{k=1}^{K_j} \rho_{jk} \leq P_{\max}^{\text{DL}}, \quad j = 1, \dots, L$   

        if feasible then  

             $\gamma^{\text{lower}} \leftarrow \gamma^{\text{candidate}}$   

             $\rho_{jk}^{\text{opt}} \leftarrow \rho_{jk}$  for all  $j, k$ , based on the solution to the  

            feasibility problem  

        else  

             $\gamma^{\text{upper}} \leftarrow \gamma^{\text{candidate}}$   

while  $\gamma^{\text{upper}} - \gamma^{\text{lower}} > \epsilon$ 

```

which is a geometric program.

Proof. Inserting the max product SINR utility function into (7.6) yields

$$\begin{aligned} & \underset{\rho_{11} \geq 0, \dots, \rho_{LK_L} \geq 0}{\text{maximize}} \quad \prod_{j=1}^L \prod_{k=1}^{K_j} \frac{\rho_{jk} a_{jk}}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} b_{lijk} + \sigma_{\text{DL}}^2} \\ & \text{subject to} \quad \sum_{k=1}^{K_j} \rho_{jk} \leq P_{\text{max}}^{\text{DL}}, \quad j = 1, \dots, L \end{aligned} \quad (7.9)$$

from which (7.8) follows by introducing the auxiliary variables c_{jk} such that

$$c_{jk} \left(\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} b_{lijk} + \sigma_{\text{DL}}^2 \right) \leq \rho_{jk} a_{jk} \quad (7.10)$$

and by maximizing $\prod_{j=1}^L \prod_{k=1}^{K_j} c_{jk}$ instead. Finally, we notice that the objective function and the constraint functions in (7.8) are posynomials, thus it is a geometric program. \square

This theorem shows that the max product SINR problem can also be solved efficiently, but this time as a geometric program. The notion of a geometric program is defined in Appendix B.6 on p. 575. Geometric programs can also be transformed into convex programs by a change of variable; see [66, 81] or Appendix B.6 for details.

The optimal power allocation solutions in Theorems 7.1 and 7.2 are straightforward to obtain using either general-purpose solvers of linear/geometric programs or dedicated implementations. Alternatively, distributed algorithms can be developed by using decomposition methods [250]. The considered utility maximization problems have a total power constraint per BS, but other types of constraints can be also included; for example, to limit the minimum or maximum power allocated per UE or to limit the average interference that is caused to a particular UE in the network. The algorithmic solutions provided by Theorems 7.1 and 7.2 can still be applied, as long as the additional constraints are linear or geometric, respectively.

Remark 7.1 (Per-antenna power constraints). One type of additional constraint that cannot be treated in the framework described above is per-antenna (peak) power constraints, which are required in practice to handle the limited dynamic range of the individual power amplifiers. If all BS antennas exhibit roughly the same large-scale fading value to a UE (in terms of the diagonal elements of the spatial correlation matrix being roughly the same), the precoding will divide the transmit power almost equally over the antennas. The small-scale fading creates variations within a coherence block, but these are averaged out since the signals from many different coherence blocks separated in the frequency domain are transmitted simultaneously. The fact that the signals to different UEs are spatially multiplexed also contribute to this effect. In these cases, the per-antenna constraints can be reasonably neglected. Otherwise, we need to find an alternative solution. One approach is to take the solution from one of the optimization problems above and then heuristically reduce the total transmit power of each BS so that all per-antenna constraints are satisfied. It is also possible to optimize the precoding vectors jointly with the power allocation, as done in [346, 367, 46], but then the number of optimization variables is proportional to the number of antennas and, thus, very large in Massive MIMO.

Example of DL SE with Different Utility Functions

To illustrate the effect of different power allocation schemes on the individual SEs of the UEs, we continue the running example that was defined in Section 4.1.3 on p. 288. We consider $M = 100$, $K = 10$, $f = 2$, and spatially correlated channels based on the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. Moreover, we assume that the total DL power per BS is $P_{\max}^{\text{DL}} = 30 \text{ dBm}$. Except for pilots, all samples in each coherence block are used for DL data transmission.

Figure 7.2 shows CDF curves of the UEs' individual DL SEs, where the randomness is due to the UE locations and shadow fading realizations. We consider MR, RZF, and M-MMSE precoding and compare three power allocation schemes: *i*) Equal power allocation of 20 dBm per UE; *ii*) Max-min fairness; *iii*) Max product SINR. The latter two are implemented using Theorems 7.1 and 7.2.

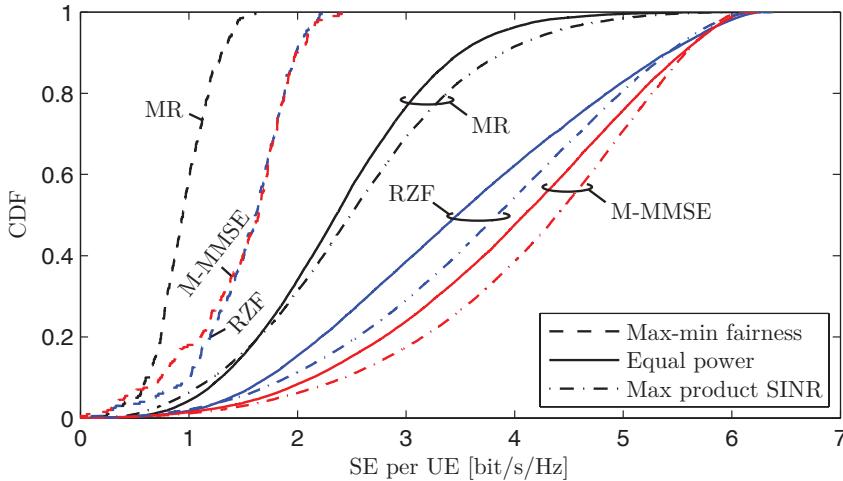


Figure 7.2: CDF of the DL SE per UE for the running example with $M = 100$, $K = 10$, $f = 2$, and using the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. We compare three power allocation schemes applied with MR, RZF, or M-MMSE precoding.

The general observation from Figure 7.2 is that, irrespective of the precoding scheme, the CDF curve with max product SINR is to the right of the equal power allocation curve, which in turn is to the right of the max-min fairness curve. This basically means that every UE will, statistically speaking, achieve better performance with max product SINR power allocation than with the other schemes.

If we look at the tails of the CDF curves, we notice that this general observation is not fully accurate. The UEs with the 10% strongest channels could achieve slightly higher SE using equal power allocation. Similarly, a small fraction of the UEs is better off with max-min fairness. For a UE at a given location, the simulation data shows that there is around 5% chance that this UE will achieve its highest SE with max-min fairness. However, for a UE that is placed at a random location and that requires a certain SE level, the chance that it will be supported is always larger with max product SINR power allocation than with max-min fairness—because the CDF curves with max product SINR in Figure 7.2 are to the right of the max-min fairness curves, except at SE levels very close to zero. Table 7.1 shows the SE level that is guaranteed to 95% of

Scheme	Max-min fairness	Equal power	Max product SINR
M-MMSE	0.40	1.67	1.83
RZF	0.67	1.38	1.44
MR	0.49	1.05	0.91

Table 7.1: SE in bit/s/Hz that can be guaranteed to 95% of the UEs, at random locations, with $M = 100$, $K = 10$, and $f = 2$. The results correspond to the 0.05 percentile in Figure 7.2. The largest value for each precoding scheme is in bold face.

the UEs and we note that max-min fairness provides the smallest SEs. This non-intuitive behavior is explained by the fact that the max-min fairness allocation makes every UE operate at the SE achieved by the weakest UE in the entire network. Statistically speaking, everyone gets a higher performance with max product SINR power allocation.

The three precoding schemes behave as expected, with M-MMSE giving the highest SEs and MR giving the lowest SEs. However, RZF actually works slightly better than M-MMSE with max-min fairness power allocation, as can be seen from both the CDF curves and Table 7.1. This is due to the equal power allocation assumption made for the channel estimation in this example. Since the precoding vectors depend on the UL powers, the result is that M-MMSE overemphasizes on interference suppression in the DL. This changes if we also optimize the UL powers, as described in Section 7.1.2 below.

In summary, DL power allocation can greatly affect the SE distribution among the UEs. The max product SINR utility provides a good balance between sum SE and fairness. In this simulation, it provides around $2.5 \times$ higher sum SE than max-min fairness with almost negligible SE losses for the statistically weakest UEs. We will compare these power allocation schemes again in the case study of Section 7.7. The results are similar, but one key difference is that the max-min fairness utility provides some benefits for the weakest UEs.

7.1.2 Uplink Power Control

It is more complicated to optimize the transmit powers in the UL than in the DL because the UL powers affect not only the UL data

transmission but also the quality of the channel estimates and indirectly the combining vectors. This makes the power control in practical systems (that operate with imperfect CSI) different from the classic power control algorithms studied in [369, 360, 81], which assume that the channels are perfectly known. The SE expressions for the UL in Theorem 4.1 on p. 276 and Theorem 4.4 on p. 302 appear to be too complicated for finding the optimal transmit powers by convex optimization. However, in the special case of i.i.d. Rayleigh fading channels, one can obtain closed-form SE expressions with MR and ZF that can be utilized for tractable power optimization [178, 359, 80].

We consider the general case with arbitrary spatial correlation matrices and combining schemes and assume that each UE has a maximum UL transmit power of $P_{\max}^{\text{UL}} > 0$. To pave the way for the use of low-resolution ADCs at the BSs (see Section 6.4.1 on p. 442), it is important to limit the power differences between the intra-cell signals that are received at the BS. Recall that $\beta_{jk}^j = \frac{1}{M_j} \text{tr}(\mathbf{R}_{jk}^j)$ denotes the average channel gain from UE k in cell j to any antenna at BS j . In a cellular network, there are commonly channel gain differences of up to 50 dB between the UEs in a cell, but when using low-resolution ADCs it is beneficial to substantially reduce these differences to avoid that weak signals drown in the quantization distortion from stronger signals.

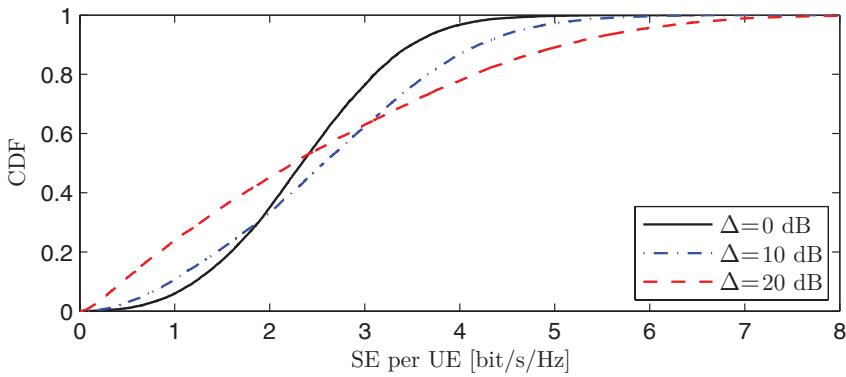
In LTE systems, this issue is managed by letting UEs at the cell edge transmit at maximum power and by gradually reducing the power for UEs that are located closer to the BS [288]. Inspired by this principle, we define a maximum received power ratio $\Delta \geq 0$ dB and consider the heuristic power control policy

$$p_{jk} = \begin{cases} P_{\max}^{\text{UL}} & \Delta > \frac{\beta_{jk}^j}{\beta_{j,\min}^j} \\ P_{\max}^{\text{UL}} \Delta \frac{\beta_{j,\min}^j}{\beta_{jk}^j} & \Delta \leq \frac{\beta_{jk}^j}{\beta_{j,\min}^j} \end{cases} \quad (7.11)$$

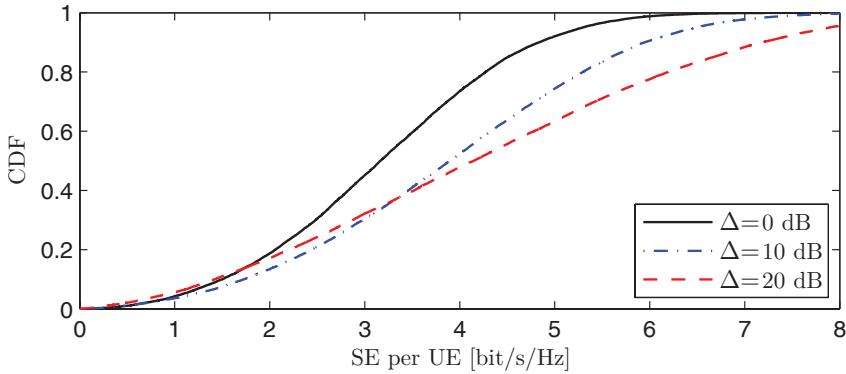
for $k = 1, \dots, K_j$ in cell j , where $\beta_{j,\min}^j = \min_{i=1, \dots, K_j} \beta_{ji}^j$ is the average channel gain of the weakest UE in cell j . We refer to (7.11) as power control, rather than power allocation, since no power can be reallocated between the UEs in the UL.

With the power control policy in (7.11), the UE with the smallest channel gain in the cell will use the full transmit power P_{\max}^{UL} . The UEs with slightly larger channel gains that satisfy $\beta_{jk}^j < \beta_{j,\min}^j \Delta$ will also use maximum power, while UEs with better channel conditions will use less power so that the received signal power becomes at most Δ times larger than the received signal of the weakest UE. This heuristic policy is applied independently in every cell and does not take the spatial channel characteristics of the UEs into account, thus there is certainly room for improvements. The next example shows how the choice of Δ affects the SE of the UEs.

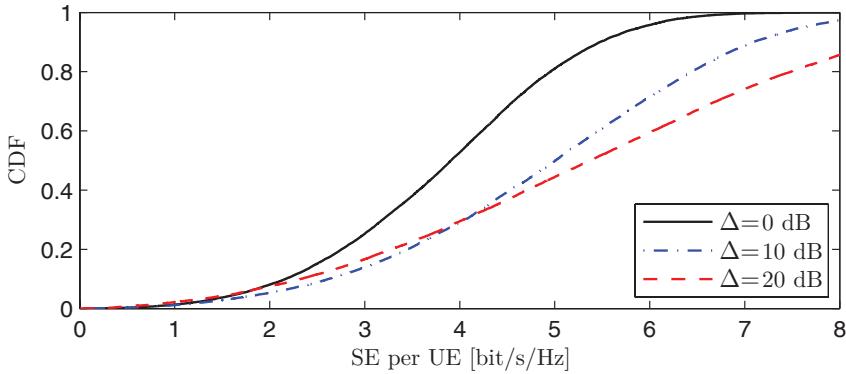
Remark 7.2 (Unequal pilot and data powers). We have assumed that the UL transmit power used for pilot and data are equal, since having large differences in power between consecutive UL samples is generally not feasible due to hardware constraints, but some modulation schemes allow for it. By relaxing this constraint, we can assign unequal powers for UL data and pilot transmission. This gives greater flexibility to the power control. It can be used to set a fixed pilot power level and only optimize the data power, which leads to optimization problems with a similar structure as the DL utility maximization problem in (7.6). This approach has been adopted in [193, 210], among others, and the resulting power control solutions build on the heritage from utility maximization with perfect CSI [81]. Alternatively, the pilot and data powers can be treated as separate variables and optimized jointly, as carried out in [135, 247, 80] for some different utility functions. In particular, under the assumption of i.i.d. Rayleigh fading and MR or ZF combining, the utility maximization may lead to geometric programs that can be solved to global optimality. The gain from having an unequal pilot and data powers are large for UEs with weak channel conditions, which can then allocate more power for channel estimation and, thus, enable better receive combining (e.g., a larger array gain and better interference suppression).



(a) MR combining with the heuristic power control policy in (7.11).



(b) RZF combining with the heuristic power control policy in (7.11).



(c) M-MMSE combining with the heuristic power control policy in (7.11).

Figure 7.3: CDF of the UL SE per UE with $M = 100$, $K = 10$, $f = 2$, and using the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$.

Example of UL SE with Heuristic Power Control

We will now exemplify the SE distribution achieved by the heuristic power control scheme in (7.11) with MR, RZF, and M-MMSE combining. We continue the running example that was defined in Section 4.1.3 on p. 288. We consider $M = 100$, $K = 10$, $f = 2$, and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. The maximum UL transmit power is $P_{\max}^{\text{UL}} = 20 \text{ dBm}$ per UE. Except for pilots, all samples in each coherence block are used for UL data transmission.

Figure 7.3 shows CDF curves of the UEs' individual SE, where the randomness is due to the UE locations and shadow fading realizations. For each scheme, we consider $\Delta \in \{0, 10, 20\} \text{ dB}$.

A first observation is that there are large variations in SE, even for the case of $\Delta = 0 \text{ dB}$ in which all UEs in a cell adapt their transmit power to reach the same received signal power. There are several reasons for this. First, it is the UE with the worst channel condition in a cell that determines the SE level of all UEs in the cell and its location and shadow fading realization vary substantially. Second, the interference from other cells varies as well, depending on their UE distribution and power control. Third, the spatial channel correlation creates further variations, since the UEs with the smallest channel gain may have similar spatial characteristics as an interfering UE and the UE with the best channel gain may have spatial characteristics that are different from all interfering UEs. When running the same simulation with uncorrelated fading, the CDF curves are compressed around their mean values, in the sense that the lower and upper tails are smaller.

The choice of receive combining scheme impacts the shape of the CDF curves. With MR, the weakest UEs greatly benefit from using a small Δ , since the intra-cell interference from UEs with good channel conditions would be large unless the power control reduces their power. UEs with good channel conditions prefer a larger Δ , since they can then obtain a higher SE. With RZF or M-MMSE, the intra-cell interference is suppressed to such an extent that the UEs with good channels can transmit their signals to get a larger received signal power, while the detrimental effect on the weakest UEs is almost negligible. This shows that, by having a larger received power, the increased quality of the

channel estimate improves the receive combining to such an extent that it almost entirely counteracts the increased interference level that it creates.

7.2 Spatial Resource Allocation

There can be thousands of UEs residing in each cell of a cellular network, but only a small fraction of them are in general active in a given coherence block. The intermittent activity is both created by the end user and by the bursty nature of packet-based data traffic. While the number of active UEs can change rapidly, the number of pilot sequences τ_p is essentially fixed in practice—one BS cannot decide to reduce the number of pilots if another BS needs all the pilots. The number can change over the course of the day, to adapt to the substantial long-term variations in average traffic load that occur in practice [26], but it is not practical to adapt τ_p to short-term traffic variations.

In this section, we outline some key considerations in the allocation of spatial resources, which includes the pilot assignment, the interplay between spatial multiplexing and time-frequency scheduling, and the impact of traffic load variations on the sum SE. The number of pilot sequences and how these are allocated to the UEs will play a key role.

7.2.1 Pilot Assignment

In cellular networks, every UE that intends to connect to a BS must go through a network entry procedure. This refers to all the functions that a UE goes through in order to establish a communication link with the BS for data transmission and reception. In the LTE standard, this procedure is called random access (RA) and relies on a contention-based protocol. The development of RA procedures for Massive MIMO systems is still in its infancy and is therefore not treated in this monograph, but a few details are given in Remark 7.4 later. Next, we assume that the RA procedure has been successfully completed, and we will only deal with the pilot assignment problem. Once a UE is connected to a cell, it is assigned to one of the pilot sequences that are available in that cell. The pilot assignment implicitly determines which other

UEs in the network will cause pilot contamination to the UE. The assignment thereby impacts the SE that the UE will achieve. Joint optimization of the pilot assignment across cells is a way to empower the weakest UEs [192, 363, 378], which are the ones most susceptible to the reduced estimation quality and the coherent interference caused by pilot contamination. The pilot assignment is a combinatorial problem in which BS j should select K_j pilots from the set of τ_p pilots and assign them properly to its K_j UEs. The computational complexity grows very rapidly with the number of UEs and is therefore of limited practical use.

There are two categories of heuristic solutions in the literature. The first category contains greedy algorithms [192, 363, 378] that assign and reassign the pilots to the UEs in an iterative manner to improve a utility function. For example, if a UE at the cell edge is greatly affected by pilot contamination, it can switch pilot with a UE in the cell-center, motivated by the fact that UEs with strong channel conditions are less affected by pilot contamination. The second category consists of predetermined pilot reuse patterns [154, 358, 49], where the τ_p pilots are divided into f groups with τ_p/f pilots each. The integer f is called the pilot reuse factor. Each cell is associated with one of these disjunct pilot groups, according to a predetermined pattern. This approach was taken in the running example, for which Figure 4.4b shows three different pilot reuse patterns in a symmetric network. An example with $f = 3$ in an asymmetric cellular network is provided in Figure 7.4, where each color/pattern represents one pilot group. The classical four-color theorem proves that $f = 4$ is sufficient to make sure that every cell belongs to another cell group than its immediate neighbors [130]. Note that the use of pilot reuse patterns resembles the frequency reuse patterns in GSM, but in contrast to legacy systems the reuse patterns are only applied to the pilot transmission while all cells send payload data in parallel over the full bandwidth (i.e., universal frequency reuse). The reuse pattern should be designed to automatically give a large spatial separation between UEs that use the same pilot, to alleviate the need for further coordination between cells. Note that the two categories described above can also be used jointly, by having pilot reuse patterns

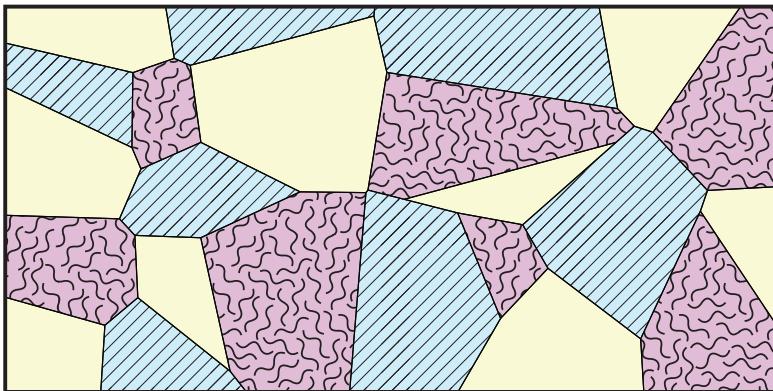


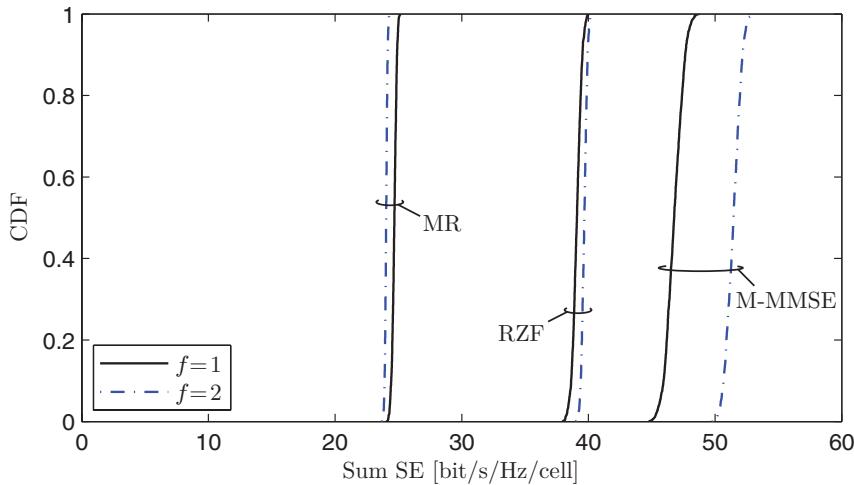
Figure 7.4: Illustration of the division of an asymmetric cellular network into $f = 3$ disjoint pilot groups, where f is called the pilot reuse factor. Each group is indicated with a distinct color and pattern.

and then apply a greedy algorithm for pilot assignment within each pilot group.

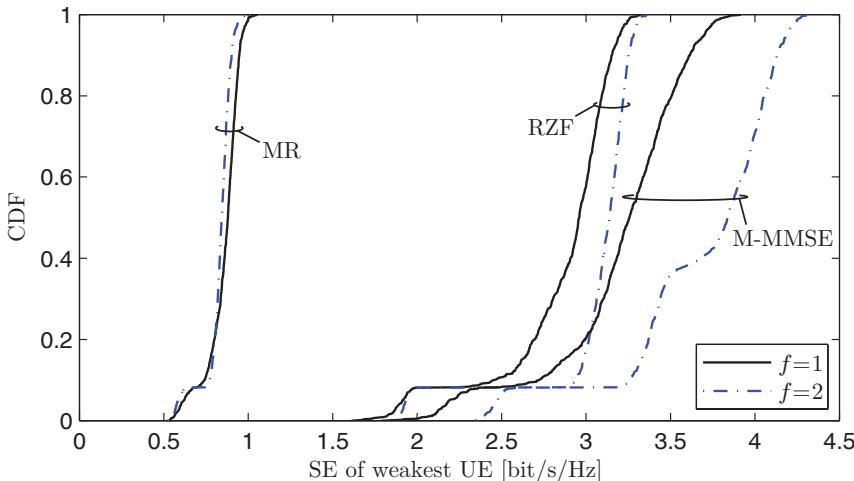
A pilot assignment that is good for a UE in the UL, in terms of giving little coherent interference, is not necessarily good for the UE in the DL. For example, the spatial channel correlation can be very different depending on who is transmitting, as illustrated in Figure 4.16. Channel gain differences can also create asymmetry since the UL interference is sent from UEs in other cells and the DL interference is sent from BSs in other cells.

To investigate the impact of pilot assignment on the SEs, we revisit the running example defined in Section 4.1.3 on p. 288. We consider one snapshot of the network with $K = 10$ UEs at fixed locations in each cell. We focus on the UL with $M = 100$ antennas, 20 dBm transmit power per UE, and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. Except for the pilots, all samples in the coherence blocks are used for UL data transmission.

In order to show the impact that different types of pilot assignment can have on the SE, we consider the pilot reuse factors $f = 1$ and $f = 2$, as illustrated in Figure 4.4b. We uniformly randomize the pilot assignment in every cell and pilot group, and we plot CDF curves of the resulting SE variations. Figure 7.5a shows the sum SE, averaged



(a) Impact of pilot assignment on the average sum SE.



(b) Impact of pilot assignment on the SE of the weakest UE in a random cell.

Figure 7.5: CDFs of the average UL sum SE among the cells and the weakest UE's SE in an arbitrary cell for different random pilot assignments. The running example is considered for different pilot reuse factors f , $M = 100$, $K = 10$, and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$.

over the cells, and Figure 7.5b shows the SE of the UE with the weakest channel (smallest average channel gain) in an arbitrarily selected cell. Results are provided for MR, RZF, and M-MMSE combining.

The sum SE in Figure 7.5a is greatly affected by the choice of combining scheme and slightly affected by the choice of pilot reuse factor. More importantly, all the CDF curves are almost vertical. The largest value is only 3%–9% higher than the smallest value, with the largest gain achieved with M-MMSE. Hence, there is little to gain in sum SE from optimizing the pilot assignment (e.g., using a greedy algorithm).

If we look instead at the SE in Figure 7.5b of the UE with the weakest channel in a randomly selected cell, we observe that it fluctuates substantially. The curves are not smooth since the pilot assignment problem is combinatorial. In particular, we notice that there is one interfering UE that provides particularly high pilot contamination, which results in the discontinuity at 0.1 probability (which is the probability that two particular UEs use the same pilot when assigning 10 pilots uniformly at random). The absolute SE variations are small with MR, since its performance is mainly limited by non-coherent intra-cell interference, while the weakest UE has much to gain from optimizing the pilot assignment when using RZF or M-MMSE combining.

In summary, the sum SE can be optimized by selecting an appropriate pilot reuse factor. The exact pilot assignment within a cell has little impact on the sum SE, but can greatly affect how the sum SE is divided among the UEs. A reasonable goal for the pilot assignment is to improve the fairness, by making sure that UEs with bad channel conditions are assigned pilots that give little pilot contamination.

Remark 7.3 (Joint spatial-division and multiplexing). An interesting transmission scheme that allows pilot reuse within a cell is joint spatial-division and multiplexing (JSDM) [7, 235]. The key idea behind JSDM is to capitalize on spatial channel correlation of large antenna arrays. This allows us to partition a cell into geographical regions which are characterized by spatial correlation matrices that are almost spatially orthogonal (see Definition 4.2 on p. 341). Suppose we can partition the K UEs in a cell into G groups consisting of K_g UEs, $g = 1, \dots, G$, respectively, and denote by $\mathbf{h}_{gk} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R}_{gk})$ the channel of the k th

UE in the g th group. The UE grouping is done in such a way that the correlation matrices of the UEs in each group are very similar, in the sense that $\frac{1}{\text{tr}(\mathbf{R}_{gk})}\mathbf{R}_{gk} \approx \frac{1}{\text{tr}(\mathbf{R}_{gl})}\mathbf{R}_{gl}$ for all g, k, l , but almost orthogonal to the correlation matrices of the UEs in other groups: $\text{tr}(\mathbf{R}_{gk}\mathbf{R}_{hl}) \approx 0$ for all k, l and $g \neq h$. As there is little interference between the groups by design, we can reuse the same pilot sequences for each group without causing prohibitive amounts of pilot contamination. Due to the intra-cell pilot reuse, a larger number of UEs can be served per cell for a given τ_p . Although JSMD is a theoretically very appealing concept, it is at the time of writing this monograph unclear if practical channels have the necessary orthogonality properties.

Remark 7.4 (Random access in Massive MIMO). The RA is a contention-based procedure by which a UE can connect to a cell in the network, either as an initial connection to the network or to switch cell. This procedure is used in LTE and, in its basic form, operates as follows: Each accessing UE acquires basic synchronization from the BS (e.g., determining LTE parameters, frequency synchronization, and frame timing) and makes use of the so-called random access channel (RACH)¹ to transmit a randomly selected pilot-like sequence. Since the accessing UEs are not coordinated in selecting their sequences, collisions may occur. After detecting the selected sequences and trying to identify the UEs that are using a given sequence (collision resolution), the BS broadcasts a response message, informing the identified UEs that the RA procedure has been successful while providing physical parameters (e.g., timing adjustments). The detected UEs send connection requests to further specify the resources needed for data transmission. On the other hand, the undetected UEs repeat the RA procedure after a random waiting time, until successful notification. As exemplified by this monograph, the benefits of Massive MIMO in terms of SE, EE and HE are nowadays well understood. On the other hand, the potential impact that Massive MIMO has on the network access functionalities has received less attention so far. Recent attempts in this direction can be found in [76, 41, 286, 305]. In [41], the authors exploit the

¹The RACH is typically composed of a specified set of consecutive OFDM symbols and adjacent subcarriers.

channel hardening and spatial resolution of Massive MIMO to resolve collisions distributively. This approach can be used as an add-on to conventional resolution mechanisms. A coded RA protocol leveraging the channel hardening properties of Massive MIMO is presented in [305]. The large number of antennas were used in [286] to obtain accurate estimates of the timing misalignments of the accessing UEs, which were used to develop an RA procedure that resolves collisions with high probability.

7.2.2 Scheduling

While conventional BSs manage intra-cell interference by scheduling the active UEs in different coherence blocks, Massive MIMOs greatly alleviates this issue since the interference is mitigated spatially, by receive combining and transmit precoding. There is no fundamental upper limit on how many UEs can be served per cell in a given coherence block—it is even practically feasible to serve more UEs than there are BS antennas² and to assign a pilot sequence to multiple UEs within a cell (cf. Remark 7.3). As we increase the number of UEs, for a fixed number of BS antennas, the sum SE first increases, then reaches a maximum, and finally starts to decay again. As long as the increased multiplexing gain (i.e., the number of SEs that are summed up) outweighs the larger pilot overhead and extra interference (i.e., the reduction of the individual SEs), the sum SE increases. It was proved in [49] that one should not use more than half the samples in a coherence block for pilots, as $M \rightarrow \infty$, but in practice, the number of pilots can be optimized based on the anticipated number of UEs.

These behaviors are illustrated in Figure 7.6. The horizontal axis represents the number of single-antenna UEs or, equivalently, the number of data streams that are transmitted or received by the BS. The vertical axis illustrates the sum SE in a scenario that can either represent the UL or DL. Three operating regimes are identified in the figure. The first is the *multiplexing regime* in which the sum SE grows almost linearly with the number of served UEs, because the increased multiplexing gain

²All receive combining and transmit precoding schemes described in Section 4, except ZF, can be applied with an arbitrary number of UEs.

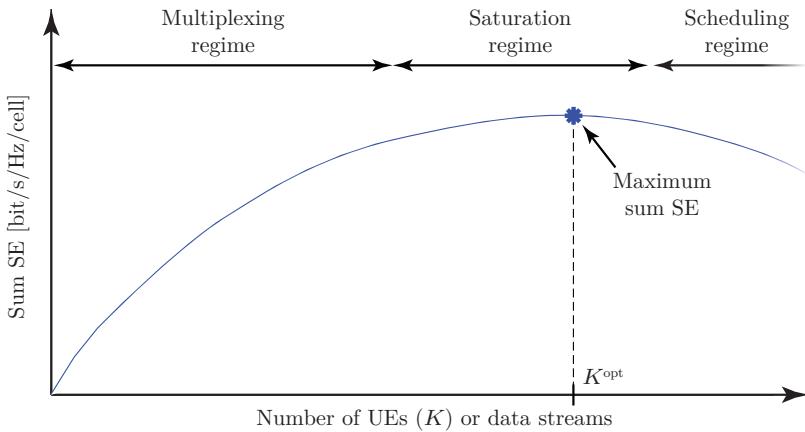


Figure 7.6: Illustration of how the number of UEs affect the sum SE, of either the UL or DL. There is a maximum at some K^{opt} . In order to maximize the area throughput, the network should serve all K UEs over the entire bandwidth when $K \leq K^{\text{opt}}$, while scheduling should be used to bring down the effective number of UEs to K^{opt} when $K > K^{\text{opt}}$.

dominates over the extra interference. The growth slows down in the *saturation regime*, where we eventually reach a number of UEs that provides the maximum sum SE. Beyond this point, for every additional UE, the sum SE decreases. A well-dimensioned network operates in the saturation regime and occasionally in the multiplexing regime (when the data traffic is low). The scheduling of UEs is then trivial because all time-frequency resources are allocated to all active UEs in an effort to maximize the throughput of the network.

The regime beyond the maximum point is called the *scheduling regime*. As the sum SE decreases, it does not pay off to serve all of them in parallel by spatial multiplexing. Time-frequency scheduling can then be used to reduce the effective number of UEs per coherence block to operate close to the maximum sum SE point. Conventional cellular networks with 1–8 antennas reach the scheduling regime as soon as there are more than a few active UEs in the cell, while the next example illustrates that Massive MIMO can handle many tens of UEs before reaching the scheduling regime.

Due to the natural traffic variations over the day, from rush hours in the mornings and evenings to low-traffic hours in the late nights, cellular networks are often dimensioned for the highest anticipated data traffic load. Consequently, during parts of the day, there are not sufficiently many active UEs to reach the maximum sum SE by sending one data stream per UE. It is then important to recall that we have focused on the single-stream transmission to single-antenna UEs in this monograph, while multiantenna UEs have only been briefly discussed in Remark 1.4 on p. 203. The key benefit of having multiple UE antennas is that we can spatially multiplex several data streams to/from the UE, for example, by treating each antenna as a separate UE in the signal processing. In a practical scenario where some UEs are equipped with multiple antennas, it is the scheduling algorithm that determines how many streams should be assigned to each UE. Recall that the horizontal axis in Figure 7.6 represents the number of data streams that are simultaneously spatially multiplexed by the BS, which not necessarily equals the number of UEs. When there are few active UEs in a cell—fewer than what is needed to reach the saturation regime by single-stream transmission—we can increase the sum SE by sending multiple data streams to some of the UEs. In contrast, when there are sufficiently many UEs to reach the saturation regime by sending only one data stream per UE, this is the preferred choice [194, 52]. In principle, we could instead schedule a subset of the UEs and send multiple streams to each of them. But since the channels to the different antennas of a UE can be strongly spatially correlated, the sum SE is typically larger when we schedule a larger number of UEs and send only one stream to each of them.

We will now exemplify the multiplexing, saturation, and scheduling regimes by revisiting the running example that was defined in Section 4.1.3 on p. 288. We consider $M = 100$ antennas, 20 dBm transmit power per UE in both UL and DL, and the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. Figure 7.7 shows the DL sum SE as a function of the number of active UEs, for two different channel coherence block lengths: $\tau_c = 200$ and $\tau_c = 400$. In each case, $\tau_p = fK$ samples are used for pilots and the remaining samples are used for DL data transmission. For each number of antennas, each coherence block length,

and each of MR, RZF, and M-MMSE precoding, we use the pilot reuse factor $f \in \{1, 2\}$ and bounding technique that maximizes the sum SE.

The simulation results match well with the schematic behavior described in Figure 7.6. The sum SE grows steeply with K for the first 20 UEs. It then keeps increasing linearly with K when continuing to increase the number of UEs, but with a smaller slope. Hence, the SE per UE decays while the sum SE grows. The pilot reuse factor $f = 2$ is beneficial when K is small, while $f = 1$ gives the largest sum SE as K increases (since it becomes important to reduce the pilot overhead). The location of the maximum value depends on the length of the channel coherence block and precoding scheme. With $\tau_c = 200$, the maximum sum SE is achieved at around $K = 50$ for all precoding schemes. All SE values increase when using $\tau_c = 400$ instead since the relative size of the pilot overhead is reduced. The maximum sum SE is now achieved at $K = 60$ for RZF, while MR and M-MMSE can handle 80 UEs without reaching the maximum point.

The numerical comparison between different precoding schemes in Section 4.3 on p. 316 focused on $K = 10$. Figure 7.7 shows that the differences between the precoding schemes grow as more UEs are added. M-MMSE and RZF can achieve more than twice the SE of MR. RZF is competitive with M-MMSE for $K \leq 40$, but the performance gap is substantial for larger K .

In summary, Massive MIMO leads to a paradigm shift when it comes to scheduling, which becomes the last resort to deal with peak traffic loads, instead of the main method for resource allocation—as in conventional networks. The maximum sum SE in this example is achieved when scheduling tens of UEs and having an antenna-UE ratio of around $M/K = 100/50 = 2$. The SE per UE is not particularly large at the sum SE maximizing point. With $K = 50$, the average SE per UE is 1.7–2.0 bit/s/Hz with M-MMSE and 0.8–0.9 bit/s/Hz with MR. This can be achieved in practice using 4-QAM modulation and different channel codes. The average throughput per UE is still substantial because there is no time/frequency scheduling, so every UE enjoys the full bandwidth. Recall that, in the running example, we consider 20 MHz bandwidth, which with M-MMSE leads to an average

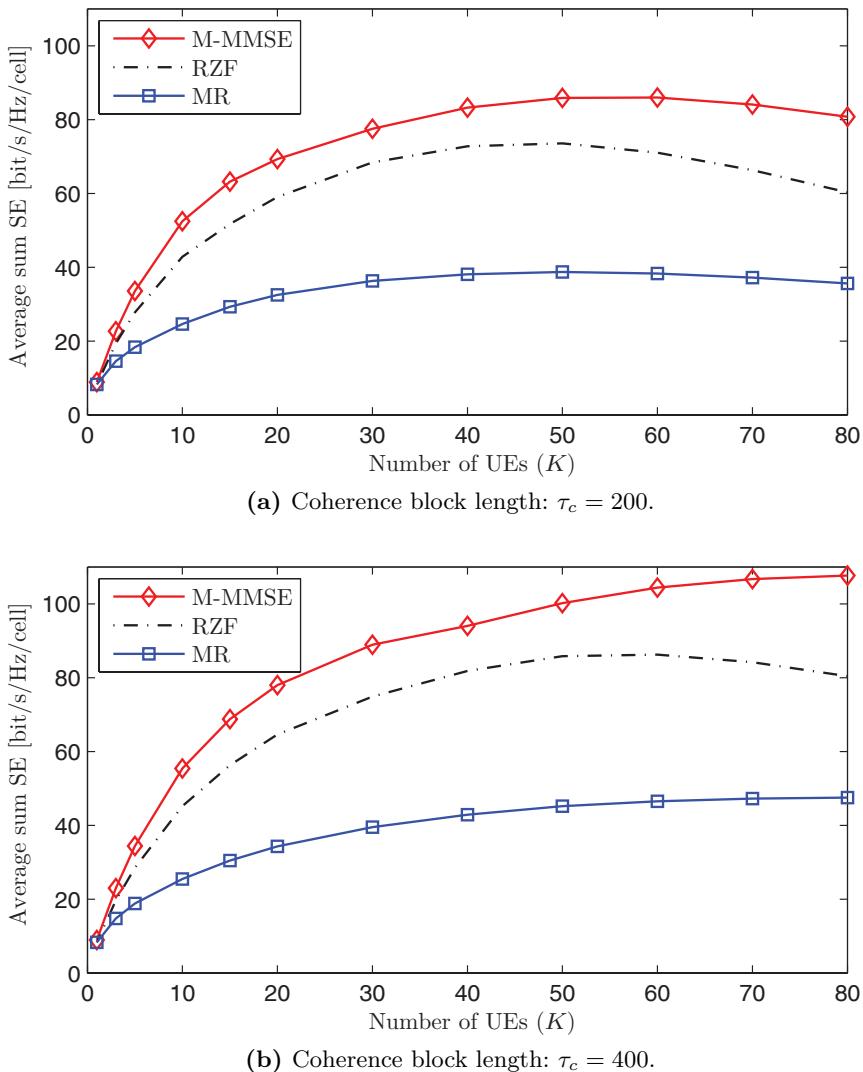


Figure 7.7: Average DL sum SE as a function of the number of UEs, for $M = 100$ and different precoding schemes. We consider the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$. The pilot reuse factor is optimized for each point on the curves.

of 34–40 Mbit/s for each of the 50 UEs and a total of 1.7–2.0 Gbit/s/cell. The latter number can be compared with the 64 Mbit/s/cell that is achieved by a contemporary LTE system for an equivalent bandwidth (see Remark 4.1 on p. 291). Hence, the simulated Massive MIMO setup provides more than an order-of-magnitude higher cell throughput.

7.2.3 Impact of Traffic Load Variations

As mentioned in the beginning of Section 7.2, the number of pilots τ_p is typically constant in practice, while the number of UEs that have data to transmit/receive can change rapidly due to user behavior and the bursty nature of packet transmission. The number of active UEs in an arbitrary coherence block can be treated as a random variable and Poisson distributions are commonly used to model such traffic variations. For example, $K^{\text{active}} \sim \text{Po}(K)$ is a random integer with mean K and standard deviation \sqrt{K} . If we use time/frequency scheduling to only serve τ_p UEs when $K^{\text{active}} > \tau_p$, we can use $\min(K^{\text{active}}, \tau_p)$ as a way to randomly generate the number of active UEs in an arbitrary coherence block. This distribution is shown in Figure 7.8 for $K \in \{1, 10, 20, 40\}$ and $\tau_p = 40$. The figure shows that the load distribution has a Gaussian-like shape and the variations grow with K , so in a cell with more UEs, there will also be larger traffic variations. Note that when $K = \tau_p$, it happens frequently that $K^{\text{active}} > \tau_p$ and then scheduling is often needed to handle the traffic load variations.

Both the load variations and the concepts of multiplexing/saturation regimes, illustrated in Figure 7.6, are important when dimensioning the number of pilots in practice. When the traffic load is high, τ_p can be selected based on what is needed to reach the maximum sum SE point. When the traffic load is low, τ_p can be selected based on the distribution in Figure 7.8 to balance between having a low probability of pilot shortage with $K^{\text{active}} > \tau_p$ and keeping the pilot overhead low.

As long as there are sufficiently many pilots, each UE should connect to the BS that provides the best channel conditions [82], which can be measured as having the largest trace of the spatial correlation matrix. However, when there is a pilot shortage in a cell, which calls for time-

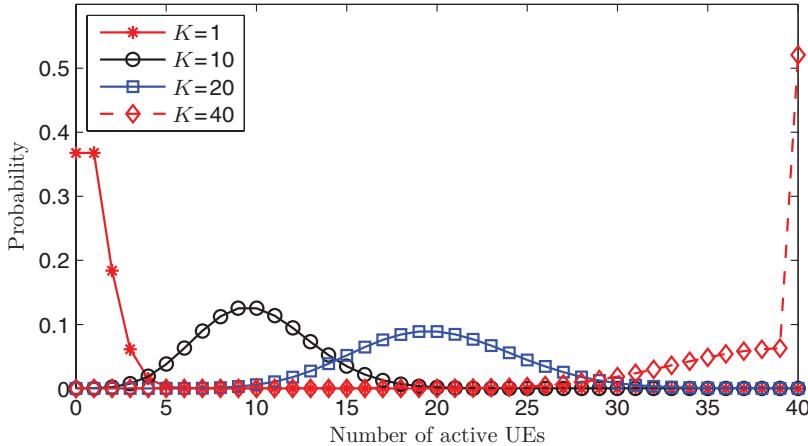


Figure 7.8: Distribution of the number of active UEs in a cell, modeled as $\min(K^{\text{active}}, \tau_p)$ where $K^{\text{active}} \sim \text{Po}(K)$ and $\tau_p = 40$. Note that the probability of $K^{\text{active}} > \tau_p$ is substantial for $K = 40$, which leads to the large probability of having 40 active UEs in that case.

frequency scheduling, it can happen that a few UEs get higher SE by connecting to neighboring cells, which currently have lower traffic load [35]. This type of load balancing is important in conventional networks, which mainly rely on time-frequency scheduling, but is less critical in Massive MIMO. For example, if $K^{\text{active}} > \tau_p$, then each UE can be scheduled in a fraction τ_p/K^{active} of all coherence blocks. If there are 10 more UEs than pilots, the fraction becomes 0.09 in a network with $\tau_p = 1$ and 0.67 in a Massive MIMO network with $\tau_p = 20$.

We will now illustrate how the sum SE of a cell is affected by having a varying number of UEs per cell. To this end, we continue the running example that was defined in Section 4.1.3 on p. 288. We consider $M = 100$ antennas, 20 dBm transmit power per UE in both UL and DL, and spatially correlated channels based on the Gaussian local scattering model with ASD $\sigma_\varphi = 10^\circ$. We denote the average number of UEs per cell as K and consider two scenarios. In the first scenario, the number of UEs is exactly K in every coherence block. In the second scenario, the number of UEs is independently randomized in each cell as $\min(K^{\text{active}}, \tau_p)$ where $K^{\text{active}} \sim \text{Po}(K)$. The number of pilot sequences

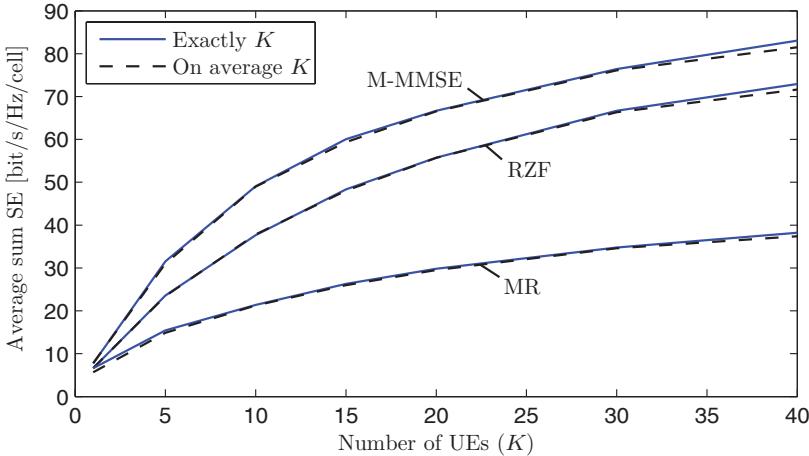


Figure 7.9: Average DL sum SE as a function of the average number of UEs for $M = 100$, $\tau_p = 40$, and different precoding schemes. In each coherence block, the number of UEs in a cell is either exactly equal to K or computed as $\min(K^{\text{active}}, \tau_p)$, where $K^{\text{active}} \sim \text{Po}(K)$. We consider the local scattering model with Gaussian angular distribution and ASD $\sigma_\varphi = 10^\circ$.

is $\tau_p = 40$ and each BS independently selects a random subset of pilots so that its own UEs use orthogonal pilots, but the pilot contamination across cells is random.

The average DL sum SE is shown in Figure 7.9 for different K , using M-MMSE, RZF, or MR precoding. These schemes behave as expected from previous examples. The key observation is that the two scenarios give basically the same sum SE in all cases. This is explained by the fact that the sum SE grows almost linearly with the number of UEs, thus constantly serving 10 UEs gives roughly the same average sum SE as switching between serving 8 and 12 UEs with equal probability. The largest deviation between the curves occurs when K is close to τ_p , in which case the sum SE is slightly smaller when having a random number of UEs. This loss is due to the scheduling because $\mathbb{E}\{\min(K^{\text{active}}, \tau_p)\} < K$ in these cases, although $\mathbb{E}\{K^{\text{active}}\} = K$.

In summary, the number of active UEs varies between coherence blocks in practice, due to the bursty traffic demand. Substantial changes can occur over a few tens of coherence times, but we need to select a

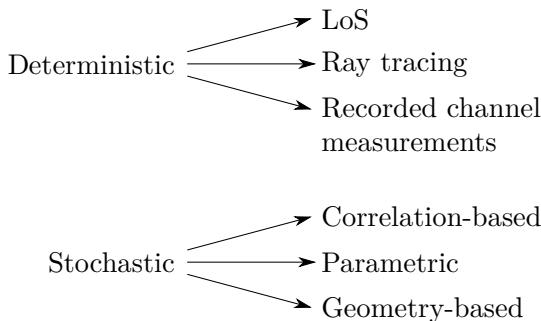


Figure 7.10: Taxonomy of wireless channel models.

fixed pilot length to accommodate different loads in the cells at different times. When quantifying the sum SE for network planning, it is not necessary to randomize the number of UEs per cell, since the results are almost the same as when we have a fixed number of UEs that equals the average number of UEs.

7.3 Channel Modeling

Realistic performance assessment of Massive MIMO systems requires the use of a channel model that reflects the main characteristics of large antenna arrays. Such a model must account at least for the array geometry, the correlation between the channel responses of different antennas, and the physical location and orientation of BSs and UEs. Our goal here is not to provide an introduction to channel modeling for MIMO systems, but rather to convey some insights from simple and analytically tractable models, which are commonly used in the research literature. For further details, the interested reader is referred to classical textbooks and tutorial papers on channel modeling for MIMO systems, such as [255, 225, 12, 88].

The taxonomy of channel models for wireless communications is provided in Figure 7.10. In a nutshell, channel models are either deterministic or stochastic. Deterministic models depend on a given environment with fixed locations of transmitters, receivers, scatterers, reflectors, etc. Examples of such models are ray tracing based on 3D-building models and recorded channel measurements. Also, the previously introduced

LoS model for a horizontal ULA in (1.23) is an example of a very simple deterministic channel model. We will provide an extension of this model to three dimensions and arbitrary array geometries in Section 7.3.1. A drawback of deterministic models is that they are only valid for a specific scenario and, consequently, do not allow for far-reaching conclusions. Moreover, the results cannot be easily reproduced by others as there are very few openly accessible databases of channel measurements and 3D-building models. However, deterministic models can provide very accurate performance predictions for their specific scenarios.

Stochastic channel models are independent of a particular environment and can be used to generate an essentially unlimited amount of channel realizations with the desired statistical properties. These models can be roughly separated into correlation-based, parametric, and geometry-based channel models. The correlated Rayleigh fading channel model, as introduced in (2.1), is an example of a correlation-based model. The channel responses are all Gaussian distributed with zero mean and entirely defined through the correlation matrices. However, it is a frequency flat-fading channel model since all multipath components are assumed to arrive with the same delay (or to be irresolvable with the sampling frequency of the system). For most of the analysis in this monograph, this is sufficient as we focus only on communication over a coherence block in which the channel is assumed to be constant (see Definition 2.2 on p. 219). In contrast, parametric channel models define stochastic distributions of the number of multipath clusters and the delay, power, angle of arrival (AoA), and angle of departure (AoD) of the individual multipath components. Examples include the Saleh-Valenzuela model [283] and extensions thereof [338, 85]. Since parametric models are independent of the geometry of the propagation environment, they can generally not be used for system-level simulations with time-evolution of the channel, caused by movements of the transmitters or receivers. Lastly, stochastic geometry-based models define a distribution of the physical location of scatterers around the transmitters and receivers. Once the locations of all scatterers are chosen, individual propagation paths are modeled in a quasi-deterministic manner. Such models are frequently adopted by standardization bodies such as the

3GPP or the Institute of Electrical and Electronics Engineers (IEEE) since they are easy to simulate, agree very well with measurements, and enable time-evolution. Geometry-based models can be seen as a balance between the two extremes of purely stochastic and purely deterministic channel modeling. We will discuss the 3GPP 3D MIMO channel model [1] as an example in Section 7.3.3.

In (1.23), we introduced a deterministic LoS channel model for a horizontal ULA in a two-dimensional setting. The only parameters of this model are the azimuth angle of the incoming wave and the channel gain. It is important to remember that this model is based on a plane-wave assumption, that is only valid if the UE is located within the far-field of the antenna arrays.³ We will now extend this channel model to three dimensions and arbitrary antenna array geometries.

Antenna arrays come in arbitrary shapes and sizes depending on the use case, carrier frequency, and the area to be covered. The most widely used arrays in cellular communications are either linear or planar, but also cylindrical arrays find applications in ground-based military communications. Figure 7.11 shows some common antenna array architectures, namely the horizontal and vertical ULA as well as the planar rectangular and cylindrical array. While a horizontal (vertical) ULA is only capable of separating UEs in the azimuth (elevation) domain, a planar or cylindrical array is capable of separating UEs in both azimuth and elevation domains. This aspect becomes very important in metropolitan areas with high-rising buildings, where UEs are located on different floors. For this reason, it is important that a channel model for Massive MIMO systems captures the 3D nature of the propagation environment. We will discuss the impact of the array geometry in detail in Section 7.4.

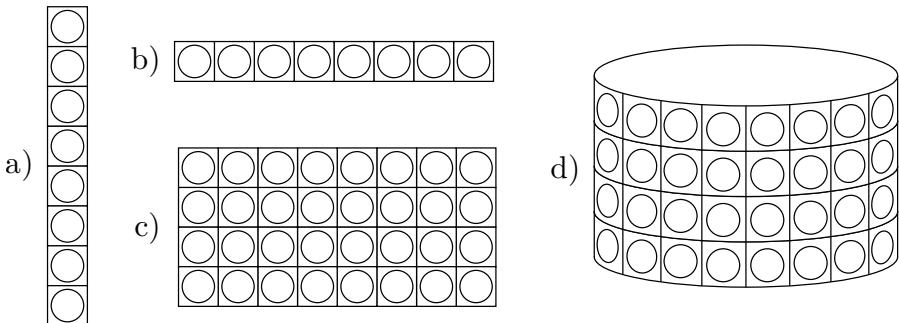


Figure 7.11: Examples of different antenna array geometries: a) linear vertical; b) linear horizontal; c) planar; d) cylindrical. Each circle represents one antenna.

7.3.1 3D LoS Model with Arbitrary Array Geometry

To proceed, we need to define the wave vector $\mathbf{k}(\varphi, \theta) \in \mathbb{R}^3$ of a plane wave with wavelength λ that impinges on the antenna array under the azimuth angle $\varphi \in [-\pi, \pi)$ and the elevation angle $\theta \in [-\pi/2, \pi/2)$:

$$\mathbf{k}(\varphi, \theta) = \frac{2\pi}{\lambda} \begin{bmatrix} \cos(\theta) \cos(\varphi) \\ \cos(\theta) \sin(\varphi) \\ \sin(\theta) \end{bmatrix}. \quad (7.12)$$

The wave vector $\mathbf{k}(\varphi, \theta)$ describes the phase variation of a plane wave with respect to the three Cartesian coordinates (see Figure 7.12). Thus, the wave observed at location $\mathbf{u} \in \mathbb{R}^3$ experiences a phase shift of $\mathbf{k}(\varphi, \theta)^T \mathbf{u}$ with respect to the origin. Consequently, the LoS channel response $\mathbf{h} \in \mathbb{C}^M$ of an antenna array with M antennas, respectively placed at the locations $\mathbf{u}_m \in \mathbb{R}^3$, $m = 1, \dots, M$, is given by

$$\mathbf{h} = \sqrt{\beta} \underbrace{\left[e^{j\mathbf{k}(\varphi, \theta)^T \mathbf{u}_1}, \dots, e^{j\mathbf{k}(\varphi, \theta)^T \mathbf{u}_M} \right]^T}_{\triangleq \mathbf{a}(\varphi, \theta)} \quad (7.13)$$

where β describes the macroscopic large-scale fading and is assumed to be the same for all antennas (i.e., the array aperture is small as

³In the far-field of an antenna in free space, the power intensity of the EM radiation is inversely proportional to the squared distance. The far-field region generally begins at a few wavelengths away from an antenna, while the far-field region of an array begins at a distance that grows with the square of the array aperture.

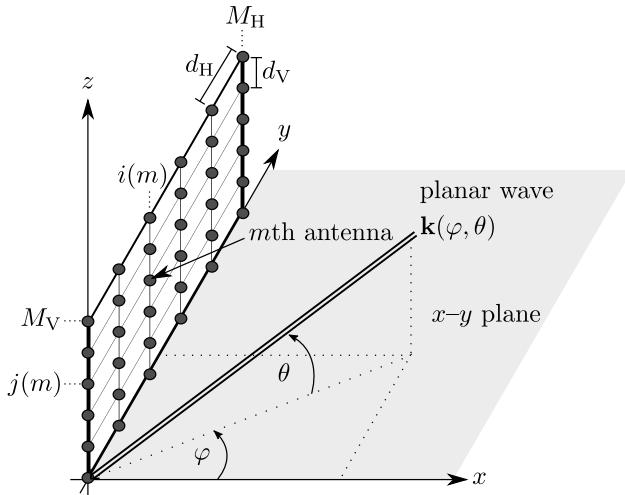


Figure 7.12: Planar rectangular antenna array with an impinging plane wave from elevation angle θ and azimuth angle φ . The $M = M_V M_H$ antennas are horizontally and vertically equally spaced with spacing d_H and d_V , respectively. The m th antenna has the horizontal index $i(m)$ and the vertical index $j(m)$.

compared to the propagation distance). The vector $\mathbf{a}(\varphi, \theta) \in \mathbb{C}^M$ is the so called *array response* or steering vector. For a horizontal ULA along the y -axis with antenna spacing d_H (in multiples of the wavelength) and waves arriving only from directions within the x - y plane (i.e., $\theta = 0$ and $\mathbf{u}_m = [0, \lambda(m-1)d_H, 0]^T$), it is easy to see that (7.13) coincides with (1.23).

Figure 7.12 shows a planar array in the y - z -plane consisting of M_V horizontal rows with M_H antennas each. The antennas are uniformly spaced with horizontal and vertical spacing d_H and d_V (measured in multiples of the wavelength), respectively. The antennas are consecutively indexed row-by-row by $m \in [1, M]$, $M = M_V M_H$, so that the location of the m th antenna can be described as

$$\mathbf{u}_m = \begin{bmatrix} 0 \\ i(m)d_H\lambda \\ j(m)d_V\lambda \end{bmatrix} \quad (7.14)$$

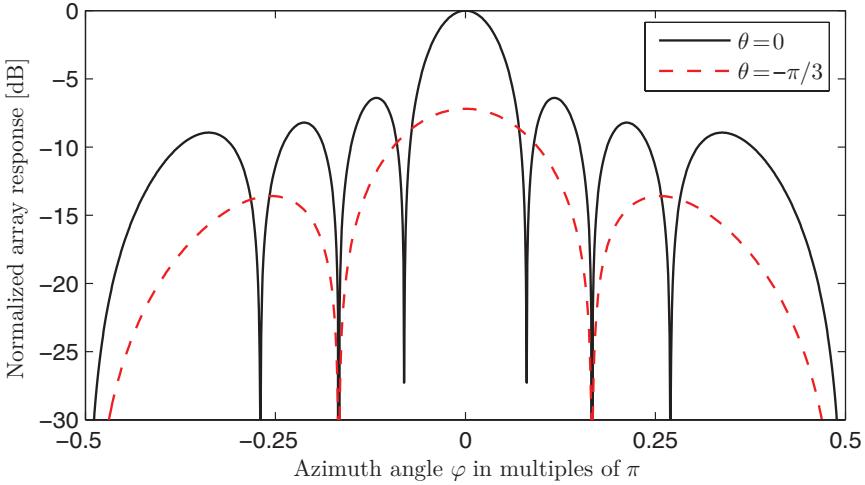


Figure 7.13: Normalized array response $|\frac{1}{M} \mathbf{a}(0, 0)^T \mathbf{a}(\varphi, \theta)|$ as a function of the azimuth angle φ . The array geometry is given by $M_V = 4$, $M_H = 8$ and $d_V = d_H = 0.5$.

where

$$i(m) = \text{mod}(m - 1, M_H) \quad (7.15)$$

$$j(m) = \lfloor (m - 1)/M_H \rfloor \quad (7.16)$$

are the horizontal and vertical index of antenna m , respectively. Figure 7.13 shows the normalized array response $|\frac{1}{M} \mathbf{a}(0, 0)^T \mathbf{a}(\varphi, \theta)|$ of the planar array defined in (7.14) with $M_V = 4$, $M_H = 8$, and $d_V = d_H = 0.5$ as function of φ for different elevation angles θ . The figure shows how much an interfering signal arriving from azimuth angle φ is attenuated when receiving a signal from $\varphi = \theta = 0$. The resolution of the array is better when the main lobe is narrower because a smaller angular difference between the desired and interfering signals is sufficient to get a certain attenuation of the interfering signal. We observe that the horizontal resolution of the array is significantly reduced for lower elevation angles (e.g., $\theta = -\pi/3$).

7.3.2 3D Local Scattering Model with Arbitrary Array Geometry

The local scattering model was introduced in Section 2.2 on p. 235 for a horizontal ULA in a two-dimensional scenario. It provides an easy

way to compute the correlation matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$ of a channel vector $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R})$ as a function of the distribution of the AoA of the incoming waves. With the definition of the array response in (7.13), we can now extend the local scattering model in (2.23) to three dimensions and arbitrary array geometries. If we redefine (2.20) as $\mathbf{a}_n = g_n \mathbf{a}(\varphi_n, \theta_n)$ and follow the same subsequent steps as in Section 2.2, we obtain a new expression for the elements of \mathbf{R} :

$$[\mathbf{R}]_{m,l} = \beta \iint e^{j\mathbf{k}(\varphi, \theta)^T (\mathbf{u}_m - \mathbf{u}_l)} f(\varphi, \theta) d\varphi d\theta \quad (7.17)$$

where $f(\varphi, \theta)$ is the joint PDF of the azimuth and elevation angle and the integration is over all angles. For the case of the planar array as shown in Figure 7.12, this expression simplifies to

$$\begin{aligned} & [\mathbf{R}]_{m,l} \\ &= \beta \iint \underbrace{e^{j2\pi d_V[j(m) - j(l)] \sin(\theta)}}_{\text{Vertical correlation}} \underbrace{e^{j2\pi d_H[i(m) - i(l)] \cos(\theta) \sin(\varphi)}}_{\text{Horizontal correlation}} f(\varphi, \theta) d\varphi d\theta. \end{aligned} \quad (7.18)$$

The terms $d_V[j(m) - j(l)]$ and $d_H[i(m) - i(l)]$ represent the vertical and horizontal distance between antennas m and l , respectively. Thus, the two complex exponential terms in (7.18) can be interpreted as the vertical and horizontal correlation of the array. Consider the set of antenna pairs (m, l) such that either $i(l) = i(m)$ or $j(l) = j(m)$; that is, the columns and rows of the antenna array. For these pairs, the correlation matrix has the following values:

$$[\mathbf{R}]_{m,l} = \begin{cases} \beta \int e^{j2\pi d_V[j(m) - j(l)] \sin(\theta)} f(\theta) d\theta & i(l) = i(m) \\ \beta \iint e^{j2\pi d_H[i(m) - i(l)] \cos(\theta) \sin(\varphi)} f(\varphi, \theta) d\varphi d\theta & j(l) = j(m). \end{cases} \quad (7.19)$$

Interestingly, while the correlation of the vertical columns coincides with the ULA in (2.23), this is not the case for the correlation of the horizontal rows. The larger the absolute value of the elevation angle θ , the smaller is the effective horizontal antenna spacing $d_H(i(m) - i(l)) \cos(\theta)$ of the array. Thus, the correlation between the antennas of the horizontal rows depends on the elevation angle. The same effect has already been observed in Figure 7.13 for a pure LoS scenario.

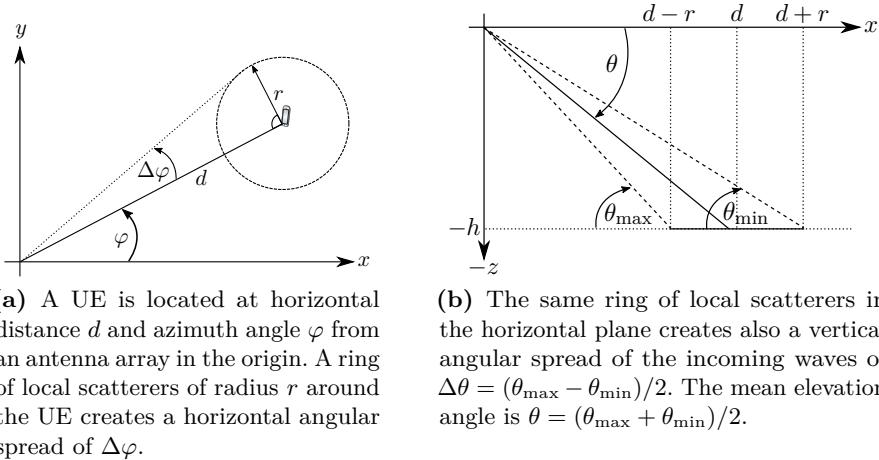


Figure 7.14: Derivation of the angular spread of the 3D one-ring model.

The local scattering model for the planar array is entirely defined by the joint PDF of θ and φ . We will now provide a way to define this distribution for a somewhat realistic system, which takes into account the height at which the antenna array is installed and the distance between the UE and array [7]. As shown in Figure 7.14, we consider a UE being located at azimuth angle φ , horizontal distance d , and height $z = -h$ below a planar antenna array. We assume that the UE is surrounded by a horizontal ring of local scatterers of radius r .⁴ As shown in Figure 7.14a, the radius r defines the horizontal *angular spread* $\Delta\varphi = \tan^{-1}(r/d)$. The same scatterers give also rise to a vertical angular spread which can be computed as follows. As can be seen from Figure 7.14b, the maximum elevation angle $\theta_{\max} = \tan^{-1}(h/(d - r))$ is achieved by a scatterer at distance $d - r$. Similarly, the minimum elevation angle $\theta_{\min} = \tan^{-1}(h/(d + r))$ is achieved by a scatterer at distance $d + r$. Assuming that the elevation angles are uniformly distributed in the interval $[\theta_{\min}, \theta_{\max}]$, the mean elevation angle is computed as $\theta = (\theta_{\max} + \theta_{\min})/2$ and the vertical spread is $\Delta\theta =$

⁴The scatterer ring is only a means to define reasonable values of the angular spread, but has no other physical meaning.

$(\theta_{\max} - \theta_{\min})/2$.⁵ Thus, the spatial correlation matrix is finally given as

$$[\mathbf{R}]_{m,l} = \frac{\beta}{4\Delta\varphi\Delta\theta} \int_{\theta-\Delta\theta}^{\theta+\Delta\theta} \int_{\varphi-\Delta\varphi}^{\varphi+\Delta\varphi} e^{jk(\varphi,\theta)^T(\mathbf{u}_m - \mathbf{u}_l)} d\varphi d\theta. \quad (7.20)$$

Note that the angular spread with this model is rather small. For example, for a BS at a height of $h = 25$ m, a UE located at a distance of $d = 200$ m with a scattering radius of $r = 50$ m sees a horizontal spread of $\Delta\varphi = 14^\circ$ and a vertical spread of less than $\Delta\theta = 2^\circ$. To avoid angular distributions with small finite support, we can alternatively use the Gaussian [4, 373, 313, 363] or Laplacian distribution [225, Section 7.4.2], [161] rather than the uniform distribution in (7.20) (cf. Section 2.6 on p. 235). Recall that Gaussian distributions were used in the running example of previous sections. The mean angles and angular spreads derived above can be used as values for the ASD and mean of these distributions. In the remainder of this monograph, we will sometimes refer to the model in (7.20) as the 3D one-ring model or simply the one-ring model.

Dominant Eigenspace and Chordal Distance

As anticipated in Section 2.6 on p. 235 for the local scattering model, correlation matrices generally have a mix of many weak and a few strong eigendirections. To further exemplify this, Figure 7.15 shows the eigenvalues in decreasing order when using the 3D one-ring model in (7.20) with an 8×8 antenna array and different scatter radii. We can clearly see that the smaller the scatter radius, the more energy is packed into a few strong eigendirections of the channel; 99% of the energy is contained in 5 out of the 64 eigenvalues when $r = 50$ m and 20 eigenvalues when $r = 200$ m. Strictly speaking, the correlation matrix might still have full rank, but we can define the following notion of p -dominant eigenspace to capture the eigenspace that contains most of the energy.

Definition 7.1 (p -Dominant eigenspace). Let $\mathbf{X} \in \mathbb{C}^{M \times M}$ be a Hermitian matrix with eigenvalue decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}^H$, where $\mathbf{U} =$

⁵This is a crude approximation. A ring of uniformly distributed scatterers would not lead to uniformly distributed azimuth and elevation angles.

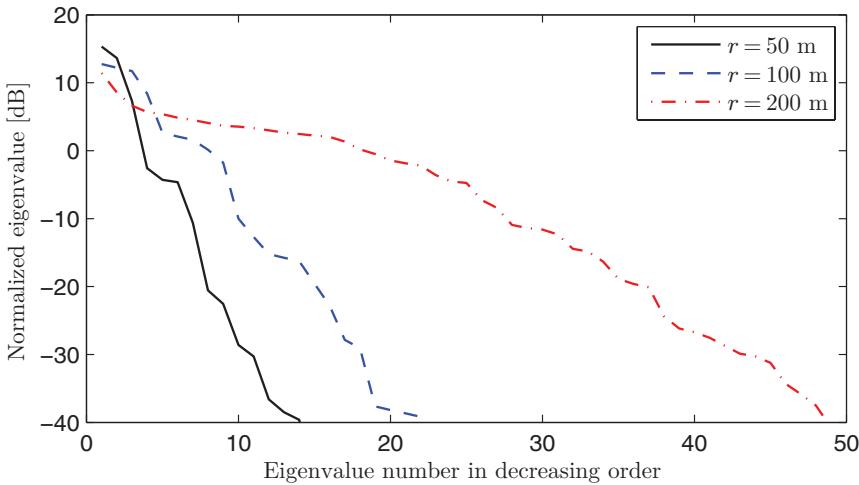


Figure 7.15: Eigenvalues of the spatial correlation matrix \mathbf{R} when using the 3D one-ring model in (7.20) for an 8×8 antenna array and different scatter-radii r . The other model parameters are $d = 200$ m, $\varphi = 23.5^\circ$ and $h = 23.5$ m.

$[\mathbf{u}_1 \dots \mathbf{u}_M] \in \mathbb{C}^{M \times M}$ is unitary and $\mathbf{D} = \text{diag}(\lambda_1, \dots, \lambda_M) \in \mathbb{R}^{M \times M}$ consists of the non-negative eigenvalues of \mathbf{X} in non-increasing order (i.e., $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M \geq 0$). The p -dominant eigenspace $\text{eig}_p(\mathbf{X}) \in \mathbb{C}^{M \times p}$ is defined as $\text{eig}_p(\mathbf{X}) = [\mathbf{u}_1 \dots \mathbf{u}_p]$.

The p -dominant eigenspace of a Hermitian matrix is the (tall) unitary matrix composed of the p eigenvectors belonging to its p largest eigenvalues. In order to define a metric for measuring orthogonality between two eigenspaces, we consider the chordal distance (see, e.g., [180, 235]) which is defined for arbitrary matrices as follows.⁶

Definition 7.2 (Chordal distance). The *chordal distance* $d_C(\mathbf{X}, \mathbf{Y})$ between two matrices \mathbf{X} and \mathbf{Y} is defined as

$$d_C(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X}\mathbf{X}^H - \mathbf{Y}\mathbf{Y}^H\|_F^2. \quad (7.21)$$

For two (tall) unitary matrices $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{C}^{M \times p}$, the chordal distance

⁶Other distance or orthogonality metrics between correlation matrices can be also defined. We refer the interested reader to [88, Sec. 3.1.1].

takes the form

$$\begin{aligned}
 d_C(\mathbf{U}_1, \mathbf{U}_2) &= \|\mathbf{U}_1 \mathbf{U}_1^H - \mathbf{U}_2 \mathbf{U}_2^H\|_F^2 \\
 &= \text{tr}((\mathbf{U}_1 \mathbf{U}_1^H - \mathbf{U}_2 \mathbf{U}_2^H)(\mathbf{U}_1 \mathbf{U}_1^H - \mathbf{U}_2 \mathbf{U}_2^H)^H) \\
 &= \text{tr}(\mathbf{U}_1 \mathbf{U}_1^H + \mathbf{U}_2 \mathbf{U}_2^H - 2\mathbf{U}_1 \mathbf{U}_1^H \mathbf{U}_2 \mathbf{U}_2^H) \\
 &= 2p - 2 \sum_{i=1}^p \sum_{j=1}^p |\mathbf{u}_{1i}^H \mathbf{u}_{2j}|^2
 \end{aligned} \tag{7.22}$$

where \mathbf{u}_{ki} denotes the i th column vector of matrix \mathbf{U}_k for $k = 1, 2$. The chordal distance can be interpreted as the number of dimensions of the subspace that can be reached by a linear combination of the column vectors of only one of the two matrices. For example, if $\mathbf{U}_1 = \mathbf{U}_2$, we have $d_C(\mathbf{U}_1, \mathbf{U}_2) = 0$. Although each matrix individually spans p dimensions, all of them can be reached through a linear combination of the column vectors of \mathbf{U}_1 or \mathbf{U}_2 . On the other hand, for $\mathbf{U}_1^H \mathbf{U}_2 = \mathbf{0}_{p \times p}$, we have $d_C(\mathbf{U}_1, \mathbf{U}_2) = 2p$ because each matrix spans a p -dimensional space which cannot be reached through a linear combination of the column vectors of the other matrix.

Figure 7.16 shows the chordal distance for different values of M between the 6-dominant eigenspaces of the correlation matrices of a UE at azimuth 0° and distance 200 m and that of a UE at the same distance but at azimuth angle φ . The 3D one-ring model in (7.20) is used with scatter radius $r = 50$ m. We can see that the chordal distance increases with M , due to the increased spatial resolution of the array, and with φ since the spatial correlation matrices become increasingly different.

7.3.3 The 3GPP 3D MIMO Channel Model

The previously presented LoS and local scattering channel models are mathematically convenient and very easy to simulate while still capturing essential characteristics of the wireless channel between BSs equipped with large antenna arrays and single-antenna UEs. In particular, we have seen the dependence of the channel (correlation) matrix on the antenna array geometry as well as on the physical location and orientation of the BS and UEs. Both channel models are *spatially consistent*; that is, the channel statistics for a given location are always the

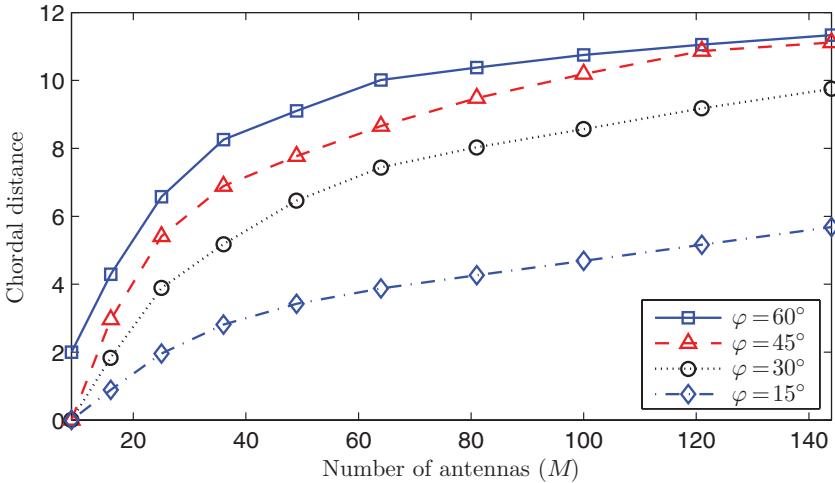


Figure 7.16: Chordal distance between the 6-dominating eigenspaces of the correlation matrices of a UE at azimuth 0° and distance $d = 200$ m and of a UE at the same distance but at azimuth angle φ . The antenna array is assumed to be quadratic with $d_H = d_V = 0.5$ and installed at height $h = 23.5$ m. The correlation matrices are generated by the 3D one-ring model in (7.20) with scatter radius $r = 50$ m.

same and do not depend on the simulation run. However, these models suffer from several shortcomings, which render them inappropriate for large-scale system-level simulations with the goal of quantifying the real-world performance of MIMO systems. In particular, it is neither realistic to assume pure LoS conditions nor to consider a single local cluster of scatterers around the UE. For this reason, the 3GPP has defined a 3D stochastic geometry-based channel model for MIMO systems [1], which explicitly accounts for multiple spatial clusters of scatterers and a mix of NLoS and LoS propagation paths. This so-called 3GPP 3D MIMO channel model is an extension of previously standardized channel models, such as Winner II [347], which assumes that all scatterers, reflectors, UEs and BSs are located in a two-dimensional plane. This extension was required as it was impossible with such models to simulate MIMO systems exploiting the elevation dimension.

The 3GPP 3D MIMO channel model is characterized by a deterministic system layout (i.e., BS and UE locations, antenna orientations, field

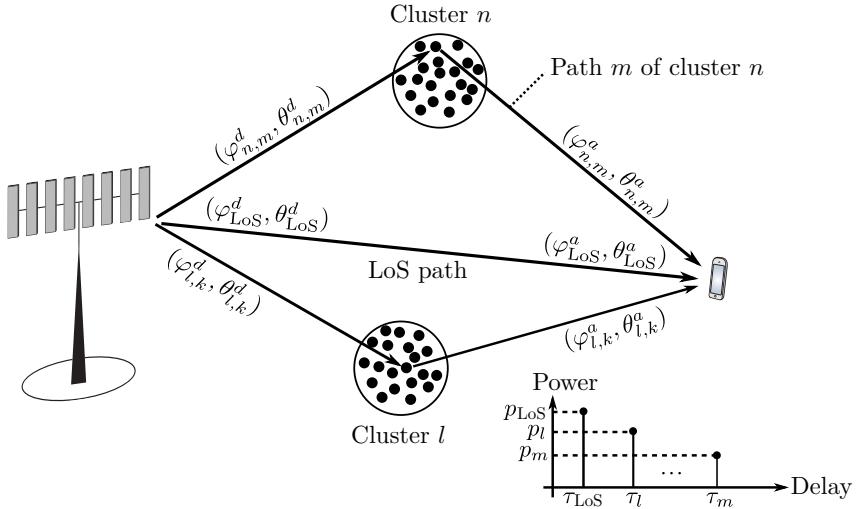


Figure 7.17: The 3GPP 3D MIMO channel model [1] is a stochastic geometry-based model. For each BS-UE link, C scattering clusters with random angles of arrival/departure are generated. Each cluster is associated with a time delay τ_l and a power p_l , for $l = 1, \dots, C$. The distributions of angles, delays, and powers depend on the chosen scenario. Each cluster generates 20 multipath components, which are assumed to be resolvable in the angular domain, but not in the time domain.

patterns⁷, and carrier frequency) and a set of random parameters (i.e., delay spread, delay values, angular spread, shadow fading, and cluster powers) which are generated from statistical distributions obtained from extensive channel measurements. The model is valid for carrier frequencies in the range 2–6 GHz and bandwidths of up to 100 MHz.

The modeling approach is as follows. For each pair of UE and BS, C scattering clusters with random AoAs, AoDs, delays, and power levels are generated. With a scenario-dependent⁸ probability, also a LoS path is present. This is schematically shown in Figure 7.17. The value of C depends on the scenario and is in the range of 12–20. Each cluster is assumed to have 20 resolvable multipath components in the angular

⁷The field or radiation pattern of an antenna describes the angular dependence of the strength of the transmitted or received radio waves.

⁸The 3GPP 3D MIMO channel model can be used to simulate different propagation scenarios (e.g., coverage tier, hotspot tier, LoS, NLoS, indoor, outdoor, outdoor-to-indoor, urban, and suburban), which are detailed in [1].

domain with random angular offsets from the cluster AoA/AoD. The angular spread of the multipath components of each cluster is rather low, within 1° – 6° , while the angular spread of the clusters themselves is large, within 20° – 90° . The multipath components of each cluster are assumed to be unresolvable in the time domain (i.e., they all arrive with the same cluster delay). The 3GPP 3D MIMO channel model considers a multitude of different scenarios with different BS and UE heights and arbitrary antenna characteristics. In particular, high-rise buildings with UEs on multiple floors can be simulated to gauge the gains of elevation precoding. Dual-polarized antennas (see Section 7.4.4), as well as time-evolution of the channel due to movement of UEs, can be also considered. An important aspect of this channel model is that it is not spatially consistent; that is, the location of scattering clusters are random and uncorrelated between different UEs. In other words, two UEs at almost the same location do not share any scatterers. Thus, it cannot be used for simulations requiring spatial consistency. In this case, one must resort to ray tracing, recorded channel measurements, or simpler channel models as presented earlier. There are also scatterer-centric channel models, such as COST2100 (e.g., [88, Sec. 4.4.5]), where a global set of scatterers is shared by all UEs. Although this type of channel model is spatially consistent, there is currently no widely accepted model to be used with large antenna arrays (see [123, 160]). Two open-source implementations of the 3GPP 3D MIMO model are currently available: the standalone QuaDRiGa channel model [159] developed by the Fraunhofer Heinrich Hertz Institute and the LTE Advanced system simulator from the Vienna University of Technology [5]. The QuaDRiGa model will be used for the case study in Section 7.7.

7.3.4 Observations from Channel Measurements

Several channel measurement campaigns with large antenna arrays have been conducted during the last years. Among the first were [120, 150] that provided the first verifications of favorable propagation (see Definition 2.5 on p. 233) in practice.⁹ In [120], a cylindrical array of

⁹The authors of [300] built the first 64-antenna Massive MIMO prototype and confirmed the expected tremendous performance. However, their publication does

128 antennas was used for indoor-to-outdoor measurements, while the authors of [150] conducted outdoor measurements with a virtual planar array of 112 antennas. Both papers quantify the level of favorable propagation (see Section 2.5.2 on p. 233) by computing the correlation metric

$$\delta_{\text{corr}} = \mathbb{E} \left\{ \frac{|\mathbf{h}_1^H \mathbf{h}_2|^2}{\|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2} \right\} \quad (7.23)$$

between the channels of two randomly chosen measurement locations. This metric is similar to (2.19) and essentially describes the variance of the inner products of the normalized channels. The two papers conclude that close to ideal results (i.e., i.i.d. Rayleigh fading channels for which the average correlation can be shown to be $\delta_{\text{corr}} = 1/M$) can be achieved in practice. However, the larger the number of antennas, the higher was the observed gap to the ideal case, giving rise to the conclusion that some saturation effect appears and that the marginal gain of additional antennas rapidly diminishes. Similar measurements were conducted in [124] and different antenna architectures (horizontal, planar, vertical) were compared. Figure 7.18 shows the average correlation over randomly picked UE locations within a cell sector. The measurement results are reproduced from [124]¹⁰ and simulation results are based on the 3GPP 3D MIMO channel model as implemented in the QuaDRiGa software [159]. We can clearly see that a horizontal array provides a decorrelation similar to i.i.d. channels, which implies that the UEs' channels will be nearly orthogonal. In contrast, the planar and vertical arrays exhibit higher δ_{corr} due to the lower vertical angular spread between UEs; that is, it is easier to separate UEs in the azimuth domain than in the elevation domain. If we would instead consider a scenario with UEs located at different floors of high-rise buildings, the results would be more favorable towards planar (and vertical) arrays as the elevation angular spreads are then higher. The differences between the results for simulated and measured channels can be partially attributed to a reduced number of measurement locations and the lack of spatial consistency of the 3GPP 3D MIMO channel model.

not report specifically on channel measurements.

¹⁰We would like to thank Marc Gauger for recomputing some of the results.

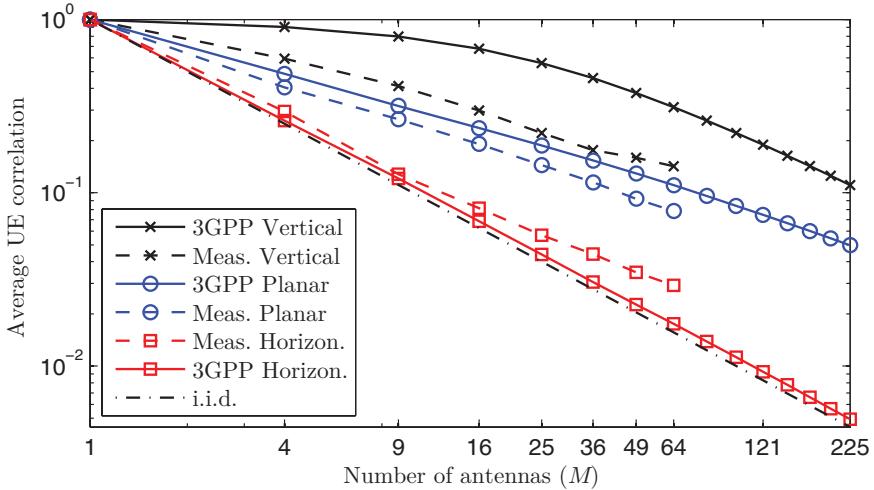


Figure 7.18: Average UE correlation δ_{corr} in (7.23) between the channel vectors of two UEs at randomly chosen locations within a cell sector as a function of the number of antennas. Horizontal, vertical and quadratic planar antenna geometries are considered. The measurement results are reproduced from [124]. The simulation results are obtained for the urban macro LoS scenario of the 3GPP 3D MIMO channel model.

Several papers investigate the resolvability of closely spaced UEs under LoS conditions. Even though we will see in the next section, from simple analytical models, that low angular separation is detrimental for Massive MIMO, the results of [207, 115] indicate that arrays with a very large aperture can resolve UEs even in close proximity. This has been confirmed for indoor [207] as well as outdoor [115] channels.

The subsequent works [121, 122] reported on refined outdoor channel measurements using a 7.4 m long virtual ULA as well as a more compact cylindrical array with 128 antennas. These works demonstrate that SEs similar to those predicted by closed-form expressions obtained for i.i.d. Rayleigh fading channels can be achieved on measured channels, although the channels have very different characteristics. Moreover, the authors make the interesting observation that substantial variations of the received power over the array exist. In other words, antenna arrays with a large aperture can experience antenna-dependent shadow fading. We have shown in Section 4.4.1 on p. 338 that this phenomenon leads

to linearly independent correlation matrices, which is a key property in the asymptotic performance analysis. The antenna-dependent shadow fading was observed for the ULA as well as for the compact cylindrical array. However, the explanation for this phenomenon is different in these cases. Different parts of a ULA “see” different parts of the propagation environment such that obstacles might only block certain parts of the array. A similar effect arises in distributed antenna systems, where most antennas have independent propagation paths to a UE. On the other hand, the antennas of a cylindrical array point in different directions, where different scattering clusters or obstacles are relevant. The authors of [122] also observed that shadowing over the array can give rise to the effect that some antennas contribute more to the overall channel than others. Thus, some antennas might be dynamically turned on/off to save energy and/or processing power. This phenomenon is further discussed in the following example.

Large-Scale Fading Variations over the Antenna Array

Consider the scenario in Figure 7.19 where two obstacles block two different parts of the antenna array in two different cell regions A and B . In region A , the first M' antennas of the array exhibit strong shadowing with attenuation $\beta \in [0, 1]$, while the remaining $M - M'$ antennas have an unobstructed path to the UEs with attenuation $(M - \beta M') / (M - M')$. This value is chosen such that the average per-antenna energy of the channel is independent of β and equal to one. In region B , the situation is the opposite, which means that the last M' antennas are obstructed. We neglect other types of spatial channel correlation in this example to focus exclusively on the effects of antenna-dependent shadowing. The channel vectors of UEs in region A and B are distributed as $\mathbf{h}_A \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_A)$ and $\mathbf{h}_B \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_B)$, respectively, where¹¹

$$\mathbf{R}_A = \begin{bmatrix} \beta \mathbf{I}_{M'} & \mathbf{0} \\ \mathbf{0} & \frac{M - \beta M'}{M - M'} \mathbf{I}_{M - M'} \end{bmatrix} \quad (7.24)$$

$$\mathbf{R}_B = \begin{bmatrix} \frac{M - \beta M'}{M - M'} \mathbf{I}_{M - M'} & \mathbf{0} \\ \mathbf{0} & \beta \mathbf{I}_{M'} \end{bmatrix}. \quad (7.25)$$

¹¹For brevity, we did not write out the dimensions of the all-zero matrices.

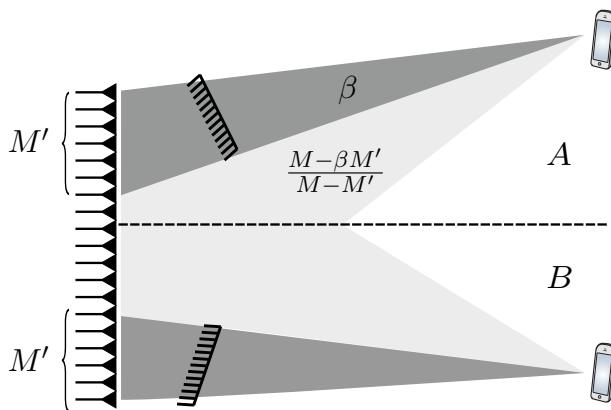


Figure 7.19: Obstacles which only partially block propagation paths towards a UE can lead to large-scale fading variations over the antenna array.

The effect of shadowing in this example is as follows. Two UEs located in the same region see a channel with an effectively reduced number of antennas. For instance, with $\beta = 0$, only $M - M'$ antennas are visible. Figure 7.20 illustrates the effect with $M = 64$, $M' = 30$, and different values of β . As can be seen from Figure 7.20, shadowing increases the average correlation δ_{corr} as defined in (7.23) between the UEs' channels in the same region case. On the other hand, for two UEs in different regions, shadowing decreases the correlation since two parts of the antenna array receive more energy from one UE than from the other. Thus, antenna-dependent shadowing can have a positive or negative effect depending on which UEs are simultaneously scheduled in a cell.

One can also consider shadowing over the antenna array as a particular form of spatial channel correlation which could be exploited in a manner similar to the previously discussed JSDM in Remark 7.3 on p. 472. In this case, the dominating eigenspace of a group would be simply the subset of antennas of the array over which most energy is received. In the example in Figure 7.19, the UEs in region *A* and *B* would be respectively served by the lower and upper part of the antenna array so that the present obstacles would naturally reduce the interference between the groups.

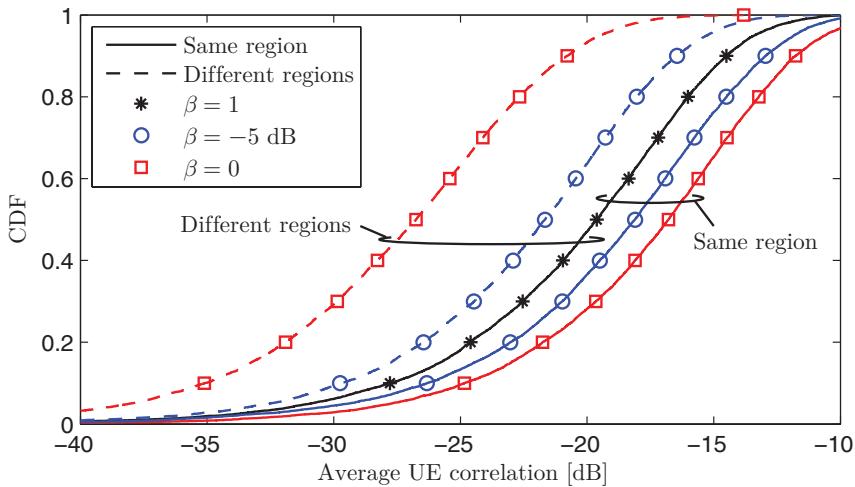


Figure 7.20: Average UE correlation δ_{corr} in (7.23) of the channels of two UEs in either the same ($A - A$) or different ($A - B$) regions of Figure 7.19. Depending on the scenario, antenna-dependent shadowing can either increase or decrease the correlation.

7.4 Array Deployment

Until now, we have characterized antenna arrays only indirectly through their spatial channel correlation. We will now shift focus to the arrangement of the individual antennas; that is, the array geometry. In this section, we will explore the effects of different antenna array geometries and have an in-depth look at the effects of antenna spacing and polarization. For an introduction to the design of BS antennas and antenna arrays for cellular communications, we refer to [79].

The most important factors of an antenna array are the antenna spacing and its total size (relative to the wavelength), which is known as the aperture. The size determines the array's directivity; that is, its ability to focus the radiated energy towards certain directions, while the number of antennas determines the radiated/received energy. Some of these aspects will be discussed in more detail in Section 7.4.2. Each individual antenna consists of one or more radiating elements¹² that have

¹²A radiating element can of course also receive energy.

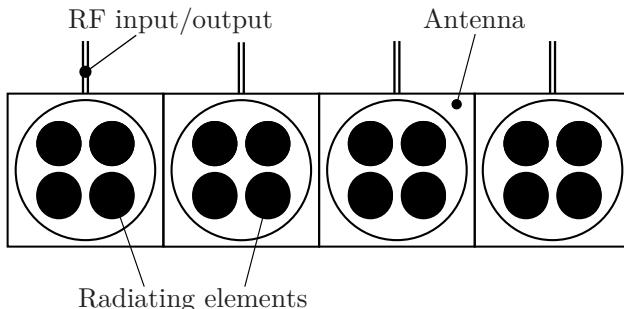


Figure 7.21: Antenna array consisting of four antennas each of which is composed of four radiating elements. The radiating elements of an antenna share the same RF input and output.

a fixed size that depends on the wavelength λ of the carrier frequency. For example, a half-wave dipole, is in essence, a piece of wire of length $\lambda/2$. For a carrier frequency of 2.6 GHz, such a dipole has a length of 5.8 cm. Since the size of a dipole cannot simply be made larger, multiple dipoles or other radiating elements need to be connected together if the captured energy of an antenna shall be increased. The following definition, which is also visualized in Figure 7.21, makes the relation between the terms radiating element, antenna, and antenna array clear.

Definition 7.3 (Radiating element, antenna, antenna array). An *antenna* consists of one or more *radiating elements* (e.g., dipoles) which are fed by the same RF signal. An *antenna array* is composed of multiple antennas with individual RF chains.

In legacy mobile communication systems (e.g., GSM, UMTS), a BS with a single antenna and single RF chain needs to provide coverage for an entire cell sector with a horizontal width of 120° and a radius of up to several kilometers. At the same time, it must not radiate energy to neighboring cell sectors. Such BS antennas have typically an azimuth beamwidth of 65° and an elevation beamwidth of $3^\circ\text{--}15^\circ$ [79, Sec. 2.2].¹³ The resulting coverage gap between the 65° sectors is generally filled by

¹³The *beamwidth* is defined as the angular distance between the half-power points of the main lobe of the antenna's radiation pattern. It is also referred to as the half-power beamwidth or 3-dB beamwidth. The beamwidth is generally different in the azimuth and elevation directions.

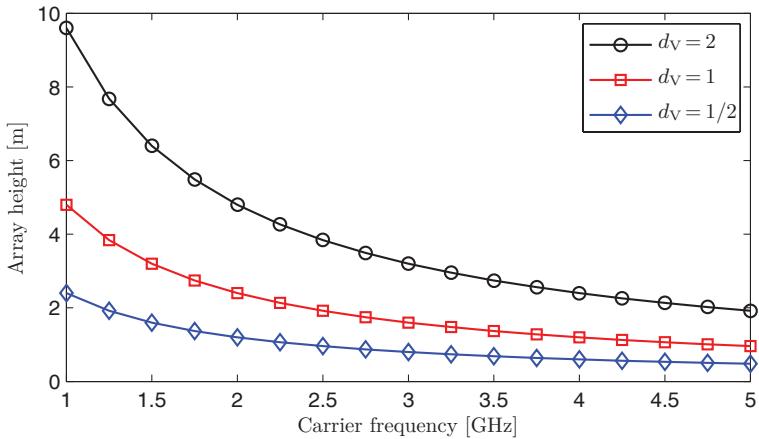


Figure 7.22: Height of a planar antenna array as a function of the carrier frequency for $M_V = 16$ antennas per vertical column and different vertical antenna spacings $d_V = [\frac{1}{2}, 1, 2]$.

neighboring BSs. As we will see later, to achieve such a focused radiation of energy in the elevation domain, the antenna must be composed of multiple vertically stacked radiating elements. Common antenna heights are 8–16 wavelengths [79, Sec. 2.2.1.3]. With a typical element spacing of 0.8λ , a single BS antenna consists of 10–20 radiating elements. This number is doubled since dual-polarized co-located antennas are the norm for cellular communications (see Section 7.4.4). Thus, a legacy BS covering three cell sectors is already equipped with 60–120 radiating elements but has only three RF chains, one per sector. In contrast, a Massive MIMO system has a similar number of radiating elements but an individual RF chain for each of them. The practical challenge of Massive MIMO is therefore not to accommodate a large number of radiating elements but to process their individual RF signals.

7.4.1 Preliminaries on Physical Array Size

Consider the planar antenna array shown in Figure 7.12. The horizontal width and vertical height of this array are respectively given by $M_H d_H \lambda$ and $M_V d_V \lambda$, where d_H and d_V are the horizontal and vertical antenna spacing (in multiples of the wavelength). In Figure 7.22, the array height

is exemplified as a function of the carrier frequency for $M_V = 16$ and different values of d_V . Clearly, the higher the carrier frequency, the smaller is the array form factor. This makes arrays with a large number of antennas especially attractive at mmWave frequencies. This aspect will be discussed in more detail in Section 7.5. As Figure 7.22 shows, antenna arrays can easily become bulky for frequencies below 5 GHz. For example, a 16×16 planar array with $\lambda/2$ spacing at a carrier frequency of 2.5 GHz has the physical dimensions of $1\text{ m} \times 1\text{ m}$. Thus, in order to optimally design large antenna arrays, it is important to understand the role of the antenna spacing as well as the array height and width.

7.4.2 Physical Array Size and Antenna Spacing

To get some simple insight into the array design problem, we will revisit the analysis of the two-UE LoS uplink channel of (1.25) in a 3D setting with a planar array. To this end, we focus on the LoS channel model in (7.13) and assume that the two UEs are respectively located at the azimuth and elevation angles (φ_1, θ_1) and (φ_2, θ_2) . Our goal is to study how the correlation of the array responses $|\frac{1}{M}\mathbf{a}(\varphi_2, \theta_2)^H\mathbf{a}(\varphi_1, \theta_1)|$ behaves as a function of the array geometry and the angular separation of the UEs.¹⁴ For ease of notation, we introduce the quantities

$$\Omega = \sin(\theta_1) - \sin(\theta_2) \quad (7.26)$$

$$\Psi = \cos(\theta_1)\sin(\varphi_1) - \cos(\theta_2)\sin(\varphi_2). \quad (7.27)$$

Since the azimuth and elevation angles are confined within the interval $[-\pi/2, \pi/2]$, it follows that $\Omega, \Psi \in [-2, 2]$. The distributions of these quantities are in general not uniform as they depend on the distribution of the UEs in the cell sector and on the height at which the antenna array is installed. This has an impact on the ranges of values of Ω and Ψ that are of practical importance. Using the definition of $\mathbf{a}(\varphi, \theta)$ in

¹⁴A detailed discussion for the case of a ULA can be found in [314, Sec. 7.2.4].

(7.13) and (7.14), it is easy to show that

$$\begin{aligned}
& \left| \frac{1}{M} \mathbf{a}(\varphi_2, \theta_2)^H \mathbf{a}(\varphi_1, \theta_1) \right| \\
&= \left| \frac{1}{M} \sum_{m=1}^M e^{j2\pi(d_V j(m)\Omega + d_H i(m)\Psi)} \right| \\
&= \underbrace{\left| \frac{1}{M_V} \sum_{k=0}^{M_V-1} e^{j2\pi d_V k \Omega} \right|}_{\triangleq S(\Omega)} \underbrace{\left| \frac{1}{M_H} \sum_{l=0}^{M_H-1} e^{j2\pi d_H l \Psi} \right|}_{\triangleq T(\Psi)}. \tag{7.28}
\end{aligned}$$

Thus, the correlation can be represented as the product of two functions $S(\Omega)$ and $T(\Psi)$. Note that $S(\Omega)$ only depends on the angular separation of the UEs in the elevation domain while $T(\Psi)$ depends on all azimuth and elevation angles. As in the proof of Lemma 1.5 in Appendix C.1.4 on p. 583, we can rely on the identity $\sum_{n=0}^{N-1} q^n = (1 - q^N)/(1 - q)$ for $q \neq 1$, to simplify $S(\Omega)$ as follows:

$$\begin{aligned}
S(\Omega) &= \left| \frac{1}{M_V} \frac{1 - e^{j2\pi d_V M_V \Omega}}{1 - e^{j2\pi d_V \Omega}} \right| \\
&= \left| \frac{e^{j\pi d_V (M_V-1)\Omega}}{M_V} \frac{e^{-j\pi d_V M_V \Omega} - e^{j\pi d_V M_V \Omega}}{e^{-j\pi d_V \Omega} - e^{j\pi d_V \Omega}} \right| \\
&= \left| \frac{\sin(\pi L_V \Omega)}{M_V \sin(\pi d_V \Omega)} \right| \tag{7.29}
\end{aligned}$$

where $L_V = M_V d_V$ is the normalized height of the array. It is easy to verify that $S(\Omega) = \sqrt{g(\theta_1, \theta_2)/M_V}$, where $g(\theta_1, \theta_2)$ was defined in Lemma 1.5 on p. 184 (using $d_H = d_V$ and $M = M_V$). Following the same steps and defining the normalized array width $L_H = M_H d_H$, we can express $T(\Psi)$ as

$$T(\Psi) = \left| \frac{\sin(\pi L_H \Psi)}{M_H \sin(\pi d_H \Psi)} \right|. \tag{7.30}$$

We can make the following observations from (7.29) and (7.30):

- $S(\Omega)$ and $T(\Psi)$ are periodic with period $\frac{1}{d_V}$ and $\frac{1}{d_H}$, respectively;
- $S\left(\frac{k}{L_V}\right) = 0$ for $k = 1, \dots, M_V - 1$, and $T\left(\frac{k}{L_H}\right) = 0$ for $k = 1, \dots, M_H - 1$;

- Both functions peak at $\Omega = \Psi = 0$ with $S(0) = T(0) = 1$.¹⁵

The function $S(\Omega)$ is shown for different values of M_V and d_V in Figure 7.23. The behavior of $T(\Psi)$ is identical. Both functions have main-lobes around the origin with a width of $\frac{2}{L_V}$ and $\frac{2}{L_H}$, respectively.¹⁶ The maximum values of all other lobes are much smaller. This implies that the array is not able to separate two UEs whenever for some integers k and l we have

$$\left| \Omega - \frac{k}{d_V} \right| \ll \frac{1}{L_V} \quad \text{and} \quad \left| \Psi - \frac{l}{d_H} \right| \ll \frac{1}{L_H}. \quad (7.31)$$

These conditions simply say that whenever the signals of two UEs arrive from similar azimuth and elevation angles they will interfere strongly with each other. To get a feeling of what “ \ll ” means in (7.31), we can similarly to (1.31) rely on the approximation $\sin(\pi z) \approx \pi z$ for $|z| < 0.2$ to show that $S(\Omega) \approx 1$ whenever $|\Omega - k/d_V| < 0.2/L_V$ (the same holds analogously for $T(\Psi)$). The quantities k, l in the conditions above arise because $S(\Omega)$ and $T(\Psi)$ are periodic and, depending on the antenna spacing, more than one period fits into the interval $[-2, 2]$ on which both functions are defined. The conditions in (7.31) also reveal that the angular resolution of the array in the elevation and azimuth dimension is respectively defined through its normalized height L_V and width L_H . Note that this resolution is not affected by a change of the number of antennas as long as they are distributed over the same area. For example, doubling the number of antennas while halving their distances does not improve the spatial resolution of the antenna array (although the array gain is increased). Similarly, the width of the main-lobes depends solely on the physical size of the antenna array but not on the number of antennas. For this reason, we require physically large antenna arrays to achieve a high directivity (i.e., narrow main-lobes).

Remark 7.5 (Relation to time-sampling of a band-limited signal). There is an interesting analogy between the time-sampling of a band-limited

¹⁵This can be most easily seen from the definition of $S(\Omega)$ and $T(\Psi)$ in (7.28).

¹⁶Since both functions are periodic, they also have such lobes around multiples of $\frac{1}{d_V}$ and $\frac{1}{d_H}$, respectively. These lobes are also called main-lobes, while the smaller lobes in between are called side-lobes.

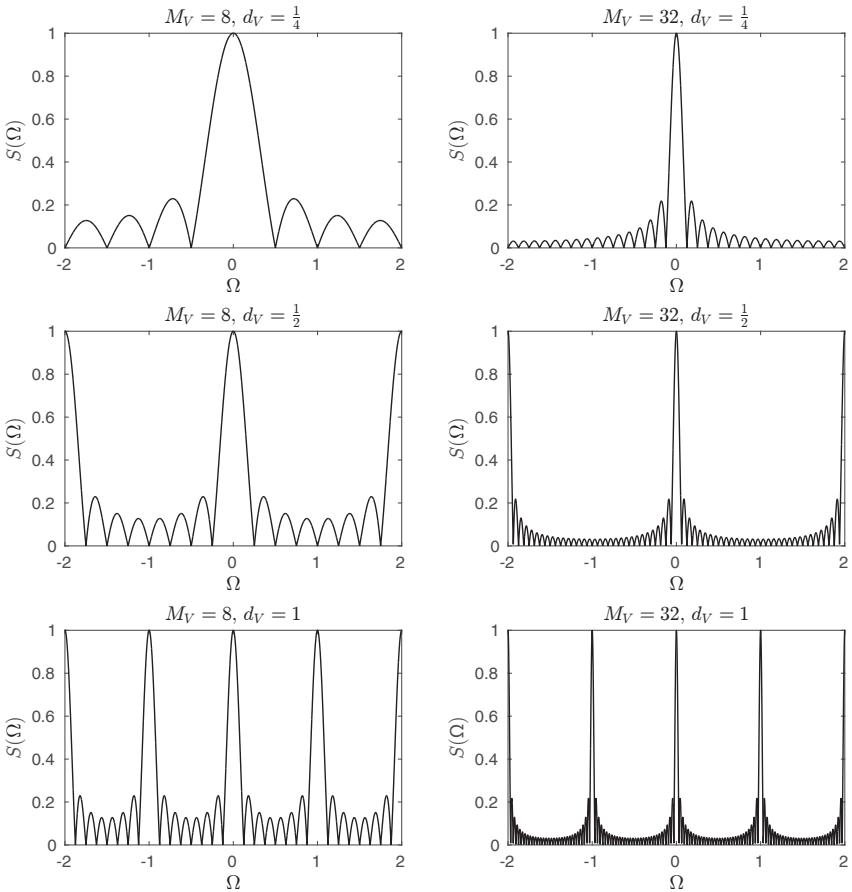


Figure 7.23: The function $S(\Omega)$ in (7.29) for different values of M_V and d_V .

signal and the spatial-sampling of an EM field through an antenna array [268] (see also [314, Sec. 7.3.3]). A passband signal of bandwidth B allows the receiver to resolve two multipath components only if their arrival times are separated by more than $\frac{1}{B}$. Similar to the observation that increasing the number of antennas for a fixed array size does not improve the spatial resolution, increasing the sampling rate does not help to distinguish more multipath components (or to increase the time-resolution). Another analogy can be made with respect to OFDM. A symbol time of T_s creates subcarriers with frequency spacing $1/T_s$. Thus, the frequency resolution is defined by the symbol length and not

by the number of samples per symbol. Increasing the sampling rate will increase the bandwidth but not the frequency resolution.

Even though the beamwidth and the angular resolution are independent of the antenna spacing (or, equivalently, of the number of antennas), this parameter has a very important influence on the characteristics of the antenna array. We have noted earlier that $S(\Omega)$ and $T(\Psi)$ are defined on the interval $[-2, 2]$, that they are periodic with period $\frac{1}{d_V}$ and $\frac{1}{d_H}$, respectively, and that they have a main-lobe around the origin. Thus, whenever $d_V \geq \frac{1}{2}$ or $d_H \geq \frac{1}{2}$, one or more additional main-lobes appear in the interval of interest. This implies that two UEs whose signals arrive from azimuth and elevation angles such that $\Omega = \frac{k}{d_V}$ or $\Psi = \frac{k}{d_H}$, for some integer k , cannot be separated by the array. For this reason, we often consider antenna arrays with *critical spacing* $d_V = \frac{1}{2}$ and $d_H = \frac{1}{2}$.¹⁷ Antenna arrays with larger/smaller antenna spacing are called super-/sub-critically spaced. Similar to the fact that undersampling of a signal in the time domain creates aliasing (i.e., higher frequency components are distinguishable from lower frequency components), spatial undersampling creates aliasing in the angular domain since multiple distinct directions cannot be distinguished by the array.

In some cases, it is desirable to have a high angular resolution or directivity, but one cannot afford to have a physically long/high array with a large number of critically spaced antennas. As explained above, super-critical spacing will lead to multiple main-lobes. However, if the additional main-lobes appear at values outside the feasible range of Ω and Ψ it does not matter. For example, in a typical cell sector, the outdoor UEs are located on the ground below the antenna array such that θ is limited to the range $[-\pi/2, 0]$. As a consequence, Ω is limited to the interval $[-1, 1]$ and a spacing of $d_V = 1$ is sufficient to avoid the negative effect of having multiple main-lobes. Due to this reason, antenna arrays for cellular communications have often super-critical vertical antenna spacing. In some other scenarios, where separability is not an important criterion, multiple main-lobes can even be desired to illuminate better certain areas of a cell-sector.

¹⁷For critical spacing, the second main-lobe appears only in the pathological case where $\theta_1 = \pi/2$ and $\theta_2 = -\pi/2$ or $\theta_1 = \theta_2 = 0$, $\varphi_1 = \pi/2$ and $\varphi_2 = -\pi/2$.

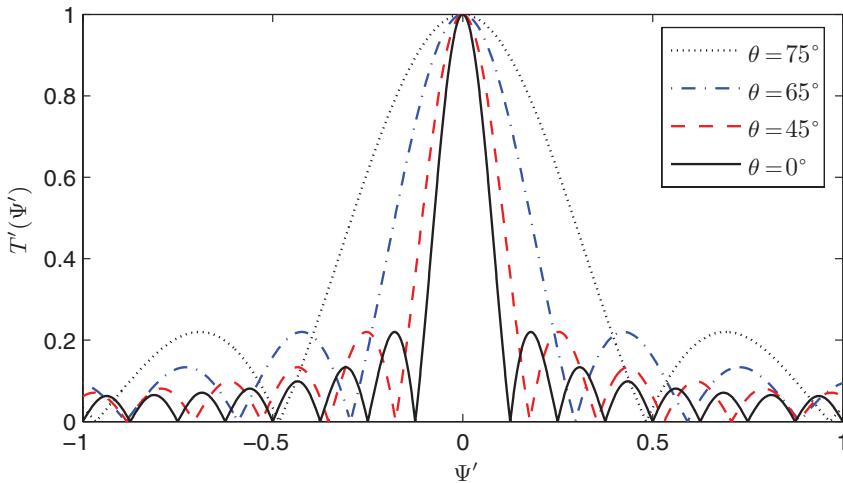


Figure 7.24: The function $T'(\Psi')$ for $M_H = 16$, $d_H = \frac{1}{2}$ and different values of θ .

In Section 7.3.1, we have already observed that the azimuth angular resolution of an array depends on the elevation angle. Using (7.30), this observation can now be mathematically explained. Consider two UEs whose signals arrive from the same elevation angle $\theta_1 = \theta_2 = \theta$ but from different azimuth angles: $\varphi_1 \neq \varphi_2$. This means $\Omega = 0$ and $\Psi = \cos(\theta)(\sin(\varphi_1) - \sin(\varphi_2)) = \cos(\theta)\Psi'$. The function $T'(\Psi') = T(\cos(\theta)\Psi')$ is periodic with period $\frac{1}{\cos(\theta)d_H}$ and has nulls for $\Psi' = k/(\cos(\theta)d_H M_H)$. This is equivalent to the behavior of an array with the same number of antennas but with reduced horizontal antenna spacing $|\cos(\theta)|d_H \leq d_H$. In other words, with increasing elevation angle, the spatial resolution in the azimuth domain is reduced. For two UEs with the same elevation angle, the array behaves like a ULA with antenna spacing $\cos(\theta)d_H$. This effect is visualized in Figure 7.24, which shows the function $T'(\Psi')$ for different values of θ . The main-lobe is less wide for smaller values of θ , which improves the resolution.

Although the above discussion is entirely based on a LoS channel model, it also provides important insights into the representation of NLoS channels. As described in Section 7.3.3, the channel response of a NLoS channel is composed of reflections from different scattering clusters which arrive with different delays from different angular directions. While the resolvability of such multipath components in the time domain

depends on the bandwidth, the angular resolvability depends on the array geometry. One can thus decompose the channel response into spatially separable components that essentially determine the degrees of freedom of the wireless channel. For example, a horizontal ULA is able to sample the angular domain with a resolution of $1/L_H$. The number of non-empty “angular bins” determines hence the available degrees of freedom of the channel. In a pure LoS channel, all signals arrive from the same direction and fall consequently in the same angular bin. The angular domain representation of wireless channels was explored in [289] and is discussed in great detail in [314, Sec. 7.3.3-7].

7.4.3 Cell-Free Systems

A radically different way of deploying large antenna arrays, and even cellular networks in general, is by distributing subarrays, connected through optical fibers, over a large geographic area as illustrated in Figure 7.25. This idea is sometimes referred to as *cell-free* Massive MIMO [240, 236], although it is conceptually very close to coordinated multipoint (CoMP) with joint transmission or network MIMO [126, 325, 156, 46] with the difference that the number of deployed antennas per UE is assumed to be large, as in Massive MIMO. Expressed with the parameters of the canonical system model in Definition 2.1 on p. 217, we have a single cell (i.e., $L = 1$) in which a BS consisting of S distributed antenna arrays with M_s antennas, $s = 1, \dots, S$, respectively, is deployed and serves K UEs. The total number of antennas is $M = \sum_{s=1}^S M_s$. The channel $\mathbf{h}_k^s \in \mathbb{C}^{M_s \times 1}$ from UE k to subarray s has the spatial correlation matrix $\mathbf{R}_k^s \in \mathbb{C}^{M_s \times M_s}$, so that the overall channel $\mathbf{h}_k \in \mathbb{C}^{M \times 1}$ of UE k is modeled as $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_k)$ where the block-diagonal correlation matrix is

$$\mathbf{R}_k = \text{diag} \left(\mathbf{R}_k^1, \dots, \mathbf{R}_k^S \right). \quad (7.32)$$

The SE of such a system is then given by the results in Section 4. However, the problems of power control and pilot allocation have a quite different flavor in a setting with distributed arrays and reuse of pilots within the same cell. Thus, the algorithms described in Sections 7.1 and 7.2 are generally not applicable.

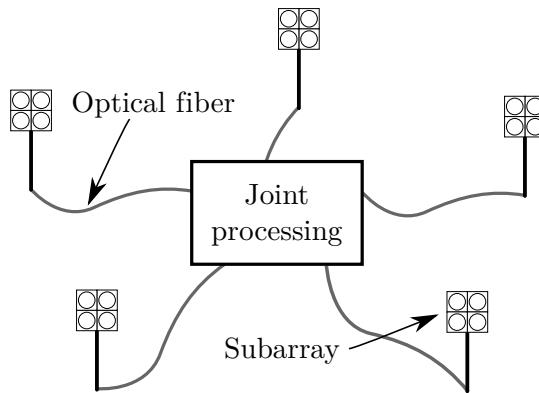


Figure 7.25: A cell-free Massive MIMO system consisting of five spatially distributed subarrays with four antennas each.

The key motivation behind cell-free Massive MIMO is to trade-off channel hardening (as defined in Section 2.5.1 on p. 231) against macro-diversity.¹⁸ Since a UE is likely to have strong channels to only a few subarrays, its overall channel vector is dominated by a few strong components [78]; that is, a few of the matrices \mathbf{R}_k^s have much larger eigenvalues than the others. As a result of this strong spatial channel correlation, the effect of channel hardening is dramatically reduced and SE expressions that rely on hardening can vastly underestimate the achievable performance [78]. However, the UL expression in Theorem 4.1 on p. 276 and the DL expression in Theorem 4.9 on p. 326 are still usable. The increased macro-diversity in cell-free systems can lead to significant power gains and a more uniform distribution of the received signal power, especially in urban scenarios with strong shadow fading.

Cell-free systems employ TDD to exploit channel reciprocity and signal processing relying on only locally available CSI at each subarray, to avoid the costly exchange of CSI and precoding/combining vectors. MR is one suitable scheme due to its distributed nature, but subarrays

¹⁸Since the subarrays are sufficiently separated, their average channel gain coefficients to a randomly located UE can be viewed as uncorrelated random variables (depending on how the subarrays are distributed). This effect is called “macro-diversity” in contrast to “micro-diversity” which relates to uncorrelated small-scale fading coefficients of sufficiently spaced antennas of the same subarray.

with multiple antennas can also process the signals jointly within the array [61]. Joint power control over all subarrays has been shown to be beneficial [241, 237]. The price to pay for cell-free Massive MIMO systems is a high traffic load on the optical fiber connections because payload data for all UEs must be sent to all subarrays. Given the high deployment cost of optical fiber and the technical challenges related to the synchronization and calibration of a large number of distributed antennas arrays, it is unclear at the time of writing of this monograph if cell-free systems will be used in practice. One step towards reducing the backhaul traffic is to serve each UE only by the few subarrays that have the strongest average channel gains [45]. One alternative but less radical approach is to have multiple cells with distributed subarrays in each of them [53].

We will now demonstrate how the effect of channel hardening is reduced in a cell-free Massive MIMO system. To this end, we consider a single UE located at the center of a cell of fixed radius in which S subarrays with M_S antennas each are distributed uniformly at random. We vary the values of S and M_S in such a way that the total number of antennas $M = SM_S = 256$ remains constant. We ignore spatial channel correlation at the subarrays and focus only on large-scale fading. Thus, the UE's channel $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R})$ has the correlation matrix

$$\mathbf{R} = \text{diag}\left(\beta_1 \mathbf{I}_{M/S}, \dots, \beta_S \mathbf{I}_{M/S}\right) \quad (7.33)$$

where β_s is the average channel gain of subarray s . The channel gain coefficients are modeled according to (2.3) as $\beta_s = -148.1 - 37.6 \log_{10}(d_s/1 \text{ km}) + F_s$, where d_s is the distance between the UE and subarray s and $F_s \sim \mathcal{N}(0, 8)$ accounts for shadow fading. Recall from (2.17) that the variance of $\|\mathbf{h}\|^2 / \text{tr}(\mathbf{R})$ is a suitable measure of channel hardening, where a smaller variance corresponds to more hardening. Figure 7.26 shows this variance as a function of S for a cell radius of either 100 m or 350 m. We assumed that the subarrays are uniformly distributed within a disc of radius 100 m or 350 m around the UE while keeping a minimum distance of 10 m. The results are calculated over 1000 random drops of subarrays. One can clearly see that the variance increases with the number of subarrays, which means that the channel hardening is reduced. The smaller the density of subarrays (i.e., the

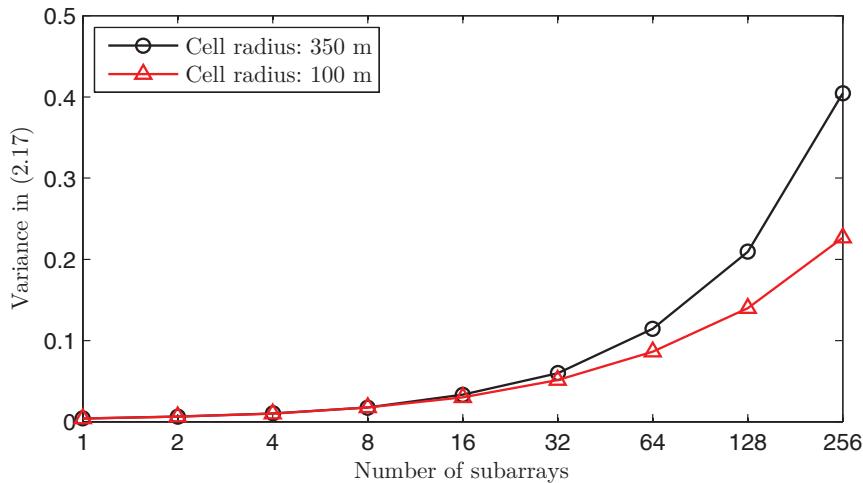


Figure 7.26: Variance in (2.17) of $\|\mathbf{h}\|^2 / \text{tr}(\mathbf{R})$, which measures the closeness to channel hardening, as a function of the number of subarrays S for a cell radius of 100 m and 350 m. The variance is computed with respect to random subarray locations and small-scale fading realizations.

larger the cell radius), the more pronounced is this effect since there are fewer dominating subarray links.

Although increasing S reduces the desirable effect of channel hardening, it increases the macro-diversity. This can be seen from Figure 7.27, which shows the CDF of the average channel gain $\beta = \frac{1}{M} \text{tr}(\mathbf{R})$ for different numbers of subarrays. Interestingly, the expected average channel gain $\mathbb{E}\{\beta\}$ is independent of S (where the expectation is taken with respect to the subarray locations as well as the shadow fading realizations) owing to the i.i.d. large-scale fading coefficients of the individual subarrays. However, going from $S = 1$ to $S = 16$ improves the median of β by around 12.5 dB; increasing this number to $S = 256$ adds another 9.5 dB. Thus, cell-free systems reduce substantially the fluctuations in average channel gain among the UEs, which is important when the goal is to guarantee a certain service quality to the UEs, uniformly over the coverage area.

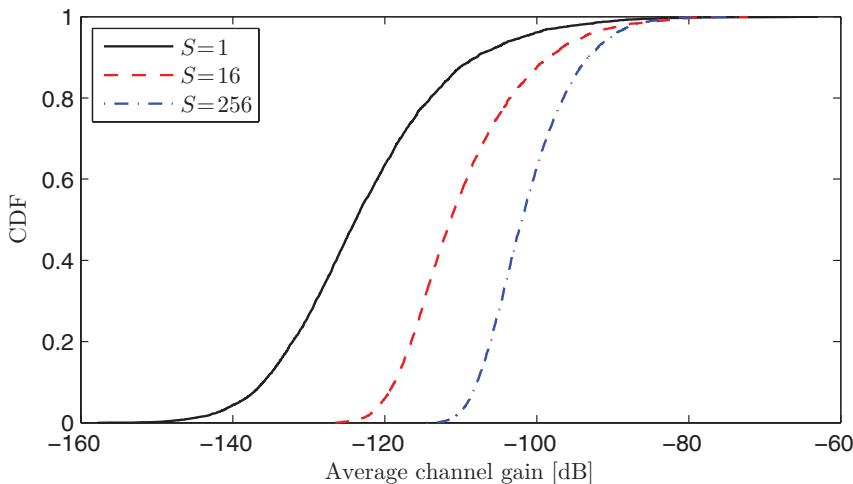


Figure 7.27: CDF of the average channel gain $\beta = \frac{1}{M} \text{tr}(\mathbf{R})$ in a cell-free system for different numbers of subarrays $S \in \{1, 16, 256\}$.

7.4.4 Polarization

One of the aspects related to antenna design we have not been discussed is the polarization of the EM waves radiated by the UEs or BSs. If one tracks at a fixed location the movement of the tip of the electric field vector over time, one obtains a curve called the polarization ellipse. One can then classify the polarization of an EM wave according to the shape of this ellipse which can be either linear, circular, or elliptical. In cellular communications, linearly polarized antennas are most commonly used. The direction of a linear polarization is defined by the tilt angle of the polarization ellipse; for example, 90° (vertical), 0° (horizontal), and $\pm 45^\circ$ (slant). Linear polarization always come in two orthogonal pairs; for example, horizontal and vertical or $\pm 45^\circ$ slant. An example of a horizontally and a vertically polarized wave is shown in Figure 7.28. Importantly, any linear polarization can be obtained from a superposition of two orthogonal polarizations. Antennas which radiate (or respond to) EM waves with only one polarization direction are called *uni-polarized*, while antennas that create (or respond to) field components in two orthogonal polarization directions are called *dual-polarized*. Generally, UEs have uni-polarized antennas or antennas with

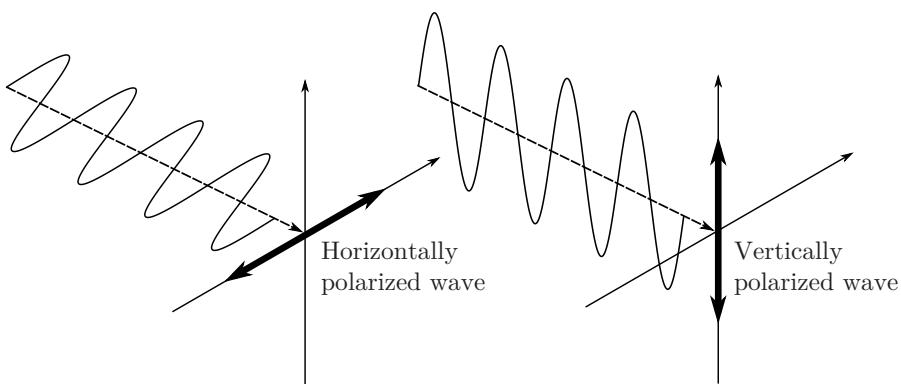


Figure 7.28: Polarization ellipse for a horizontally and a vertically polarized wave.

a strong dominating polarization.¹⁹ However, one needs to keep in mind that the effective polarization depends on the physical orientation of the UE which is generally random (e.g., depending on how a smartphone is held). For example, a UE with a vertically polarized antenna can create (or respond to) an EM wave with vertical and horizontal polarization components if it is rotated by a few degrees. In order to prevent that radiated energy in one of the polarization directions is lost, BSs are almost exclusively equipped with dual-polarized antennas.

Ideally, a horizontally polarized EM wave should not be received by a vertically polarized antenna and vice versa. However, there is generally some form of cross-talk between both polarization directions due to imperfect cross-polar isolation (XPI) of the antennas and imperfect cross-polar discrimination (XPD) of the channel. XPI is a property of the antenna alone and describes how a (supposedly) linearly polarized antenna responds to cross-polar wave components; that is, the components of an EM wave in the orthogonal polarization direction. XPD describes the phenomenon that EM waves can change their polarization when going through a scattering medium. This effect is also called (channel) depolarization. In general, the reflection coefficients for each polarization are different so that the phases of the two orthogonal polarizations undergo different changes for each reflection. Thus, in a sufficiently rich

¹⁹We use the terms “polarization” and “polarization direction” interchangeably.

scattering propagation environment, the received polarization would be independent of the transmitted polarization. However, in practice, we do not have sufficient scattering to observe full depolarization and some correlation exists [323].

Remark 7.6 (Dual-polarized antenna arrays). When we speak about a dual-polarized antenna array with M antennas in this monograph, we consider an antenna array composed of $M/2$ uni-polarized antennas for each polarization direction. For space reasons, the antennas for both polarization directions are generally co-located. This means that an antenna array with dual-polarized co-located antennas accommodates twice the number of antennas as a uni-polarized array of the same physical dimensions.

Channel Model with Dual-polarized Antennas

We will now present a general channel model for a BS antenna array consisting of M dual-polarized co-located antennas and a UE with a uni-polarized, but randomly oriented, antenna, as shown in Figure 7.29.²⁰ This model is inspired by [89, 171]. The interested reader is also referred to [88, Sec. 3.3] and [249, 293] for a more detailed discussion of polarization modeling for MIMO systems. Denote by $h_{i,V}$ and $h_{i,H}$ the channel coefficients from the UE to the i th vertically and horizontally polarized antenna, respectively. Then, for $i = 1, \dots, \frac{M}{2}$,

$$\mathbf{h}_i \triangleq \begin{bmatrix} h_{i,V} \\ h_{i,H} \end{bmatrix} = \mathbf{F}_{\text{BS}} \mathbf{Z}_i \mathbf{m}_{\text{UE}}(\theta_r) \quad (7.34)$$

where

- $\mathbf{F}_{\text{BS}} \in \mathbb{C}^{2 \times 2}$ is the deterministic polarization matrix of the BS's antennas which has orthonormal columns and can be modeled as

$$\mathbf{F}_{\text{BS}} = \frac{1}{\sqrt{1 + \chi_a}} \begin{bmatrix} 1 & \sqrt{\chi_a} e^{j\phi_a} \\ -\sqrt{\chi_a} e^{j\phi_a} & 1 \end{bmatrix} \quad (7.35)$$

where χ_a^{-1} is the XPI and ϕ_a a phase offset depending on the antenna characteristics;

²⁰We assume isotropic radiating antennas. Thus, only the polarization but not the radiation pattern is affected by rotations.

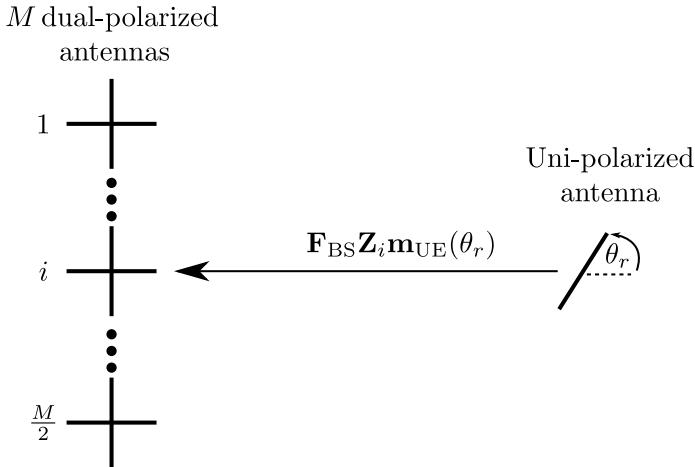


Figure 7.29: A BS antenna array with M dual-polarized co-located antennas and a UE with a uni-polarized, but randomly oriented, antenna. The 2×1 channel from the UE to the i th antenna location is denoted $\mathbf{F}_{\text{BS}} \mathbf{Z}_i \mathbf{m}_{\text{UE}}(\theta_r)$.

- $\mathbf{m}_{\text{UE}}(\theta_r) \in \mathbb{C}^{2 \times 1}$ is the unit-norm polarization vector of the UE's antenna, which is modeled as²¹

$$\mathbf{m}_{\text{UE}}(\theta_r) = \frac{1}{\sqrt{1 + \chi_b}} \begin{bmatrix} \cos(\theta_r) & -\sin(\theta_r) \\ \sin(\theta_r) & \cos(\theta_r) \end{bmatrix} \begin{bmatrix} 1 \\ \sqrt{\chi_b} e^{j\phi_b} \end{bmatrix} \quad (7.36)$$

where $\theta_r \in [0, 2\pi)$ is the rotation angle of the polarization directions, χ_b^{-1} is the XPI, and ϕ_b a phase offset depending on the antenna characteristics;

- $\mathbf{Z}_i \in \mathbb{C}^{2 \times 2}$ is the random channel matrix describing the propagation of the vertical and horizontal EM wave components, which is defined as

$$\mathbf{Z}_i = \begin{bmatrix} z_{i,\text{VV}} & z_{i,\text{VH}} \\ z_{i,\text{HV}} & z_{i,\text{HH}} \end{bmatrix}. \quad (7.37)$$

Somewhat surprisingly, the 2×1 channel \mathbf{h}_i from the UE to two co-located dual-polarized antennas depends on four complex random quantities in \mathbf{Z}_i which we will characterize next. The XPD describes the

²¹The matrix in (7.36) is a rotation matrix.

ability of the channel to separate vertical and horizontal polarizations and is defined as the ratio (independent of i)

$$\text{XPD} = \frac{\mathbb{E}\{|z_{i,\text{VV}}|^2\}}{\mathbb{E}\{|z_{i,\text{HV}}|^2\}} = \frac{\mathbb{E}\{|z_{i,\text{HH}}|^2\}}{\mathbb{E}\{|z_{i,\text{VH}}|^2\}} = \frac{1 - q_{\text{XPD}}}{q_{\text{XPD}}} \quad (7.38)$$

that depends on the parameter $q_{\text{XPD}} = 1/(1 + \text{XPD})$, which satisfies $0 < q_{\text{XPD}} \leq 1$. In this definition, we have used the following assumptions

$$\begin{aligned} \mathbb{E}\{|z_{i,\text{VV}}|^2\} &= \mathbb{E}\{|z_{i,\text{HH}}|^2\} = 1 - q_{\text{XPD}} \\ \mathbb{E}\{|z_{i,\text{HV}}|^2\} &= \mathbb{E}\{|z_{i,\text{VH}}|^2\} = q_{\text{XPD}} \end{aligned} \quad (7.39)$$

which additionally imply that $\mathbb{E}\{\text{tr}(\mathbf{Z}_i \mathbf{Z}_i^H)\} = 2$. This means that the total radiated power is transferred through the channel without any losses, whatever the value of XPD. The higher the XPD (or the smaller q_{XPD}), the smaller is the power that leaks from one polarization to the other. Due to insufficient scattering, it is reasonable to assume that the components of \mathbf{Z}_i are correlated. We adopt here the separable correlation model

$$\mathbf{Z}_i = \left(\mathbf{C}_{r_p}^{\frac{1}{2}} \mathbf{G}_i \mathbf{C}_{t_p}^{\frac{1}{2}} \right) \odot \boldsymbol{\Sigma} \quad (7.40)$$

where \odot is the Hadamard (or elementwise) product,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{1 - q_{\text{XPD}}} & \sqrt{q_{\text{XPD}}} \\ \sqrt{q_{\text{XPD}}} & \sqrt{1 - q_{\text{XPD}}} \end{bmatrix} \quad (7.41)$$

is the XPD matrix,

$$\mathbf{C}_{r_p} = \begin{bmatrix} 1 & r_p \\ r_p^* & 1 \end{bmatrix}, \quad \mathbf{C}_{t_p} = \begin{bmatrix} 1 & t_p \\ t_p^* & 1 \end{bmatrix} \quad (7.42)$$

are the receive and transmit polarization correlation matrices, respectively, and

$$\mathbf{G}_i = \begin{bmatrix} g_{i,\text{VV}} & g_{i,\text{VH}} \\ g_{i,\text{HV}} & g_{i,\text{HH}} \end{bmatrix} \quad (7.43)$$

contains the small-scale fading channel coefficients (whose distributions also account for the average channel gain β , defined in (2.2)). Since the horizontally and vertically polarized antennas are co-located, there is no

additional spatial correlation on top of the polarization correlation between the elements of \mathbf{G}_i . However, the vectors $\mathbf{g}_{xy} = [g_{1,xy}, \dots, g_{\frac{M}{2},xy}]^T$ for $x, y \in \{\text{V}, \text{H}\}$ consisting of the stacked xy -components of each of the $M/2$ channel matrices \mathbf{G}_i are spatially correlated with the same correlation matrix $\mathbf{R} \in \mathbb{C}^{\frac{M}{2} \times \frac{M}{2}}$; that is, $\mathbf{g}_{xy} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{\frac{M}{2}}, \mathbf{R})$ for $x, y \in \{\text{V}, \text{H}\}$. One can show after some simple algebra that the entire channel matrix $\mathbf{Z} = [\mathbf{Z}_1^H, \dots, \mathbf{Z}_{\frac{M}{2}}^H]^H \in \mathbb{C}^{M \times 2}$ can be modeled as

$$\mathbf{Z} = \left((\mathbf{R} \otimes \mathbf{C}_{r_p})^{\frac{1}{2}} \mathbf{W} \mathbf{C}_{t_p}^{\frac{1}{2}} \right) \odot \left(\mathbf{1}_{\frac{M}{2}} \otimes \boldsymbol{\Sigma} \right) \quad (7.44)$$

where $\mathbf{W} \in \mathbb{C}^{M \times 2}$ has i.i.d. $\mathcal{N}_{\mathbb{C}}(0, 1)$ elements and \otimes denotes the Kronecker product. The final channel vector $\mathbf{h} = [\mathbf{h}_1^H, \dots, \mathbf{h}_{\frac{M}{2}}^H]^H \in \mathbb{C}^{M \times 1}$ is then given as

$$\mathbf{h} = \left(\mathbf{I}_{\frac{M}{2}} \otimes \mathbf{F}_{\text{BS}} \right) \mathbf{Z} \mathbf{m}_{\text{UE}}(\theta_r). \quad (7.45)$$

Measurements have shown that the absolute values of the transmit and receive polarization correlation are rather small (i.e., $|t_p|, |r_p| < 0.2$ [108]) and can be sometimes even fully neglected [25]. The XPD has also been found to be rather strong in cellular systems, lying in the range from 5 dB to 15 dB [25]. We can additionally assume in simulations that the UE's polarization is uniformly randomly distributed: $\theta_r \sim U[0, 2\pi]$.

It is difficult to say if dual-polarized antennas are advantageous for Massive MIMO systems, apart from allowing for more compact arrays. To exemplify this issue, consider a BS with a horizontal ULA of either M critically-spaced uni-polarized or dual-polarized—but not co-located (i.e., neighboring antennas have orthogonal polarizations)—antennas, as shown in Figure 7.30. The array has the same number of antennas and the same length in both cases and hence the same angular resolution (see (7.31)). We assume that there are three UEs in the cell at azimuth angles $\varphi_1 = 30^\circ$, $\varphi_2 = 25^\circ$, and $\varphi_3 = -10^\circ$, respectively. Figure 7.31 shows the average correlation $\mathbb{E} \{ |\mathbf{h}_1^H \mathbf{h}_i|^2 / (\|\mathbf{h}_1\| \|\mathbf{h}_i\|)^2 \}$ for $i = 2, 3$ of the channels between UE 1 and UE 2 and between UE 1 and UE 3. We use the local scattering model with Gaussian angular distribution and ASD of 5° . The polarization parameters are $r_p = t_p = 0.2$, XPD = 7 dB, and both the BS and the UE have antennas with perfect XPI (i.e.,

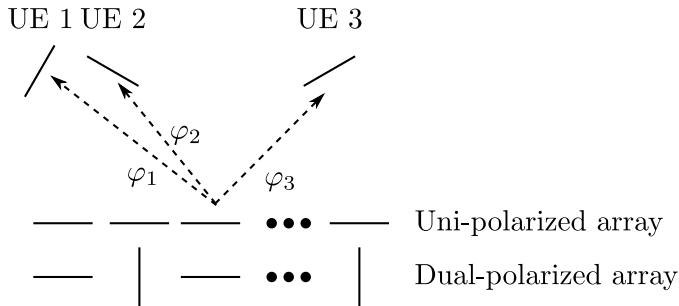


Figure 7.30: A horizontal uni-polarized or dual-polarized (but not co-located) ULA of M critically-spaced antennas communicating with three UEs.

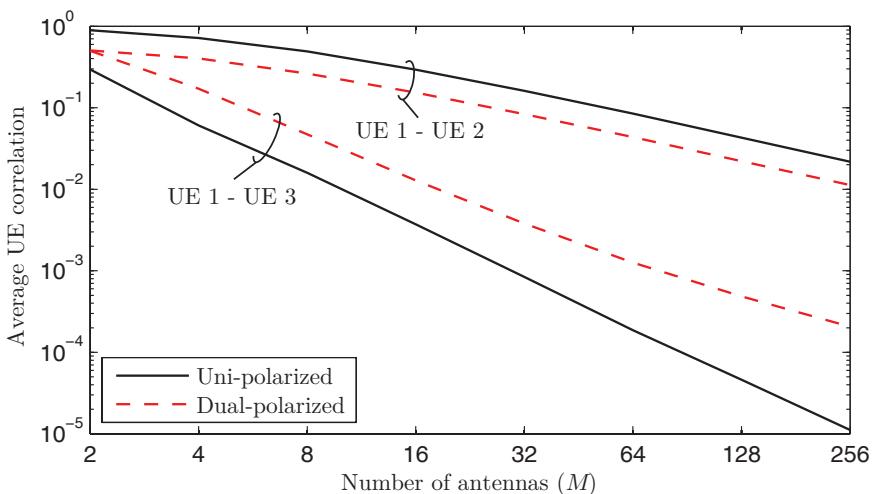


Figure 7.31: Average correlation between the channels of UEs 1 and 2 and between the channels of UEs 1 and 3, as a function of M (in logarithmic scale), for uni- and dual-polarized antenna arrays.

$\chi_a = \chi_b = 0$). The UEs' polarization directions are uniformly randomly distributed.

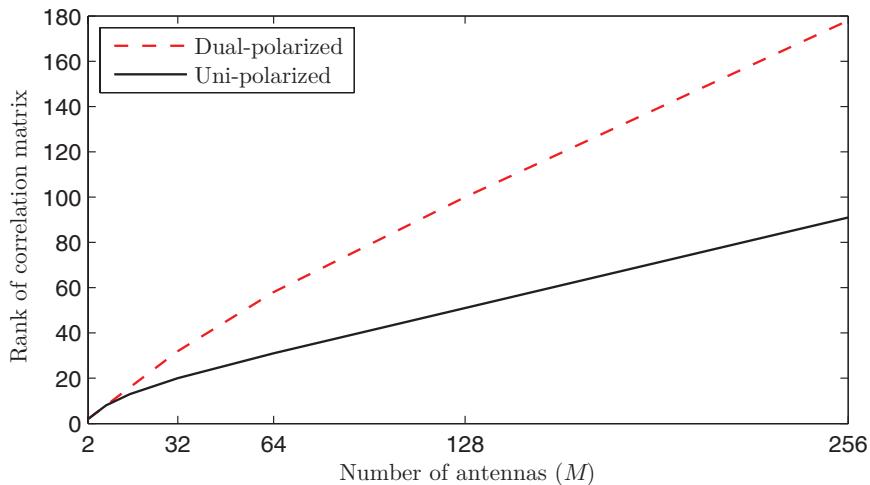
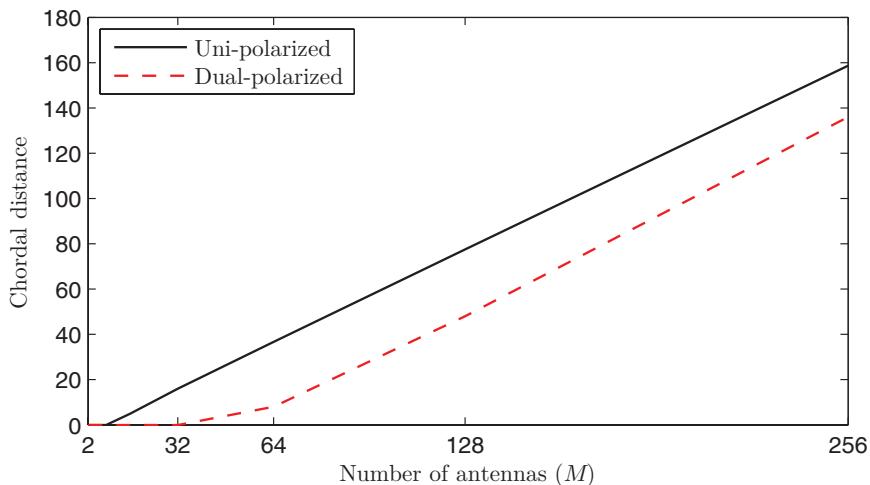
We can observe that the dual-polarized array achieves a better separation of UE 1 and UE 2 than the uni-polarized array. The opposite is true for UE 1 and UE 3. The reason for this behavior is as follows. UEs 1 and 2 are located very closely together and their channels have

similar correlation matrices. We have seen in Figure 7.15 that the dimension p of the dominating eigenspace (which contains almost all energy) can be much smaller than M under spatial channel correlation. Since the UEs essentially share the available degrees of freedom of the same p -dimensional subspace, the average correlation of their channels decays not as $1/M$ but as $1/p$. Now, since the orthogonal polarization components are almost uncorrelated, the dual-polarized array almost doubles the number of degrees of freedom compared to the uni-polarized array. Thus, whenever a system is limited by spatial channel correlation, polarization diversity helps. Figure 7.32a shows the rank of the spatial correlation matrix of UE 1 for the uni- and dual-polarized arrays as a function of M . This demonstrates that the use of a dual-polarized array almost doubles the degrees of freedom. This effect can be also explained mathematically from (7.45). Since $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A})\text{rank}(\mathbf{B})$ for two matrices \mathbf{A} , \mathbf{B} with suitable dimensions, it follows that

$$\text{rank}(\mathbf{R} \otimes \mathbf{C}_{r_p}) = \text{rank}(\mathbf{R}) \text{rank}(\mathbf{C}_{r_p}) = 2 \text{rank}(\mathbf{R}). \quad (7.46)$$

On the other hand, UEs 1 and 3 have correlation matrices whose eigenspaces are very different. Thus, even if each of the UEs' spatial correlation matrices had only rank one, their average correlation would be small because the correlation-eigenspaces are almost orthogonal. In this case, dual-polarized arrays are disadvantageous because they increase the dimensions of the subspaces to which the channel vectors are confined. This can be seen from Figure 7.32b, which shows the chordal distance (as defined in (7.21)) of the eigenspaces of the correlation matrices of UE 1 and 3 for the uni- and dual-polarized arrays as a function of M . The chordal distance for the uni-polarized array is much larger than that for the dual-polarized array, indicating that it achieves a better decorrelation of the UEs' channels.

Due to the rather involved distribution of \mathbf{h} in (7.45) with dual-polarized antennas, we have refrained from explicitly considering polarization in the analysis carried out in this monograph. However, since we have considered arbitrary correlation matrices in most sections, the results apply also to correlated Rayleigh fading with dual-polarized antennas. The interested reader is referred to the rather small number of publications dealing with dual-polarized Massive MIMO systems,

(a) Rank of the spatial correlation matrix of UE 1 as a function of M .(b) Chordal distance between the eigenspaces of the spatial correlation matrices of UE 1 and UE 3 as a function of M .**Figure 7.32:** Comparison of spatial correlation matrices with uni- and dual-polarization.

such as [252, 253, 354]. The 3GPP 3D MIMO model [1] and its open-source implementations [159, 5] support the simulation of dual-polarized antenna arrays, which will be used in the case study in Section 7.7.

7.5 Millimeter Wavelength Communications

Most wireless communication systems today make use of the 300 MHz–6 GHz frequency range while the spectrum from 6–300 GHz is comparatively empty. This is mainly due to the very advantageous propagation characteristics at low frequencies which allow the radio waves to penetrate buildings, reflect multiple times, and bend around corners. The mmWave band refers to the frequency range from 30–300 GHz with wavelength ranging from 1–10 mm. MmWaves suffer from high atmospheric absorption, rain and foliage attenuation, strong penetration and reflection losses, and little diffraction, which essentially restrict their use to LoS outdoor-to-outdoor or indoor-to-indoor communications over relatively short distances. The current main use cases are wireless backhaul in the unlicensed 60 GHz band, as a cost-efficient alternative to wired solutions, and WLANs based on the IEEE 802.11ad standard. Nevertheless, recent theoretical considerations and measurement campaigns have provided evidence that outdoor SCs with up to 200 m cell radii are viable if the transmitters and receivers are equipped with sufficiently “large” antenna arrays (in a sense that we will define shortly) to compensate for the otherwise prohibitive propagation losses [259, 275].

To understand why large antenna arrays or—to be more precise—antennas composed of a large number of radiating elements (see Definition 7.3) are needed for mmWave communications, let us have a look at Friis’ transmission formula [118] which describes the relation between the received signal power P_r and transmitted signal power P_t for two antennas separated by distance d under ideal conditions and free-space propagation:

$$\frac{P_r}{P_t} = G_r G_t \left(\frac{\lambda}{4\pi d} \right)^2 = \frac{A_r A_t}{(d\lambda)^2} \quad (7.47)$$

where λ is the wavelength, G_t, G_r and A_t, A_r are the gains²² and the effective areas²³ of the transmitter and receiver antennas, respectively. Note that (7.47) assumes that both antennas are perfectly aligned and neglects that G_t, G_r are generally angle-dependent. The first part of the equation says that for fixed G_t, G_r , the pathloss P_t/P_r is proportional to λ^{-2} , while the second part says that for fixed A_t, A_r , the pathloss is proportional to λ^2 . The seeming contradiction between the two results is resolved once we understand that a dipole (or any other radiating element) has a frequency-independent gain but an effective area that shrinks with the carrier frequency. For example, a half-wavelength dipole has an effective area of approximately $0.125 \lambda^2$, but a fixed gain of 0.5π over an (hypothetical) isotropic antenna [118] with effective area $\lambda^2/(4\pi)$. The effective area of an antenna can therefore only be kept constant as λ decreases if more and more radiating elements are connected together. The number of radiating elements fitting into a given area is proportional to λ^{-2} and so is the resulting antenna gain. The discussion above implies that we can even achieve a net gain in the received signal power proportional to λ^{-2} , given that the antennas at the transmitter and the receiver are composed of a large number of radiating elements that scales at the same speed. However, in order to keep a constant pathloss, it is sufficient to scale the number of elements such that $A_r A_t \sim \lambda^2$. For example, the receiver antenna could have a single radiating element while the number of elements of the transmit antenna scales as λ^{-2} or vice versa. Alternatively, the number of elements at both antennas could be scaled such that their effective areas are linear in λ . For cellular communications, one could hence have UEs with comparatively few radiating elements, while BSs with antennas composed of many radiating elements compensate for the vast majority of the propagation loss.

²²The antenna gain is the ratio of the maximum power density radiated by the antenna in any direction to that of an ideal isotropic antenna. It includes the antenna's efficiency, i.e., the ratio of the input power to the total radiated power.

²³The effective area of an antenna is equal to the area, oriented perpendicular to the incoming wave, which would collect the same amount of power as was actually received by the antenna.

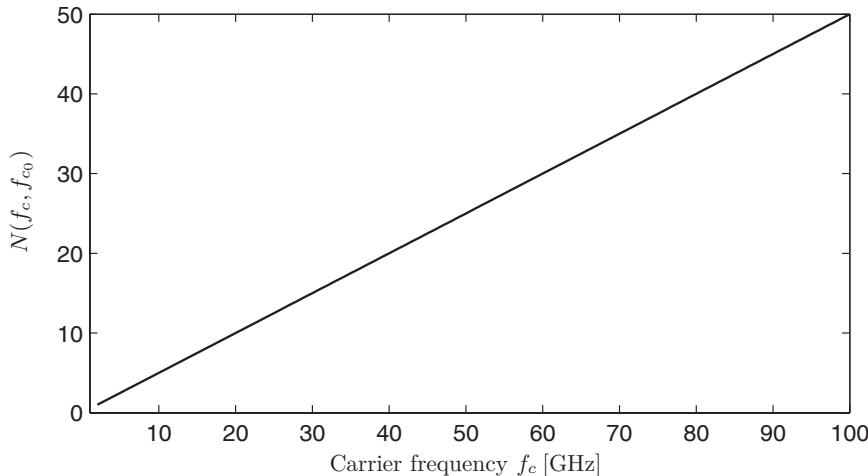


Figure 7.33: The function $N(f_c, f_{c0})$ versus carrier frequency f_c for a fixed reference carrier frequency $f_{c0} = 2\text{ GHz}$. $N(f_c, f_{c0})$ tells us how many times more half-wavelength dipoles are needed at the transmit and receive antennas at carrier frequency f_c to maintain the same pathloss as observed at f_{c0} .

In order to get an idea of the required number of radiating elements at mmWave frequencies, denote by $A(\lambda) = 0.125\lambda^2$ the effective area of a half-wavelength dipole. Consider now a communication channel at carrier frequency f_{c0} (with wavelength λ_0) between two antennas consisting of one (or multiple) such dipoles. According to (7.47), in order to achieve the same pathloss at carrier frequency $f_c > f_{c0}$ (with wavelength λ), the antennas at the transmitter and the receiver must consist of

$$N(f_c, f_{c0}) = \frac{\lambda}{\lambda_0} \frac{A(\lambda_0)}{A(\lambda)} = \frac{f_c}{f_{c0}} \quad (7.48)$$

times more dipoles. As a rule of thumb, we can therefore say that doubling the carrier frequency requires twice the number of radiating elements at both the transmitter and receiver to maintain the same received signal strength. Figure 7.33 shows $N(f_c, f_{c0})$ as a function of f_c for $f_{c0} = 2\text{ GHz}$. However, one must not forget that the main motivation of going to higher frequencies is that the available bandwidth is dramatically increased, but since the total transmit power is generally fixed (due to hardware or regulatory constraints), the SNR

is inversely proportional to the bandwidth. Thus, a communication system operating at $f_c = 60$ GHz with 100 MHz of bandwidth would require transmit and receive antennas consisting of $\sqrt{10} \times 30 \approx 95$ times more radiating elements to maintain the same SNR as a communication system operating at $f_{c_0} = 2$ GHz with 10 MHz of bandwidth.²⁴ In a cellular setting, this picture might change as the very focused wave propagation at mmWave frequencies creates essentially isolated communication channels. Hence, achieving the same SINR at higher frequencies as at 2 GHz would probably be easier than maintaining the same SNR. The exact comparison is outside the scope of this monograph.

In the discussion above, we considered a single antenna with multiple radiating elements, according to Definition 7.3. But one must not forget that, although an antenna with many radiating elements has a very high gain, which is necessary to overcome the pathloss at mmWave frequencies, the orientation of its radiation pattern is static and cannot be controlled with a single RF input. For this reason, subsets of the radiating elements need to be provided with individual RF inputs to form an antenna array that has a dynamically controllable array response.

Although it is quite encouraging that mmWave cellular communication is rendered possible by large antenna arrays, one must be aware of the following caveats.

First, we illustrated in Section 7.4.2 that the beamwidth of an antenna array in the elevation (azimuth) domain is inversely proportional to its normalized vertical height L_V (horizontal width L_H) measured in multiples of the wavelength. For a fixed effective area, L_V and L_H scale linearly with λ so that the beamwidth is proportional to λ^{-1} . This means that, although large antenna arrays enable communications at mmWaves due to their high gain, they only do so if the narrow transmit and receive beams are well aligned. As a side effect, this implies that mmWave communication suffers strongly from blocking of the strongest (LoS) path since most other paths depart/arrive at angles which are not

²⁴The factor 30 is needed to counteract the propagation loss (see Figure 7.33) while the additional factor $\sqrt{10}$ is required to compensate for the increased level of noise due to a ten-fold bandwidth extension.

aligned with the beam directions of the arrays. For these reasons, it is challenging to provide coverage over a large area and to support highly mobile UEs with mmWaves. A quickly growing body of literature deals with the problems of beam training and refinement (see [141, 355] and references therein).

Second, equipping each radiating element of a large antenna with an individual RF chain is impossible with today's technology at mmWave frequencies, because the size, cost, and power consumption of the required hardware is prohibitive for use in UEs [274, 141, 355]. In particular, the PAs and DACs/ADCs are very power consuming at mmWaves (if the bandwidth is increased) and so is the parallel processing of a large number of data streams with billions of samples per second. For this reason, alternative approaches such as analog and hybrid analog-digital beamforming²⁵ [11] as well as low-resolution DACs/ADCs [222] are subjects of current research.

Third, as discussed in Remark 2.1 on p. 221, λ has a substantial impact on τ_c , which is the size of the coherence block of the wireless channel. Since the channel coherence time T_c is proportional to λ and the coherence bandwidth B_c is inversely proportional to the delay spread T_d , the coherence block $\tau_c = T_c B_c$ satisfies the proportionality

$$\tau_c \sim \frac{\lambda}{T_d}. \quad (7.49)$$

This equation implies that more frequent pilot signals are required in the mmWave bands than in the sub-6 GHz band. However, since mmWave communication systems are foreseen to have rather small cell radii and support only slowly moving UEs, the delay spread is reduced and the coherence time increased which counterbalances the linear scaling in λ to some extent. For example, assuming 60 GHz carrier frequency (i.e., $\lambda = 5$ mm), delay spread $T_d = 500$ ns corresponding to a path length difference of 150 m, and $v = 3$ km/h mobility, we obtain with the help of Remark 2.1 a channel with $B_c = 1$ MHz and $T_c = 1.5$ ms, resulting in a coherence block of $\tau_c = 1500$ samples.

²⁵Analog beamforming introduces phase shifts to the signals going from a single RF chain to the individual radiating elements. Hybrid analog-digital beamforming refers to an antenna array in which each antenna consists of multiple radiating elements whose phases can be individually controlled via analog beamforming.

In the light of the constraints and challenges outlined above, it appears that mmWave communications are best suited for SC hotspot deployments with the goal of providing high throughput to slowly moving UEs. Due to the rather low number of RF chains compared to the number of radiating elements and a relatively small number of UEs per SC (which are likely not to be served simultaneously on the same time-frequency resource), we do not consider mmWave communications as Massive MIMO systems according to Definition 2.1. The interested reader is referred to the textbook [274] and one of the numerous overview papers (e.g., [272, 141]) for a more detailed introduction into the topic.

7.6 Heterogeneous Networks

We have argued in Section 1.3 on p. 173 that a combination of cell size shrinking (i.e., adding more BSs) and increased spatial multiplexing (i.e., adding more antennas to each BS to serve more UEs simultaneously) is needed to satisfy the area throughput requirements of next generation cellular networks. With both techniques, the network is densified, meaning that the number of antennas per unit area is increased. Massive MIMO can be seen as a concentrated form of network densification, while BSs with a small coverage area and only a few antennas, so-called small-cell base stations (SBSs), are a distributed form of network densification. From a pure capacity point of view, there is some theoretical evidence that one should distribute the available antennas as much as possible [104]. This means that if we would have the freedom to distribute M antennas over L BSs, we should choose $L = M$ and spread out the BSs as widely as possible over the coverage area. However, such a network densification quickly reaches its practical limits because with antennas located below the rooftops and a cell radius of less than 50 m, supporting highly mobile UEs and providing seamless coverage over large areas become increasingly difficult. Backhaul provisioning also becomes prohibitively costly. Cell-free Massive MIMO, as described in Section 7.4.3, is one approach to tackle these challenges, but how well it performs under practical conditions remains to be demonstrated. On the other hand, conventional Massive MIMO is particularly suited to provide area coverage and to support high mobility. Thus, a simple

network architecture, integrating the complimentary benefits of Massive MIMO and SBSs, consists of a coverage tier with Massive MIMO BSs overlaid with a hotspot tier of SBSs. In this architecture, the Massive MIMO BSs ensure coverage and serve highly mobile UEs while the SBSs provide high capacity for indoor and outdoor hotspots. There are two main challenges related to such an architecture:

1. How to avoid cross-tier interference if BSs and SBSs share the same spectrum?
2. How to provide backhaul to a large number of SBSs?

It turns out that the significant amount of excess antennas²⁶ at the Massive MIMO BSs can be used to tackle both of these challenges. This will be discussed in the remainder of this section.

7.6.1 Massive MIMO for Cross-Tier Interference Mitigation

Consider a two-tier network consisting of a canonical Massive MIMO system (according to Definition 2.1 on p. 217) overlaid with a hotspot tier of SBSs. An example of such a network is shown in Figure 7.34. The only differences between SBSs and BSs are that the former has a smaller number of antennas, serve fewer UEs, have less transmit power (and hence a smaller coverage area), and are generally deployed at a lower height. We can, therefore, model such a two-tier network according to Definition 2.1 by choosing appropriate (small) values of M_j and K_j for all indices j that correspond to SBSs. Thus, all results on SE that have been developed in Section 4 can be directly applied. Note, however, that due to the rather small number of antennas, there is little channel hardening, so that the UatF bound (Theorem 4.4 on p. 302) for the UL SE and the hardening bound (Theorem 4.6 on p. 317) for the DL SE are less tight. In this section, we refer to the UEs served by the BSs and SBSs as macro user equipments (MUEs) and small-cell user equipments (SUEs), respectively.

By Definition 2.1, the BSs and SBSs are assumed to operate according to a synchronized TDD protocol and hence share the same spectrum.

²⁶We say that a BS with M antennas serving K UEs has $M - K$ excess antennas.

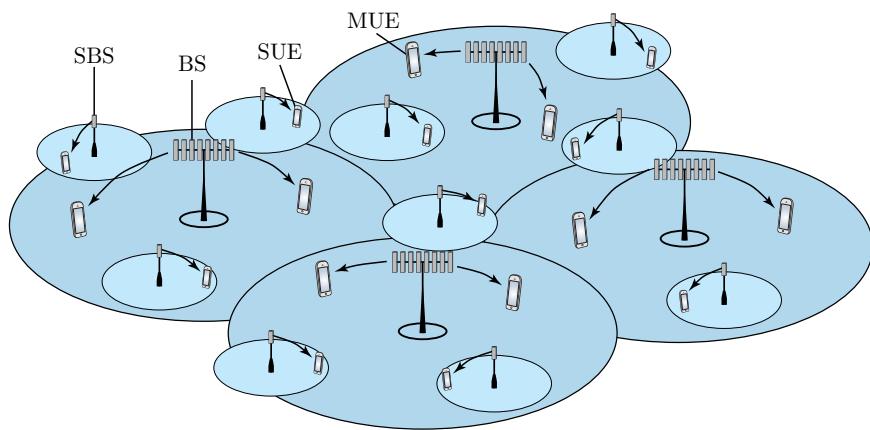


Figure 7.34: A heterogeneous network based on a two-tier deployment with Massive MIMO BSs in the coverage tier and SBSs in the hotspot tier. The BSs serve the MUEs while the SBSs serve the SUEs.

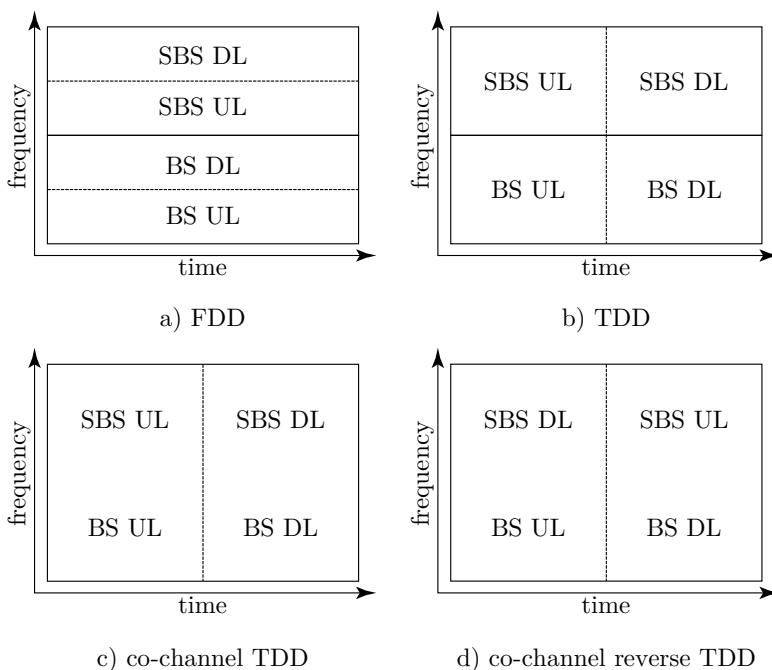


Figure 7.35: Operating principles of different duplexing schemes for heterogeneous networks.

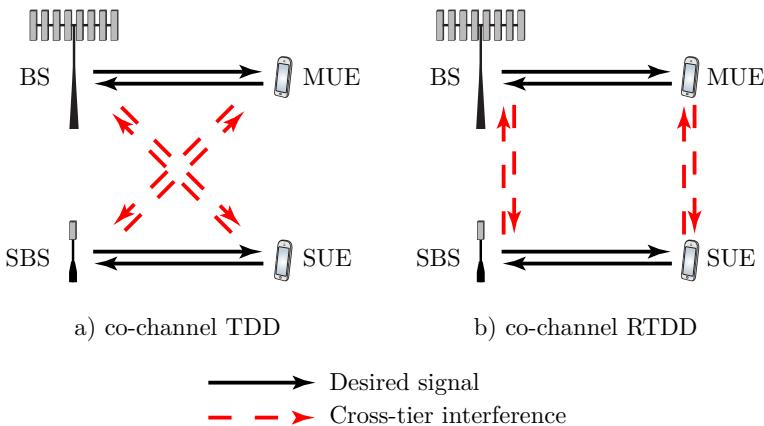


Figure 7.36: Interfering links in co-channel TDD and co-channel RTDD.

We refer to this as co-channel TDD. However, there are several other possible duplexing modes that could be employed to communicate over the available bandwidth: TDD, FDD, and co-channel reverse time-division duplex (RTDD). The operating principles of these different duplexing modes are illustrated in Figure 7.35. In both the FDD and TDD schemes, the BS and SBS tiers operate on non-overlapping frequency bands, while UL and DL transmissions are duplexed in either frequency (FDD) or time (TDD). Thus, transmissions do not interfere across the tiers. Unlike the aforementioned schemes, with co-channel TDD and co-channel RTDD, both tiers share the entire bandwidth. While the UL and DL transmissions are synchronized in both tiers with co-channel TDD, their order is reversed in one of the tiers with co-channel RTDD; that is, the BSs are in DL mode while the SBSs operate in UL mode, and vice versa. The duplexing mode determines which groups of devices interfere with each other. For example, in co-channel TDD, the SUEs interfere with the MUEs in the UL while the BSs interfere with the SBSs in the DL. This fact is shown in Figure 7.36.

A network-wide synchronized TDD protocol (i.e., co-channel TDD or co-channel RTDD) and the resulting channel reciprocity have two important advantages. First, as explained in Section 3.1 on p. 244, the BSs and SBSs can estimate the DL channels from UL pilots sent by the MUEs and SUEs. Second, they can estimate the correlation matrix of

the received signals, which is not only useful for signal detection but also for the design of interference-aware precoding/combining which does not require explicit knowledge of the interfering channels. To see this, recall the M-MMSE combining vector from (4.4):

$$\mathbf{v}_{jk} = p_{jk} \left(\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \left(\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \mathbf{C}_{li}^j \right) + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right)^{-1} \hat{\mathbf{h}}_{jk}^j. \quad (7.50)$$

The inverse matrix in this expression is $\mathbb{E} \left\{ \mathbf{y}_j \mathbf{y}_j^H | \{\hat{\mathbf{h}}_{li}^j\} \right\}$, which is actually an estimate of the conditional correlation matrix $\mathbb{E} \left\{ \mathbf{y}_j \mathbf{y}_j^H | \{\mathbf{h}_{li}^j\} \right\}$ of the received signal \mathbf{y}_j at BS j in (2.5). However, rather than computing this matrix based on the channel estimates of all UEs, which requires knowledge of the pilot sequences and spatial correlation matrices (cf. Section 3.1), it can be alternatively estimated from the sample correlation matrix $\hat{\mathbf{Q}}_j$, defined as

$$\hat{\mathbf{Q}}_j = \frac{1}{\tau_u} \sum_{n=1}^{\tau_u} \mathbf{y}_j[n] \mathbf{y}_j^H[n] \quad (7.51)$$

where $\mathbf{y}_j[n]$ denotes the n th sample of the received UL data signal at BS j and τ_u is the total number of UL data symbols within the current coherence block. The approximate receive combining vector $\hat{\mathbf{v}}_{jk}$ based on the sample correlation matrix is then given as

$$\hat{\mathbf{v}}_{jk} = p_{jk} \hat{\mathbf{Q}}_j^{-1} \hat{\mathbf{h}}_{jk}^j. \quad (7.52)$$

Note that BS j still needs to estimate the channels $\hat{\mathbf{h}}_{jk}^j$ of the served UEs. Moreover, the quality of $\hat{\mathbf{Q}}_j$ depends strongly on τ_u . We have discussed different aspects of correlation matrix estimation in Section 3.3.3 on p. 260. Thanks to the channel reciprocity, (7.52) can be also used as an approximate M-MMSE downlink precoder (cf. (4.37)).

Since the correlation matrix of the received UL signal completely describes the subspace from which the interference was received, the BSs and SBSs can simply precode their transmitted data such that less energy is radiated towards those directions. One can think about this as if every node sacrifices some of its degrees of freedom (or antennas) to reduce the interference it creates. This is an especially attractive

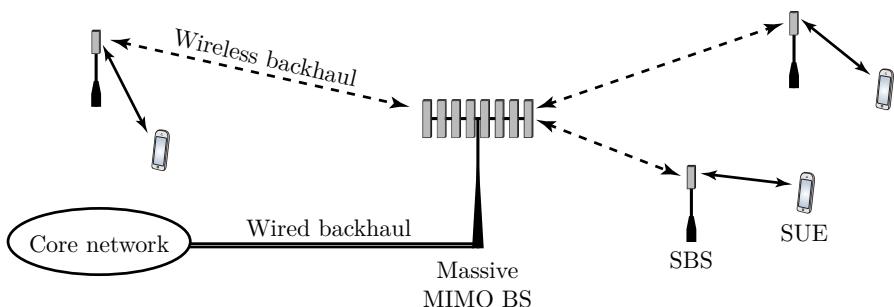


Figure 7.37: A Massive MIMO BS providing wireless backhaul to multiple SBSs.

option for BSs having a large number of excess antennas. Surprisingly, there is a huge gain in SE when also the SBSs are equipped with a few antennas and apply this precoding scheme. For more details and a critical discussion, we refer to [151, 51]. Another possibility to reduce cross-tier interference with the help of large antenna arrays is spatial (angular) blanking of certain cell areas [9, 361]. Similar to the concept of almost-blank subframes in time (as introduced in the LTE Release 10 [14]), this technique creates interference-free spatial zones for SC communications.

7.6.2 Massive MIMO for Wireless Backhaul

With a Massive MIMO infrastructure in place, it is possible to provide point-to-multipoint wireless backhaul from the BSs to a fraction of the SBSs on either the same or a different frequency band. An example is provided in Figure 7.37. In contrast to what is common practice, it would be advantageous to use cellular frequencies (≤ 6 GHz) for backhaul signaling and higher (mmWave) frequencies for the SBSs' data communications. This would have the benefit that LoS backhaul links are not necessary and the inter-cell interference between the SBSs is reduced due to high propagation losses in mmWave bands (cf. Section 7.5). Such a solution would have the following additional advantages:

- As neither standardization nor backward-compatibility is necessary for backhauling, manufacturers can use proprietary solutions and rapidly integrate technological innovations.

- The BS–SBS channels vary very slowly with time, due to the fixed deployment. Thus, complex (cooperative) transmission/detection schemes could be implemented which are prohibitive for BS–UE communications due to practical constraints on channel coherence, latency, and complexity. For example, multiple BSs could jointly provide wireless backhaul to a large number of SBSs in a network MIMO fashion. Also, the utilization of FDD bands might be feasible as the CSI feedback rate is smaller compared to what is necessary for BS–UE communication over fast-fading channels.
- Backhaul can be provided dynamically where it is needed; for example, the backhaul links to empty SBSs can be turned off such that more resources are available for highly loaded cells. This would avoid over-provisioning of backhaul capacity and could also allow for energy savings compared to traditional fiber-optical links whose energy consumption is independent of the traffic.
- With wireless backhaul links, the SBSs only require a power connection to be operational. Thus, they can be installed wherever and whenever needed with a minimum amount of manual configuration (e.g., antenna adjustment and wired power supply). This further reduces the capital and operational expenditures related to their deployment.

A relevant question in this context is how many antennas we would need to satisfy a certain backhaul rate (measured in bit/s) with a given transmit power budget. As the wireless backhaul channels can be seen as quasi-static (or slowly fading), it is not unreasonable to assume that CSI regarding the channels to all neighboring SBSs is available at the BSs and that these can coordinate their transmissions to some extent. Full sharing of UE data among the BSs might be infeasible as this would impose an extremely high traffic load on the wired backhaul network. Therefore, cooperative schemes relying on multicell CSI and only some additional data exchange are preferable.

Consider the canonical Massive MIMO system from Definition 2.1 on p. 217 and think of UEs as SBSs. We make the additional simplifying assumption that all BSs have M antennas and serve S SBSs each.

Moreover, we ignore channel estimation and assume that perfect CSI is available. We will exemplify a power minimization algorithm from [100] that allows us to fix a desired SINR target γ_{js} for each backhaul link in each cell and to find the precoding vectors \mathbf{w}_{js} and transmit powers ρ_{js} that achieve the minimum necessary total transmit power. In other words, our goal is to solve the optimization problem

$$\begin{aligned} & \underset{\{\rho_{js}, \mathbf{w}_{js}\}}{\text{minimize}} \quad \sum_{j=1}^L \sum_{s=1}^S \rho_{js} \\ & \text{subject to} \quad \text{SINR}_{js}^{\text{DL}} \geq \gamma_{js} \quad j = 1, \dots, L, s = 1, \dots, S \\ & \quad \|\mathbf{w}_{js}\| = 1 \quad j = 1, \dots, L, s = 1, \dots, S \end{aligned} \quad (7.53)$$

where $\{\rho_{js}, \mathbf{w}_{js}\}$ denotes the set of transmit powers and precoding vectors and

$$\text{SINR}_{js}^{\text{DL}} = \frac{\rho_{js} \left| \mathbf{w}_{js}^H \mathbf{h}_{js}^j \right|^2}{\sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,s)}}^S \rho_{li} \left| \mathbf{w}_{li}^H \mathbf{h}_{js}^l \right|^2 + \sigma_{\text{DL}}^2} \quad (7.54)$$

is the instantaneous SINR of SBS s in cell j . The solution to (7.53) is provided by the following theorem.

Theorem 7.3 ([100]). The solution to (7.53), if it exists, is given by ρ_{js}^* and $\mathbf{w}_{js}^* = \mathbf{v}_{js}^* / \|\mathbf{v}_{js}^*\|$ for $j = 1, \dots, L, s = 1, \dots, S$, where

$$\mathbf{v}_{js}^* = \left(\sum_{l=1}^L \sum_{i=1}^S \lambda_{li}^* \mathbf{h}_{li}^j \left(\mathbf{h}_{li}^j \right)^H + \mathbf{I}_M \right)^{-1} \mathbf{h}_{js}^j \quad (7.55)$$

with λ_{js}^* being the unique solutions to the set of fixed-point equations

$$\lambda_{js}^* = \frac{\left(1 + \frac{1}{\gamma_{js}} \right)^{-1}}{\left(\mathbf{h}_{js}^j \right)^H \left(\sum_{l=1}^L \sum_{i=1}^S \lambda_{li}^* \mathbf{h}_{li}^j \left(\mathbf{h}_{li}^j \right)^H + \mathbf{I}_M \right)^{-1} \mathbf{h}_{js}^j} \quad (7.56)$$

for $j = 1, \dots, L$ and $s = 1, \dots, S$, where ρ_{js}^* are the unique solutions to

the set of equations

$$\begin{aligned} \frac{\rho_{js}^*}{\gamma_{js}} \left| \left(\mathbf{w}_{js}^* \right)^H \mathbf{h}_{js}^j \right|^2 - \sum_{\substack{i=1 \\ i \neq s}}^S \rho_{ji}^* \left| \left(\mathbf{w}_{ji}^* \right)^H \mathbf{h}_{js}^j \right|^2 \\ - \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^S \rho_{li}^* \left| \left(\mathbf{w}_{li}^* \right)^H \mathbf{h}_{js}^l \right|^2 = \sigma_{\text{DL}}^2. \end{aligned} \quad (7.57)$$

The solution to (7.56) can be computed by a standard fixed-point algorithm which iteratively updates λ_{js}^* starting from some random initial values. Equation (7.57), can be written in a matrix form and solved through matrix inversion. Note, that if (7.53) is infeasible, no solution to (7.56) is found (i.e., the fixed-point algorithm does not converge).

To exemplify the backhauling, we consider an extension of the running example, described in Section 4.1.3 on p. 288, in which $S = 81$ SBSs are distributed on a regular grid within each cell. The channels from the BSs to the SBSs are modeled in the same way as the UE channels; that is, 20 MHz of bandwidth is used, uncorrelated Rayleigh fading is considered, and the shadow fading has $\sigma_{\text{sf}} = 10$. We first fix a maximum transmit power per BS and a desired DL SINR target for each backhaul link. Then using Theorem 7.3, we start with $M = 1$ and find the minimum transmit power necessary to achieve all SINR targets. If an SINR target is infeasible or the necessary transmit power is too high, we increase the number of antennas M until the problem becomes feasible and the transmit power is below the desired level. As the algorithm in (7.56) converges slowly for high SINR targets and it is computationally expensive for large systems, we resort to an asymptotic approximation (assuming very large values of M and S) of this algorithm instead [187, 285]. It only utilizes the average channel gains of all channels and its complexity is independent of the number of antennas. For medium- to large-sized systems (i.e., $M, S > 20$), the difference between the exact and approximate algorithms is generally very small.

In Figure 7.38, we show the minimum number of BS antennas necessary to provide a desired DL backhaul rate to either all, 40, or 20 randomly selected SBSs in each cell with a maximum power budget of

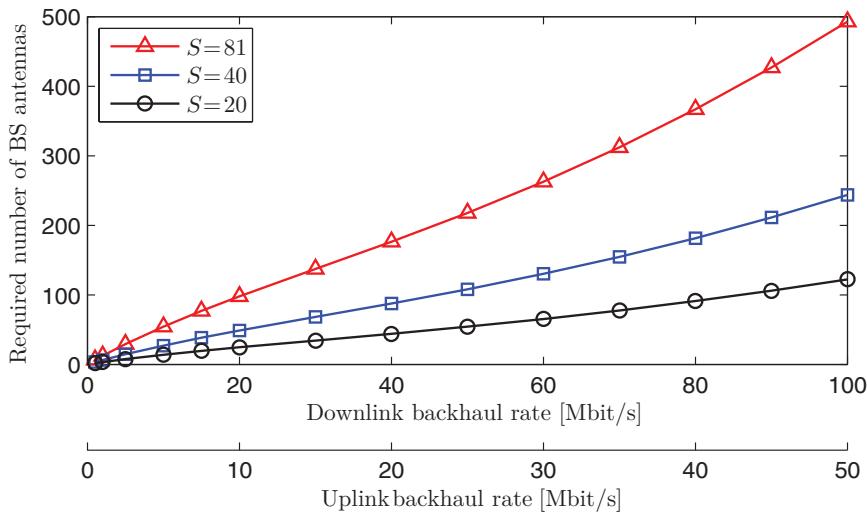


Figure 7.38: Required number of BS antennas M versus the DL/UL backhaul rates for different numbers of randomly selected SBSs $S \in \{20, 40, 81\}$ and a maximum average transmit power of 46 dBm per BS.

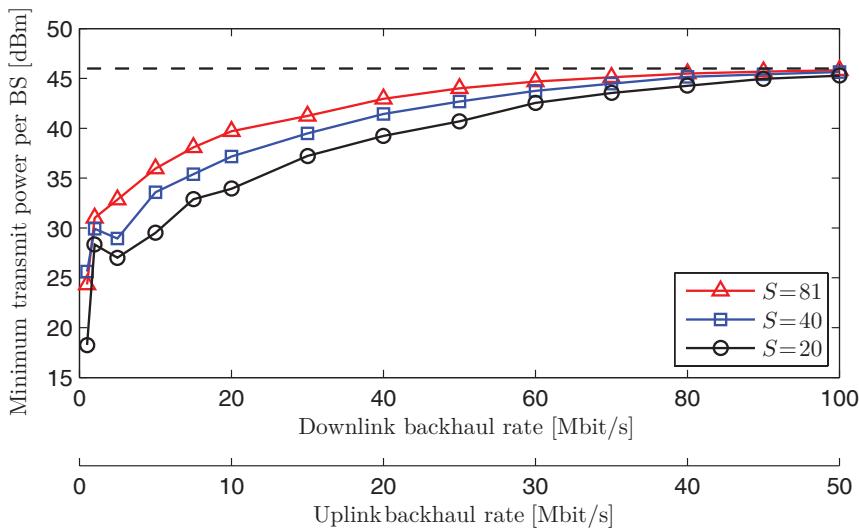


Figure 7.39: Minimum required transmit power per BS versus the DL/UL backhaul rates for different numbers of randomly selected SBSs $S \in \{20, 40, 81\}$. For each target rate, the smallest possible number of antennas is chosen according to Figure 7.38. The dashed line indicates the maximum transmit power of 46 dBm.

46 dBm per BS. Due to the classic UL-DL duality [370, 63, 335, 163], it is possible to achieve the same UL SINR using the precoding vectors \mathbf{w}_{js}^* as receive combining vectors (although the power allocation must be recomputed). By fixing the UL-DL transmission ratio to $\tau_u/\tau_d = 2/3$, the UL rates are 50% smaller than the DL rates. We remark that the UL-DL duality only ensures that the sum of the transmit powers of the BSs and SBSs are equal, but do not respect any individual power constraints. One can see from the figure that the number of antennas increases approximately linearly with the target rate and the number of simultaneously served SBSs. Serving 20 SBSs at a rate of 100 Mbit/s requires $M = 122$ antennas per BS. If the number of SBSs (and hence the aggregated rate) is doubled, 244 antennas are necessary, which becomes 493 if all 81 SBSs are simultaneously provided with backhaul. In the same context, Figure 7.39 shows the average transmit power per BS that is needed to achieve a certain backhaul rate using the smallest possible number of antennas provided in Figure 7.38. As one would expect, the smaller the number of SBSs (and hence the aggregate sum rate which must be delivered by each BS), the smaller is the necessary transmit power. For large target rates, the entire power budget is needed, independently of the number of SBSs. The curves are not entirely smooth since each point on the curves uses a different number of antennas.

In summary, Figures 7.38 and 7.39 provide some evidence that high-speed backhaul provisioning via Massive MIMO BSs is possible up to a very large number of SBSs per cell. In our example, in order to provide backhaul with 100 Mbit/s in the DL and 50 Mbit/s in the UL to 81 SBSs per cell using 20 MHz of bandwidth, which corresponds to an aggregate area throughput of 8.1 Gbit/s/km², around 500 antennas and 46 dBm of transmit power per BS are necessary.

7.7 Case Study

We conclude this monograph with a case study that will jointly analyze some of the practical deployment aspects and tradeoffs that have been described earlier in this section. The purpose is to give a baseline for the anticipated throughput of Massive MIMO, using MR or RZF pro-

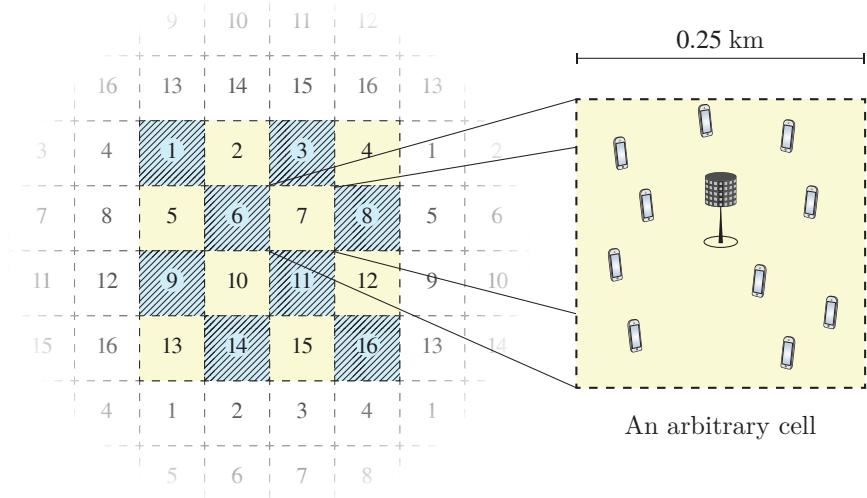


Figure 7.40: The case study considers 16 BSs on a square grid with a wrap-around topology. Note that, due to large-scale fading, the coverage area of each BS will not be the square area around the BS.

cessing, LS channel estimation, and some previously described resource allocation schemes. We stress that higher throughput can potentially be achieved by optimizing the array deployment (based on the propagation environment) and by exploiting MMSE channel estimation and M-MMSE processing.

7.7.1 Network Configuration and Parameters

We consider the network setup illustrated in Figure 7.40. Similar to the running example, which was described in Section 4.1.3 on p. 288, the case study considers $L = 16$ cells that are distributed on a 4×4 square grid with 250 m inter-BS distance. The channels are generated based on the 3GPP 3D UMi NLoS channel model from [1], using the open-source QuaDRiGa implementation developed by the Fraunhofer Heinrich Hertz Institute [159].²⁷ Within the 250 m \times 250 m geographical area closest to each BS, we distribute 10 UEs uniformly at random (at

²⁷The simulation results were generated using QuaDRiGa version 1.4.8-571.

Parameter	Value
Network layout	Square pattern (wrap-around)
Number of cells	$L = 16$
Inter-BS distance	250 m
UE dropping	$K = 10$ UEs in 250 m \times 250 m area around each BS, with 35 m minimum distance
Channel model	3GPP 3D urban microcell (UMi)
BS array configurations	Cylindrical arrays: $10 \times 5 \times 2$ ($M = 100$), $20 \times 5 \times 1$ ($M = 100$), $20 \times 5 \times 2$ ($M = 200$)
BS height	25 m
UE height	1.5 m
Carrier frequency	2 GHz
Bandwidth	$B = 20$ MHz
Number of subcarriers	2000
Subcarrier bandwidth	10 kHz
Maximum UE transmit power	20 dBm
Maximum BS transmit power	30 dBm
Receiver noise power	-94 dBm
Cyclic prefix overhead	5%
Frame dimensions	$B_c = 50$ kHz, $T_c = 4$ ms
Subcarriers per frame	5
Useful samples per frame	$\tau_c = B_c T_c / 1.05 \approx 190$
Pilot reuse factor	$f = 2$
Number of pilot sequences	$\tau_p = 30$
Channel estimation	LS
Combining and precoding	RZF or MR

Table 7.2: Network parameters in the case study.

distances larger than 35 m from the BS). The UEs are at outdoor NLoS locations, 1.5 m above the ground. Each UE is associated with the BS that provides the strongest average channel gain, which results in an

uneven number of UEs per cell due to the shadow fading characteristics of the channel model. Note that one could, potentially, obtain better performance by also taking the spatial channel correlation into account in the UE-BS association.

Each BS is deployed at an elevated location, 25 m above the ground. We consider a cylindrical array that covers an entire cell (i.e., 360°), without the need for cell sectorization. The BS operates at a 2 GHz carrier frequency and we assume an antenna spacing of $\lambda/2$ in the horizontal and vertical direction. We describe the antenna configuration as “horizontal \times vertical \times polarization”, which refers to the number of antennas on each circle, number of circles in the vertical direction, and number of polarizations (see Remark 7.6 on p. 515), respectively. We consider three configurations: $10 \times 5 \times 2$, $20 \times 5 \times 1$, and $20 \times 5 \times 2$. The first and second configurations require $M = 100$ RF chains, while the third one requires $M = 200$ RF chains. The array is 37.5 cm high in all cases, while the diameter is 23.9 cm in the first configuration and 47.7 cm in second and third configurations. Note that by considering dual-polarized co-located antennas, we can effectively double the number of antennas (and RF chains) compared to having uni-polarized antennas, without increasing the array size; see Section 7.4 for further details.

We consider an OFDM system with a bandwidth of $B = 20$ MHz, which is divided into 2000 subcarriers that are each 10 kHz wide. The transmission protocol is based on dividing the time-frequency resources into frames of $B_c = 50$ kHz and $T_c = 4$ ms. All UEs are assumed to have channel coherence blocks that are equal or larger than the frame dimension. Hence, the channel responses are fixed and identical over the five subcarriers in a frame. The cyclic prefix (which combats inter-symbol interference) is assumed to increase the OFDM symbol duration by 5%, thus there are $\tau_c = B_c T_c / 1.05 \approx 190$ useful samples per frame. In each frame, $\tau_p = 30$ samples are used for pilots. We consider a pilot reuse factor of $f = 2$, which results in $\tau_p/f = 15$ pilots per BS. These pilots are distributed uniformly at random to the UEs in the cell, and some remain unused when there are fewer than 15 UEs in the cell. In the unlikely event that more than 15 UEs connect to a BS, a random subset of exactly 15 UEs is scheduled. The remaining 160 samples per

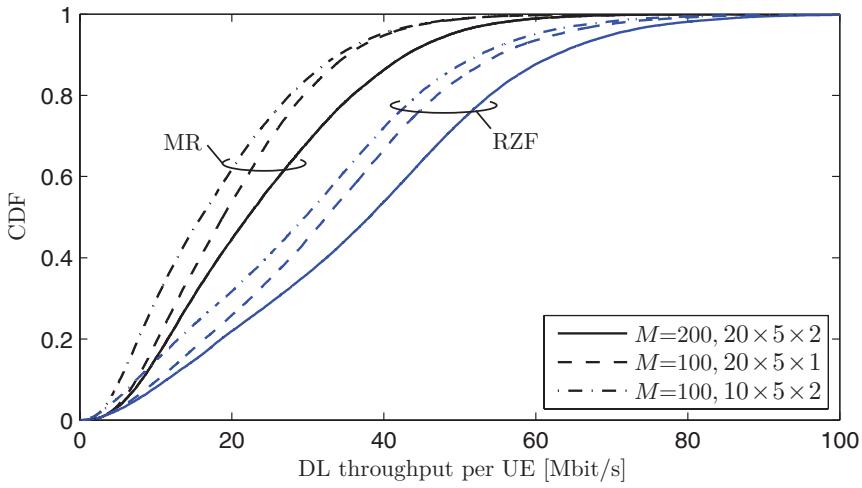
frame are used for payload data, whereof 1/3 is used for UL and 2/3 is used for DL. All active UEs are scheduled over all subcarriers and the maximum UL transmit power is 20 dBm. The maximum DL transmit power per BS is 30 dBm. Note that both numbers are at least an order of magnitude smaller than in current LTE systems (cf. Remark 4.1 on p. 291), but we will anyway achieve higher throughput in Massive MIMO thanks to the array gain and spatial multiplexing. The receiver noise power is -94 dBm.

To showcase the baseline throughput that can be achieved without requiring knowledge of channel statistics, we consider LS channel estimation. Since M-MMSE combining/precoding does not work well under LS estimation (cf. Figures 4.14 and 4.20), we only consider MR and RZF combining/precoding in this case study.

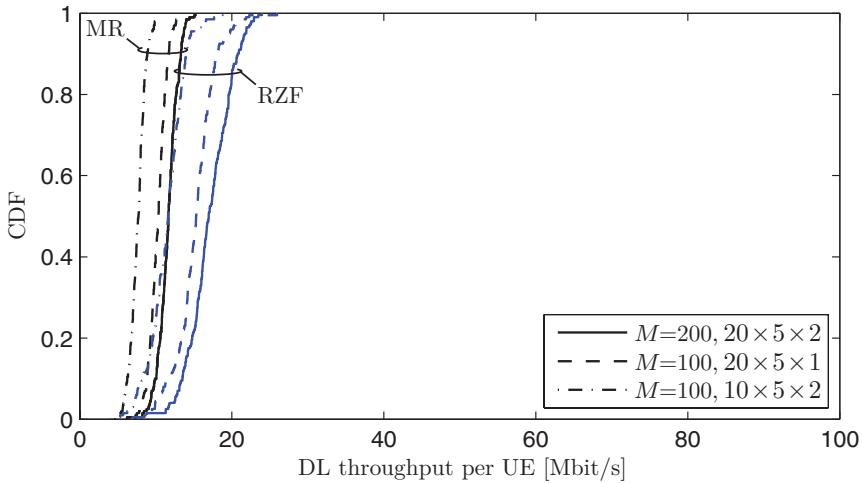
7.7.2 Simulation Results

We consider the DL and UL throughput of the UEs with different power allocation/control schemes and for different random user drops. Figure 7.41a and Figure 7.41b show the CDFs of the DL throughput with max product SINR and max-min fairness power allocation, respectively.²⁸ These power allocations are obtained using the optimization algorithms described in Section 7.1.1. The first observation is that the choice of antenna configuration has a great impact on the throughput. The dual-polarized configuration $20 \times 5 \times 2$ with $M = 200$ provides higher throughput, for every UE, than the uni-polarized configuration $20 \times 5 \times 1$ with $M = 100$. The uni-polarized configuration is, however, preferable compared with the dual-polarized configuration $10 \times 5 \times 2$, which also has $M = 100$. The conclusion is that, for arrays with a fixed physical size, dual-polarized arrays are beneficial due to the higher array gain that is obtained by having twice the number of antennas. If instead the number of RF chains is the limiting factor in the implementation, a physically larger array with uni-polarization (or dual-polarization but not co-located antennas) is preferable due to the higher spatial resolution.

²⁸Equal maximum UL power is assumed for the pilot transmission when evaluated the DL.



(a) Max product SINR power allocation.



(b) Max-min fairness power allocation.

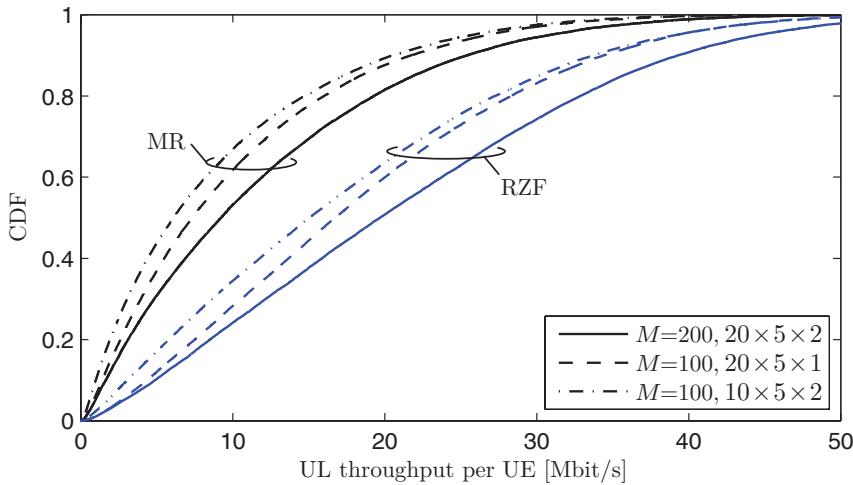
Figure 7.41: CDF of the DL throughput per UE in the case study. We compare different array configurations, transmit precoding schemes, and power allocation schemes.

Scheme	95% likely	Median	5% likely
Max product SINR (MR)	6.0 Mbit/s	22.1 Mbit/s	48.7 Mbit/s
Max product SINR (RZF)	7.5 Mbit/s	38.1 Mbit/s	70.1 Mbit/s
Max-min fairness (MR)	9.3 Mbit/s	11.7 Mbit/s	13.7 Mbit/s
Max-min fairness (RZF)	12.6 Mbit/s	17.0 Mbit/s	21.7 Mbit/s

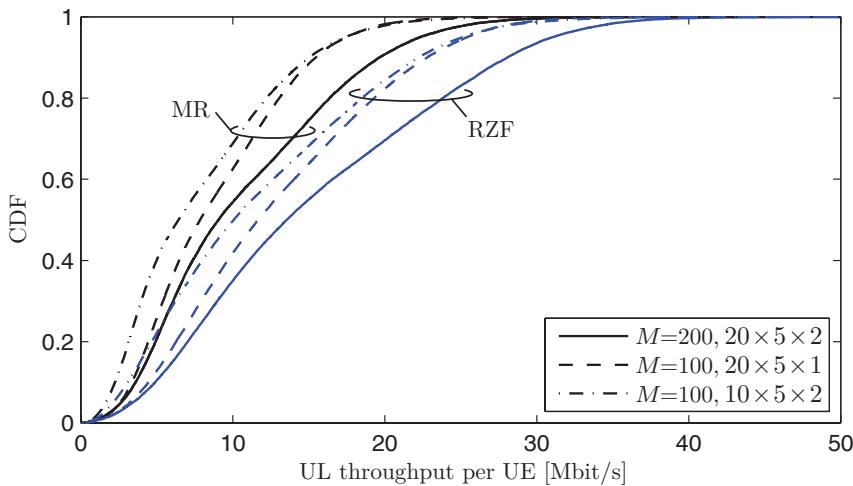
Table 7.3: DL throughput per UE in the case study, with different power allocation and precoding schemes. The $20 \times 5 \times 2$ antenna configuration is considered.

We now concentrate on the preferable $20 \times 5 \times 2$ antenna configuration and compare the different power allocation and precoding schemes. With max product SINR power allocation, the average UE throughput is 24.1 Mbit/s with MR and 37.7 Mbit/s with RZF, but there are substantial variations between the UEs. With max-min fairness power allocation, the average throughput is reduced to 11.6 Mbit/s with MR and to 17.1 Mbit/s with RZF, but there is greater fairness in the sense of smaller variations and higher throughput for the UEs with the weakest channel conditions. These observations are further provided in Table 7.3, which shows the 95% likely, median, and 5% likely UE throughput. The throughput that is guaranteed to 95% of the UEs is around 70% higher with max-min fairness, compared with max product SINR. However, the median is 85%–105% higher with max product SINR power allocation, and the UEs with the 5% best channels achieve even higher gains with this scheme.

The CDF of the UL throughput is shown in Figure 7.42. We consider the power control framework described in Section 7.1.2, where the largest received power ratio within a cell is limited to Δ . We consider $\Delta = 20$ dB in Figure 7.42a and $\Delta = 0$ dB in Figure 7.42b. The UL results confirm many of the observations made in the DL. For a given number of RF chains, it is beneficial to have a physically larger uni-polarized array. For a given physical array size, it is instead beneficial to use dual polarization to squeeze in more antennas. There are substantial variations in throughput between the UEs, particularly with $\Delta = 20$ dB, but also with $\Delta = 0$ dB since the heuristic UL power control is performed independently in every cell—in contrast to the DL power allocation,



(a) Heuristic power control policy in (7.11) with $\Delta = 20$ dB.



(b) Heuristic power control policy in (7.11) with $\Delta = 0$ dB.

Figure 7.42: CDF of the UL throughput per UE in the case study. We compare different array configurations, receive combining schemes, and power control schemes.

Scheme	95% likely	Median	5% likely
$\Delta = 20 \text{ dB (MR)}$	0.9 Mbit/s	9.2 Mbit/s	30.8 Mbit/s
$\Delta = 20 \text{ dB (RZF)}$	2.8 Mbit/s	19.7 Mbit/s	44.6 Mbit/s
$\Delta = 0 \text{ dB (MR)}$	2.7 Mbit/s	9.1 Mbit/s	22.3 Mbit/s
$\Delta = 0 \text{ dB (RZF)}$	3.4 Mbit/s	13.6 Mbit/s	31.0 Mbit/s

Table 7.4: UL throughput per UE in the case study, with different combining schemes and different values of the maximum received power ratio Δ that is used for power control. The $20 \times 5 \times 2$ antenna configuration is considered.

where max-min fairness power allocation attempts to give the same throughput to every UE in the entire network.

Focusing on the preferable $20 \times 5 \times 2$ antenna configuration and $\Delta = 20 \text{ dB}$, the average UE throughput is 11.6 Mbit/s with MR and 21.1 Mbit/s with RZF. In case of $\Delta = 0 \text{ dB}$, the average UE throughput is reduced to 10.6 Mbit/s with MR and 15.3 Mbit/s with RZF, but the 95% likely throughput is higher than with $\Delta = 20 \text{ dB}$. The 95% likely, median, and 5% likely UE throughput are provided in Table 7.4. A smaller Δ improves the baseline throughput that is guaranteed to 95% of the UEs, at the cost of reducing the median throughput and also the throughput of the UEs with the strongest channels. Generally speaking, the throughput differences are larger in the UL than in the DL due to the power control scheme.

In summary, the average DL sum throughput per cell can be as large as 373 Mbit/s over a 20 MHz channel, which corresponds to an area throughput of 6.0 Gbit/s/km². The average UL sum throughput per cell can be as large as 209 Mbit/s, which corresponds to an area throughput of 3.3 Gbit/s/km². Note that the substantial differences between the throughput in DL and UL are caused by the fact that twice as many samples per frame are used for DL data transmission as for UL data transmission — the average SE is approximately the same in both directions. To put the area throughput values into context, we can compare them with a contemporary LTE system, as described in Remark 4.1 on p. 291. Such a system would deliver 263 Mbit/s/km² in the DL and 115 Mbit/s/km² in the UL in a corresponding scenario. We conclude that, in this case study, Massive MIMO delivers 20–30 times higher throughput.

7.8 Summary of Key Points in Section 7

- The UEs in a cellular network have conflicting performance goals due to the interference and shared power budgets. Power allocation can be used to find a tradeoff between aggregate throughput and fairness between the UEs. Maximizing the product of the effective SINRs leads to a reasonable tradeoff. Thanks to channel hardening, the SINRs only depend on large-scale fading and the same solution can be used over many channel coherence blocks.
- Despite the large number of UEs, resource allocation is fairly simple since the UEs are separated spatially and, thus, every UE can use the full bandwidth whenever needed. When Massive MIMO is deployed with a high antenna-UE ratio, time-frequency scheduling is only needed to cope with traffic peaks, when there are more UEs than pilots or insufficient spatial resolution. This is a paradigm shift from conventional networks, which constantly rely on time-frequency scheduling to serve the UEs. The assignment of pilots to UEs and traffic load variations have little impact on the average sum SE, but can affect the SE of particular UEs.
- The spatial channel correlation depends mainly on the antenna array geometry (aperture, antenna spacing) and the angular spread, and must be accounted for by any realistic channel model. Channel measurements have confirmed favorable propagation in practice.
- The angular resolution of an antenna array depends on its aperture and not on the number of antennas. Critically spaced antennas are needed to avoid aliasing in the angular domain. Dual-polarized co-located antennas can significantly reduce the size of an array, but the theoretical performance analysis with polarization is difficult. Modern geometry-based

channel models support polarization and can be used in simulations.

- MmWave frequencies suffer from high propagation loss—that can be overcome by large antenna arrays—and are best suited for hotspots and low mobility scenarios. Due to a small number of RF chains compared to the number of radiating elements and a small number of UEs per cell, mmWave communication systems differ fundamentally from the definition of Massive MIMO in Definition 2.1 on p. 217.
- Massive MIMO plays a key role in heterogeneous networks to ensure coverage over large areas and to serve fast-moving UEs. A two-tier network consisting of Massive MIMO BSs and SBSs together with a synchronized TDD protocol allows the BSs to use their excess antennas to reduce intra-tier interference. Massive MIMO is also a promising solution for wireless backhaul provisioning to a large number of SBSs, without the need for LoS links.
- We have provided a case study based on state-of-the-art channel models, optimized resource allocation, and conservative assumptions on the selection of signal processing schemes. It demonstrates that Massive MIMO can deliver tens of Mbit/s per UE in both UL and DL over a channel of 20 MHz bandwidth. With around 10 UEs per cell, this results in an area throughput of several Gbit/s/km².

Acknowledgements

We would like to thank the editor Robert W. Heath Jr. for organizing the review of this monograph and the anonymous reviewers for their constructive and detailed comments. We are grateful for the feedback provided by our proof-readers Alessio Zappone (University of Cassino and Southern Lazio), Maximilian Arnold (University of Stuttgart), Andrea Pizzo (University of Pisa), Daniel Verenzuela, Hei Victor Cheng, Giovanni Interdonato, Marcus Karlsson, Antzela Kosta, Özgecan Özdogan (Linköping university), and Zahid Aslam (Siradel).

Emil Björnson has been supported by ELLIIT, CENIIT, and the Swedish Foundation for Strategic Research.

Luca Sanguinetti has been supported by the ERC Starting Grant 305123 MORE.

Appendices

A

Notation and Abbreviations

Mathematical Notation

Upper-case boldface letters are used to denote matrices (e.g., \mathbf{X}, \mathbf{Y}), while column vectors are denoted with lower-case boldface letters (e.g., \mathbf{x}, \mathbf{y}). Scalars are denoted by lower/upper-case italic letters (e.g., x, y, X, Y) and sets by calligraphic letters (e.g., \mathcal{X}, \mathcal{Y}).

The following mathematical notations are used:

$\mathbb{C}^{N \times M}$	The set of complex-valued $N \times M$ matrices
$\mathbb{R}^{N \times M}$	The set of real-valued $N \times M$ matrices
$\mathbb{C}^N, \mathbb{R}^N$	Short forms of $\mathbb{C}^{N \times 1}$ and $\mathbb{R}^{N \times 1}$ for vectors
\mathbb{R}_+^N	The set of non-negative members of \mathbb{R}^N
$x \in \mathcal{S}$	x is a member of the set \mathcal{S}
$x \notin \mathcal{S}$	x is not a member of the set \mathcal{S}
$\{x \in \mathcal{S} : P\}$	The subset of \mathcal{S} containing all members that satisfy a property P
$a_n \asymp b_n$	This denotes $a_n - b_n \rightarrow_{n \rightarrow \infty} 0$ almost surely, for two infinite sequences of random variables a_n, b_n
$[\mathbf{x}]_i$	The i th element of a vector \mathbf{x}
$[\mathbf{X}]_{ij}$	The (i, j) th element of a matrix \mathbf{X}

$\text{diag}(\cdot)$	$\text{diag}(x_1, \dots, x_N)$ is a diagonal matrix with the scalars x_1, \dots, x_N on the diagonal,
$\text{diag}(\mathbf{X}_1, \dots, \mathbf{X}_N)$	a block diagonal matrix with the matrices $\mathbf{X}_1, \dots, \mathbf{X}_N$ on the diagonal
\mathbf{X}^*	The complex conjugate of \mathbf{X}
\mathbf{X}^t	The transpose of \mathbf{X}
\mathbf{X}^h	The conjugate transpose of \mathbf{X}
\mathbf{X}^{-1}	The inverse of a square matrix \mathbf{X}
$\mathbf{X}^{\frac{1}{2}}$	The square-root of a square matrix \mathbf{X}
$\Re(x)$	Real part of x
$\Im(x)$	Imaginary part of x
j	The imaginary unit
$ x $	Absolute value (or magnitude) of a scalar variable x
$\lfloor x \rfloor$	Closest integer smaller or equal to x
$\lceil x \rceil$	Closest integer greater or equal to x
e	Euler's number ($e \approx 2.718281828$)
$\max(x, y)$	The maximum of x and y
$\text{mod}(x, y)$	The modulo operation, i.e., the remainder of the Euclidean division of x by y
$\log_a(x)$	Logarithm of x using the base $a \in \mathbb{R}_+$
$E_1(x)$	The exponential integral function, defined as $E_1(x) = \int_1^\infty \frac{e^{-xu}}{u} du$
$\sin(x)$	The sine function of x
$\cos(x)$	The cosine function of x
$\tan^{-1}(x)$	The inverse tangent function of x , also known as $\arctan(x)$
$W(x)$	The Lambert function, see Definition B.2 on p. 567
$x!$	The factorial function for positive integers x , defined as $x! = x(x-1) \dots \cdot 1$
$\text{tr}(\mathbf{X})$	Trace of a square matrix \mathbf{X}
$\det(\mathbf{X})$	Determinant of a square matrix \mathbf{X}
$\mathbf{X} \odot \mathbf{Y}$	Hadamard (elementwise) product of \mathbf{X}, \mathbf{Y}
$\mathbf{X} \otimes \mathbf{Y}$	Kronecker product of \mathbf{X}, \mathbf{Y}
$\text{rank}(\mathbf{X})$	Rank of \mathbf{X} (i.e., number of non-zero singular values)
$\mathcal{N}(\mathbf{x}, \mathbf{R})$	The real Gaussian distribution

	with mean \mathbf{x} and covariance matrix \mathbf{R}
$\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$	The circularly symmetric complex Gaussian distribution with zero mean and correlation matrix \mathbf{R} , where circular symmetry means that if $\mathbf{y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$ then $e^{j\phi}\mathbf{y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$ for any given ϕ
$\mathcal{N}_{\mathbb{C}}(\mathbf{x}, \mathbf{R})$	The generalization of the circularly symmetric complex Gaussian distribution to a non-zero mean \mathbf{x} , where circular symmetry holds if the mean is subtracted
$\text{Exp}(x)$	Exponential distribution with rate $x > 0$
$\chi^2(x)$	Chi-squared distribution with x degrees of freedom
$\text{Lap}(\mu, \sigma)$	Laplace distribution with mean μ and standard deviation σ , with PDF $f(x) = \frac{1}{\sqrt{2\sigma}} e^{-\frac{\sqrt{2} x-\mu }{\sigma}}$
$U[a, b]$	Uniform distribution between a and b
$\text{Po}(\lambda)$	Poisson distribution with mean value and variance λ
$\mathbb{E}\{x\}$	The expectation of a random variable x
$\mathbb{V}\{x\}$	The variance of a random variable x
$\ \mathbf{x}\ $	The L_2 -norm $\ \mathbf{x}\ = \sqrt{\sum_i \ [\mathbf{x}]_i\ ^2}$ of a vector \mathbf{x}
$\ \mathbf{X}\ _F$	The Frobenius norm $\ \mathbf{X}\ _F = \sqrt{\sum_{i,j} \ [\mathbf{X}]_{ij}\ ^2}$ of \mathbf{X}
$\ \mathbf{X}\ _2$	The spectral norm of \mathbf{X}
	(i.e., the largest singular value)
$\text{eig}_p(\mathbf{X})$	The p -dominant eigenspace of a Hermitian matrix \mathbf{X} , see Definition 7.1 on p. 490
\mathbf{I}_M	The $M \times M$ identity matrix
$\mathbf{1}_N$	The $N \times 1$ matrix (i.e., vector) with only ones
$\mathbf{1}_{N \times M}$	The $N \times M$ matrix with only ones
$\mathbf{0}_M$	The $M \times 1$ matrix (i.e., vector) with only zeros
$\mathbf{0}_{N \times M}$	The $N \times M$ matrix with only zeros

System Model Notation

The following symbols are commonly used in the system model of this monograph:

$\mathbf{a}(\varphi, \theta)$	Spatial signature of an antenna array
\mathbf{A}_{li}^j	Matrix used as $\mathbf{A}_{li}^j \mathbf{y}_{jli}^p$ in an arbitrary linear estimator of \mathbf{h}_{li}^j

α	Pathloss exponent in the large-scale fading model of (2.3)
B	Total bandwidth used for communication [Hz]
B_c	Channel coherence bandwidth [Hz]
β_{li}^j	Average channel gain $\text{tr}(\mathbf{R}_{li}^j)/M_j$ per antenna of the channel \mathbf{h}_{li}^j between BS j and UE i in cell l
\mathbf{C}_{li}^j	Estimation error correlation matrix for channel between BS j and UE i in cell l
D	Average cell density [cells/km ²]
d_H, d_V	Horizontal and vertical antenna spacing of a uniform planar array measured in multiples of the wavelength λ
d_{li}^j	Distance in km between BS j and UE i in cell l
Δ	Maximum received power difference in the power control policy in (7.11)
f_c	Carrier frequency [Hz]
g_{jk}	Precoded channel $(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}$ to UE k in cell j from its BS
\mathbf{h}_{li}^j	Channel response between BS j and UE i in cell l
$\hat{\mathbf{h}}_{li}^j$	Estimate of the channel \mathbf{h}_{li}^j between BS j and UE i in cell l
$\tilde{\mathbf{h}}_{li}^j$	Estimation error defined as $\tilde{\mathbf{h}}_{li}^j = \mathbf{h}_{li}^j - \hat{\mathbf{h}}_{li}^j$
$\hat{\mathbf{H}}_l^j$	Matrix defined in (3.12) with the channel responses between BS j and all UEs in cell l
j, l, l'	Used as cell indices
k, i, i'	Used as UE indices
$\mathbf{k}(\varphi, \theta)$	Wave vector of a planar wave
K	Number of UEs per cell (if it is the same in all cells)
K_j	Number of UEs in cell j
κ_t^{BS}	Transmitter hardware quality of a BS
κ_r^{BS}	Receiver hardware quality of a BS
κ_t^{UE}	Transmitter hardware quality of a UE
κ_r^{UE}	Receiver hardware quality of a UE
λ	Wavelength [m]
L	Number of cells
L_{BS}	Computational efficiency of the BS (in flops/Watt)
M	Number of BS antennas (if it is the same in all cells)
M_j	Number of BS antennas in cell j
\mathcal{P}_{jk}	Set of UEs using the same pilot as UE k in cell j

P_{\max}^{DL}	Maximum DL transmit power per BS
P_{\max}^{UL}	Maximum UL transmit power per UE
p_{jk}	UL transmit power used by UE k in cell j
ρ_{jk}	DL transmit power allocated to UE k in cell j
φ	Azimuth angle
ϕ_{jk}	Pilot sequence associated with UE k in cell j
Φ	Pilot book with τ_p mutually orthogonal sequences
Ψ_{li}^j	Inverse of the correlation matrix in the estimation of the channel between BS j and UE i in cell l ; Defined in (3.10) with ideal hardware and in (6.24) with hardware impairments
P_{BS}	Power per BS antennas for transceiver hardware
P_{BT}	Power per (bit/s) for backhaul traffic
P_{COD}	Power per (bit/s) for data encoding
P_{DEC}	Power per (bit/s) for data decoding
P_{FIX}	Fixed power of a BS, which is independent of traffic load, number of BS antennas, and number of UEs
P_{LO}	Power per LO
P_{UE}	Power per UE for transceiver hardware
\mathbf{R}_{li}^j	Correlation matrix of the channel between BS j and UE i in cell l
SNR_{jk}^p	Effective SNR in (3.13) in the pilot transmission of UE k in cell j
σ_{DL}^2	Noise variance in the DL
σ_{UL}^2	Noise variance in the UL
σ_φ	ASD in the local scattering model, see Section 2.6 on p. 235
σ_{sf}	Standard deviation of the shadow fading in the large-scale fading model defined in (2.3)
θ	Elevation angle
T_c	Channel coherence time [s]
T_d	Delay spread [s]
τ_c	Number of samples per coherence block (equals $B_c T_c$)
τ_d	DL data samples per coherence block
τ_p	Number of samples allocated for pilots per coherence block
τ_u	UL data samples per coherence block

$U(\cdot)$	Utility function in power allocation optimization
Υ	Median channel gain at a reference distance of 1 km in the large-scale fading model of (2.3)
\mathbf{v}_{jk}	Receive combining vector for UE k in cell j
\mathbf{w}_{jk}	Transmit precoding vector for UE k in cell j
\mathbf{y}_{jli}^p	Processed received pilot signal, defined in (3.2)

Abbreviations

The following acronyms and abbreviations are used in this monograph:

3GPP	3rd Generation Partnership Project
ACLR	Adjacent-Channel Leakage Ratio
ADC	Analog-to-Digital Converter
AoA	Angle of Arrival
AoD	Angle of Departure
ASD	Angular Standard Deviation
ATP	Area Transmit Power
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
BPSK	Binary Phase-Shift Keying
BS	Base Station
CDF	Cumulative distribution function
CDMA	Code-Division Multiple Access
CoMP	Coordinated Multipoint
CP	Circuit Power
CSI	Channel State Information
DAB	Digital Audio Broadcast
DAC	Digital-to-Analog Converter
DFT	Discrete Fourier Transform
DL	Downlink
EE	Energy Efficiency
EM	Electromagnetic
ETP	Effective Transmit Power
EVM	Error Vector Magnitude
EW-MMSE	Element-Wise MMSE

FBMC	Filter Bank Multi-Carrier
FDD	Frequency-Division Duplex
GSM	Global System for Mobile Communications
HE	Hardware Efficiency
i.i.d.	Independent and Identically Distributed
I/Q	In-Phase/Quadrature
IEEE	Institute of Electrical and Electronics Engineers
JSDM	Joint Spatial-Division and Multiplexing
LMMSE	Linear MMSE
LO	Local Oscillator
LoS	Line-of-Sight
LS	Least-Squares
LTE	Long Term Evolution
MISO	Multiple-Input Single-Output
M-MMSE	Multicell Minimum Mean-Squared Error
MIMO	Multiple-Input Multiple-Output
MMSE	Minimum Mean-Squared Error
mmWave	Millimeter Wavelength
MR	Maximum Ratio
MSE	Mean-Squared Error
MUE	Macro User Equipment
NLoS	Non-Line-of-Sight
NMSE	Normalized MSE
OFDM	Orthogonal Frequency-Division Multiplexing
PA	Power Amplifier
PC	Power Consumption
PDF	Probability Density Function
PSK	Phase-Shift Keying
QAM	Quadrature Amplitude Modulation
RA	Random Access
RACH	Random Access Channel
RF	Radio Frequency
RTDD	Reverse Time-Division Duplex
RZF	Regularized Zero-Forcing

SBS	Small-Cell Base Station
S-MMSE	Single-Cell Minimum Mean-Squared Error
SC	Small Cell
SDMA	Space-Division Multiple Access
SE	Spectral Efficiency
SIMO	Single-Input Multiple-Output
SINR	Signal-to-Interference-and-Noise Ratio
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
SUE	Small-Cell User Equipment
TDD	Time-Division Duplex
UatF	Use-and-then-Forget
UE	User Equipment
UL	Uplink
ULA	Uniform Linear Array
UMi	Urban Microcell
UMTS	Universal Mobile Telecommunications System
WLAN	Wireless Local Area Network
XPD	Cross-Polar Discrimination
XPI	Cross-Polar Isolation
ZF	Zero-Forcing

B

Standard Results

This appendix provides a brief overview of some theoretical results and methodologies that lay the foundation on which this monograph rests. This includes matrix analysis, random vectors, estimation theory, information theory, and optimization.

B.1 Matrix Analysis

B.1.1 Computational Complexity of Matrix Operations

Basic linear algebra operations, such as matrix-matrix multiplications and matrix inversions, have a well-defined structure and can thus be implemented efficiently in hardware. Nevertheless, the computational complexity can be a bottleneck when large matrices need to be manipulated every millisecond. The exact complexity of a matrix operation depends strongly on the hardware implementation, including the bit width (i.e., the number of binary digits used to represent a number) and the data type (e.g., floating point or fixed point). In this section, we provide first-order approximations by counting the number of complex multiplications and divisions that are needed, while the complexity of additions/subtractions is neglected since these operations are much

easier to implement in hardware.

Lemma B.1. Consider the matrices $\mathbf{A} \in \mathbb{C}^{N_1 \times N_2}$ and $\mathbf{B} \in \mathbb{C}^{N_2 \times N_3}$. The matrix-matrix multiplication \mathbf{AB} requires $N_1 N_2 N_3$ complex multiplications. The multiplication \mathbf{AA}^H only requires $\frac{N_1^2 + N_1}{2} N_2$ complex multiplications, by utilizing the Hermitian symmetry.

Proof. There are $N_1 N_3$ elements in \mathbf{AB} and the computation of each element involves N_2 multiplications (the elements of one row in \mathbf{A} are multiplied by the corresponding elements of one column in \mathbf{B}). In the case of $\mathbf{B} = \mathbf{A}^H$, the Hermitian symmetry is utilized to only compute $\frac{N_1^2 + N_1}{2}$ elements, which represent the main diagonal and half of the off-diagonal elements. \square

When the inverse of a matrix is multiplied by another matrix, the \mathbf{LDL}^H decomposition can be used to achieve an efficient hardware implementation, both in terms of computations and memory usage [183]. The matrix \mathbf{L} is a lower triangular matrix with ones on the main diagonal and \mathbf{D} is a diagonal matrix.

Lemma B.2. Consider the Hermitian positive semi-definite matrix $\mathbf{A} \in \mathbb{C}^{N_1 \times N_1}$ and the matrix $\mathbf{B} \in \mathbb{C}^{N_1 \times N_2}$. The \mathbf{LDL}^H decomposition of \mathbf{A} can be computed using $\frac{N_1^3 - N_1}{3}$ complex multiplications. The matrix $\mathbf{A}^{-1}\mathbf{B}$ can be computed using $N_1^2 N_2$ complex multiplications and N_1 complex divisions if the \mathbf{LDL}^H decomposition of \mathbf{A} is known.

Proof. Efficient algorithms for computing the \mathbf{LDL}^H decomposition are reviewed in [155, 183], and the number of multiplications can be found in [155, Table I]. Note that $\mathbf{A}^{-1}\mathbf{B}$ can be computed by solving N_2 linear systems of equations. If the \mathbf{LDL}^H decomposition is known, it can be exploited to solve the systems of equations by forward-backward substitution, as described in [67, Appendix C.2], which requires N_1^2 multiplications per system. N_1 additional divisions are required to compute \mathbf{D}^{-1} . \square

B.1.2 Matrix Identities

The following identity is key to manipulating matrix inverses.

Lemma B.3 (Matrix inversion lemma). Consider the matrices $\mathbf{A} \in \mathbb{C}^{N_1 \times N_1}$, $\mathbf{B} \in \mathbb{C}^{N_1 \times N_2}$, $\mathbf{C} \in \mathbb{C}^{N_2 \times N_2}$, and $\mathbf{D} \in \mathbb{C}^{N_2 \times N_1}$. The following identity holds if all the involved inverses exist:

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1}. \quad (\text{B.1})$$

By utilizing Lemma B.3, we can obtain the following identity for rank-one perturbations of an inverse.

Lemma B.4. Consider the invertible Hermitian matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ and some vector $\mathbf{x} \in \mathbb{C}^N$. It holds that

$$(\mathbf{A} + \mathbf{x}\mathbf{x}^H)^{-1} = \mathbf{A}^{-1} - \frac{1}{1 + \mathbf{x}^H\mathbf{A}^{-1}\mathbf{x}}\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^H\mathbf{A}^{-1} \quad (\text{B.2})$$

$$(\mathbf{A} + \mathbf{x}\mathbf{x}^H)^{-1}\mathbf{x} = \frac{1}{1 + \mathbf{x}^H\mathbf{A}^{-1}\mathbf{x}}\mathbf{A}^{-1}\mathbf{x}. \quad (\text{B.3})$$

Proof. The identity in (B.2) follows from Lemma B.3 with $\mathbf{B} = \mathbf{x}$, $\mathbf{C} = 1$, and $\mathbf{D} = \mathbf{x}^H$. We obtain (B.3) from (B.2) by multiplying with \mathbf{x} from the right and then simplifying the expression. \square

The following identities are commonly used for matrix manipulation.

Lemma B.5. For matrices $\mathbf{A} \in \mathbb{C}^{N_1 \times N_2}$ and $\mathbf{B} \in \mathbb{C}^{N_2 \times N_1}$, it holds that

$$(\mathbf{I}_{N_1} + \mathbf{AB})^{-1}\mathbf{A} = \mathbf{A}(\mathbf{I}_{N_2} + \mathbf{BA})^{-1} \quad (\text{B.4})$$

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}). \quad (\text{B.5})$$

The first identity is used to prove the following result.

Lemma B.6. Let $\mathbf{A} \in \mathbb{C}^{N \times N}$ and $\mathbf{B} \in \mathbb{C}^{N \times N}$ be two positive semi-definite Hermitian matrices that satisfy $\mathbf{BA} = \mathbf{0}_{N \times N}$. It then holds that

$$(\mathbf{I}_N + \mathbf{A} + \mathbf{B})^{-1}\mathbf{A} = (\mathbf{I}_N + \mathbf{A})^{-1}\mathbf{A}. \quad (\text{B.6})$$

Proof. The left-hand side of (B.6) can be rewritten as

$$\begin{aligned}
& (\mathbf{I}_N + \mathbf{A} + \mathbf{B})^{-1} \mathbf{A} \\
&= (\mathbf{I}_N + \mathbf{A} + \mathbf{B})^{-1} \mathbf{A} (\mathbf{I}_N + \mathbf{A}) (\mathbf{I}_N + \mathbf{A})^{-1} \\
&\stackrel{(a)}{=} (\mathbf{I}_N + \mathbf{A} + \mathbf{B})^{-1} (\mathbf{I}_N + \mathbf{A}) \mathbf{A} (\mathbf{I}_N + \mathbf{A})^{-1} \\
&\stackrel{(b)}{=} (\mathbf{I}_N + \mathbf{A} + \mathbf{B})^{-1} (\mathbf{I}_N + \mathbf{A} + \mathbf{B}) \mathbf{A} (\mathbf{I}_N + \mathbf{A})^{-1} \\
&= \mathbf{A} (\mathbf{I}_N + \mathbf{A})^{-1} \\
&\stackrel{(c)}{=} (\mathbf{I}_N + \mathbf{A})^{-1} \mathbf{A}
\end{aligned} \tag{B.7}$$

where (a) exploits the fact that the matrices $(\mathbf{I}_N + \mathbf{A})$ and \mathbf{A} commute and (b) utilizes the assumption that $\mathbf{B}\mathbf{A} = \mathbf{0}_{N \times N}$ to add this term to the expressions. Lastly, (c) utilizes the matrix identity in (B.4). \square

The following matrix identities are commonly used for lower/upper bounding of matrix expressions.

Lemma B.7. Consider an arbitrary matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ and a positive semi-definite matrix $\mathbf{B} \in \mathbb{C}^{N \times N}$. It holds that

$$|\text{tr}(\mathbf{AB})| \leq \|\mathbf{A}\|_2 \text{tr}(\mathbf{B}). \tag{B.8}$$

If \mathbf{A} is a positive semi-definite matrix, it further holds that

$$\text{tr}(\mathbf{AB}) \leq \|\mathbf{A}\|_2 \text{tr}(\mathbf{B}). \tag{B.9}$$

Proof. Let $\sigma_i(\cdot)$ denote the i th decreasingly ordered singular value of a matrix. It generally holds that $|\text{tr}(\mathbf{AB})| \leq \sum_{i=1}^N \sigma_i(\mathbf{A}) \sigma_i(\mathbf{B})$. Since $\sigma_i(\mathbf{A}) \leq \sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$, we further have $|\text{tr}(\mathbf{AB})| \leq \|\mathbf{A}\|_2 \sum_{i=1}^N \sigma_i(\mathbf{B}) = \|\mathbf{A}\|_2 \text{tr}(\mathbf{B})$ since \mathbf{B} is a positive semi-definite matrix and thus the singular values are also the eigenvalues. When \mathbf{A} is positive semi-definite, (B.9) follows from (B.8) since $|\text{tr}(\mathbf{AB})| = \text{tr}(\mathbf{AB})$. \square

Lemma B.8. Consider the positive definite matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ and the positive semi-definite matrix $\mathbf{B} \in \mathbb{C}^{N \times N}$. It holds that

$$\text{tr}(\mathbf{A}^{-1} \mathbf{B}) \geq \frac{1}{\|\mathbf{A}\|_2} \text{tr}(\mathbf{B}). \tag{B.10}$$

Proof. Let $\mathbf{C} = \mathbf{A}^{-1}$, which is a positive definite matrix with strictly positive eigenvalues. The left hand side of (B.10) can be rewritten as

$$\begin{aligned}\text{tr}(\mathbf{CB}) &= \text{tr}(\lambda_{\min}(\mathbf{C})\mathbf{I}_N\mathbf{B}) + \text{tr}((\mathbf{C} - \lambda_{\min}(\mathbf{C})\mathbf{I}_N)\mathbf{B}) \\ &\geq \lambda_{\min}(\mathbf{C})\text{tr}(\mathbf{B}) = \frac{1}{\|\mathbf{C}^{-1}\|_2}\text{tr}(\mathbf{B})\end{aligned}\quad (\text{B.11})$$

where $\lambda_{\min}(\mathbf{C}) > 0$ denotes the smallest eigenvalue of \mathbf{C} . The inequality follows from the fact that $\mathbf{C} - \lambda_{\min}(\mathbf{C})\mathbf{I}_N$ and \mathbf{B} are positive semi-definite matrices, thus $\text{tr}((\mathbf{C} - \lambda_{\min}(\mathbf{C})\mathbf{I}_N)\mathbf{B}) \geq 0$. Finally, we note that the smallest eigenvalue of \mathbf{C} equals the largest eigenvalue of $\mathbf{C}^{-1} = \mathbf{A}$, which is equivalent to $\|\mathbf{A}\|_2$. \square

Lemma B.9 (Cauchy-Schwarz inequality). Consider the vectors $\mathbf{a} \in \mathbb{C}^N$ and $\mathbf{b} \in \mathbb{C}^N$. It holds that

$$|\mathbf{a}^H \mathbf{b}|^2 \leq \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \quad (\text{B.12})$$

with equality if and only if \mathbf{a} and \mathbf{b} are linearly dependent.

Lemma B.10 (Generalized Rayleigh quotient). Consider the fixed vector $\mathbf{a} \in \mathbb{C}^N$ and the Hermitian positive definite matrix $\mathbf{B} \in \mathbb{C}^{N \times N}$. It then holds that

$$\max_{\mathbf{v} \in \mathbb{C}^N} \frac{|\mathbf{v}^H \mathbf{a}|^2}{\mathbf{v}^H \mathbf{B} \mathbf{v}} = \mathbf{a}^H \mathbf{B}^{-1} \mathbf{a} \quad (\text{B.13})$$

where the maximum is attained by $\mathbf{v} = \mathbf{B}^{-1} \mathbf{a}$.

Proof. The matrix square root $\mathbf{C} = \mathbf{B}^{\frac{1}{2}}$ of \mathbf{B} exists since \mathbf{B} is positive definite. We begin by making the change of variable $\bar{\mathbf{v}} = \mathbf{C}\mathbf{v}$, leading to the equivalent optimization problem

$$\max_{\bar{\mathbf{v}} \in \mathbb{C}^N} \frac{|\bar{\mathbf{v}}^H (\mathbf{C}^{-1})^H \mathbf{a}|^2}{\|\bar{\mathbf{v}}\|^2}. \quad (\text{B.14})$$

Next, we note that $|\bar{\mathbf{v}}^H (\mathbf{C}^{-1})^H \mathbf{a}|^2 \leq \|\bar{\mathbf{v}}\|^2 \|(\mathbf{C}^{-1})^H \mathbf{a}\|^2$ according to the Cauchy-Schwarz inequality (see Lemma B.9) with equality if and only if $\bar{\mathbf{v}}$ and $(\mathbf{C}^{-1})^H \mathbf{a}$ are equal up to a scalar factor. By inserting this achievable upper bound in (B.14), we get the maximum value as $\|(\mathbf{C}^{-1})^H \mathbf{a}\|^2 = \mathbf{a}^H \mathbf{B}^{-1} \mathbf{a}$. We note that the scaling of $\bar{\mathbf{v}}$ does not affect the result, thus we can set $\bar{\mathbf{v}} = (\mathbf{C}^{-1})^H \mathbf{a}$ which leads to $\mathbf{v} = \mathbf{C}^{-1} (\mathbf{C}^{-1})^H \mathbf{a} = \mathbf{B}^{-1} \mathbf{a}$. \square

B.2 Random Vectors and Matrices

This section provides some standard results that are useful when manipulating random scalars, vectors, and matrices.

Definition B.1 (Complex Gaussian random variable). An N -dimensional circularly symmetric complex Gaussian random vector \mathbf{x} with mean value $\boldsymbol{\mu} \in \mathbb{C}^N$ and the positive definite covariance matrix $\mathbf{R} \in \mathbb{C}^{N \times N}$ has the PDF

$$f(\mathbf{x}) = \frac{e^{-(\mathbf{x}-\boldsymbol{\mu})^H \mathbf{R}^{-1} (\mathbf{x}-\boldsymbol{\mu})}}{\pi^N \det(\mathbf{R})}. \quad (\text{B.15})$$

This is denoted as $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \mathbf{R})$. The circular symmetry implies that $\mathbf{x} - \boldsymbol{\mu}$ and $e^{j\phi}(\mathbf{x} - \boldsymbol{\mu})$ are identically distributed for any given ϕ .¹

If the covariance matrix \mathbf{R} has rank $r < N$, then $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \mathbf{R})$ instead means that

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{U} \mathbf{D}^{\frac{1}{2}} \begin{bmatrix} \mathbf{g} \\ \mathbf{0}_{N-r} \end{bmatrix} \quad (\text{B.16})$$

where $\mathbf{g} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_r, \mathbf{I}_r)$ has the PDF $\frac{e^{-\|\mathbf{g}\|^2}}{\pi^r}$ and $\mathbf{R} = \mathbf{U} \mathbf{D} \mathbf{U}^H$ is the eigenvalue decomposition, with the diagonal matrix \mathbf{D} containing the eigenvalues in decaying order.

If $N = 1$ and the mean value is zero, then $x \sim \mathcal{N}_{\mathbb{C}}(0, q)$ has the PDF

$$f(x) = \frac{e^{-|x|^2/q}}{\pi q}. \quad (\text{B.17})$$

Lemma B.11 (Jensen's inequality). Consider a scalar real-valued integrable² random variable x and a scalar function $g(\cdot)$. It holds that

$$g(\mathbb{E}\{x\}) \leq \mathbb{E}\{g(x)\} \quad (\text{B.18})$$

if the function is convex and

$$g(\mathbb{E}\{x\}) \geq \mathbb{E}\{g(x)\} \quad (\text{B.19})$$

if the function is concave. Equality holds in (B.18) and (B.19) if and only if x is either deterministic or $g(\cdot)$ is a linear function.

¹Strictly speaking, only a random variable with zero mean can be circularly symmetric, but we consider the extension of this notion as defined in this sentence.

²More precisely, the expectation $\mathbb{E}\{g(x)\}$ must be finite.

The following is a special case of the strong law of large numbers.

Lemma B.12. Let x_1, \dots, x_M be a sequence of non-negative i.i.d. random scalars with $\mathbb{E}\{x_m\} = a$ and $\mathbb{E}\{x_m^4\} < \infty$, then

$$\frac{1}{M} \sum_{m=1}^M x_m \rightarrow a \quad (\text{B.20})$$

almost surely as $M \rightarrow \infty$.

Proof. This is a special case of [93, Theorem 3.4]. \square

The following trace lemma is also a consequence of the strong law of large numbers.

Lemma B.13. Let $\mathbf{R}_1, \mathbf{R}_2, \dots$ be a sequence of matrices, with $\mathbf{R}_M \in \mathbb{C}^{M \times M}$, that satisfies $\limsup_M \|\mathbf{R}_M\|_2 < \infty$. Let $\mathbf{x}_1, \mathbf{x}_2, \dots$ be a sequence of random vectors with $\mathbf{x}_M \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{I}_M)$. It holds that

$$\frac{1}{M} \mathbf{x}_M^H \mathbf{R}_M \mathbf{x}_M - \frac{1}{M} \text{tr}(\mathbf{R}_M) \rightarrow 0 \quad (\text{B.21})$$

almost surely as $M \rightarrow \infty$.

Proof. This is a special case of [93, Theorem 3.4] where we limit the scope to complex Gaussian vectors. \square

Lemma B.14. Consider the vector $\mathbf{a} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{A})$, with covariance matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$, and any diagonalizable matrix $\mathbf{B} \in \mathbb{C}^{N \times N}$. It holds that

$$\mathbb{E}\{|\mathbf{a}^H \mathbf{B} \mathbf{a}|^2\} = |\text{tr}(\mathbf{B} \mathbf{A})|^2 + \text{tr}(\mathbf{B} \mathbf{A} \mathbf{B}^H \mathbf{A}). \quad (\text{B.22})$$

Proof. Note that $\mathbf{a} = \mathbf{A}^{\frac{1}{2}} \mathbf{w}$ for $\mathbf{w} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{I}_N)$, thus we can write

$$\mathbb{E}\{|\mathbf{a}^H \mathbf{B} \mathbf{a}|^2\} = \mathbb{E}\{|\mathbf{w}^H (\mathbf{A}^H)^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}} \mathbf{w}|^2\}. \quad (\text{B.23})$$

Next, let $\mathbf{U} \Lambda \mathbf{U}^H = (\mathbf{A}^H)^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}$ denote an eigenvalue decomposition with $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$ and define $\bar{\mathbf{w}} = [\bar{w}_1 \dots \bar{w}_N]^T = \mathbf{U}^H \mathbf{w}$ ~

$\mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{I}_N)$. It then follows that

$$\begin{aligned}
& \mathbb{E}\{|w^H(\mathbf{A}^H)^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}w|^2\} = \mathbb{E}\{|\bar{w}^H\Lambda\bar{w}|^2\} \\
&= \mathbb{E}\left\{\left|\sum_{n=1}^N |\bar{w}_n|^2 \lambda_n\right|^2\right\} = \sum_{n_1=1}^N \sum_{n_2=1}^N \mathbb{E}\{|\bar{w}_{n_1}|^2 |\bar{w}_{n_2}|^2\} \lambda_{n_1} \lambda_{n_2}^* \\
&\stackrel{(a)}{=} \sum_{n_1=1}^N \sum_{\substack{n_2=1 \\ n_2 \neq n_1}}^N \lambda_{n_1} \lambda_{n_2}^* + \sum_{n_1=1}^N 2|\lambda_{n_1}|^2 \\
&= \sum_{n_1=1}^N \lambda_{n_1} \sum_{n_2=1}^N \lambda_{n_2}^* + \sum_{n_1=1}^N |\lambda_{n_1}|^2 \\
&\stackrel{(b)}{=} |\text{tr}(\Lambda)|^2 + \text{tr}(\Lambda\Lambda^H) \\
&= |\text{tr}((\mathbf{A}^H)^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})|^2 + \text{tr}((\mathbf{A}^H)^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}(\mathbf{A}^H)^{\frac{1}{2}}\mathbf{B}^H\mathbf{A}^{\frac{1}{2}}) \\
&= |\text{tr}(\mathbf{B}\mathbf{A})|^2 + \text{tr}(\mathbf{B}\mathbf{A}\mathbf{B}^H\mathbf{A})
\end{aligned} \tag{B.24}$$

where (a) utilizes the independence between the elements of \bar{w} and also the moments $\mathbb{E}\{|\bar{w}_n|^2\} = 1$ and $\mathbb{E}\{|\bar{w}_n|^4\} = 2$. In (b), we write the sum of the eigenvalues as the trace and then we reintroduce the unitary matrices from the eigenvalue decomposition. The final equality follows from the fact that $\text{tr}(\mathbf{C}_1\mathbf{C}_2) = \text{tr}(\mathbf{C}_2\mathbf{C}_1)$ for any matrices $\mathbf{C}_1, \mathbf{C}_2$ such that \mathbf{C}_1 and \mathbf{C}_2^T have the same dimensions. \square

Lemma B.15. Consider the random vectors $\mathbf{a} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{N_a}, \mu_a \mathbf{I}_{N_a})$ and $\mathbf{b} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{N_b}, \mu_b \mathbf{I}_{N_b})$ with $\mu_a \neq \mu_b$. The scalar

$$y = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \tag{B.25}$$

has the PDF

$$\begin{aligned}
f(x) &= \sum_{m=1}^{N_a} \frac{x^{N_a-m} e^{-\frac{x}{\mu_a}} (-1)^{m+1} \binom{N_b+m-2}{m-1}}{\mu_a^{N_a} \mu_b^{N_b} (N_a - m)! \left(\frac{1}{\mu_b} - \frac{1}{\mu_a}\right)^{N_b+m-1}} \\
&+ \sum_{m=1}^{N_b} \frac{x^{N_b-m} e^{-\frac{x}{\mu_b}} (-1)^{m+1} \binom{N_a+m-2}{m-1}}{\mu_a^{N_a} \mu_b^{N_b} (N_b - m)! \left(\frac{1}{\mu_a} - \frac{1}{\mu_b}\right)^{N_a+m-1}}
\end{aligned} \tag{B.26}$$

for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Moreover, it holds that

$$\begin{aligned} & \mathbb{E} \left\{ \log_2 \left(1 + \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \right) \right\} \\ &= \sum_{m=1}^{N_a} \sum_{l=0}^{N_a-m} \frac{\binom{N_b+m-2}{m-1} (-1)^{N_a-l+1}}{\left(\frac{1}{\mu_b} - \frac{1}{\mu_a} \right)^{N_b+m-1}} \frac{\left(e^{\frac{1}{\mu_a}} E_1 \left(\frac{1}{\mu_a} \right) + \sum_{n=1}^l \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{j! \mu_a^j} \right)}{(N_a - m - l)! \mu_a^{N_a-l-1} \mu_b^{N_b} \log_e(2)} \\ &+ \sum_{m=1}^{N_b} \sum_{l=0}^{N_b-m} \frac{\binom{N_a+m-2}{m-1} (-1)^{N_b-l+1}}{\left(\frac{1}{\mu_a} - \frac{1}{\mu_b} \right)^{N_a+m-1}} \frac{\left(e^{\frac{1}{\mu_b}} E_1 \left(\frac{1}{\mu_b} \right) + \sum_{n=1}^l \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{j! \mu_b^j} \right)}{(N_b - m - l)! \mu_a^{N_a} \mu_b^{N_b-l-1} \log_e(2)}. \end{aligned} \quad (\text{B.27})$$

Proof. The squared absolute value of a $\mathcal{N}_{\mathbb{C}}(0, v)$ -distributed random variable has the exponential distribution $\text{Exp}(1/v)$, thus $\|\mathbf{a}\|^2$ is the sum of N_a independent $\text{Exp}(1/\mu_a)$ -distributed random variables. Similarly, $\|\mathbf{b}\|^2$ is the sum of N_b independent $\text{Exp}(1/\mu_b)$ -distributed random variables. When $\mu_a \neq \mu_b$, the PDF $f(x)$ in (B.26) is obtained from the general PDF formula derived in [13].

The expectation in (B.27) is computed by expanding it as

$$\begin{aligned} & \mathbb{E} \left\{ \log_2 \left(1 + \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \right) \right\} = \int_0^\infty \log_2(1+x) f(x) dx \\ &= \sum_{m=1}^{N_a} \frac{\int_0^\infty \log_2(1+x) x^{N_a-m} e^{-\frac{x}{\mu_a}} dx (-1)^{m+1} \binom{N_b+m-2}{m-1}}{\mu_a^{N_a} \mu_b^{N_b} (N_a - m)! \left(\frac{1}{\mu_b} - \frac{1}{\mu_a} \right)^{N_b+m-1}} \\ &+ \sum_{m=1}^{N_b} \frac{\int_0^\infty \log_2(1+x) x^{N_b-m} e^{-\frac{x}{\mu_b}} dx (-1)^{m+1} \binom{N_a+m-2}{m-1}}{\mu_a^{N_a} \mu_b^{N_b} (N_b - m)! \left(\frac{1}{\mu_a} - \frac{1}{\mu_b} \right)^{N_a+m-1}} \end{aligned} \quad (\text{B.28})$$

and the final expression is obtained by using the integral identity

$$\begin{aligned} & \int_0^\infty \log_2(1+x) x^{N-m} e^{-\frac{x}{\mu}} dx \\ &= \sum_{l=0}^{N-m} \frac{(N-m)!}{(N-m-l)!} \frac{\mu^{l+1} (-1)^{N-m-l}}{\log_e(2)} \left(e^{\frac{1}{\mu}} E_1 \left(\frac{1}{\mu} \right) + \sum_{n=1}^l \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{j! \mu^j} \right) \end{aligned} \quad (\text{B.29})$$

from [55, Theorem 2] to compute the remaining expectations. \square

B.3 Properties of the Lambert W Function

The Lambert W function appears frequently in EE analysis and is defined as follows.

Definition B.2. The Lambert W function is denoted by $W(x)$ and defined by the equation $x = W(x)e^{W(x)}$ for any $x \in \mathbb{R}$ with e being Euler's number.

The function can be lower and upper bounded as follows.

Lemma B.16. The Lambert W function $W(x)$ is an increasing function for $x \geq 0$ and satisfies the inequalities

$$e \frac{x}{\log_e(x)} \leq e^{W(x)+1} \leq (1 + e) \frac{x}{\log_e(x)} \quad \text{for } x \geq e \quad (\text{B.30})$$

with e being Euler's number.

The above lemma follows from the results and inequalities in [146] and implies that $e^{W(x)+1}$ is approximately equal to e for small x (i.e., when $\log_e(x) \approx x$) whereas it increases almost linearly with x when x is large (i.e., when $\log_e(x)$ is almost constant).

B.4 Basic Estimation Theory

This section provides some basic results on the estimation of unknown variables that are utilized in this monograph. We refer to textbooks such as [175] for further details and explanations.

The purpose of estimation is to obtain an approximate value of an unknown variable based on measurements. We are particularly interested in Bayesian estimation where the unknown variable is a realization of a random variable having a known, or partially known, statistical distribution. We have the following basic definition.

Definition B.3 (Bayesian estimator). Consider a random variable $\mathbf{x} \in \mathbb{C}^N$ with support in \mathcal{X} and let $\hat{\mathbf{x}}(\mathbf{y})$ denote an arbitrary estimator of \mathbf{x} based on the observation $\mathbf{y} \in \mathbb{C}^T$. For a given loss function $\ell(\cdot, \cdot)$, the estimator that minimizes the expected loss

$$\mathbb{E}\{\ell(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{y}))\} \quad (\text{B.31})$$

is called a *Bayesian estimator*.

There are many potential loss functions, but the squared difference is of particular interest in this monograph since its expectation corresponds to the estimation error variance.

Definition B.4 (Minimum mean-squared error estimator). The MMSE estimator is given by the loss function $\ell(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{y})) = \|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\|^2$ and thus minimizes the MSE

$$\mathbb{E} \left\{ \|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\|^2 \right\}. \quad (\text{B.32})$$

It is further computed as

$$\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{y}) = \mathbb{E}\{\mathbf{x}|\mathbf{y}\} = \int_{\mathcal{X}} \mathbf{x} f(\mathbf{x}|\mathbf{y}) d\mathbf{x} \quad (\text{B.33})$$

where $f(\mathbf{x}|\mathbf{y})$ is the conditional PDF of \mathbf{x} given the observation \mathbf{y} .

The MMSE estimator of a complex Gaussian random variable from an observation that is corrupted by independent additive complex Gaussian noise (and interference) is of particular interest in this monograph.

Lemma B.17. Consider estimation of the N -dimensional vector $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(\bar{\mathbf{x}}, \mathbf{R})$, from the observation $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \in \mathbb{C}^L$. The covariance matrix \mathbf{R} is positive definite, $\mathbf{A} \in \mathbb{C}^{L \times N}$ is a known matrix, and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\bar{\mathbf{n}}, \mathbf{S})$ is an L -dimensional independent noise/interference vector with a positive definite covariance matrix.

The MMSE estimator of \mathbf{x} based on \mathbf{y} is

$$\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{y}) = \bar{\mathbf{x}} + \mathbf{R}\mathbf{A}^H(\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1}(\mathbf{y} - \mathbf{A}\bar{\mathbf{x}} - \bar{\mathbf{n}}). \quad (\text{B.34})$$

The error covariance matrix $\mathbf{C}_{\text{MMSE}} = \mathbb{E}\{(\mathbf{x} - \hat{\mathbf{x}}_{\text{MMSE}})(\mathbf{x} - \hat{\mathbf{x}}_{\text{MMSE}})^H\}$ is

$$\mathbf{C}_{\text{MMSE}} = \mathbf{R} - \mathbf{R}\mathbf{A}^H(\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1}\mathbf{A}\mathbf{R} \quad (\text{B.35})$$

and the MSE, $\text{MSE} = \mathbb{E}\{\|\mathbf{x} - \hat{\mathbf{x}}_{\text{MMSE}}\|^2\}$, is

$$\text{MSE} = \text{tr} \left(\mathbf{R} - \mathbf{R}\mathbf{A}^H(\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1}\mathbf{A}\mathbf{R} \right). \quad (\text{B.36})$$

Proof. We begin by computing the conditional PDF $f(\mathbf{x}|\mathbf{y})$. To this end, we notice that

$$f(\mathbf{x}) = \frac{e^{-(\mathbf{x}-\bar{\mathbf{x}})^H \mathbf{R}^{-1} (\mathbf{x}-\bar{\mathbf{x}})}}{\pi^N \det(\mathbf{R})} \quad (\text{B.37})$$

$$f(\mathbf{y}) = \frac{e^{-(\mathbf{y}-\mathbf{A}\bar{\mathbf{x}}-\bar{\mathbf{n}})^H (\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1} (\mathbf{y}-\mathbf{A}\bar{\mathbf{x}}-\bar{\mathbf{n}})}}{\pi^L \det(\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})} \quad (\text{B.38})$$

$$f(\mathbf{y}|\mathbf{x}) = \frac{e^{-(\mathbf{y}-\mathbf{A}\mathbf{x}-\bar{\mathbf{n}})^H \mathbf{S}^{-1} (\mathbf{y}-\mathbf{A}\mathbf{x}-\bar{\mathbf{n}})}}{\pi^L \det(\mathbf{S})} \quad (\text{B.39})$$

are the PDF of \mathbf{x} , the PDF of \mathbf{y} , and the conditional PDF of \mathbf{y} given \mathbf{x} , respectively. Using Bayes' formula, we can compute $f(\mathbf{x}|\mathbf{y})$ as

$$\begin{aligned} f(\mathbf{x}|\mathbf{y}) &= \frac{f(\mathbf{y}|\mathbf{x})f(\mathbf{x})}{f(\mathbf{y})} = \frac{\frac{e^{-(\mathbf{y}-\mathbf{A}\mathbf{x}-\bar{\mathbf{n}})^H \mathbf{S}^{-1} (\mathbf{y}-\mathbf{A}\mathbf{x}-\bar{\mathbf{n}})} e^{-(\mathbf{x}-\bar{\mathbf{x}})^H \mathbf{R}^{-1} (\mathbf{x}-\bar{\mathbf{x}})}}{\pi^L \det(\mathbf{S})} \frac{e^{-(\mathbf{x}-\bar{\mathbf{x}})^H \mathbf{R}^{-1} (\mathbf{x}-\bar{\mathbf{x}})}}{\pi^N \det(\mathbf{R})}}{\frac{e^{-(\mathbf{y}-\mathbf{A}\bar{\mathbf{x}}-\bar{\mathbf{n}})^H (\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1} (\mathbf{y}-\mathbf{A}\bar{\mathbf{x}}-\bar{\mathbf{n}})}}{\pi^L \det(\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})}} \\ &= \frac{e^{-(\mathbf{x}-\mathbf{m})^H (\mathbf{R}^{-1} + \mathbf{A}^H \mathbf{S}^{-1} \mathbf{A})(\mathbf{x}-\mathbf{m})}}{\pi^N \det((\mathbf{R}^{-1} + \mathbf{A}^H \mathbf{S}^{-1} \mathbf{A})^{-1})} \end{aligned}$$

where

$$\begin{aligned} \mathbf{m} &= \bar{\mathbf{x}} + (\mathbf{R}^{-1} + \mathbf{A}^H \mathbf{S}^{-1} \mathbf{A})^{-1} \mathbf{A}^H \mathbf{S}^{-1} (\mathbf{y} - \mathbf{A}\bar{\mathbf{x}} - \bar{\mathbf{n}}) \\ &= \bar{\mathbf{x}} + \mathbf{R}\mathbf{A}^H (\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1} (\mathbf{y} - \mathbf{A}\bar{\mathbf{x}} - \bar{\mathbf{n}}) \end{aligned} \quad (\text{B.40})$$

after some straightforward algebra (including the use of Lemma B.3). We identify $f(\mathbf{x}|\mathbf{y})$ as a circularly symmetric complex Gaussian distribution with mean value \mathbf{m} and covariance matrix $(\mathbf{R}^{-1} + \mathbf{A}^H \mathbf{S}^{-1} \mathbf{A})^{-1}$. By definition, the MMSE estimator is $\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{y}) = \mathbb{E}\{\mathbf{x}|\mathbf{y}\} = \mathbf{m}$ and the estimation error covariance matrix is

$$\mathbf{C}_{\text{MMSE}} = (\mathbf{R}^{-1} + \mathbf{A}^H \mathbf{S}^{-1} \mathbf{A})^{-1} = \mathbf{R} - \mathbf{R}\mathbf{A}^H (\mathbf{A}\mathbf{R}\mathbf{A}^H + \mathbf{S})^{-1} \mathbf{A}\mathbf{R} \quad (\text{B.41})$$

where the second equality follows from Lemma B.3. Finally, we notice that the MSE is $\text{tr}(\mathbf{C}_{\text{MMSE}})$. \square

The MMSE estimator can be applied even when the covariance matrix of the unknown variable is rank-deficient, as shown by the next corollary. For brevity, we only consider zero-mean variables, because that is the focus of this monograph.

Corollary B.18. Consider estimation of the N -dimensional vector $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{R})$, with a positive semi-definite covariance/correlation matrix \mathbf{R} , from the observation $\mathbf{y} = \mathbf{x}q + \mathbf{n} \in \mathbb{C}^N$. The pilot signal $q \in \mathbb{C}$ is known and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{S})$ is an independent noise/interference vector with a positive definite covariance/correlation matrix.

The MMSE estimator of \mathbf{x} is

$$\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{y}) = q^* \mathbf{R} \left(|q|^2 \mathbf{R} + \mathbf{S} \right)^{-1} \mathbf{y}. \quad (\text{B.42})$$

The estimation error correlation matrix is

$$\mathbf{C}_{\text{MMSE}} = \mathbf{R} - |q|^2 \mathbf{R} \left(|q|^2 \mathbf{R} + \mathbf{S} \right)^{-1} \mathbf{R} \quad (\text{B.43})$$

and the MSE is

$$\text{MSE} = \text{tr} \left(\mathbf{R} - |q|^2 \mathbf{R} \left(|q|^2 \mathbf{R} + \mathbf{S} \right)^{-1} \mathbf{R} \right). \quad (\text{B.44})$$

Proof. Similar to Remark 2.2 on p. 224, we let $\mathbf{R} = \mathbf{UDU}^H$, where $\mathbf{D} \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing the $r = \text{rank}(\mathbf{R})$ positive non-zero eigenvalues of \mathbf{R} and $\mathbf{U} \in \mathbb{C}^{N \times r}$ consists of the associated eigenvectors. Using this notation, we can express $\mathbf{x} = \mathbf{Ug}$ where $\mathbf{g} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_r, \mathbf{D})$ and notice that $\mathbf{y} = \mathbf{Ug}q + \mathbf{n}$. Since \mathbf{g} is the variable to be estimated and it has a full-rank correlation/covariance matrix \mathbf{D} , we can apply Lemma B.34 with \mathbf{g} as the unknown variable, \mathbf{Ug} as the known matrix \mathbf{A} , and \mathbf{n} as the noise/interference vector. It then follows that

$$\begin{aligned} \mathbb{E}\{\mathbf{x}|\mathbf{y}\} &= \mathbf{U}\mathbb{E}\{\mathbf{g}|\mathbf{y}\} \\ &= q^* \mathbf{UDU}^H (|q|^2 \mathbf{UDU}^H + \mathbf{S})^{-1} \mathbf{y} \\ &= q^* \mathbf{R} (|q|^2 \mathbf{R} + \mathbf{S})^{-1} \mathbf{y} \end{aligned} \quad (\text{B.45})$$

which yields the estimator in (B.42). The estimation error correlation matrix and the MSE can be obtained accordingly. \square

In many cases with non-Gaussian unknown variables, the MMSE estimator is hard to compute, either because of analytical intractability or since the full statistical characterization of the unknown variable cannot be obtained in practice. Linear Bayesian estimators are then useful because only the mean value and covariance matrix of the variable are needed.

Definition B.5 (Linear minimum mean-squared error estimator). The LMMSE estimator is the Bayesian estimator that minimizes the MSE

$$\mathbb{E} \left\{ \| \mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}) \|^2 \right\} \quad (\text{B.46})$$

under the additional constraint that the estimator is a linear (or affine) function of the observation. More precisely,

$$\hat{\mathbf{x}}_{\text{LMMSE}}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{b} \quad (\text{B.47})$$

where \mathbf{A} and \mathbf{b} are selected jointly to minimize the MSE.

The LMMSE estimator can be computed in closed form.

Lemma B.19. Consider estimation of the vector \mathbf{x} from the observation \mathbf{y} . The LMMSE estimator is

$$\hat{\mathbf{x}}_{\text{LMMSE}}(\mathbf{y}) = \mathbb{E}\{\mathbf{x}\} + \mathbf{C}_{\mathbf{xy}}\mathbf{C}_{\mathbf{yy}}^{-1}(\mathbf{y} - \mathbb{E}\{\mathbf{y}\}) \quad (\text{B.48})$$

if $\mathbf{C}_{\mathbf{yy}}$ is invertible, and achieves the MSE $\text{tr}(\mathbf{C}_{\text{LMMSE}})$, where the estimation error covariance matrix $\mathbf{C}_{\text{LMMSE}}$ is

$$\mathbf{C}_{\text{LMMSE}} = \mathbf{C}_{\mathbf{xx}} - \mathbf{C}_{\mathbf{xy}}\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{xy}}^H \quad (\text{B.49})$$

and

$$\mathbf{C}_{\mathbf{xy}} = \mathbb{E} \left\{ (\mathbf{x} - \mathbb{E}\{\mathbf{x}\})(\mathbf{y} - \mathbb{E}\{\mathbf{y}\})^H \right\} \quad (\text{B.50})$$

$$\mathbf{C}_{\mathbf{xx}} = \mathbb{E} \left\{ (\mathbf{x} - \mathbb{E}\{\mathbf{x}\})(\mathbf{x} - \mathbb{E}\{\mathbf{x}\})^H \right\} \quad (\text{B.51})$$

$$\mathbf{C}_{\mathbf{yy}} = \mathbb{E} \left\{ (\mathbf{y} - \mathbb{E}\{\mathbf{y}\})(\mathbf{y} - \mathbb{E}\{\mathbf{y}\})^H \right\}. \quad (\text{B.52})$$

Proof. The LMMSE estimator has the form $\hat{\mathbf{x}}_{\text{LMMSE}}(\mathbf{y}) = \mathbf{Ay} + \mathbf{b}$ by definition and minimizes

$$\begin{aligned} \mathbb{E} \left\{ \| \mathbf{x} - \mathbf{Ay} - \mathbf{b} \|^2 \right\} &= \text{tr}(\mathbb{E} \left\{ (\mathbf{x} - \mathbf{Ay} - \mathbf{b})(\mathbf{x}^H - \mathbf{y}^H \mathbf{A}^H - \mathbf{b}^H) \right\}) \\ &= \text{tr}(\mathbb{E} \left\{ (\mathbf{x} - \mathbf{Ay})(\mathbf{x}^H - \mathbf{y}^H \mathbf{A}^H) \right\}) \\ &\quad + \text{tr}(\mathbf{b}\mathbf{b}^H - \mathbb{E}\{\mathbf{x}\}\mathbf{b}^H - \mathbf{b}\mathbb{E}\{\mathbf{x}^H\} + \mathbf{A}\mathbb{E}\{\mathbf{y}\}\mathbf{b}^H + \mathbf{b}\mathbb{E}\{\mathbf{y}^H\}\mathbf{A}^H) \\ &= \mathbb{E} \left\{ \| (\mathbf{x} - \mathbb{E}\{\mathbf{x}\}) - \mathbf{A}(\mathbf{y} - \mathbb{E}\{\mathbf{y}\}) \|^2 \right\} + \| \mathbf{b} - \mathbb{E}\{\mathbf{x}\} + \mathbf{A}\mathbb{E}\{\mathbf{y}\} \|^2 \end{aligned} \quad (\text{B.53})$$

where the last equality follows from completing the squares and then doing some algebra. Note that \mathbf{b} only appears in (B.53) as $\| \mathbf{b} - \mathbb{E}\{\mathbf{x}\} + \mathbf{A}\mathbb{E}\{\mathbf{y}\} \|^2$.

$\mathbf{A}\mathbb{E}\{\mathbf{y}\}\|^2$, which implies that (B.53) is minimized by $\mathbf{b}_{\min} = \mathbb{E}\{\mathbf{x}\} - \mathbf{A}\mathbb{E}\{\mathbf{y}\}$. By substituting this value back into (B.53), we obtain

$$\begin{aligned} & \mathbb{E} \left\{ \|(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}) - \mathbf{A}(\mathbf{y} - \mathbb{E}\{\mathbf{y}\})\|^2 \right\} \\ &= \text{tr} (\mathbf{C}_{\mathbf{xx}} - \mathbf{AC}_{\mathbf{yx}} - \mathbf{C}_{\mathbf{xy}}\mathbf{A}^H + \mathbf{AC}_{\mathbf{yy}}\mathbf{A}^H) \\ &= \text{tr} \left((\mathbf{AC}_{\mathbf{yy}} - \mathbf{C}_{\mathbf{xy}})\mathbf{C}_{\mathbf{yy}}^{-1}(\mathbf{AC}_{\mathbf{yy}} - \mathbf{C}_{\mathbf{xy}})^H \right) \quad (\text{B.54}) \\ &+ \text{tr} (\mathbf{C}_{\mathbf{xx}}) - \text{tr} \left(\mathbf{C}_{\mathbf{xy}}\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{xy}}^H \right) \end{aligned}$$

where the last equality follows from completing the squares and exploiting the fact that $\mathbf{C}_{\mathbf{yx}} = \mathbf{C}_{\mathbf{xy}}^H$. The expression in (B.54) is minimized by $\mathbf{A}_{\min} = \mathbf{C}_{\mathbf{xy}}\mathbf{C}_{\mathbf{yy}}^{-1}$. The final estimator expression in (B.48) is obtained by using \mathbf{A}_{\min} and \mathbf{b}_{\min} . Substituting \mathbf{A}_{\min} back into (B.54) gives the MSE and error covariance matrix. \square

Notice that the LMMSE estimator in Lemma B.19 depends on the first-order moments ($\mathbb{E}\{\mathbf{x}\}$, $\mathbb{E}\{\mathbf{y}\}$) and the second-order moments ($\mathbf{C}_{\mathbf{xx}}$, $\mathbf{C}_{\mathbf{yy}}$, $\mathbf{C}_{\mathbf{xy}}$) of the unknown variable \mathbf{x} and the observation \mathbf{y} . However, the exact distributions of these random variables are not needed. This makes the LMMSE estimator particularly suitable for practical implementations, where these moments can be estimated relatively easily, while the full PDF is very hard to estimate since it may not follow any known distribution. Note that the MMSE estimator of a Gaussian random variable that is observed in independent Gaussian noise, which was considered in Lemma B.17, is a linear estimator and thus equals the LMMSE estimator; in other words, there exist no better non-linear Bayesian estimators in this special case.

B.5 Basic Information Theory

This section provides some basic information-theoretic definitions and results that are used in this monograph. We refer to textbooks such as [94] for further details and explanations.

The classic channel capacity result in Theorem 1.1 on p. 167 considers a discrete memoryless channel, where one input symbol is sent at a time and each output symbol only depends on the current input. These concepts are defined as follows.

Definition B.6 (Discrete channel). A discrete channel is a channel where one input symbol x is sent to receive one output symbol y . The channel is characterized by an input alphabet \mathcal{X} , an output alphabet \mathcal{Y} , and the transition PDF $f(y|x)$ for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$.

We can now define a memoryless channel as a collection of independent discrete channels.

Definition B.7 (Discrete memoryless channel). A discrete memoryless channel is a collection of N discrete channels, having input x_n and output y_n for $n = 1, \dots, N$, with joint transition PDF $f(y_1, \dots, y_N | x_1, \dots, x_N)$ satisfying

$$f(y_1, \dots, y_N | x_1, \dots, x_N) = \prod_{n=1}^N f(y_n | x_n). \quad (\text{B.55})$$

If $f(y_n | x_n)$ is the same for all n , we drop the index n for brevity and let the notation x , y , and $f(y|x)$ be used for an arbitrary instance of the channel.

Next, we define the entropy of a continuous random variable.

Definition B.8 (Differential entropy). The differential entropy of a continuous random variable y with support in \mathcal{Y} and PDF $f(y)$ is

$$\mathcal{H}(y) = - \int_{\mathcal{Y}} \log_2 (f(y)) f(y) dy. \quad (\text{B.56})$$

If the random variable y is given and the conditional PDF is $f(y|x)$, then the conditional differential entropy is

$$\mathcal{H}(y|x) = - \int_{\mathcal{Y}} \log_2 (f(y|x)) f(y|x) dx. \quad (\text{B.57})$$

The differential entropy $\mathcal{H}(y)$ measures the surprisal when observing a realization of the random variable y , which can be interpreted as the amount of information that the variable conveys. The differential entropy can take any value from $-\infty$ to $+\infty$, where a larger value implies larger surprisal. Similarly, $\mathcal{H}(y|x)$ measures the amount of additional information that we obtain by observing y , if we already know x . It holds that $\mathcal{H}(y) \geq \mathcal{H}(y|x)$ since observing x cannot increase the

surprisal when we later observe y , but it can potentially reduce the surprisal. The difference $\mathcal{I}(x; y) = \mathcal{H}(x) - \mathcal{H}(x|y)$ quantifies the mutual information of x and y . The differential entropy of a zero-mean circularly symmetric complex Gaussian variable is of particular importance in this monograph.

Lemma B.20. The differential entropy of $x \sim \mathcal{N}_{\mathbb{C}}(0, q)$ is

$$\mathcal{H}(x) = \log_2(e\pi q). \quad (\text{B.58})$$

Proof. The PDF $f(x)$ of x is given in (B.17) and has support in \mathbb{C} . According to the definition of differential entropy, we have

$$\begin{aligned} \mathcal{H}(x) &= - \int_{\mathbb{C}} \log_2 \left(\frac{e^{-|x|^2/q}}{\pi q} \right) \frac{e^{-|x|^2/q}}{\pi q} dx \\ &= \frac{\log_2(e)}{q} \underbrace{\int_{\mathbb{C}} |x|^2 \frac{e^{-|x|^2/q}}{\pi q} dx}_{=q} + \log_2(\pi q) \underbrace{\int_{\mathbb{C}} \frac{e^{-|x|^2/q}}{\pi q} dx}_{=1} \\ &= \log_2(e\pi q) \end{aligned} \quad (\text{B.59})$$

by identifying the first integral on the second line as being the definition of the variance of $x \sim \mathcal{N}_{\mathbb{C}}(0, q)$ and the second integral as being the total probability. \square

A deterministic variable has $-\infty$ as differential entropy, while the next lemma proves that Gaussian random variables have the highest differential entropy among all variables with a given power.

Lemma B.21. Consider any continuous random variable $z \in \mathbb{C}$ with $\mathbb{E}\{|z|^2\} = q$. The differential entropy of z is upper bounded as

$$\mathcal{H}(z) \leq \log_2(e\pi q) \quad (\text{B.60})$$

with equality if and only if $z \sim \mathcal{N}_{\mathbb{C}}(0, q)$.

Proof. Denote the PDF of z as $g(z)$ and consider $x \sim \mathcal{N}_{\mathbb{C}}(0, q)$ with

the PDF $f(x)$. Notice that

$$\begin{aligned}
 \mathcal{H}(z) - \mathcal{H}(x) &= \int_{\mathbb{C}} \log_2(f(x))f(x)dx - \int_{\mathbb{C}} \log_2(g(z))g(z)dz \\
 &\stackrel{(a)}{=} \int_{\mathbb{C}} \log_2(f(x))g(x)dx - \int_{\mathbb{C}} \log_2(g(z))g(z)dz \\
 &\stackrel{(b)}{=} \int_{\mathbb{C}} \log_2\left(\frac{f(z)}{g(z)}\right)g(z)dz \\
 &\stackrel{(c)}{\leq} \log_2\left(\int_{\mathbb{C}} \frac{f(z)}{g(z)}g(z)dz\right) = \log_2(1) = 0
 \end{aligned} \tag{B.61}$$

where (a) follows from the fact that

$$\log_2(f(x)) = -\log_2(\pi q) + \frac{\log_2(e)}{q}|x|^2 \tag{B.62}$$

and thus $\mathbb{E}\{\log_2(f(x))\}$ has the same value when taking the expectation over any distribution of x having $\mathbb{E}\{|x|^2\} = q$, including the distribution given by the PDF $g(x)$. We obtain (b) by changing the name of the integration variable from x to z in the first integral. Next, (c) follows from applying Jensen's inequality in Lemma B.11 to the concave logarithmic function, which only gives equality if $f(x) = g(x)$. Finally, (B.60) follows from the expression for $\mathcal{H}(x)$ in Lemma B.20. \square

Motivated by this lemma, Gaussian distributed variables are often considered in communications, to convey a maximum amount of information under a given power constraint.

B.6 Basic Optimization Theory

This section provides the basic terminology, definitions, and classification of optimization problems. We refer to textbooks and survey articles such as [34, 66, 67] for further details and explanations.

The main purpose of optimization is to analyze a set of feasible solutions to a problem and determine which one that is most preferable, in terms of maximizing a given utility function. Let us denote the utility function as $f_0 : \mathbb{R}^V \rightarrow \mathbb{R}$, then the optimization problem can be

expressed as

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} \quad f_0(\mathbf{x}) \\ & \text{subject to } \mathbf{x} \in \mathcal{X} \end{aligned} \tag{B.63}$$

where the V -dimensional vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_V]^T \in \mathbb{R}^V$ is called the *optimization variable*. This variable can be selected from the feasible set \mathcal{X} of feasible solutions. It is usually assumed that \mathcal{X} is a compact set and that the utility function is continuously differentiable over this set. A feasible vector $\mathbf{x}^{\text{opt}} \in \mathcal{X}$ is called an *optimal solution* to (B.63) if it provides the largest utility among all feasible solutions; that is, $f_0(\mathbf{x}^{\text{opt}}) \geq f_0(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. The optimization problem (B.63) is feasible if the feasible set is non-empty, otherwise it is said to be infeasible.

To enable structured analysis and algorithmic development, it is convenient to write an optimization problem on the standard form

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} \quad f_0(\mathbf{x}) \\ & \text{subject to } f_n(\mathbf{x}) \leq 0 \quad n = 1, \dots, N \end{aligned} \tag{B.64}$$

where the N functions $f_n : \mathbb{R}^V \rightarrow \mathbb{R}$ are called the *constraint functions*. Any (constrained) optimization problem can be reformulated on the standard form [67], but the dimension of \mathbf{x} might change in the reformulation. For example, (B.64) is equivalent to (B.63) for

$$\mathcal{X} = \left\{ \mathbf{x} \in \mathbb{R}^V : f_n(\mathbf{x}) \leq 0 \quad n = 1, \dots, N \right\}. \tag{B.65}$$

The utility function and the constraint functions completely characterize an optimization problem that is on the standard form. It might not be necessary to solve (B.64) from scratch, but there are important classes of problems for which there are general-purpose algorithms that solve any instance of the class. The following are important classes in wireless communications:

- **Linear program:** f_0 and f_1, \dots, f_N are linear or affine functions. Note that a function $f_n : \mathbb{R}^V \rightarrow \mathbb{R}$ is *affine* if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^V$ and $t \in [0, 1]$, $f_n(t\mathbf{x}_1 + (1 - t)\mathbf{x}_2) = tf_n(\mathbf{x}_1) + (1 - t)f_n(\mathbf{x}_2)$.

- **Geometric program:** $-f_0$ and $f_1 - 1, \dots, f_N - 1$ are posynomial functions. Note that a function $f_n : \mathbb{R}_+^V \rightarrow \mathbb{R}$ is *posynomial* if it can be expressed as $f_n(\mathbf{x}) = \sum_{b=1}^B c_b x_1^{e_{1,b}} x_2^{e_{2,b}} \cdots x_V^{e_{V,b}}$ for some positive integer B , constants $c_b > 0$, and exponents $e_{1,b}, \dots, e_{V,b} \in \mathbb{R}$ for $b = 1, \dots, B$, where $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_V]^\top$ is a vector with non-negative elements.
- **Convex program:** $-f_0$ and f_1, \dots, f_N are convex functions. Note that a function $f_n : \mathbb{R}^V \rightarrow \mathbb{R}$ is *convex* if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^V$ and $t \in [0, 1]$, $f_n(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf_n(\mathbf{x}_1) + (1-t)f_n(\mathbf{x}_2)$.

These three classes represent successively more general conditions: every linear program is also convex and every geometric program can be transformed into a convex program by a standard change of variable. More precisely, for the optimization variable $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_V]^\top$ we can make the change of variable $x_v = e^{\bar{x}_v}$ for $v = 1, \dots, V$, which turns an arbitrary posynomial constraint $\sum_{b=1}^B c_b x_1^{e_{1,b}} x_2^{e_{2,b}} \cdots x_V^{e_{V,b}} \leq 1$ into

$$\begin{aligned} \sum_{b=1}^B c_b e^{\bar{x}_1 e_{1,b} + \bar{x}_2 e_{2,b} + \dots + \bar{x}_V e_{V,b}} &\leq 1 \\ \log_e \left(\sum_{b=1}^B c_b e^{\bar{x}_1 e_{1,b} + \bar{x}_2 e_{2,b} + \dots + \bar{x}_V e_{V,b}} \right) &\leq 0 \end{aligned} \quad (\text{B.66})$$

where the latter can be shown to be a convex constraint.

If the utility function is a constant or, equivalently, if there is no utility function at all, then the optimization problem is called a feasibility problem. The purpose of solving a feasibility problem is to find any point in the feasible set \mathcal{X} , which can be a far from trivial task.

Practical optimization problems can be difficult to classify and reformulation tricks are sometimes needed to reveal that a given optimization problem belongs to one of the three above-mentioned classes. There is no systematic way of identifying and extracting an underlying structure, but it is rather an art that includes making good changes of variables and relaxations [251]. A survey of reformulation tricks that are relevant to resource allocation in wireless communications is provided in [46].

Most optimization problems have no closed-form optimal solutions, but can still be solved numerically to any accuracy $\epsilon > 0$ on the optimal

value $f_0(\mathbf{x}^{\text{opt}})$. The problem classification enables the use of numerical algorithms developed for that particular class. For example, interior-point methods can be applied to linear, geometric, and convex programs with a polynomial worst-case complexity (under some mild conditions [67]). General-purpose implementations of interior-point methods are available in **SeDuMi** [307], **SDPT3** [316], and **MOSEK** [22]. The use of these implementations can be simplified by the high-level modeling frameworks **CVX** [132] and **Yalmip** [196]. Hence, for the purpose of this monograph, we will only classify optimization problems into one of the categories listed above and then we consider the solution to have been obtained. We used **CVX** and **MOSEK** when writing this monograph.

It is important to differentiate between the globally optimal point \mathbf{x}^{opt} (that maximizes the utility for all $\mathbf{x} \in \mathcal{X}$) and locally optimal points that provide the highest utility among the feasible points in their surroundings. Formally, a point $\bar{\mathbf{x}}$ is locally optimal if there exist $\epsilon > 0$ such that $f_0(\bar{\mathbf{x}}) \geq f_0(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ satisfying $\|\bar{\mathbf{x}} - \mathbf{x}\|_2 < \epsilon$.

As noted in [278], there is a great watershed between convex and non-convex programs; every locally optimal solution to a convex program is also globally optimal, while this is not the case for general non-convex programs [67]. Therefore, the entire feasible set \mathcal{X} basically needs to be searched when solving non-convex programs, which corresponds to a complexity that grows exponentially (or faster) with the number of optimization variables and/or constraints. Practical algorithms for non-convex programs are often designed to only search for locally optimal points, which might be achieved with manageable complexity using sequential convex approximations [206]. In case the globally optimal solution is needed, there are also algorithms that can find it. Many non-convex programs in the area of wireless communications belong to the alternative class of monotonic programs [46] and can be solved by the polyblock outer approximation algorithm [317] or the branch-reduce-and-bound algorithm [318]. These algorithms are guaranteed to find the globally optimal solution but have exponential worst-case complexity with respect to the number of optimization variables, thus they are mainly useful for small problems.

C

Collection of Proofs

C.1 Proofs in Section 1

C.1.1 Proof of Corollary 1.2

Consider the case when h is deterministic and known at the output of the channel. In the capacity expression (1.2), the conditional entropy becomes $\mathcal{H}(y|x) = \mathcal{H}(n)$ since only the additive noise is unknown at the output. From Lemma B.20 on p. 574, we have

$$\mathcal{H}(n) = \log_2(e\pi\sigma^2). \quad (\text{C.1})$$

To compute $\mathcal{H}(y)$ and take the supremum, we notice that the power of the output signal is $\mathbb{E}\{|y|^2\} = |h|^2\mathbb{E}\{|x|^2\} + \mathbb{E}\{|n|^2\} \leq |h|^2p + \sigma^2$, with equality for any input distribution $f(x)$ that satisfies $\mathbb{E}\{|x|^2\} = p$. From Lemma B.21 on p. 574, we know that the entropy of a random variable with a given power is maximized when that variable has a circularly symmetric complex Gaussian distribution with as large variance as possible. This is achieved for y by selecting $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$. For this entropy-maximizing input distribution, we have

$$\mathcal{H}(y) = \log_2 \left(e\pi\mathbb{E}\{|y|^2\} \right) = \log_2 \left(e\pi(|h|^2p + \sigma^2) \right) \quad (\text{C.2})$$

which yields

$$C = \log_2 \left(e\pi(|h|^2 p + \sigma^2) \right) - \log_2(e\pi\sigma^2) = \log_2 \left(1 + \frac{p|h|^2}{\sigma^2} \right). \quad (\text{C.3})$$

This corresponds to (1.4) and finishes the proof for the case when h is deterministic. In the case when h is a realization of the independent random variable \mathbb{H} , the output of the channel is (y, \mathbb{H}) and thus the general capacity expression in (1.2) becomes

$$\sup_{f(x)} (\mathcal{H}(y, \mathbb{H}) - \mathcal{H}(y, \mathbb{H}|x)) = \sup_{f(x)} (\mathbb{E} \{ \mathcal{H}(y, \mathbb{H} = h) - \mathcal{H}(y, \mathbb{H} = h|x) \}) \quad (\text{C.4})$$

where the equality follows from conditioning on an arbitrary realization h of \mathbb{H} and taking the expected value with respect to this realization. The expression inside the expectation considers a deterministic value of h and can thus be computed and maximized with respect to x as done earlier in this proof. Since the same input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$ is optimal irrespective of the realization of h , the capacity expression in (C.4) becomes (1.5).

C.1.2 Proof of Corollary 1.3

To prove this result, we consider the following equivalent way of expressing the channel capacity [297]:

$$C = \sup_{f(x)} (\mathcal{H}(x) - \mathcal{H}(x|y)). \quad (\text{C.5})$$

We begin by considering the case when h and p_v are deterministic. A lower bound on the capacity in (C.5) is computed by making three suboptimal assumptions. The first assumption is $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$, which might not be the optimal input distribution and gives the lower bound

$$C \geq \mathcal{H}(x) - \mathcal{H}(x|y) \quad (\text{C.6})$$

where $\mathcal{H}(x) = \log_2(e\pi p)$ for the input distribution (see Lemma B.20 on p. 572). It remains to compute the conditional differential entropy $\mathcal{H}(x|y)$. The second suboptimal assumption is that the input x is estimated from y using an LMMSE estimator as described in Lemma B.19

on p. 571:

$$\hat{x} = \frac{\mathbb{E}\{xy^*\}}{\mathbb{E}\{|y|^2\}}y \quad (\text{C.7})$$

where

$$\mathbb{E}\{xy^*\} = \mathbb{E}\{xx^*h^*\} + \mathbb{E}\{xv^*\} + \mathbb{E}\{xn^*\} = ph^* \quad (\text{C.8})$$

since the noise and input are independent (and have zero mean) and because $\mathbb{E}\{xv^*\} = 0$ by assumption. Moreover,

$$\mathbb{E}\{|y|^2\} = p|h|^2 + p_v + \sigma^2 \quad (\text{C.9})$$

by also utilizing the independence between the noise and the interference, and the assumption $\mathbb{E}\{v\} = 0$. Note that all expectations are with respect to x , v , and n . The MSE of the LMMSE estimator is

$$\text{MSE}_x = \mathbb{E}\{|x|^2\} - \frac{|\mathbb{E}\{xy^*\}|^2}{\mathbb{E}\{|y|^2\}} = p - \frac{p^2|h|^2}{p|h|^2 + p_v + \sigma^2}. \quad (\text{C.10})$$

This is exploited to upper bound the conditional differential entropy as

$$\mathcal{H}(x|y) \stackrel{(a)}{=} \mathcal{H}(x - \hat{x}|y) \stackrel{(b)}{\leq} \mathcal{H}(x - \hat{x}) \stackrel{(c)}{\leq} \log_2(e\pi\text{MSE}_x) \quad (\text{C.11})$$

where (a) follows from subtracting the known LMMSE estimate (which is a constant when y is known and thus does not change the entropy [94]) and (b) follows from removing the remaining information in y (which does not reduce the entropy). The random variable $x - \hat{x}$ has zero mean and variance MSE_x , thus (c) follows from Lemma B.21 on p. 574, which says that the largest entropy is achieved when the random variable is complex Gaussian distributed. This is this is the third suboptimal step.

In summary, we have the lower bound

$$\begin{aligned} C &\geq \log_2(e\pi p) - \log_2(e\pi\text{MSE}_x) = -\log_2\left(1 - \frac{p|h|^2}{p|h|^2 + p_v + \sigma^2}\right) \\ &= \log_2\left(\frac{p|h|^2 + p_v + \sigma^2}{p_v + \sigma^2}\right) = \log_2\left(1 + \frac{p|h|^2}{p_v + \sigma^2}\right) \end{aligned} \quad (\text{C.12})$$

which is the expression in (1.9).

Suppose h is instead a realization of the random variable \mathbb{H} and the conditional interference variance $p_v(h, u)$ depends on a realization u of

the random variable \mathbb{U} . It is assumed that \mathbb{H} and \mathbb{U} are known at the output, which means that $(y, \mathbb{H}, \mathbb{U})$ is the output of the channel. The capacity C can then be lower bounded as

$$\begin{aligned} C &\geq \mathcal{H}(x) - \mathcal{H}(x|y, \mathbb{H}, \mathbb{U}) \\ &= \log_2(e\pi p) - \mathbb{E}\{\mathcal{H}(x|y, \mathbb{H} = h, \mathbb{U} = u)\} \end{aligned} \quad (\text{C.13})$$

by first assuming the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$ and then conditioning on particular realizations of \mathbb{H} and \mathbb{U} . The expectation in (C.13) is with respect to the realizations h and u . The conditional expression inside the expectation considers given realizations h and u , respectively, thus we can apply the same bounding technique as (C.7)–(C.9), by letting all expectations be conditioned on h, u . The derivation then requires conditionally independent noise, $\mathbb{E}\{v|h, u\} = 0$, and $\mathbb{E}\{x^*v|h, u\} = 0$. Using the notation $p_v(h, u) = \mathbb{E}\{|v|^2|h, u\}$ of the conditional variance, the final expression in (1.10) follows accordingly.

C.1.3 Proof of Lemma 1.4

This lemma considers a special case of the channel in Corollary 1.3 on p. 171 with $h = h_0^0$ and $v = \sqrt{p}h_1^0$. Since the LoS channels are deterministic, an achievable SE is obtained directly from (1.9) and becomes

$$\text{SE}_0^{\text{LoS}} = \log_2 \left(1 + \frac{p\beta_0^0}{p\beta_1^0 + \sigma^2} \right). \quad (\text{C.14})$$

This turns into the SE expression in (1.17) by using the SNR definition in (1.13) and the definition of $\bar{\beta}$ in (1.12).

The NLoS channels are random and thus an achievable SE is instead obtained by (1.10) in Corollary 1.3 with $\mathbb{H} = h_0^0$, $\mathbb{U} = h_1^0$, and $p_v = p|h_i^0|^2$:

$$\mathbb{E} \left\{ \log_2 \left(1 + \frac{p|h_0^0|^2}{p|h_1^0|^2 + \sigma^2} \right) \right\}. \quad (\text{C.15})$$

This expectation is further divided as

$$\mathbb{E} \left\{ \log_2 \left(1 + \sum_{i=0}^1 \frac{p|h_i^0|^2}{\sigma^2} \right) \right\} - \mathbb{E} \left\{ \log_2 \left(1 + \frac{p|h_1^0|^2}{\sigma^2} \right) \right\} \quad (\text{C.16})$$

and both expectations need to be computed with respect to the random channel responses. Next, we will use the identity

$$\mathbb{E} \left\{ \log_2 \left(1 + \sum_{i=1}^L |b_i|^2 \right) \right\} = \sum_{i=1}^L \frac{e^{\frac{1}{\mu_i}} E_1 \left(\frac{1}{\mu_i} \right)}{\log_e(2) \prod_{\substack{l=1 \\ l \neq i}}^L \left(1 - \frac{\mu_l}{\mu_i} \right)} \quad (\text{C.17})$$

for independent variables $b_i \sim \mathcal{N}_{\mathbb{C}}(0, \mu_i)$ with distinct values of μ_i , for $i = 1, \dots, L$, from [61, Lemma 3]. By setting $\mu_1 = \text{SNR}_0 \bar{\beta}$ and $\mu_2 = \text{SNR}_0$, we can compute the two terms in (C.16) as

$$\mathbb{E} \left\{ \log_2 \left(1 + \sum_{i=0}^1 \frac{p|h_i^0|^2}{\sigma^2} \right) \right\} = \frac{e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1 \left(\frac{1}{\text{SNR}_0 \bar{\beta}} \right)}{\log_e(2) \left(1 - \bar{\beta}^{-1} \right)} + \frac{e^{\frac{1}{\text{SNR}_0}} E_1 \left(\frac{1}{\text{SNR}_0} \right)}{\log_e(2) \left(1 - \bar{\beta} \right)} \quad (\text{C.18})$$

$$\mathbb{E} \left\{ \log_2 \left(1 + \frac{p|h_1^0|^2}{\sigma^2} \right) \right\} = \frac{e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1 \left(\frac{1}{\text{SNR}_0 \bar{\beta}} \right)}{\log_e(2)}. \quad (\text{C.19})$$

By inserting (C.18) and (C.19) into (C.16), we obtain the SE expression in (1.18). Since the identity from [61, Lemma 3] only holds in the case when μ_1 and μ_2 are different, we need to exclude the case $\bar{\beta} = 1$ from this lemma.

C.1.4 Proof of Lemma 1.5

The MR processed received signal $\mathbf{v}_0^H \mathbf{y}_0$ in (1.25) is a scalar channel of the same type as (1.14), thus the SE expressions are obtained by the same methodology as in the proof of Lemma 1.4. The details are provided below.

In the deterministic LoS case, the SE expression

$$\log_2 \left(1 + \frac{p\|\mathbf{h}_0^0\|^4}{p|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2 + \sigma^2 \|\mathbf{h}_0^0\|^2} \right) = \log_2 \left(1 + \frac{p\|\mathbf{h}_0^0\|^2}{p \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} + \sigma^2} \right) \quad (\text{C.20})$$

is obtained from Corollary 1.3. We observe that $\|\mathbf{h}_0^0\|^2 = \beta_0^0 M$ since each element of the channel response in (1.23) has the squared magnitude

β_0^0 . Moreover, we notice that

$$\begin{aligned} (\mathbf{h}_0^0)^H \mathbf{h}_1^0 &= \sqrt{\beta_0^0 \beta_1^0} \sum_{m=0}^{M-1} \left(e^{2\pi j d_H (\sin(\varphi_1^0) - \sin(\varphi_0^0))} \right)^m \\ &= \begin{cases} \sqrt{\beta_0^0 \beta_1^0} \frac{1 - e^{2\pi j d_H M (\sin(\varphi_1^0) - \sin(\varphi_0^0))}}{1 - e^{2\pi j d_H (\sin(\varphi_1^0) - \sin(\varphi_0^0))}} & \text{if } \sin(\varphi_0^0) \neq \sin(\varphi_1^0) \\ \sqrt{\beta_0^0 \beta_1^0} M & \text{if } \sin(\varphi_0^0) = \sin(\varphi_1^0) \end{cases} \end{aligned} \quad (\text{C.21})$$

where the first equality uses the LoS definition in (1.23) and the second equality utilizes the geometric series formula $\sum_{m=0}^{M-1} x^m = \frac{1-x^M}{1-x}$ for $x \neq 1$ and $\sum_{m=0}^{M-1} x^m = M$ for $x = 1$. By further utilizing the fact that

$$\begin{aligned} &\left| \frac{1 - e^{2\pi j d_H M (\sin(\varphi_1^0) - \sin(\varphi_0^0))}}{1 - e^{2\pi j d_H (\sin(\varphi_1^0) - \sin(\varphi_0^0))}} \right|^2 \\ &= \left| \frac{e^{\pi j d_H M (\sin(\varphi_1^0) - \sin(\varphi_0^0))}}{e^{\pi j d_H (\sin(\varphi_1^0) - \sin(\varphi_0^0))}} \frac{\sin(\pi d_H M (\sin(\varphi_1^0) - \sin(\varphi_0^0)))}{\sin(\pi d_H (\sin(\varphi_1^0) - \sin(\varphi_0^0)))} \right|^2 \\ &= \frac{\sin^2(\pi d_H M (\sin(\varphi_1^0) - \sin(\varphi_0^0)))}{\sin^2(\pi d_H (\sin(\varphi_1^0) - \sin(\varphi_0^0)))} \end{aligned} \quad (\text{C.22})$$

by Euler's formulas, we notice that

$$\frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} = \beta_1^0 g(\varphi_0^0, \varphi_1^0) \quad (\text{C.23})$$

where the function $g(\cdot, \cdot)$ is defined in (1.28) and $\|\mathbf{h}_0^0\|^2 = \beta_0^0 M$. Inserting this expression into (C.20) and dividing all terms in the SINR by $p\beta_0^0$, the SE expression in (1.27) is obtained.

For the random NLoS channels, Corollary 1.3 provides the SE expression

$$\begin{aligned} &\mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_0^0\|^4}{p |(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2 + \sigma^2 \|\mathbf{h}_0^0\|^2} \right) \right\} \\ &= \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_0^0\|^2}{p \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} + \sigma^2} \right) \right\} \end{aligned} \quad (\text{C.24})$$

which is further separated into two expectations as

$$\begin{aligned} & \mathbb{E} \left\{ \log_2 \left(1 + \frac{p}{\sigma^2} \|\mathbf{h}_0^0\|^2 + \frac{p}{\sigma^2} \left| \frac{(\mathbf{h}_0^0)^H}{\|\mathbf{h}_0^0\|} \mathbf{h}_1^0 \right|^2 \right) \right\} \\ & - \mathbb{E} \left\{ \log_2 \left(1 + \frac{p}{\sigma^2} \left| \frac{(\mathbf{h}_0^0)^H}{\|\mathbf{h}_0^0\|} \mathbf{h}_1^0 \right|^2 \right) \right\}. \end{aligned} \quad (\text{C.25})$$

Next, we note that

$$\sqrt{\frac{p}{\sigma^2}} \frac{(\mathbf{h}_0^0)^H}{\|\mathbf{h}_0^0\|} \mathbf{h}_1^0 \sim \mathcal{N}_{\mathbb{C}}(0, \text{SNR}_0 \bar{\beta}) \quad (\text{C.26})$$

since $\mathbf{h}_0^0/\|\mathbf{h}_0^0\|$ is uniformly distributed over the unit sphere in \mathbb{C}^M and thereby merely projects the M -variate random variable $\sqrt{p/\sigma^2} \mathbf{h}_1^0 \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \text{SNR}_0 \bar{\beta} \mathbf{I}_M)$ to a single-variate random variable with the same variance and distribution. The second expectation in (C.25) can thus be computed by using the identity in (C.17) from [61, Lemma 3]:

$$\mathbb{E} \left\{ \log_2 \left(1 + \frac{p}{\sigma^2} \left| \frac{(\mathbf{h}_0^0)^H}{\|\mathbf{h}_0^0\|} \mathbf{h}_1^0 \right|^2 \right) \right\} = \frac{e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1 \left(\frac{1}{\text{SNR}_0 \bar{\beta}} \right)}{\log_e(2)}. \quad (\text{C.27})$$

Next, we utilize the fact that $y = \frac{p}{\sigma^2} \|\mathbf{h}_0^0\|^2 + \frac{p}{\sigma^2} \left| \frac{(\mathbf{h}_0^0)^H}{\|\mathbf{h}_0^0\|} \mathbf{h}_1^0 \right|^2$ is the sum of independent exponentially distributed random variables: M variables with variance SNR_0 from the first term and one variable with variance $\text{SNR}_0 \bar{\beta}$ from the second term. The first expectation in (C.25) is then obtained by using Lemma B.15 on p. 565 for the case of $N_a = M$, $\mu_a = \text{SNR}_0$, $N_b = 1$, and $\mu_b = \text{SNR}_0 \bar{\beta}$:

$$\begin{aligned} & \mathbb{E} \left\{ \log_2 \left(1 + \frac{p}{\sigma^2} \|\mathbf{h}_0^0\|^2 + \frac{p}{\sigma^2} \left| \frac{(\mathbf{h}_0^0)^H}{\|\mathbf{h}_0^0\|} \mathbf{h}_1^0 \right|^2 \right) \right\} \\ &= \sum_{m=1}^M \sum_{l=0}^{M-m} \frac{(-1)^{M-l+1}}{\left(\frac{1}{\text{SNR}_0 \bar{\beta}} - \frac{1}{\text{SNR}_0} \right)^m} \frac{\left(e^{\frac{1}{\text{SNR}_0}} E_1 \left(\frac{1}{\text{SNR}_0} \right) + \sum_{n=1}^l \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{j! \text{SNR}_0^j} \right)}{(M-m-l)! \text{SNR}_0^{M-l} \bar{\beta} \log_e(2)} \\ &+ \frac{1}{\left(\frac{1}{\text{SNR}_0} - \frac{1}{\text{SNR}_0 \bar{\beta}} \right)^M} \frac{e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1 \left(\frac{1}{\text{SNR}_0 \bar{\beta}} \right)}{\text{SNR}_0^M \log_e(2)} \end{aligned}$$

$$\begin{aligned}
&= \sum_{m=1}^M \sum_{l=0}^{M-m} \frac{(-1)^{M-m-l+1}}{\left(1 - \frac{1}{\bar{\beta}}\right)^m} \frac{\left(e^{\frac{1}{\text{SNR}_0}} E_1\left(\frac{1}{\text{SNR}_0}\right) + \sum_{n=1}^l \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{j! \text{SNR}_0^j}\right)}{(M-m-l)! \text{SNR}_0^{M-m-l} \bar{\beta} \log_e(2)} \\
&\quad + \frac{1}{\left(1 - \frac{1}{\bar{\beta}}\right)^M} \frac{e^{\frac{1}{\text{SNR}_0 \bar{\beta}}} E_1\left(\frac{1}{\text{SNR}_0 \bar{\beta}}\right)}{\log_e(2)}. \tag{C.28}
\end{aligned}$$

Substituting (C.27) and (C.28) into (C.25) yields the SE expression in (1.29).

C.1.5 Proof of Corollary 1.6

The lower bound in (1.32) is obtained by applying Jensen's inequality (see Lemma B.11 on p. 563) to the convex function $\log_2(1 + 1/x)$ as

$$\begin{aligned}
&\mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_0^0\|^4}{p |(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2 + \sigma^2 \|\mathbf{h}_0^0\|^2} \right) \right\} \\
&= \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_0^0\|^2}{p \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} + \sigma^2} \right) \right\} \\
&\geq \log_2 \left(1 + \left(\mathbb{E} \left\{ \frac{p \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} + \sigma^2}{p \|\mathbf{h}_0^0\|^2} \right\} \right)^{-1} \right) \\
&= \log_2 \left(1 + \left(\frac{p \beta_1^0 + \sigma^2}{p(M-1)\beta_0^0} \right)^{-1} \right) \tag{C.29}
\end{aligned}$$

where $x = \frac{p \frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2} + \sigma^2}{p \|\mathbf{h}_0^0\|^2}$. The last equality follows from the facts that $\frac{|(\mathbf{h}_0^0)^H \mathbf{h}_1^0|^2}{\|\mathbf{h}_0^0\|^2}$ is independent of \mathbf{h}_0^0 , has mean value β_1^0 , and that

$$\mathbb{E} \left\{ \frac{1}{\|\mathbf{h}_0^0\|^2} \right\} = \frac{1}{(M-1)\beta_0^0} \tag{C.30}$$

since $\frac{2}{\beta_0^0} \|\mathbf{h}_0^0\|^2 \sim \chi^2(2M)$ and thus the expectation in (C.30) can be computed from standard results of the inverse- χ^2 distribution; see [315, Lemma 2.10].

C.1.6 Proof of Lemma 1.7

The MR processed received signal $\mathbf{v}_{0k}^H \mathbf{y}_0$ in (1.40) for UE k is a scalar channel of the same type as (1.14), thus achievable SE expressions of this UE can be obtained from Corollary 1.3 on p. 171 (using the same methodology as in the proof of Lemma 1.4). The sum SE is the summation over the K UEs' SEs. The details are provided below.

In the deterministic LoS case, the SE expression from Corollary 1.3 becomes

$$\log_2 \left(1 + \frac{p \|\mathbf{h}_{0k}^0\|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sum_{i=1}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{1i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sigma^2} \right) \quad (\text{C.31})$$

for UE k by dividing each term in the SINR by $\|\mathbf{h}_{0k}^0\|^2$. We notice that $\|\mathbf{h}_{0k}^0\|^2 = \beta_0^0 M$ since each element of the channel response in (1.38) has squared magnitude β_0^0 . Moreover, one can prove that

$$\frac{\left|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{ji}^0\right|^2}{\|\mathbf{h}_{0k}^0\|^2} = \beta_j^0 g\left(\varphi_{0k}^0, \varphi_{ji}^0\right) \quad (\text{C.32})$$

for $j = 0, 1$ and $i = 1, \dots, K$, by following the same approach as in the proof of Lemma 1.5. By summing up (C.31) for all UEs in the cell, we arrive at (1.43).

For NLoS channels, the channels are random and thus Corollary 1.3 (with $\mathbb{H} = \|\mathbf{h}_{0k}^0\|^2$ and \mathbb{U} containing $|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{ji}^0|^2$ for all j, i) provides the achievable SE

$$\mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_{0k}^0\|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sum_{i=1}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{1i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sigma^2} \right) \right\} \quad (\text{C.33})$$

for UE k by dividing each term in the SINR by $\|\mathbf{h}_{0k}^0\|^2$. We begin by computing the expectation of the inverse of the SINR in (C.33), which

is

$$\begin{aligned} & \mathbb{E} \left\{ \frac{\sum_{i=1, i \neq k}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sum_{i=1}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sigma^2}{p \|\mathbf{h}_{0k}^0\|^2} \right\} \\ &= \frac{p(K-1)\beta_0^0 + pK\beta_1^0 + \sigma^2}{p(M-1)\beta_0^0} = \frac{(K-1) + K\bar{\beta} + \frac{1}{\text{SNR}_0}}{M-1} \end{aligned} \quad (\text{C.34})$$

by utilizing the facts that $\frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0j}^0|^2}{\|\mathbf{h}_{0k}^0\|^2}$ is independent of \mathbf{h}_{0k}^0 , has mean value β_j^0 (whenever $(j, i) \neq (0, k)$), and

$$\mathbb{E} \left\{ \frac{1}{\|\mathbf{h}_{0k}^0\|^2} \right\} = \frac{1}{(M-1)\beta_0^0} \quad (\text{C.35})$$

since $\frac{2}{\beta_0^0} \|\mathbf{h}_{0k}^0\|^2 \sim \chi^2(2M)$ and thus the expectation in (C.35) can be computed from standard results of the inverse- χ^2 distribution; see [315, Lemma 2.10].

We now utilize (C.34) to compute the lower bound in (1.44) by applying Jensen's inequality (Lemma B.11 on p. 563) to (C.33) as

$$\begin{aligned} & \sum_{k=1}^K \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \|\mathbf{h}_{0k}^0\|^2}{\sum_{i=1, i \neq k}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sum_{i=1}^K p \frac{|(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0|^2}{\|\mathbf{h}_{0k}^0\|^2} + \sigma^2} \right) \right\} \\ & \geq \sum_{k=1}^K \log_2 \left(1 + \left(\frac{(K-1) + K\bar{\beta} + \frac{1}{\text{SNR}_0}}{M-1} \right)^{-1} \right) \end{aligned} \quad (\text{C.36})$$

where we utilized the convexity of $\log_2(1+1/x)$ with respect to x (which is the inverse SINR in this case). The expression in (1.44) follows from (C.36) by noting that the SE bound is the same for all K UEs.

C.1.7 Proof of Lemma 1.8

The transmit precoding reduces the MISO channels to the effective scalar channel in (1.45). This channel is of the same type as (1.14), thus

the SE expressions are obtained by the same methodology as in the proof of Lemma 1.4 and Lemma 1.7. The details are provided below.

In the deterministic LoS case, the SE expression

$$\text{SE}_0^{\text{LoS}} = \sum_{k=1}^K \log_2 \left(1 + \frac{p \left| (\mathbf{h}_{0k}^0)^H \frac{\mathbf{h}_{0k}^0}{\|\mathbf{h}_{0k}^0\|} \right|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K p \left| (\mathbf{h}_{0k}^0)^H \frac{\mathbf{h}_{0i}^0}{\|\mathbf{h}_{0i}^0\|} \right|^2 + \sum_{i=1}^K p \left| (\mathbf{h}_{0k}^1)^H \frac{\mathbf{h}_{1i}^1}{\|\mathbf{h}_{1i}^1\|} \right|^2 + \sigma^2} \right). \quad (\text{C.37})$$

is obtained from Corollary 1.3 on p. 171 with $h = \|\mathbf{h}_{0k}^0\|$ and v being the sum of all interference terms. The expression in (1.49) is then obtained by utilizing the facts that $\|\mathbf{h}_{0k}^0\|^2 = \beta_0^0 M$ and $|(\mathbf{h}_{0k}^j)^H \mathbf{h}_{ji}^j|^2 / \|\mathbf{h}_{ji}^j\|^2 = \beta_0^j g(\varphi_{ji}^j, \varphi_{0k}^j)$ for $j = 0, 1$.

The NLoS channels are random and in this case an achievable SE is instead obtained by (1.5) in Corollary 1.3 (with $\mathbb{H} = \|\mathbf{h}_{0k}^0\|$ and \mathbb{U} containing $\frac{|(\mathbf{h}_{0k}^j)^H \mathbf{h}_{ji}^j|^2}{\|\mathbf{h}_{ji}^j\|^2}$ for all j, i):

$$\text{SE}_0^{\text{NLoS}} = \sum_{k=1}^K \mathbb{E} \left\{ \log_2 \left(1 + \frac{p \left| (\mathbf{h}_{0k}^0)^H \frac{\mathbf{h}_{0k}^0}{\|\mathbf{h}_{0k}^0\|} \right|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K p \left| (\mathbf{h}_{0k}^0)^H \frac{\mathbf{h}_{0i}^0}{\|\mathbf{h}_{0i}^0\|} \right|^2 + \sum_{i=1}^K p \left| (\mathbf{h}_{0k}^1)^H \frac{\mathbf{h}_{1i}^1}{\|\mathbf{h}_{1i}^1\|} \right|^2 + \sigma^2} \right) \right\}. \quad (\text{C.38})$$

We begin by computing the expectation of the inverse of the SINR in

(C.38), which is

$$\begin{aligned}
& \mathbb{E} \left\{ \frac{\sum_{i=1, i \neq k}^K p \left| (\mathbf{h}_{0k}^0)^H \frac{\mathbf{h}_{0i}^0}{\|\mathbf{h}_{0i}^0\|} \right|^2 + \sum_{i=1}^K p \left| (\mathbf{h}_{0k}^1)^H \frac{\mathbf{h}_{1i}^1}{\|\mathbf{h}_{1i}^1\|} \right|^2 + \sigma^2}{p \|\mathbf{h}_{0k}^0\|^2} \right\} \\
&= \sum_{i=1, i \neq k}^K \mathbb{E} \left\{ \left| \frac{(\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0}{\|\mathbf{h}_{0k}^0\| \|\mathbf{h}_{0i}^0\|} \right|^2 \right\} \\
&\quad + \left(\sum_{i=1}^K \mathbb{E} \left\{ \left| \frac{(\mathbf{h}_{0k}^1)^H \mathbf{h}_{1i}^1}{\|\mathbf{h}_{1i}^1\|} \right|^2 \right\} + \frac{\sigma^2}{p} \right) \mathbb{E} \left\{ \frac{1}{\|\mathbf{h}_{0k}^0\|^2} \right\} \\
&= \frac{K-1}{M} + \frac{K\beta_0^1 + \frac{\sigma^2}{p}}{(M-1)\beta_0^0} \tag{C.39}
\end{aligned}$$

where the first expectation is computed by utilizing the fact that $\left| (\mathbf{h}_{0k}^0)^H \mathbf{h}_{0i}^0 \right|^2 / (\|\mathbf{h}_{0k}^0\|^2 \|\mathbf{h}_{0i}^0\|^2)$ is beta-distributed with parameters 1 and M [162], which implies that the expectation is $1/M$. The second expectation follows from the fact that $\frac{(\mathbf{h}_{0k}^1)^H \mathbf{h}_{1i}^1}{\|\mathbf{h}_{1i}^1\|} \sim \mathcal{N}_{\mathbb{C}}(0, \beta_0^1)$ and the third expectation follows from (C.35).

By applying Jensen's inequality (Lemma B.11 on p. 563) in the same way as in (C.36), we obtain that (C.37) is lower bounded by

$$\begin{aligned}
& \sum_{k=1}^K \log_2 \left(1 + \left(\frac{K-1}{M} + \frac{K\beta_0^1 + \frac{\sigma^2}{p}}{(M-1)\beta_0^0} \right)^{-1} \right) \\
&= K \log_2 \left(1 + \frac{(M-1)}{(K-1)\frac{M-1}{M} + K\bar{\beta} + \frac{1}{\text{SNR}_0}} \right). \tag{C.40}
\end{aligned}$$

This is the final expression provided in (1.50).

C.2 Proofs in Section 3

C.2.1 Proof of Theorem 3.1 and Corollary 3.2

The received pilot signal in (3.1) can be separated into two terms:

$$\mathbf{Y}_j^p = \mathbf{Y}_j^p \left(\frac{1}{\tau_p} \phi_{li}^* \phi_{li}^T \right) + \mathbf{Y}_j^p \left(\mathbf{I}_{\tau_p} - \frac{1}{\tau_p} \phi_{li}^* \phi_{li}^T \right). \quad (\text{C.41})$$

The first part is the orthogonal projection onto the subspace spanned by pilot sequence of UE i in cell l and the second part is the projection onto the orthogonal complement. By utilizing the assumption of a pilot book with orthogonal sequences, the first term in (C.41) becomes

$$\begin{aligned} \mathbf{Y}_j^p \left(\frac{1}{\tau_p} \phi_{li}^* \phi_{li}^T \right) &= \frac{1}{\tau_p} \left(\sum_{(l', i') \in \mathcal{P}_{li}} \sqrt{p_{l'i'}} \tau_p \mathbf{h}_{l'i'}^j + \mathbf{N}_j^p \phi_{li}^* \right) \phi_{li}^T \\ &= \frac{1}{\tau_p} \mathbf{y}_{jli}^p \phi_{li}^T \end{aligned} \quad (\text{C.42})$$

and second term becomes

$$\begin{aligned} &\mathbf{Y}_j^p \left(\mathbf{I}_{\tau_p} - \frac{1}{\tau_p} \phi_{li}^* \phi_{li}^T \right) \\ &= \sum_{(l', i') \notin \mathcal{P}_{li}} \sqrt{p_{l'i'}} \mathbf{h}_{l'i'}^j \phi_{l'i'}^T + \mathbf{N}_j^p \left(\mathbf{I}_{\tau_p} - \frac{1}{\tau_p} \phi_{li}^* \phi_{li}^T \right). \end{aligned} \quad (\text{C.43})$$

These terms are independent random variables, since they involve disjoint subsets of the UEs' channels and orthogonal projections of the noise matrix. Since only the first term is dependent on \mathbf{h}_{li}^j , we can use [175, Theorem 5.1] to conclude that $\frac{1}{\tau_p} \mathbf{y}_{jli}^p \phi_{li}^T$ is a sufficient statistic for estimating \mathbf{h}_{li}^j . Furthermore, \mathbf{y}_{jli}^p and $\frac{1}{\tau_p} \mathbf{y}_{jli}^p \phi_{li}^T$ are related through a deterministic and nondestructive transformation, making

$$\mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^* = \underbrace{\sqrt{p_{li}} \tau_p \mathbf{h}_{li}^j}_{\text{Desired pilot}} + \underbrace{\sum_{(l', i') \in \mathcal{P}_{li} \setminus (l, i)} \sqrt{p_{l'i'}} \tau_p \mathbf{h}_{l'i'}^j}_{\text{Interfering pilots}} + \underbrace{\mathbf{N}_j^p \phi_{li}^*}_{\text{Noise}} \quad (\text{C.44})$$

a sufficient statistic as well. Next, we notice that $\mathbf{y} = \mathbf{y}_{jli}^p$ in (C.44) matches the structure in Corollary B.18 on p. 570 with $q = \sqrt{p_{li}} \tau_p$,

$\mathbf{R} = \mathbf{R}_{li}^j$, and $\mathbf{S} = \sum_{(l', i') \in \mathcal{P}_{li} \setminus (l, i)} p_{l'i'} (\tau_p)^2 \mathbf{R}_{l'i'}^j + \tau_p \sigma_{\text{UL}}^2 \mathbf{I}_{M_j}$. The MMSE estimator in (3.9) and the estimation error correlation/covariance matrix in (3.11) then follow directly from Corollary B.18. The final expressions given in Theorem 3.1 are obtained by dividing with τ_p at the inside and in front of the matrix inverse.

Since \mathbf{y}_{jli}^p is circularly symmetric complex Gaussian distributed with zero mean and correlation matrix

$$\mathbb{E}\{\mathbf{y}_{jli}^p (\mathbf{y}_{jli}^p)^H\} = \tau_p \left(\sum_{(l', i') \in \mathcal{P}_{li}} p_{l'i'} \tau_p \mathbf{R}_{l'i'}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right) \quad (\text{C.45})$$

it follows that also the estimate is zero-mean complex Gaussian distributed. The correlation matrix $\mathbf{R}_{li}^j - \mathbf{C}_{li}^j$, stated in Corollary 3.2, is obtained from direct computation of $\mathbb{E}\{\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H\}$ using (3.9) and (C.45). Next, we utilize the fact that the MMSE estimate and the estimation error are jointly Gaussian and uncorrelated, where the latter follows from the orthogonality principle [175, Chapter 12], to conclude that the estimate and the error are independent and jointly Gaussian distributed.

C.3 Proofs in Section 4

C.3.1 Proof of Theorem 4.1

The received signal in (4.1) matches the discrete memoryless channel in Corollary 1.3 on p. 171 with a random channel response $h = \mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j$, the input $x = s_{jk}$, the output $y = \mathbf{v}_{jk}^H \mathbf{y}_j$, and $u = \{\hat{\mathbf{h}}_{li}^j\}$ as the random realization that affects the conditional variance of the interference. Using the notation from Corollary 1.3, the noise term is zero (i.e., $\sigma^2 = 0$) since $\mathbf{v}_{jk}^H \mathbf{n}_j$ is not necessarily Gaussian distributed and depends on the realization of \mathbf{v}_{jk} . The interference term in the corollary is

$$v = \mathbf{v}_{jk}^H \tilde{\mathbf{h}}_{jk}^j s_{jk} + \sum_{\substack{i=1 \\ i \neq k}}^{K_j} \mathbf{v}_{jk}^H \mathbf{h}_{ji}^j s_{ji} + \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j s_{li} + \mathbf{v}_{jk}^H \mathbf{n}_j. \quad (\text{C.46})$$

The value of h and the (conditional) variance of v are constant within a coherence block, but fluctuates between coherence blocks. In particular, in a given coherence block, the receiving BS j knows the current realization of the channel estimates $\hat{\mathbf{h}}_{li}^j$ for all l and i : $u = \{\hat{\mathbf{h}}_{li}^j\}$. Note that h is known at the BS since it depends only on $\hat{\mathbf{h}}_{li}^j$ and \mathbf{v}_{jk} , where the latter is a function of the channel estimates and thus of u . The conditional variance of the zero-mean interfering signal v is

$$\begin{aligned} p_v(h, u) &= \mathbb{E} \left\{ |v|^2 \middle| \{\hat{\mathbf{h}}_{li}^j\} \right\} \\ &\stackrel{(a)}{=} \mathbb{E} \{ |s_{jk}|^2 \} \mathbb{E} \left\{ |\mathbf{v}_{jk}^H \tilde{\mathbf{h}}_{jk}^j|^2 \middle| \{\hat{\mathbf{h}}_{li}^j\} \right\} + \sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} \mathbb{E} \{ |s_{li}|^2 \} \mathbb{E} \left\{ |\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2 \middle| \{\hat{\mathbf{h}}_{li}^j\} \right\} \\ &\quad + \mathbb{E} \left\{ |\mathbf{v}_{jk}^H \mathbf{n}_j|^2 \middle| \{\hat{\mathbf{h}}_{li}^j\} \right\} \\ &\stackrel{(b)}{=} p_{jk} \mathbf{v}_{jk}^H \mathbf{C}_{jk}^j \mathbf{v}_{jk} + \sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} p_{li} \mathbf{v}_{jk}^H \left(\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \mathbf{C}_{li}^j \right) \mathbf{v}_{jk} \\ &\quad + \sigma_{\text{UL}}^2 \mathbf{v}_{jk}^H \mathbf{I}_{M_j} \mathbf{v}_{jk} \\ &= \sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} p_{li} |\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{li}^j|^2 + \mathbf{v}_{jk}^H \left(\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j + \sigma_{\text{UL}}^2 \mathbf{I}_{M_j} \right) \mathbf{v}_{jk} \quad (\text{C.47}) \end{aligned}$$

where (a) follows from the independence between each of the zero-mean signals s_{li} and the independence between signals and channels. Next, (b) follows from computing the powers of the signals $\mathbb{E}\{|s_{li}|^2\} = p_{li}$ and from utilizing the fact that $\mathbb{E}\left\{|{\mathbf{v}}_{jk}^H \hat{\mathbf{h}}_{li}^j|^2 | \{\hat{\mathbf{h}}_{li}^j\}\right\} = {\mathbf{v}}_{jk}^H \left(\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \mathbf{C}_{li}^j \right) {\mathbf{v}}_{jk}$ for arbitrary values of l and i (since the estimation error is independent of the channel estimate).

To utilize the capacity bound in Corollary 1.3, we also need to prove that the interference term has conditionally zero mean, $\mathbb{E}\{v|h, u\} = 0$, which is satisfied since the signals and the receiver noise are independent of the realizations of the channel estimates and have zero mean. The corollary also requires the interference term to be conditionally uncorrelated with the input signal, $\mathbb{E}\{x^* v|h, u\} = 0$, which is satisfied since

$$\mathbb{E}\{x^* v|h, u\} = \mathbb{E}\left\{x^* v|\{\hat{\mathbf{h}}_{li}^j\}\right\} = \mathbb{E}\{{\mathbf{v}}_{jk}^H \tilde{\mathbf{h}}_{jk}^j | \{\hat{\mathbf{h}}_{li}^j\}\} \mathbb{E}\{|s_{jk}|^2\} = 0 \quad (\text{C.48})$$

where the second equality exploits the fact that $x = s_{jk}$ is independent of all terms in (C.46) except the first one and the third equality exploits the fact that the estimation error $\tilde{\mathbf{h}}_{jk}^j$ is independent of the channel estimates and has zero mean.

We have now proved that we can utilize Corollary 1.3 to lower bound the capacity. The expression $\mathbb{E}\{\log_2(1 + \text{SINR}_{jk}^{\text{UL}})\}$ follows from (1.10) by inserting the values of h and $p_v(h, u)$ obtained above. As a last step, we note that only the fraction τ_u/τ_c of the samples are used for UL data transmission, which results in the lower bound on the capacity in (4.2) measured in bit/s/Hz.

C.3.2 Proof of Corollary 4.2

The UL instantaneous SINR in (4.3) can be expressed as

$$\text{SINR}_{jk}^{\text{UL}} = \frac{|{\mathbf{v}}_{jk}^H \mathbf{a}_{jk}|^2}{{\mathbf{v}}_{jk}^H \mathbf{B}_{jk} {\mathbf{v}}_{jk}} \quad (\text{C.49})$$

for a fixed vector $\mathbf{a}_{jk} = \sqrt{p_{jk}} \hat{\mathbf{h}}_{jk}^j$ and a fixed matrix

$$\mathbf{B}_{jk} = \sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} p_{li} \hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbf{C}_{li}^j + \sigma_{UL}^2 \mathbf{I}_{M_j}. \quad (\text{C.50})$$

The maximization of the SINR is thus a generalized Rayleigh quotient and is solved by Lemma B.10 on p. 562. The maximum SINR becomes $\mathbf{a}_{jk}^H \mathbf{B}_{jk}^{-1} \mathbf{a}_{jk}$ which gives (4.5). Furthermore, the lemma provides $\mathbf{v}_{jk} = \mathbf{B}_{jk}^{-1} \mathbf{a}_{jk}$ as one combining vector that attains the maximum. Note that

$$\mathbf{B}_{jk}^{-1} \mathbf{a}_{jk} = (1 + \mathbf{a}_{jk}^H \mathbf{B}_{jk}^{-1} \mathbf{a}_{jk}) (\mathbf{B}_{jk} + \mathbf{a}_{jk} \mathbf{a}_{jk}^H)^{-1} \mathbf{a}_{jk} \quad (\text{C.51})$$

by utilizing (B.3) in Lemma B.4 on p. 560. This vector is equivalent to (4.4) except from having another scaling factor in front of the inverse. Since the SINR expression in (4.3) does not change if we scale \mathbf{v}_{jk} by any non-zero scalar, (4.4) also maximizes the instantaneous SINR.

C.3.3 Proof of Corollary 4.3

By direct computation of the conditional expectation in (4.6), we obtain the MSE expression

$$\begin{aligned} & \mathbb{E} \left\{ |s_{jk} - \mathbf{v}_{jk}^H \mathbf{y}_j|^2 \mid \{\hat{\mathbf{h}}_{li}^j\} \right\} \\ &= p_{jk} - p_{jk} \mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j - p_{jk} (\hat{\mathbf{h}}_{jk}^j)^H \mathbf{v}_{jk} \\ &+ \mathbf{v}_{jk}^H \left(\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \left(\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \mathbf{C}_{li}^j \right) + \sigma_{UL}^2 \mathbf{I}_{M_j} \right) \mathbf{v}_{jk}. \end{aligned} \quad (\text{C.52})$$

By introducing the notation

$$\mathbf{a}_{jk} = p_{jk} \hat{\mathbf{h}}_{jk}^j \quad (\text{C.53})$$

$$\mathbf{B}_{jk} = \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \left(\hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H + \mathbf{C}_{li}^j \right) + \sigma_{UL}^2 \mathbf{I}_{M_j} \quad (\text{C.54})$$

we can write the MSE in (C.52) as

$$\begin{aligned} & p_{jk} - \mathbf{v}_{jk}^H \mathbf{a}_{jk} - \mathbf{a}_{jk}^H \mathbf{v}_{jk} + \mathbf{v}_{jk}^H \mathbf{B}_{jk} \mathbf{v}_{jk} \\ &= p_{jk} - \mathbf{a}_{jk}^H \mathbf{B}_{jk}^{-1} \mathbf{a}_{jk} + \left(\mathbf{v}_{jk} - \mathbf{B}_{jk}^{-1} \mathbf{a}_{jk} \right)^H \mathbf{B}_{jk} \left(\mathbf{v}_{jk} - \mathbf{B}_{jk}^{-1} \mathbf{a}_{jk} \right) \end{aligned} \quad (\text{C.55})$$

where the last term is non-negative since \mathbf{B}_{jk} is a positive definite matrix. The MSE is minimized with respect to \mathbf{v}_{jk} when the last term is zero, which occurs when $\mathbf{v}_{jk} = \mathbf{B}_{jk}^{-1}\mathbf{a}_{jk}$. Finally, we note that this vector is the same as the M-MMSE combining vector in (4.4).

C.3.4 Proof of Theorem 4.4

The received signal in (4.13) matches the discrete memoryless channel in Corollary 1.3 on p. 171 with the deterministic channel response $h = \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}$, the input $x = s_{jk}$, and the output $y = \mathbf{v}_{jk}^H \mathbf{y}_j$. Using the notation from that corollary, the noise term is zero (i.e., $\sigma^2 = 0$), since the processed noise $\mathbf{v}_{jk}^H \mathbf{n}_j$ might not be Gaussian distributed, and the interference term is

$$\begin{aligned} v &= (\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\})s_{jk} + \sum_{\substack{i=1 \\ i \neq k}}^{K_j} \mathbf{v}_{jk}^H \mathbf{h}_{ji}^j s_{ji} + \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j s_{li} + \mathbf{v}_{jk}^H \mathbf{n}_j \\ &= \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j s_{li} - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} s_{jk} + \mathbf{v}_{jk}^H \mathbf{n}_j. \end{aligned} \quad (\text{C.56})$$

The interference term has zero mean, $\mathbb{E}\{v\} = 0$, and is uncorrelated with the input since

$$\mathbb{E}\{x^* v\} = \underbrace{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}\}}_{=0} \mathbb{E}\{|s_{jk}|^2\} = 0 \quad (\text{C.57})$$

which are two conditions for applying the capacity bound in Corollary 1.3. The variance of the interference term is

$$\begin{aligned} p_v &= \mathbb{E}\{|v|^2\} \\ &= \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} \mathbb{E}\{|s_{li}|^2\} - |\mathbb{E}\{\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j\}|^2 \mathbb{E}\{|s_{jk}|^2\} + \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{n}_j|^2\} \\ &= \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} - p_{jk} |\mathbb{E}\{\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j\}|^2 + \sigma_{\text{UL}}^2 \mathbb{E}\{\|\mathbf{v}_{jk}^H\|^2\} \end{aligned} \quad (\text{C.58})$$

which follows from utilizing the independence between each of the zero-mean signals s_{li} and the independence between signals and channels.

The lower capacity bound in (4.14) now follows from (1.9) by inserting the values of h and p_v . As a last step, we note that only the fraction τ_u/τ_c of the samples are used for UL data transmission, which results in the lower bound on the capacity that is stated in the theorem in bit/s/Hz.

C.3.5 Proof of Corollary 4.5

The expectations are computed directly, using the properties of the MMSE estimator. The expression in (4.15) is computed as

$$\begin{aligned}\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} &= \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \mathbf{h}_{jk}^j\} \stackrel{(a)}{=} \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \tilde{\mathbf{h}}_{jk}^j\} \\ &\stackrel{(b)}{=} \text{tr}(\mathbb{E}\{\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{jk}^j)^H\}) \stackrel{(c)}{=} p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j)\end{aligned}\quad (\text{C.59})$$

where (a) follows from $\mathbf{h}_{jk}^j = \hat{\mathbf{h}}_{jk}^j + \tilde{\mathbf{h}}_{jk}^j$ and the fact that $\mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \tilde{\mathbf{h}}_{jk}^j\} = 0$ since the estimate and the estimation error are independent and have zero mean. Next, (b) follows from the matrix identity (B.5) in Lemma B.5 on p. 560 and (c) utilizes (3.14). The expression in (4.16) also becomes $p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j)$ since it is equal to the third expression in (C.59).

The interference term in (4.17) is computed differently depending on whether or not $(l, i) \in \mathcal{P}_{jk}$ (i.e., if the UEs use the same or different pilot sequences). In the case of $(l, i) \notin \mathcal{P}_{jk}$, we have

$$\begin{aligned}\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} &= \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \mathbf{h}_{li}^j (\mathbf{h}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\} \\ &\stackrel{(a)}{=} \text{tr}(\mathbb{E}\{\mathbf{h}_{li}^j (\mathbf{h}_{li}^j)^H\} \mathbb{E}\{\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{jk}^j)^H\}) \\ &\stackrel{(b)}{=} p_{jk} \tau_p \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j)\end{aligned}\quad (\text{C.60})$$

where (a) utilizes the matrix identity in (B.5) and the independence between the channel and channel estimate (due to the use of different pilots). The equality (b) follows from direct computation, using the channel statistics and (3.14).

In the case of $(l, i) \in \mathcal{P}_{jk}$, we have

$$\begin{aligned}\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} &= \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H (\hat{\mathbf{h}}_{li}^j + \tilde{\mathbf{h}}_{li}^j)(\hat{\mathbf{h}}_{li}^j + \tilde{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\} \\ &= \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\} + \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \tilde{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\}\end{aligned}\quad (\text{C.61})$$

where the last equality follows from expanding the expression and removing two cross-terms that are zero due to the independence and zero mean of the estimate and the estimation error. When computing first term in (C.61), we note that $\hat{\mathbf{h}}_{li}^j = \sqrt{p_{li}} \mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{y}_{jjk}^p$ and $\hat{\mathbf{h}}_{jk}^j = \sqrt{p_{jk}} \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{y}_{jjk}^p$, where the processed received signal can be expressed as

$$\mathbf{y}_{jjk}^p \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_{M_j}, \tau_p (\Psi_{jk}^j)^{-1} \right). \quad (\text{C.62})$$

Hence, the first term in (C.61) becomes

$$\begin{aligned} & \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \hat{\mathbf{h}}_{li}^j (\hat{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\} = p_{li} p_{jk} \mathbb{E}\{ |(\mathbf{y}_{jjk}^p)^H \Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{y}_{jjk}^p|^2\} \\ & \stackrel{(a)}{=} p_{li} p_{jk} (\tau_p)^2 \left| \text{tr} \left(\Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j (\Psi_{jk}^j)^{-1} \right) \right|^2 \\ & + p_{li} p_{jk} (\tau_p)^2 \text{tr} \left(\Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j (\Psi_{jk}^j)^{-1} \Psi_{jk}^j \mathbf{R}_{jk}^j \mathbf{R}_{li}^j \Psi_{jk}^j (\Psi_{jk}^j)^{-1} \right) \\ & \stackrel{(b)}{=} p_{li} p_{jk} (\tau_p)^2 \left| \text{tr} \left(\Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j \right) \right|^2 + p_{li} p_{jk} (\tau_p)^2 \text{tr} \left(\Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j \mathbf{R}_{li}^j \right) \\ & \stackrel{(c)}{=} p_{li} p_{jk} (\tau_p)^2 \left| \text{tr} \left(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j \right) \right|^2 + p_{jk} \tau_p \text{tr} \left((\mathbf{R}_{li}^j - \mathbf{C}_{li}^j) \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j \right) \end{aligned} \quad (\text{C.63})$$

where (a) follows from Lemma B.14 on p. 564 with $\mathbf{B} = \Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j$ and $\mathbf{A} = \tau_p (\Psi_{jk}^j)^{-1}$.¹ Multiplications of matrices and their inverses are removed in (b) and finally we obtain (c) by noting that $|\text{tr}(\Psi_{jk}^j \mathbf{R}_{li}^j \mathbf{R}_{jk}^j)| = |\text{tr}(\mathbf{R}_{jk}^j \mathbf{R}_{li}^j \Psi_{jk}^j)| = |\text{tr}(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)|$, $\mathbf{R}_{li}^j - \mathbf{C}_{li}^j = p_{li} \tau_p \mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{li}^j$, and using the matrix identity in (B.5) to obtain that expression in second trace.

The second term in (C.61) becomes

$$\begin{aligned} & \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \tilde{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H \hat{\mathbf{h}}_{jk}^j\} = \text{tr} \left(\mathbb{E}\{\tilde{\mathbf{h}}_{li}^j (\tilde{\mathbf{h}}_{li}^j)^H\} \mathbb{E}\{\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{jk}^j)^H\} \right) \\ & = p_{jk} \tau_p \text{tr} \left(\mathbf{C}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j \right) \end{aligned} \quad (\text{C.64})$$

where the first equality utilizes the matrix identity in (B.5) and the independence between the estimate and estimation error, while the second equality follows from direct computation of the expectations. By

¹The absolute value in Lemma B.14 can be replaced with regular parentheses since all matrices are positive semi-definite so the trace is positive and real-valued.

substituting (C.63) and (C.64) into (C.61), we finally obtain

$$\begin{aligned} & \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} \\ &= p_{li} p_{jk} (\tau_p)^2 \left| \text{tr} \left(\mathbf{R}_{li}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right) \right|^2 + p_{jk} \tau_p \text{tr} \left(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right). \end{aligned} \quad (\text{C.65})$$

The SE expression in (4.18) is obtained from (4.14) by inserting the closed-form expressions that were computed above, and then dividing the numerator and denominator by $p_{jk} \tau_p \text{tr}(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j)$.

In the special case of uncorrelated fading, we have

$$p_{jk} \tau_p \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j = \frac{p_{jk} \tau_p \beta_{jk}^j}{\sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \tau_p \beta_{l'i'}^j + \sigma_{\text{UL}}^2} \mathbf{I}_{M_j} \quad (\text{C.66})$$

and thus direct computation of the traces yields

$$p_{jk}^2 \tau_p \text{tr} \left(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right) = p_{jk} \beta_{jk}^j M_j \frac{p_{jk} \tau_p \beta_{jk}^j}{\sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \tau_p \beta_{l'i'}^j + \sigma_{\text{UL}}^2} \quad (\text{C.67})$$

$$p_{li} \frac{\text{tr} \left(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right)}{\text{tr} \left(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right)} = p_{li} \beta_{li}^j \quad (\text{C.68})$$

$$\frac{p_{li}^2 \tau_p \left| \text{tr} \left(\mathbf{R}_{li}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right) \right|^2}{\text{tr} \left(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right)} = p_{li} \beta_{li}^j M_j \frac{p_{li} \tau_p \beta_{li}^j}{\sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \tau_p \beta_{l'i'}^j + \sigma_{\text{UL}}^2}. \quad (\text{C.69})$$

The final expression in (4.19) is obtained by inserting (C.67)–(C.69) into (4.18) and utilizing the definition of ψ_{jk} in (4.20).

C.3.6 Proof of Theorem 4.6

The received signal in (4.25) matches the discrete memoryless channel in Corollary 1.3 on p. 171 with the deterministic channel response $h = \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}$, the input $x = \varsigma_{jk}$, and the output $y = y_{jk}$. Using the notation from that corollary, the noise term is $n = n_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{DL}}^2)$

and the interference term is

$$\begin{aligned}
v &= \left((\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \right) \varsigma_{jk} \\
&\quad + \sum_{\substack{i=1 \\ i \neq k}}^{K_j} (\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji} \varsigma_{ji} + \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li} \\
&= \sum_{l=1}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \varsigma_{jk}.
\end{aligned} \tag{C.70}$$

Note that the interference term has zero mean and is uncorrelated with the input since

$$\mathbb{E}\{x^* v\} = \underbrace{\mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}\}}_{=0} \mathbb{E}\{|\varsigma_{jk}|^2\} = 0 \tag{C.71}$$

which are two conditions for applying Corollary 1.3. Furthermore, the variance of the interference term is

$$\begin{aligned}
p_v &= \mathbb{E}\{|v|^2\} \\
&= \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbb{E}\{|(\mathbf{h}_{jk}^l)^H \mathbf{w}_{li}|^2\} \mathbb{E}\{|\varsigma_{li}|^2\} - |\mathbb{E}\{(\mathbf{h}_{jk}^l)^H \mathbf{w}_{jk}\}|^2 \mathbb{E}\{|\varsigma_{jk}|^2\} \\
&= \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \mathbb{E}\{|\mathbf{w}_{li}^H (\mathbf{h}_{jk}^l)|^2\} - \rho_{jk} |\mathbb{E}\{(\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j)\}|^2
\end{aligned} \tag{C.72}$$

by utilizing the independence between each of the zero-mean signals ς_{li} and the independence between signals and channels.

The effective SINR expression in (4.26) now follows from (1.9) by inserting the values of h and p_v . As a last step, we note that only the fraction τ_d/τ_c of the samples are used for DL data transmission, which results in the lower bound on the capacity stated in the theorem and measured in bit/s/Hz.

C.3.7 Proof of Corollary 4.7

The proof consists of computing each of the expectations in (4.26) for the case of $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}}$.

The signal term in the numerator of (4.26) is

$$\begin{aligned} \mathbb{E}\{\mathbf{w}_{jk}^H \mathbf{h}_{jk}^j\|^2 &= \frac{\mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}} \stackrel{(a)}{=} \frac{\mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \hat{\mathbf{h}}_{jk}^j\}|^2}{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}} \\ &= \mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\} \stackrel{(b)}{=} p_{jk} \tau_p \text{tr} \left(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right) \end{aligned} \quad (\text{C.73})$$

where (a) follows from the independence between the channel estimate and the estimation error and (b) follows from (C.59).

The expectation in the interference terms in the denominator of (4.26) with indices that satisfy $(l, i) \notin \mathcal{P}_{jk}$ (i.e., the UEs use different pilots) is computed as

$$\mathbb{E}\{|\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2\} = \frac{\mathbb{E}\{|(\hat{\mathbf{h}}_{li}^l)^H \mathbf{h}_{jk}^l|^2\}}{\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^l\|^2\}} \stackrel{(a)}{=} \frac{\text{tr} \left(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right)}{\text{tr} \left(\mathbf{R}_{li}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right)} \quad (\text{C.74})$$

where (a) computes the expectations using (C.59) and (C.60), and then removes their common scaling factor $p_{li} \tau_p$. Note that we need to swap the indices (j, k) and (l, i) in both equations to obtain the desired result. Similarly, in the case $(l, i) \in \mathcal{P}_{jk}$ (i.e., when the UEs use the same pilot) the expectations in the interference terms are computed as

$$\begin{aligned} \mathbb{E}\{|\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2\} &= \frac{\mathbb{E}\{|(\hat{\mathbf{h}}_{li}^l)^H \mathbf{h}_{jk}^l|^2\}}{\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^l\|^2\}} \\ &\stackrel{(a)}{=} \frac{p_{jk} \tau_p \left| \text{tr} \left(\mathbf{R}_{jk}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right) \right|^2 + \text{tr} \left(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right)}{\text{tr} \left(\mathbf{R}_{li}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right)} \end{aligned} \quad (\text{C.75})$$

where (a) follows from (C.59) and (C.65) by swapping the indices (j, k) and (l, i) .

By substituting (C.73)–(C.75) into (4.26) and noting that

$$\begin{aligned} &\sum_{(l,i) \in \mathcal{P}_{jk}} \rho_{li} \frac{p_{jk} \tau_p \left| \text{tr} \left(\mathbf{R}_{jk}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right) \right|^2}{\text{tr} \left(\mathbf{R}_{li}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right)} - \rho_{jk} p_{jk} \tau_p \text{tr} \left(\mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j \right) \\ &= \sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \frac{\rho_{li} p_{jk} \tau_p \left| \text{tr} \left(\mathbf{R}_{jk}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right) \right|^2}{\text{tr} \left(\mathbf{R}_{li}^l \boldsymbol{\Psi}_{li}^l \mathbf{R}_{li}^l \right)} \end{aligned} \quad (\text{C.76})$$

we obtain the final expression for $\underline{\text{SINR}}_{jk}^{\text{DL}}$ in (4.28).

The simplification for uncorrelated fading in (4.29) follows directly from inserting the simplified expressions in (C.67)–(C.69) and utilizing the definition of ψ_{li} in (4.20).

C.3.8 Proof of Theorem 4.8

Let $\gamma_{jk} = \underline{\text{SINR}}_{jk}^{\text{UL}}$ denote the value of the effective SINR in (4.14) for the given UL transmit power vector \mathbf{p} and receive combining vectors \mathbf{v}_{jk} , for $j = 1, \dots, L$ and $k = 1, \dots, K_j$. The goal of this proof is to establish that $\gamma_{jk} = \underline{\text{SINR}}_{jk}^{\text{DL}}$ is achievable in the DL when using $\mathbf{w}_{jk} = \mathbf{v}_{jk} / \sqrt{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}$ for all j and k . Substituting these precoding vectors into (4.14), we obtain the SINR constraints

$$\gamma_{jk} = \frac{\rho_{jk} \frac{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \frac{\mathbb{E}\{|\mathbf{v}_{li}^H \mathbf{h}_{jk}^l|^2\}}{\mathbb{E}\{\|\mathbf{v}_{li}\|^2\}} - \rho_{jk} \frac{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} + \sigma_{\text{DL}}^2} \quad (\text{C.77})$$

for $j = 1, \dots, L$ and $k = 1, \dots, K_j$, which can be rewritten as

$$\gamma_{jk} \frac{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2} = \frac{\rho_{jk}}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \frac{\mathbb{E}\{|\mathbf{v}_{li}^H \mathbf{h}_{jk}^l|^2\}}{\mathbb{E}\{\|\mathbf{v}_{li}\|^2\}} - \rho_{jk} \frac{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} + \sigma_{\text{DL}}^2}. \quad (\text{C.78})$$

Using the matrices \mathbf{B} and \mathbf{D} , defined in Theorem 4.8, the constraints in (C.78) can be expressed as

$$[\mathbf{D}_j]_{kk} = \frac{\rho_{jk}}{\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} [\mathbf{B}_{jl}]_{ki} + \sigma_{\text{DL}}^2} \quad (\text{C.79})$$

for $j = 1, \dots, L$ and $k = 1, \dots, K_j$, from which we have that

$$\sigma_{\text{DL}}^2 = \frac{\rho_{jk}}{[\mathbf{D}_j]_{kk}} - \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} [\mathbf{B}_{jl}]_{ki}. \quad (\text{C.80})$$

The K_{tot} constraints can be written in matrix form as

$$\mathbf{1}_{K_{\text{tot}}} \sigma_{\text{DL}}^2 = \mathbf{D}^{-1} \boldsymbol{\rho} - \mathbf{B} \boldsymbol{\rho} = (\mathbf{D}^{-1} - \mathbf{B}) \boldsymbol{\rho}. \quad (\text{C.81})$$

This is a linear system of equations, thus the DL transmit power vector $\rho = \rho_{\text{opt}}$ that satisfies all the SINR constraints is obtained as

$$\rho_{\text{opt}} = (\mathbf{D}^{-1} - \mathbf{B})^{-1} \mathbf{1}_{K_{\text{tot}}} \sigma_{\text{DL}}^2. \quad (\text{C.82})$$

This is a feasible power vector with positive values if the inverse exists and all elements of $(\mathbf{D}^{-1} - \mathbf{B})^{-1}$ are positive. To prove that this is the case whenever \mathbf{p} is feasible, we study the corresponding UL SINR constraints $\gamma_{jk} = \underline{\text{SINR}}_{jk}^{\text{UL}}$:

$$\gamma_{jk} = \frac{p_{jk} \frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}}}{\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{li}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} - p_{jk} \frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} + \sigma_{\text{UL}}^2} \quad (\text{C.83})$$

which can be rewritten as

$$[\mathbf{D}_j]_{kk} = \frac{p_{jk}}{\sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} [\mathbf{B}_{lj}]_{ik} + \sigma_{\text{UL}}^2} \Leftrightarrow \quad (\text{C.84})$$

$$\sigma_{\text{UL}}^2 = \frac{p_{jk}}{[\mathbf{D}_j]_{kk}} - \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} [\mathbf{B}_{lj}]_{ik} \quad (\text{C.85})$$

for $j = 1, \dots, L$ and $k = 1, \dots, K_j$. This can be expressed as the system of equations $\mathbf{1}_{K_{\text{tot}}} \sigma_{\text{UL}}^2 = \mathbf{D}^{-1} \mathbf{p} - \mathbf{B}^T \mathbf{p} = (\mathbf{D}^{-1} - \mathbf{B}^T) \mathbf{p}$, where there is a transpose on \mathbf{B} since the indices of the interference term in (C.85) are swapped as compared to (C.80). Clearly, we have $\mathbf{p} = (\mathbf{D}^{-1} - \mathbf{B}^T)^{-1} \mathbf{1}_{K_{\text{tot}}} \sigma_{\text{UL}}^2$, which implies that the inverse exists and all elements of $(\mathbf{D}^{-1} - \mathbf{B}^T)^{-1}$ are positive.² This also implies that $(\mathbf{D}^{-1} - \mathbf{B}) = (\mathbf{D}^{-1} - \mathbf{B}^T)^T$ has strictly positive eigenvalues. Consequently, if \mathbf{p} is a feasible power vector the UL, then (C.82) is a feasible power vector for

²It is intuitive but not straightforward to prove that the elements are positive. To see that this must be true, we note that $\boldsymbol{\sigma} = \mathbf{1}_{K_{\text{tot}}} \sigma_{\text{UL}}^2$ is a vector with the noise variances of all UEs. If we reduce the noise variance of some selected UEs, we increase the corresponding SINRs (if the transmit power is fixed) and we should be able to find a new feasible power vector that gives the same SINRs using less sum power. However, if $(\mathbf{D}^{-1} - \mathbf{B}^T)^{-1}$ has negative elements we can find a positive noise vector $\boldsymbol{\sigma}$ such that $\mathbf{p} = (\mathbf{D}^{-1} - \mathbf{B}^T)^{-1} \boldsymbol{\sigma}$ has negative elements and thus is infeasible. This is not reasonable, thus $(\mathbf{D}^{-1} - \mathbf{B}^T)^{-1}$ must have only positive elements. We refer to [63, 260] for detailed methods to prove this result.

the DL. This proves the main part of the theorem and (4.32) is obtained by substituting $\mathbf{1}_{K_{\text{tot}}} = \frac{1}{\sigma_{\text{UL}}^2} (\mathbf{D}^{-1} - \mathbf{B}^T) \mathbf{p}$ into (C.82).

The sum power condition in (4.33) follows from direct computation and by utilizing $\mathbf{1}_{K_{\text{tot}}}^T (\mathbf{D}^{-1} - \mathbf{B})^{-1} \mathbf{1}_{K_{\text{tot}}} = \mathbf{1}_{K_{\text{tot}}}^T (\mathbf{D}^{-1} - \mathbf{B}^T)^{-1} \mathbf{1}_{K_{\text{tot}}}$.

C.3.9 Proof of Theorem 4.9

Let $\{\varsigma_{jk}\}$ denote the set of the τ_d DL data signals transmitted to UE k in cell j in a given coherence block and let $\{y_{jk}\}$ denote the corresponding set of received signals at this UE. By assuming that $\varsigma_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \rho_{jk})$, the DL capacity of this UE is lower bounded by

$$\frac{\tau_d}{\tau_c} \frac{1}{\tau_d} \mathcal{I}(\{\varsigma_{jk}\}; \{y_{jk}\}) \quad (\text{C.86})$$

where τ_d/τ_c is the fraction of the coherence block used for DL data transmission and $\mathcal{I}(\{\varsigma_{jk}\}; \{y_{jk}\})/\tau_d$ is the mutual information per sample. Next, we use the chain rule of mutual information: $\mathcal{I}(X_1, X_2; Y) = \mathcal{I}(X_1; Y) + \mathcal{I}(X_2; Y|X_1) = \mathcal{I}(X_2; Y) + \mathcal{I}(X_1; Y|X_2)$ [94, Theorem 2.5.2]. We consider $X_1 = \{\varsigma_{jk}\}$, $Y = \{y_{jk}\}$, and we let $X_2 = \{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}$ denote the set of precoded channels $(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}$ for $i = 1, \dots, K_j$. From the chain rule we then obtain

$$\begin{aligned} \mathcal{I}(X_1; Y) &= \mathcal{I}(X_2; Y) + \mathcal{I}(X_1; Y|X_2) - \mathcal{I}(X_2; Y|X_1) \\ &\geq \mathcal{I}(X_1; Y|X_2) - \mathcal{I}(X_2; Y|X_1) \end{aligned} \quad (\text{C.87})$$

where the inequality follows from omitting the non-negative term $\mathcal{I}(X_2; Y)$. The first term in (C.87) is the conditional mutual information between the transmitted and received signals given the precoded channels for the intra-cell signals:

$$\begin{aligned} \mathcal{I}(X_1; Y|X_2) &= \mathcal{I}\left(\{\varsigma_{jk}\}; \{y_{jk}\} | \{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}\right) \\ &\geq \tau_d \mathcal{I}\left(\varsigma_{jk}; y_{jk} | \{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}\right) \\ &\geq \tau_d \mathbb{E} \left\{ \log_2 \left(1 + \text{SINR}_{jk}^{\text{DL}} \right) \right\} \end{aligned} \quad (\text{C.88})$$

where the first inequality follows from neglecting the mutual information between different samples. The second inequality follows from applying Corollary 1.3 on p. 171 with the input $x = \varsigma_{jk}$, the output $y = y_{jk}$, the

random channel response $h = (\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}$, $n = n_{jk}$, and $u = \{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}$. The interference term in the corollary is

$$v = \sum_{\substack{i=1 \\ i \neq k}}^{K_j} (\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji} \varsigma_{ji} + \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li}. \quad (\text{C.89})$$

This term has conditional zero mean (since the data signals have zero mean) and conditional variance

$$\begin{aligned} p_v(h, u) &= \mathbb{E} \left\{ |v|^2 \mid \{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \right\} \\ &= \sum_{\substack{i=1 \\ i \neq k}}^{K_j} \rho_{ji} |\mathbf{w}_{ji}^H \mathbf{h}_{jk}^j|^2 + \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} \rho_{li} \mathbb{E} \left\{ |\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2 \right\} \end{aligned} \quad (\text{C.90})$$

since we assumed that each BS computes its precoding vectors using only its own channel estimates, which implies that the precoded channels from other cells are independent of $\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}$. The corollary also requires the interference term to be conditionally uncorrelated with the input signal, $\mathbb{E} \{x^* v \mid h, u\} = 0$, which is satisfied since v is independent of x .

We also need to compute a bound on the second term in (C.87):

$$\begin{aligned} \mathcal{I}(X_2; Y|X_1) &= \mathcal{I} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}; \{y_{jk}\} \mid \{\varsigma_{jk}\} \right) \\ &= \mathcal{H} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \mid \{\varsigma_{jk}\} \right) - \mathcal{H} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \mid \{y_{jk}\}, \{\varsigma_{jk}\} \right) \\ &\leq \mathcal{H} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \mid \{\varsigma_{jk}\}, \Omega \right) - \mathcal{H} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \mid \{y_{jk}\}, \{\varsigma_{jk}\}, \Omega \right) \end{aligned} \quad (\text{C.91})$$

where the inequality follows from adding some side-information Ω that is independent of $\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}$, such that $\mathcal{H} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \mid \{\varsigma_{jk}\}, \Omega \right) = \mathcal{H} \left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} \mid \{\varsigma_{jk}\} \right)$, while the conditioning reduces the second differential entropy expression. In particular, we let Ω contain the transmitted intra-cell signals $\{\varsigma_{ji}\}$ for $i = 1, \dots, K_i$ and the realizations of the inter-cell interference $\{\sum_{l=1, l \neq j}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li}\}$ for the τ_d samples used for DL data transmission. Since the received signal and inter-cell interfer-

ence are known, we can compute

$$\check{y}_{jk} = y_{jk} - \sum_{l=1, l \neq j}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li} = \sum_{i=1}^{K_j} (\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji} \varsigma_{ji} + n_{jk} \quad (\text{C.92})$$

which only contains the intra-cell signals and noise. By utilizing this notation, we have

$$\mathcal{H}\left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} | \{y_{jk}\}, \{\varsigma_{jk}\}, \Omega\right) = \mathcal{H}\left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\} | \{\check{y}_{jk}\}, \{\varsigma_{jk}\}, \Omega\right). \quad (\text{C.93})$$

By substituting (C.93) into (C.91), we get

$$\begin{aligned} \mathcal{I}(X_2; Y | X_1) &\leq \mathcal{I}\left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}; \{\check{y}_{jk}\} | \{\varsigma_{jk}\}, \Omega\right) \\ &= \mathcal{I}\left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}; \{\check{y}_{jk}\} | \{\varsigma_{j1}\}, \dots, \{\varsigma_{jK_j}\}\right) \end{aligned} \quad (\text{C.94})$$

where the equality follows from removing the conditioning on the inter-cell interference terms in Ω , which are now independent of all other variables in the expression.

Interestingly, (C.94) can be interpreted as the sum mutual information of an uplink multiuser MIMO channel with transmitted signals $\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}$ from K_j UEs, received signals $\{\check{y}_{jk}\}$ over τ_d antennas, and known “channel coefficients” $\{\varsigma_{j1}\}, \dots, \{\varsigma_{jK_j}\}$. In the mutual information maximizing case, the UE channels are orthogonal and we obtain

$$\begin{aligned} &\mathcal{I}\left(\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}; \{\check{y}_{jk}\} | \{\varsigma_{j1}\}, \dots, \{\varsigma_{jK_j}\}\right) \\ &\leq \sum_{i=1}^{K_j} \mathbb{E} \left\{ \log_2 \left(1 + \frac{\sum_{t=1}^{\tau_d} |\varsigma_{jit}|^2 \mathbb{V}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}}{\sigma_{\text{DL}}^2} \right) \right\} \end{aligned} \quad (\text{C.95})$$

where ς_{jit} denotes the realization of ς_{ji} at the t th DL sample, for $t = 1, \dots, \tau_d$. Finally, we use Jensen’s inequality in Lemma B.11 on p. 563 to obtain

$$\begin{aligned} &\sum_{i=1}^{K_j} \mathbb{E} \left\{ \log_2 \left(1 + \frac{\sum_{t=1}^{\tau_d} |\varsigma_{ jit}|^2 \mathbb{V}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}}{\sigma_{\text{DL}}^2} \right) \right\} \\ &\leq \sum_{i=1}^{K_j} \log_2 \left(1 + \frac{\rho_{ji} \tau_d \mathbb{V}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji}\}}{\sigma_{\text{DL}}^2} \right) \end{aligned} \quad (\text{C.96})$$

since $\mathbb{E}\{|\zeta_{jli}|^2\} = \rho_{ji}$.

By substituting (C.87) into (C.86) and utilizing the closed-form bounds on the mutual information expressions that have been computed above, we finally obtain (4.38).

C.3.10 Proof of Theorems 4.10 and 4.11

The expression for $\underline{\text{SINR}}_{jk}^{\text{UL}}$ is given in (4.18). The proof begins by dividing the numerator and denominator by M_j . The signal term becomes $\frac{p_{jk}^2 \tau_p}{M_j} \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)$. This term is strictly positive as $M_j \rightarrow \infty$ due to the first condition in Assumption 1 and is finite due to the second condition. Each non-coherent interference term satisfies

$$\frac{p_{li}}{M_j} \frac{\text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)} \leq \frac{p_{li}}{M_j} \frac{\|\mathbf{R}_{li}^j\|_2 \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)}{\text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)} = \frac{p_{li}}{M_j} \|\mathbf{R}_{li}^j\|_2 \quad (\text{C.97})$$

where the inequality follows from Lemma B.7 on p. 561. These terms go to zero as $M_j \rightarrow \infty$ due to the second condition in Assumption 1. The noise term σ_{UL}^2/M_j also goes asymptotically to zero. The remaining coherent interference terms are bounded, since the trace expression in the denominator scales as M_j and the traces in the numerator cannot grow faster than M_j due to Assumption 1. Note that

$$\begin{aligned} \frac{1}{M_j} |\text{tr}(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j)| &\geq \frac{1}{M_j} \frac{\text{tr}(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j) + \text{tr}(\mathbf{R}_{jk}^j \Psi_{jk}^j \mathbf{R}_{li}^j)}{2} \\ &= \frac{1}{M_j} \frac{\text{tr}(\Psi_{jk}^j (\mathbf{R}_{jk}^j \mathbf{R}_{li}^j + \mathbf{R}_{li}^j \mathbf{R}_{jk}^j))}{2} \\ &\geq \frac{1}{\|(\Psi_{jk}^j)^{-1}\|_2} \frac{1}{M_j} \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j) \end{aligned} \quad (\text{C.98})$$

by first removing the imaginary part and then applying Lemma B.8 on p. 561 with $\mathbf{A} = (\Psi_{jk}^j)^{-1}$ and $\mathbf{B} = (\mathbf{R}_{jk}^j \mathbf{R}_{li}^j + \mathbf{R}_{li}^j \mathbf{R}_{jk}^j)$. Note that $1/\|(\Psi_{jk}^j)^{-1}\|_2 \leq 1/\sigma_{\text{UL}}^2 < \infty$ due to Assumption 1. Hence, if $\frac{1}{M_j} \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j)$ has a non-zero limit for some $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$, then the coherent interference term approaches a finite non-zero limit. The dif-

ference in (4.49) approaches zero asymptotically, since the non-coherent interference terms and noise vanish.

However, if $\frac{1}{M_j} \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j) \rightarrow 0$ for all $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$, then we can use Lemma B.7 to prove that

$$\frac{1}{M_j} \left| \text{tr}(\mathbf{R}_{li}^j \Psi_{jk}^j \mathbf{R}_{jk}^j) \right| \leq \frac{1}{M_j} \|\Psi_{jk}^j\|_2 \text{tr}(\mathbf{R}_{li}^j \mathbf{R}_{jk}^j) \rightarrow 0 \quad (\text{C.99})$$

since the spectral norm of Ψ_{jk}^j is bounded according to the second condition in Assumption 1. This happens exactly when \mathbf{R}_{jk}^j is asymptotically spatially orthogonal to \mathbf{R}_{li}^j for all $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$. Since all the terms in the denominator approaches zero in this case, while the numerator approaches a non-zero limit, we conclude that $\underline{\text{SINR}}_{jk}^{\text{UL}}$ grows without bound as $M_j \rightarrow \infty$. This finishes the proof for the UL.

In the DL, the expression for $\underline{\text{SINR}}_{jk}^{\text{DL}}$ contains the same matrix expressions as $\underline{\text{SINR}}_{jk}^{\text{UL}}$, except that the indices (l, i) and (j, k) are swapped in the interference terms. If we divide the numerator and denominator by M , then the signal term approaches a finite non-zero limit, while the noise and non-coherent interference terms go to zero. By a similar argument as in the UL, the coherent interference terms at UE k in cell l approach a non-zero limit if $\text{tr}(\mathbf{R}_{li}^l \mathbf{R}_{jk}^l)/M_j$ has a non-zero limit for at least one $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$. This makes the difference in (4.50) go to zero asymptotically. If none of the coherent interference terms have a non-zero limit, then $\underline{\text{SINR}}_{jk}^{\text{DL}}$ grows without bound instead. This happens when \mathbf{R}_{jk}^l and \mathbf{R}_{li}^l are spatially orthogonal for all $(l, i) \in \mathcal{P}_{jk} \setminus (j, k)$.

C.4 Proofs in Section 5

C.4.1 Proof of Lemma 5.1

We begin by substituting $p_{jk} = \bar{P}/M^{\varepsilon_1}$ and $\rho_{jk} = \underline{P}/M^{\varepsilon_2}$ into all the terms in the deterministic expression for $\underline{\text{SINR}}_{jk}^{\text{DL}}$ given in (5.4). Multiplying and dividing the signal term in the numerator by M leads to

$$\frac{\underline{P}\bar{P}\tau_p}{M^{\varepsilon_1+\varepsilon_2-1}}\frac{1}{M}\text{tr}\left(\mathbf{R}_{jk}^j\Psi_{jk}^j\mathbf{R}_{jk}^j\right) \quad (\text{C.100})$$

with

$$\Psi_{jk}^j = \left(\sum_{(l',i') \in \mathcal{P}_{jk}} \frac{\bar{P}}{M^{\varepsilon_1}} \tau_p \mathbf{R}_{l'i'}^j + \sigma_{\text{UL}}^2 \mathbf{I}_M \right)^{-1}. \quad (\text{C.101})$$

As $M \rightarrow \infty$, we have that $\Psi_{jk}^j - 1/\sigma_{\text{UL}}^2 \mathbf{I}_M \rightarrow \mathbf{0}_{M \times M}$ and consequently

$$\frac{1}{M}\text{tr}\left(\mathbf{R}_{jk}^j\Psi_{jk}^j\mathbf{R}_{jk}^j\right) - \frac{1}{\sigma_{\text{UL}}^2}\frac{1}{M}\text{tr}\left(\mathbf{R}_{jk}^j\mathbf{R}_{jk}^j\right) \rightarrow 0. \quad (\text{C.102})$$

Note that $\frac{1}{M}\text{tr}\left(\mathbf{R}_{jk}^j\mathbf{R}_{jk}^j\right)$ is strictly positive as $M \rightarrow \infty$ due to the first condition in Assumption 1 on p. 337 and is also finite due to the second condition. Therefore, as $M \rightarrow \infty$, (C.100) goes to zero if $\varepsilon_1 + \varepsilon_2 > 1$, while it grows without bound for $\varepsilon_1 + \varepsilon_2 < 1$.

By applying Lemma B.7 on p. 561 to each non-coherent interference term in (5.4), it follows that

$$\begin{aligned} \frac{\underline{P}}{M^{\varepsilon_2}} \frac{\text{tr}\left(\mathbf{R}_{jk}^l\mathbf{R}_{li}^l\Psi_{li}^l\mathbf{R}_{li}^l\right)}{\text{tr}\left(\mathbf{R}_{li}^l\Psi_{li}^l\mathbf{R}_{li}^l\right)} &\leq \frac{\underline{P}}{M^{\varepsilon_2}} \frac{\|\mathbf{R}_{jk}^l\|_2 \text{tr}\left(\mathbf{R}_{li}^l\Psi_{li}^l\mathbf{R}_{li}^l\right)}{\text{tr}\left(\mathbf{R}_{li}^l\Psi_{li}^l\mathbf{R}_{li}^l\right)} \\ &= \frac{\underline{P}}{M^{\varepsilon_2}} \|\mathbf{R}_{jk}^l\|_2 \end{aligned} \quad (\text{C.103})$$

and, hence, these terms go asymptotically to zero as $M \rightarrow \infty$, due to the second condition in Assumption 1. The remaining coherent interference terms can be written as

$$\frac{\underline{P}\bar{P}\tau_p}{M^{\varepsilon_1+\varepsilon_2-1}} \frac{\left|\frac{1}{M}\text{tr}\left(\mathbf{R}_{jk}^l\Psi_{li}^l\mathbf{R}_{li}^l\right)\right|^2}{\frac{1}{M}\text{tr}\left(\mathbf{R}_{li}^l\Psi_{li}^l\mathbf{R}_{li}^l\right)}. \quad (\text{C.104})$$

Similar to (C.102), as $M \rightarrow \infty$,

$$\frac{\left| \frac{1}{M} \text{tr} \left(\mathbf{R}_{jk}^l \Psi_{li}^l \mathbf{R}_{li}^l \right) \right|^2}{\frac{1}{M} \text{tr} \left(\mathbf{R}_{li}^l \Psi_{li}^l \mathbf{R}_{li}^l \right)} - \frac{1}{\sigma_{\text{UL}}^2} \frac{\left(\frac{1}{M} \text{tr} \left(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l \right) \right)^2}{\frac{1}{M} \text{tr} \left(\mathbf{R}_{li}^l \mathbf{R}_{li}^l \right)} \rightarrow 0 \quad (\text{C.105})$$

where all terms are bounded, since $\frac{1}{M} \text{tr} \left(\mathbf{R}_{li}^l \mathbf{R}_{li}^l \right)$ and $\frac{1}{M} \text{tr} \left(\mathbf{R}_{jk}^l \mathbf{R}_{li}^l \right)$ are strictly positive and finite as $M \rightarrow \infty$, due to Assumption 1. Therefore, as $M \rightarrow \infty$, (C.104) goes to zero if $\varepsilon_1 + \varepsilon_2 > 1$, while it grows without bound for $\varepsilon_1 + \varepsilon_2 < 1$. Putting the above results together, Lemma 5.1 follows.

C.4.2 Proof of Equation (5.18)

To obtain (5.18), we begin by rewriting (5.17) as

$$2^{\text{SE}^*} (\text{SE}^* \log_e(2) - 1) = \frac{M-1}{\nu_0} P_{\text{FIX}} - 1 \quad (\text{C.106})$$

which can be transformed via the substitution $x = \text{SE}^* \log_e(2) - 1$ into

$$xe^x = \frac{(M-1)P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e}. \quad (\text{C.107})$$

The solution of the above equation takes the form

$$x^* = W \left(\frac{(M-1)P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e} \right). \quad (\text{C.108})$$

from which (5.18) easily follows since $x^* = \text{SE}^* \log_e(2) - 1$.

C.4.3 Proof of Corollary 5.2

Using the inequalities on the Lambert W function that are reported in Lemma B.16 on p. 567, SE^* and EE^* can be lower bounded as

$$\begin{aligned} \text{SE}^* &\geq \frac{1}{\log_e(2)} \log_e \left(\frac{\frac{(M-1)P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e}}{\log_e \left(\frac{(M-1)P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e} \right)} \right) \\ &= \log_2 \left(\frac{\frac{(M-1)P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e}}{\log_e \left(\frac{(M-1)P_{\text{FIX}}}{\nu_0 e} - \frac{1}{e} \right)} \right) \end{aligned} \quad (\text{C.109})$$

by using $e^{\frac{x}{\log_e(x)}} \leq e^{W(x)+1}$ and

$$\begin{aligned}\text{EE}^* &\geq \frac{(M-1)B}{\nu_0(1+e)\log_e(2)} \frac{\log_e\left(\frac{(M-1)P_{\text{FIX}}}{\nu_0e} - \frac{1}{e}\right)}{\frac{(M-1)P_{\text{FIX}}}{\nu_0e} - \frac{1}{e}} \\ &= \frac{(M-1)B}{\nu_0(1+e)} \frac{\log_2\left(\frac{(M-1)P_{\text{FIX}}}{\nu_0e} - \frac{1}{e}\right)}{\frac{(M-1)P_{\text{FIX}}}{\nu_0e} - \frac{1}{e}}\end{aligned}\quad (\text{C.110})$$

by exploiting the fact that $e^{W(x)+1} \leq (1+e)^{\frac{x}{\log_e(x)}}$. From the above expressions, the approximations in (5.20) and (5.21) follow by assuming that M and/or P_{FIX} are large and neglecting small terms.

C.4.4 Proof of Corollary 5.3

The setup is the same as in the proof of Corollary 5.2, except that P_{FIX} is replaced by $P_{\text{FIX}} + MP_{\text{BS}}$. It then follows from (C.109) that

$$\text{SE}^* \geq \log_2 \left(\frac{\frac{(M-1)(P_{\text{FIX}}+MP_{\text{BS}})}{\nu_0e} - \frac{1}{e}}{\log_e\left(\frac{(M-1)(P_{\text{FIX}}+MP_{\text{BS}})}{\nu_0e} - \frac{1}{e}\right)} \right) \quad (\text{C.111})$$

and from (C.110) that

$$\text{EE}^* \geq \frac{(M-1)B}{\nu_0(1+e)} \frac{\log_2\left(\frac{(M-1)(P_{\text{FIX}}+MP_{\text{BS}})}{\nu_0e} - \frac{1}{e}\right)}{\frac{(M-1)(P_{\text{FIX}}+MP_{\text{BS}})}{\nu_0e} - \frac{1}{e}} \quad (\text{C.112})$$

The approximations in (5.24) and (5.25) follow by assuming that M , P_{FIX} , and/or P_{BS} are large and neglecting small terms.

C.5 Proofs in Section 6

C.5.1 Proof of Theorem 6.1

We want to estimate \mathbf{h}_{li}^j based on the observation

$$\begin{aligned} \mathbf{y}_{jli}^p = \mathbf{Y}_j^p \phi_{li}^* &= \sum_{(l', i') \in \mathcal{P}_{li}} \sqrt{p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \tau_p \mathbf{h}_{l'i'}^j \\ &\quad + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \sqrt{\kappa_r^{\text{BS}}} \mathbf{h}_{l'i'}^j \left(\boldsymbol{\eta}_{l'i'}^{\text{UE}} \right)^T \phi_{li}^* + \mathbf{G}_j^{\text{BS}} \phi_{li}^* + \mathbf{N}_j^p \phi_{li}^*. \end{aligned} \quad (\text{C.113})$$

A general LMMSE estimator expression is provided by Lemma B.19 on p. 571. In our case, $\mathbf{x} = \mathbf{h}_{li}^j$ and $\mathbf{y} = \mathbf{y}_{jli}^p$, and we notice that $\mathbb{E}\{\mathbf{x}\} = \mathbb{E}\{\mathbf{y}\} = \mathbf{0}_{M_j}$. Hence, the LMMSE estimator becomes

$$\hat{\mathbf{h}}_{li}^j = \mathbb{E}\{\mathbf{h}_{li}^j (\mathbf{y}_{jli}^p)^H\} \left(\mathbb{E}\{\mathbf{y}_{jli}^p (\mathbf{y}_{jli}^p)^H\} \right)^{-1} \mathbf{y}_{jli}^p. \quad (\text{C.114})$$

There are two expectations in (C.114) that need to be computed. The first one is

$$\begin{aligned} \mathbb{E}\{\mathbf{h}_{li}^j (\mathbf{y}_{jli}^p)^H\} &= \sum_{(l', i') \in \mathcal{P}_{li}} \sqrt{p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \tau_p \mathbb{E}\{\mathbf{h}_{li}^j (\mathbf{h}_{l'i'}^j)^H\} \\ &\quad + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \sqrt{\kappa_r^{\text{BS}}} \mathbb{E}\{\mathbf{h}_{li}^j (\mathbf{h}_{l'i'}^j)^H\} \underbrace{\mathbb{E}\{\phi_{li}^T \left(\boldsymbol{\eta}_{l'i'}^{\text{UE}} \right)^*\}}_{=0} \\ &\quad + \mathbb{E}\{\mathbf{h}_{li}^j \phi_{li}^T (\mathbf{G}_j^{\text{BS}})^H\} + \underbrace{\mathbb{E}\{\mathbf{h}_{li}^j\}}_{=\mathbf{0}_{M_j}} \underbrace{\phi_{li}^T \mathbb{E}\{(\mathbf{N}_j^p)^H\}}_{=\mathbf{0}_{\tau_p \times M_j}} \\ &= \sqrt{p_{li} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \tau_p \mathbf{R}_{li}^j \end{aligned} \quad (\text{C.115})$$

where the last equality follows from the fact that $\mathbb{E}\{\mathbf{h}_{li}^j (\mathbf{h}_{l'i'}^j)^H\} = \mathbf{0}_{M_j \times M_j}$ if $(l, i) \neq (l', i')$ and from

$$\mathbb{E}\{\mathbf{h}_{li}^j \phi_{li}^T (\mathbf{G}_j^{\text{BS}})^H\} = \mathbb{E}\left\{\underbrace{\mathbb{E}\{\mathbf{h}_{li}^j \phi_{li}^T (\mathbf{G}_j^{\text{BS}})^H | \{\mathbf{h}\}\}}_{=\mathbf{0}_{M_j \times M_j}}\right\} = \mathbf{0}_{M_j \times M_j}. \quad (\text{C.116})$$

Notice that we reached this result by conditioning on a set of channel realizations $\{\mathbf{h}\}$, to utilize that the conditional distribution of the distortion term has zero mean. The same approach can be used to prove that

the expectation of all cross-terms between \mathbf{h}_{li}^j and the noise/distortion terms are zero. This is utilized to compute the second expectation in (C.114) as

$$\begin{aligned} \mathbb{E}\{\mathbf{y}_{jli}^p(\mathbf{y}_{jli}^p)^H\} &\stackrel{(a)}{=} \sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\tau_p)^2 \mathbb{E}\{\mathbf{h}_{l'i'}^j(\mathbf{h}_{l'i'}^j)^H\} \\ &+ \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \kappa_r^{\text{BS}} \mathbb{E}\{\mathbf{h}_{l'i'}^j(\mathbf{h}_{l'i'}^j)^H\} \mathbb{E}\left\{\left|\left(\boldsymbol{\eta}_{l'i'}^{\text{UE}}\right)^T \boldsymbol{\phi}_{li}^*\right|^2\right\} \\ &+ \mathbb{E}\{\mathbf{G}_j^{\text{BS}} \boldsymbol{\phi}_{li}^* \boldsymbol{\phi}_{li}^T (\mathbf{G}_j^{\text{BS}})^H\} + \mathbb{E}\{\mathbf{N}_j^p \boldsymbol{\phi}_{li}^* \boldsymbol{\phi}_{li}^T (\mathbf{N}_j^p)^H\} \\ &\stackrel{(b)}{=} \sum_{(l',i') \in \mathcal{P}_{li}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\tau_p)^2 \mathbf{R}_{l'i'}^j + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \kappa_r^{\text{BS}} \mathbf{R}_{l'i'}^j \tau_p (1 - \kappa_t^{\text{UE}}) p_{l'i'} \\ &+ \tau_p (1 - \kappa_r^{\text{BS}}) \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} \mathbf{D}_{\mathbf{R}_{l'i'}^j} + \sigma_{\text{UL}}^2 \tau_p \mathbf{I}_{M_j} \end{aligned} \quad (\text{C.117})$$

where (a) follows from identifying the cross-terms that are zero and (b) follows from direct computation of the expectations. The only complicated computation is

$$\begin{aligned} \mathbb{E}\{\mathbf{G}_j^{\text{BS}} \boldsymbol{\phi}_{li}^* \boldsymbol{\phi}_{li}^T (\mathbf{G}_j^{\text{BS}})^H\} &= \mathbb{E}\left\{\mathbb{E}\{\mathbf{G}_j^{\text{BS}} \boldsymbol{\phi}_{li}^* \boldsymbol{\phi}_{li}^T (\mathbf{G}_j^{\text{BS}})^H | \{\mathbf{h}\}\}\right\} \\ &= \mathbb{E}\{\tau_p \mathbf{D}_{j,\{\mathbf{h}\}}\} = \tau_p (1 - \kappa_r^{\text{BS}}) \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} \mathbf{D}_{\mathbf{R}_{l'i'}^j} \end{aligned} \quad (\text{C.118})$$

where we condition on a set of channel realizations $\{\mathbf{h}\}$ to utilize the conditional distribution $\mathbf{G}_j^{\text{BS}} \boldsymbol{\phi}_{jk}^* | \{\mathbf{h}\} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{M_j}, \tau_p \mathbf{D}_{j,\{\mathbf{h}\}})$ of the receiver distortion term. We identify that $\boldsymbol{\Psi}_{li}^j = (\mathbb{E}\{\mathbf{y}_{jli}^p(\mathbf{y}_{jli}^p)^H\})^{-1} \tau_p$ and thus the LMMSE estimator in (C.114) becomes (6.23) when inserting (C.115) and (C.117). Finally, the estimation error correlation matrix is obtained from Lemma B.19 as

$$\mathbf{C}_{li}^j = \mathbb{E}\{\mathbf{h}_{li}^j(\mathbf{h}_{li}^j)^H\} - \mathbb{E}\{\mathbf{h}_{li}^j(\mathbf{y}_{jli}^p)^H\} \left(\mathbb{E}\{\mathbf{y}_{jli}^p(\mathbf{y}_{jli}^p)^H\}\right)^{-1} \left(\mathbb{E}\{\mathbf{h}_{li}^j(\mathbf{y}_{jli}^p)^H\}\right)^H \quad (\text{C.119})$$

which becomes (6.26) by inserting (C.115)–(C.117) and utilizing the fact that $\mathbb{E}\{\mathbf{h}_{li}^j(\mathbf{h}_{li}^j)^H\} = \mathbf{R}_{li}^j$.

C.5.2 Proof of Theorem 6.2

The received signal in (6.32) matches the discrete memoryless channel in Corollary 1.3 on p. 171 with the deterministic channel response $h = \sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}$, the input $x = s_{jk}$, and the output $y = \mathbf{v}_{jk}^H \mathbf{y}_j$. Using the notation from that corollary, the noise term is zero (i.e., $\sigma^2 = 0$) since the processed noise term $\mathbf{v}_{jk}^H \mathbf{n}_j$ might not be Gaussian distributed. The interference term is

$$\begin{aligned} v &= \sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \left(\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} \right) s_{jk} + \sqrt{\kappa_r^{\text{BS}}} \mathbf{v}_{jk}^H \mathbf{h}_{jk}^j \eta_{jk}^{\text{UE}} \\ &\quad + \sqrt{\kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j \left(\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}} \right) + \mathbf{v}_{jk}^H \boldsymbol{\eta}_j^{\text{BS}} + \mathbf{v}_{jk}^H \mathbf{n}_j \\ &= \sqrt{\kappa_r^{\text{BS}}} \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbf{v}_{jk}^H \mathbf{h}_{li}^j \left(\sqrt{\kappa_t^{\text{UE}}} s_{li} + \eta_{li}^{\text{UE}} \right) - \sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\} s_{jk} \\ &\quad + \mathbf{v}_{jk}^H \boldsymbol{\eta}_j^{\text{BS}} + \mathbf{v}_{jk}^H \mathbf{n}_j. \end{aligned} \tag{C.120}$$

Note that the interference term has zero mean and is uncorrelated with the input since

$$\mathbb{E}\{x^* v\} = \sqrt{\kappa_t^{\text{UE}} \kappa_r^{\text{BS}}} \underbrace{\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j - \mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}\} \mathbb{E}\{|s_{jk}|^2\}}_{=0} = 0 \tag{C.121}$$

which are two of the conditions for applying the capacity bound in Corollary 1.3. We further note that

$$\begin{aligned} p_v &= \mathbb{E}\{|v|^2\} \\ &= \kappa_r^{\text{BS}} \sum_{l=1}^L \sum_{i=1}^{K_l} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} \left(\kappa_t^{\text{UE}} \mathbb{E}\{|s_{li}|^2\} + \mathbb{E}\{|\eta_{li}^{\text{UE}}|^2\} \right) \\ &\quad - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} |\mathbb{E}\{\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j\}|^2 \mathbb{E}\{|s_{jk}|^2\} + \mathbb{E}\{|\mathbf{v}_{jk}^H \boldsymbol{\eta}_j^{\text{BS}}|^2\} + \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{n}_j|^2\} \\ &= \kappa_r^{\text{BS}} \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\} - \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} p_{jk} |\mathbb{E}\{\mathbf{v}_{jk}^H \hat{\mathbf{h}}_{jk}^j\}|^2 \\ &\quad + (1 - \kappa_r^{\text{BS}}) \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\} + \sigma_{\text{UL}}^2 \mathbb{E}\{\|\mathbf{v}_{jk}\|^2\} \end{aligned}$$

by utilizing the fact that the zero-mean signals s_{li} , the channels, and the transmitter distortion terms are all mutually independent. Moreover, the receiver distortion term η_j^{BS} is uncorrelated with all the other terms, which is easily proved by computing the conditional expectation for a given channel realization. Moreover, we utilized the distribution in (6.12) for the receiver distortion term to make the simplification

$$\begin{aligned}\mathbb{E}\{|\mathbf{v}_{jk}^H \boldsymbol{\eta}_j^{\text{BS}}|^2\} &= (1 - \kappa_r^{\text{BS}}) \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbb{E}\{|\mathbf{v}_{jk}^H (\mathbf{h}_{li}^j \odot \bar{\boldsymbol{\eta}}_{jli}^{\text{BS}})|^2\} \\ &= (1 - \kappa_r^{\text{BS}}) \sum_{l=1}^L \sum_{i=1}^{K_l} p_{li} \mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\}. \quad (\text{C.122})\end{aligned}$$

The lower capacity bound in (6.34) now follows from (1.9) by inserting the values of h and p_v provided above and then dividing all terms by $\kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}$. As a last step, we note that only the fraction τ_u/τ_c of the samples are used for UL data transmission, which yields the lower bound on the capacity that is stated in the theorem in bit/s/Hz.

C.5.3 Proof of Corollary 6.3

The expectations are computed by utilizing the statistics of the LMMSE estimate in Theorem 6.2 on p. 420. Specifically, (6.35) is computed as

$$\begin{aligned}\frac{|\mathbb{E}\{\mathbf{v}_{jk}^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} &= \frac{|\mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \mathbf{h}_{jk}^j\}|^2}{\mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \hat{\mathbf{h}}_{jk}^j\}} \stackrel{(a)}{=} \mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \hat{\mathbf{h}}_{jk}^j\} \\ &\stackrel{(b)}{=} \text{tr}(\mathbb{E}\{\hat{\mathbf{h}}_{jk}^j (\hat{\mathbf{h}}_{jk}^j)^H\}) \\ &\stackrel{(c)}{=} p_{jk} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j \quad (\text{C.123})\end{aligned}$$

where (a) follows from $\mathbf{h}_{jk}^j = \hat{\mathbf{h}}_{jk}^j + \tilde{\mathbf{h}}_{jk}^j$ and the fact that the estimate and the estimation error are uncorrelated (i.e., $\mathbb{E}\{(\hat{\mathbf{h}}_{jk}^j)^H \tilde{\mathbf{h}}_{jk}^j\} = 0$). Next, (b) follows from the matrix identity (B.5) in Lemma B.5 on p. 560 and (c) utilizes (6.27), which becomes $p_{jk} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\beta_{jk}^j)^2 \tau_p \psi_{jk} \mathbf{I}_{M_j}$ for spatially uncorrelated channels. Let $A_{jk} = \sqrt{p_{jk} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \beta_{jk}^j \psi_{jk}}$ and note that

$$\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\} = p_{jk} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} (\beta_{jk}^j)^2 \tau_p \psi_{jk} M_j = \frac{A_{jk}^2 \tau_p M_j}{\psi_{jk}} \quad (\text{C.124})$$

with this notation. By utilizing the fact that $\hat{\mathbf{h}}_{jk}^j = A_{jk}\mathbf{y}_{jjk}^p$, the interference term in (6.36) can be expressed as

$$\begin{aligned}
& \frac{\mathbb{E}\{|\mathbf{v}_{jk}^H \mathbf{h}_{li}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} = A_{jk}^2 \frac{\mathbb{E}\{(|\mathbf{y}_{jjk}^p)^H \mathbf{h}_{li}^j|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} = \frac{\psi_{jk}}{\tau_p M_j} \mathbb{E}\{(|(\mathbf{y}_{jjk}^p)^H \mathbf{h}_{li}^j|^2\} \\
& \stackrel{(a)}{=} \sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \tau_p \frac{\psi_{jk}}{M_j} \mathbb{E} \left\{ \left| (\mathbf{h}_{l'i'}^j)^H \mathbf{h}_{li}^j \right|^2 \right\} \\
& + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \kappa_r^{\text{BS}} \frac{\psi_{jk}}{\tau_p M_j} \mathbb{E} \left\{ \left| (\mathbf{h}_{l'i'}^j)^H \mathbf{h}_{li}^j \right|^2 \right\} \mathbb{E} \left\{ \left| (\boldsymbol{\eta}_{l'i'}^{\text{UE}})^T \boldsymbol{\phi}_{jk}^* \right|^2 \right\} \\
& + \frac{\psi_{jk}}{\tau_p M_j} \left(\mathbb{E}\{ |(\mathbf{G}_j^{\text{BS}} \boldsymbol{\phi}_{jk}^*)^H \mathbf{h}_{li}^j|^2 \} + \mathbb{E}\{ |(\mathbf{N}_j^p \boldsymbol{\phi}_{jk}^*)^H \mathbf{h}_{li}^j|^2 \} \right) \\
& \stackrel{(b)}{=} \frac{\psi_{jk}}{M_j} \left(\sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \tau_p \mathbb{E} \left\{ \left| (\mathbf{h}_{l'i'}^j)^H \mathbf{h}_{li}^j \right|^2 \right\} \right. \\
& + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} (1 - \kappa_t^{\text{UE}}) \kappa_r^{\text{BS}} \mathbb{E} \left\{ \left| (\mathbf{h}_{l'i'}^j)^H \mathbf{h}_{li}^j \right|^2 \right\} \\
& \left. + (1 - \kappa_r^{\text{BS}}) \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} \mathbb{E}\{ \|\mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j\|^2 \} + \beta_{li}^j M_j \sigma_{\text{UL}}^2 \right) \quad (\text{C.125})
\end{aligned}$$

where (a) follows from (6.22) by utilizing the fact that all cross-terms are zero (a consequence of the circular symmetry of the Gaussian variables). By computing the expectations with respect to the distortion terms and noise, conditioned on the channel realizations, we further obtain (b). It remains to compute the expectations with respect to the channels, which are obtained as

$$\mathbb{E} \left\{ \left| (\mathbf{h}_{l'i'}^j)^H \mathbf{h}_{li}^j \right|^2 \right\} = \begin{cases} (M_j^2 + M_j)(\beta_{li}^j)^2 & (l', i') = (l, i) \\ M_j \beta_{l'i'}^j \beta_{li}^j & (l', i') \neq (l, i) \end{cases} \quad (\text{C.126})$$

$$\begin{aligned}
\mathbb{E}\{ \|\mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j\|^2 \} &= \sum_{m=1}^{M_j} \mathbb{E}\{ |[\mathbf{h}_{l'i'}^j]_m|^2 |[\mathbf{h}_{li}^j]_m|^2 \} \\
&= \begin{cases} 2M_j(\beta_{li}^j)^2 & (l', i') = (l, i) \\ M_j \beta_{l'i'}^j \beta_{li}^j & (l', i') \neq (l, i) \end{cases} \quad (\text{C.127})
\end{aligned}$$

by utilizing Lemma B.14 on p. 564 to compute fourth-order moments of the Gaussian random variables. Substituting (C.126) and (C.127) into (C.125) yields the expression in (6.36) after some algebra (which includes identifying that many of the terms within the parenthesis add up to $\psi_{jk}^{-1}\beta_{li}^j$). Note that the first term in (C.125) is different depending on whether $(l, i) \in \mathcal{P}_{jk}$ or not (i.e., if the UEs are using the same pilot sequence or not).

The derivation of (6.37) follows along similar lines:

$$\begin{aligned}
& \frac{\mathbb{E}\{\|\mathbf{v}_{jk} \odot \mathbf{h}_{li}^j\|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} = A_{jk}^2 \frac{\mathbb{E}\{\|\mathbf{y}_{jjk}^p \odot \mathbf{h}_{li}^j\|^2\}}{\mathbb{E}\{\|\mathbf{v}_{jk}\|^2\}} = \frac{\psi_{jk}}{\tau_p M_j} \mathbb{E}\{\|\mathbf{y}_{jjk}^p \odot \mathbf{h}_{li}^j\|^2\} \\
&= \sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \tau_p \frac{\psi_{jk}}{M_j} \mathbb{E} \left\{ \left\| \mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j \right\|^2 \right\} \\
&\quad + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} \kappa_r^{\text{BS}} \frac{\psi_{jk}}{\tau_p M_j} \mathbb{E} \left\{ \left\| \mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j \right\|^2 \right\} \mathbb{E} \left\{ \left| \left(\boldsymbol{\eta}_{l'i'}^{\text{UE}} \right)^T \boldsymbol{\phi}_{jk}^* \right|^2 \right\} \\
&\quad + \frac{\psi_{jk}}{\tau_p M_j} \left(\mathbb{E} \left\{ \left\| (\mathbf{G}_j^{\text{BS}} \boldsymbol{\phi}_{jk}^*)^* \odot \mathbf{h}_{li}^j \right\|^2 \right\} + \mathbb{E} \left\{ \left\| (\mathbf{N}_j^p \boldsymbol{\phi}_{jk}^*) \odot \mathbf{h}_{li}^j \right\|^2 \right\} \right) \\
&= \frac{\psi_{jk}}{M_j} \left(\sum_{(l', i') \in \mathcal{P}_{jk}} p_{l'i'} \kappa_t^{\text{UE}} \kappa_r^{\text{BS}} \tau_p \mathbb{E} \left\{ \left\| \mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j \right\|^2 \right\} \right. \\
&\quad \left. + \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} (1 - \kappa_t^{\text{UE}}) \kappa_r^{\text{BS}} \mathbb{E} \left\{ \left\| \mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j \right\|^2 \right\} \right. \\
&\quad \left. + (1 - \kappa_r^{\text{BS}}) \sum_{l'=1}^L \sum_{i'=1}^{K_{l'}} p_{l'i'} \mathbb{E} \left\{ \left\| \mathbf{h}_{l'i'}^j \odot \mathbf{h}_{li}^j \right\|^2 \right\} + \beta_{li}^j M_j \sigma_{\text{UL}}^2 \right) \quad (\text{C.128})
\end{aligned}$$

which becomes (6.37) by using (C.127) and some algebra (e.g., identifying that many of the terms within the parenthesis add up to $\psi_{jk}^{-1}\beta_{li}^j$).

The SE expression in (6.39) is finally obtained from (6.34) by inserting the closed-form expressions for the expectations that were computed above and then performing some algebra that involves identifying \bar{F}_{li}^{jk} and \bar{G}_j in the expressions.

C.5.4 Proof of Theorem 6.5

The received signal in (6.44) matches the discrete memoryless channel in Corollary 1.3 on p. 171 with the deterministic channel response $h = \sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}$, the input $x = \varsigma_{jk}$, and the output $y = y_{jk}$. Using the notation from that corollary, the noise term is $n = n_{jk} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{DL}}^2)$ and the interference term is

$$\begin{aligned}
 v &= \sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \left((\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \right) \varsigma_{jk} + \sqrt{\kappa_r^{\text{UE}}} \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \boldsymbol{\mu}_l^{\text{BS}} \\
 &\quad + \sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \sum_{\substack{i=1 \\ i \neq k}}^{K_j} (\mathbf{h}_{jk}^j)^H \mathbf{w}_{ji} \varsigma_{ji} + \sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \sum_{l=1}^L \sum_{\substack{i=1 \\ l \neq j}}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li} + \mu_{jk}^{\text{UE}} \\
 &= \sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \left(\sum_{l=1}^L \sum_{i=1}^{K_l} (\mathbf{h}_{jk}^l)^H \mathbf{w}_{li} \varsigma_{li} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\} \varsigma_{jk} \right) \\
 &\quad + \mu_{jk}^{\text{UE}} + \sqrt{\kappa_r^{\text{UE}}} \sum_{l=1}^L (\mathbf{h}_{jk}^l)^H \boldsymbol{\mu}_l^{\text{BS}}. \tag{C.129}
 \end{aligned}$$

Note that the interference term has zero mean and is uncorrelated with the input since

$$\mathbb{E}\{x^* v\} = \sqrt{\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}} \underbrace{\mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} - \mathbb{E}\{(\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk}\}\} \mathbb{E}\{|\varsigma_{jk}|^2\}}_0 = 0 \tag{C.130}$$

which are two of the conditions for applying Corollary 1.3. We further note that

$$\begin{aligned}
p_v &= \mathbb{E} \left\{ |v|^2 \right\} \\
&= \kappa_t^{\text{BS}} \kappa_r^{\text{UE}} \left(\sum_{l=1}^L \sum_{i=1}^{K_l} \mathbb{E} \{ |(\mathbf{h}_{jk}^l)^H \mathbf{w}_{li}|^2 \} \mathbb{E} \{ |\varsigma_{li}|^2 \} - |\mathbb{E} \{ (\mathbf{h}_{jk}^j)^H \mathbf{w}_{jk} \}|^2 \mathbb{E} \{ |\varsigma_{jk}|^2 \} \right) \\
&\quad + \mathbb{E} \{ |\mu_{jk}^{\text{UE}}|^2 \} + \kappa_r^{\text{UE}} \sum_{l=1}^L \mathbb{E} \{ |(\mathbf{h}_{jk}^l)^H \boldsymbol{\mu}_l^{\text{BS}}|^2 \} \\
&= \kappa_t^{\text{BS}} \kappa_r^{\text{UE}} \left(\sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \mathbb{E} \{ |\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2 \} - \rho_{jk} \mathbb{E} \{ \mathbf{w}_{jk}^H \mathbf{h}_{jk}^j \} \right)^2 \\
&\quad + (1 - \kappa_r^{\text{UE}}) \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \left(\kappa_t^{\text{BS}} \mathbb{E} \{ |(\mathbf{h}_{jk}^l)^H \mathbf{w}_{li}|^2 \} + (1 - \kappa_t^{\text{BS}}) \mathbb{E} \{ \|\mathbf{w}_{li} \odot \mathbf{h}_{jk}^l\|^2 \} \right) \\
&\quad + \kappa_r^{\text{UE}} (1 - \kappa_t^{\text{BS}}) \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \mathbb{E} \{ \|\mathbf{w}_{li} \odot \mathbf{h}_{jk}^l\|^2 \} \\
&= \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} \kappa_t^{\text{BS}} \mathbb{E} \{ |\mathbf{w}_{li}^H \mathbf{h}_{jk}^l|^2 \} - \rho_{jk} \kappa_t^{\text{BS}} \kappa_r^{\text{UE}} |\mathbb{E} \{ \mathbf{w}_{jk}^H \mathbf{h}_{jk}^j \}|^2 \\
&\quad + \sum_{l=1}^L \sum_{i=1}^{K_l} \rho_{li} (1 - \kappa_t^{\text{BS}}) \mathbb{E} \{ \|\mathbf{w}_{li} \odot \mathbf{h}_{jk}^l\|^2 \} \tag{C.131}
\end{aligned}$$

by utilizing the fact that the zero-mean signals ς_{li} and the channels are mutually independent, while the transmitter and receiver distortion terms are uncorrelated with all other terms, which is easily proved by computing the conditional expectation for a given channel realization. The second equality follows from applying the conditional correlation matrix expressions from (6.16) and (6.18) for the distortion terms, while the last equality follows from noting that the same terms appear several times with different factors in front.

The SINR expression in (6.46) now follows from (1.9) by inserting the values of h and p_v determined above, and then dividing all terms by $\kappa_t^{\text{BS}} \kappa_r^{\text{UE}}$. As a last step, we note that only the fraction τ_d/τ_c of the samples are used for DL data transmission, which yields the lower bound on the capacity stated in the theorem in bit/s/Hz.

C.5.5 Proof of Corollary 6.6

This corollary follows from computing the expectations in Theorem 6.5 for average-normalized MR precoding with $\mathbf{w}_{jk} = \hat{\mathbf{h}}_{jk}^j / \sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_{jk}^j\|^2\}}$. These expectations are the same as those computed in (6.35)–(6.37), except that the indices (l, i) and (k, j) are swapped in all the interference terms. In particular, (6.36) becomes

$$\begin{aligned} \frac{\mathbb{E}\{|\hat{\mathbf{h}}_{li}^l|^2\}}{\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^l\|^2\}} &= \beta_{jk}^l + p_{jk}(\beta_{jk}^l)^2 \psi_{li} \left(1 - \kappa_r^{\text{BS}} + (1 - \kappa_t^{\text{UE}})\kappa_r^{\text{BS}} M_l\right) \\ &+ \begin{cases} p_{jk}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}(\beta_{jk}^l)^2\tau_p\psi_{li}M_l & (l, i) \in \mathcal{P}_{jk} \\ 0 & (l, i) \notin \mathcal{P}_{jk} \end{cases} \quad (\text{C.132}) \end{aligned}$$

and (6.37) becomes

$$\begin{aligned} \frac{\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^l \odot \mathbf{h}_{jk}^l\|^2\}}{\mathbb{E}\{\|\hat{\mathbf{h}}_{li}^l\|^2\}} &= \beta_{jk}^l + p_{jk}(\beta_{jk}^l)^2 \psi_{li} \left(1 - \kappa_t^{\text{UE}}\kappa_r^{\text{BS}}\right) \\ &+ \begin{cases} p_{jk}\kappa_t^{\text{UE}}\kappa_r^{\text{BS}}(\beta_{jk}^l)^2\tau_p\psi_{li} & (l, i) \in \mathcal{P}_{jk} \\ 0 & (l, i) \notin \mathcal{P}_{jk} \end{cases} \quad (\text{C.133}) \end{aligned}$$

where we have also utilized the facts that $\psi_{li} = \psi_{jk}$ and $\mathcal{P}_{li} = \mathcal{P}_{jk}$ whenever $(l, i) \in \mathcal{P}_{jk}$. Inserting (6.35), (C.132), and (C.133) into (6.46) and simplifying the expressions, including identifying \underline{F}_{jk}^{li} and \underline{G}_l , yields the final expression in (6.39).

References

- [1] 3GPP TR 36.873. 2015. “Study on 3D channel model for LTE”. *Tech. rep.*
- [2] 3GPP TS 25.213. 2006. “Universal Mobile Telecommunications System (UMTS); Spreading and modulation (FDD)”. *Tech. rep.*
- [3] Abramowitz, M. and I. Stegun. 1965. *Handbook of mathematical functions*. Dover Publications.
- [4] Adachi, F., M. T. Feeney, J. D. Parsons, and A. G. Williamson. 1986. “Crosscorrelation between the envelopes of 900 MHz signals received at a mobile radio base station site”. *IEE Proc. F - Commun., Radar and Signal Process.* 133(6): 506–512.
- [5] Ademaj, F., M. Tarantza, and M. Rupp. 2016. “3GPP 3D MIMO channel model: A holistic implementation guideline for open source simulation tools”. *EURASIP J. Wirel. Commun. Netw.* (55): 1–14.
- [6] Adhikary, A., A. Ashikhmin, and T. L. Marzetta. 2017. “Uplink interference reduction in Large Scale Antenna Systems”. *IEEE Trans. Commun.* 65(5): 2194–2206.
- [7] Adhikary, A., J. Nam, J.-Y. Ahn, and G. Caire. 2013. “Joint spatial division and multiplexing—The large-scale array regime”. *IEEE Trans. Inf. Theory.* 59(10): 6441–6463.

- [8] Adhikary, A., E. A. Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch. 2014. “Joint spatial division and multiplexing for mm-wave channels”. *IEEE J. Sel. Areas Commun.* 32(6): 1239–1255.
- [9] Adhikary, A., H. S. Dhillon, and G. Caire. 2015. “Massive-MIMO meets HetNet: Interference coordination through spatial blanking”. *IEEE J. Sel. Areas Commun.* 33(6): 1171–1186.
- [10] Alexanderson, E. F. W. 1919. “Transatlantic radio communication”. *Trans. American Institute of Electrical Engineers.* 38(2): 1269–1285.
- [11] Alkhateeb, A., O. El Ayach, G. Leus, and R. W. Heath. 2013. “Hybrid precoding for millimeter wave cellular systems with partial channel knowledge”. In: *Proc. IEEE ITA*. IEEE. 1–5.
- [12] Almers, P., E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, D. Laurenson, G. Matz, et al. 2007. “Survey of channel and radio propagation models for wireless MIMO systems”. *EURASIP J. Wirel. Commun. Netw.* (1): 1–19.
- [13] Amari, S. V. and R. B. Misra. 1997. “Closed-form expressions for distribution of sum of exponential random variables”. *IEEE Trans. Rel.* 46(4): 519–522.
- [14] Anderson, N. 2009. “Paired and unpaired spectrum”. In: *LTE - The UMTS Long Term Evolution: From Theory to Practice*. Ed. by S. Sesia, I. Toufik, and M. Baker. Wiley. Chap. 23. 551–583.
- [15] Anderson, S., U. Forssen, J. Karlsson, T. Witzschel, P. Fischer, and A. Krug. 1996. “Ericsson/Mannesmann GSM field-trials with adaptive antennas”. In: *Proc. IEE Colloquium on Advanced TDMA Techniques and Applications*.
- [16] Anderson, S., B. Hagerman, H. Dam, U. Forssen, J. Karlsson, F. Kronestedt, S. Mazur, and K. J. J. Molnar. 1999. “Adaptive antennas for GSM and TDMA systems”. *IEEE Personal Commun.* 6(3): 74–86.
- [17] Anderson, S., M. Millnert, M. Viberg, and B. Wahlberg. 1991. “An adaptive array for mobile communication systems”. *IEEE Trans. Veh. Technol.* 40(1): 230–236.

- [18] Andrews, J. G., F. Baccelli, and R. K. Ganti. 2011. “A tractable approach to coverage and rate in cellular networks”. *IEEE Trans. Commun.* 59(11): 3122–3134.
- [19] Andrews, J. G., X. Zhang, G. D. Durgin, and A. K. Gupta. 2016. “Are we approaching the fundamental limits of wireless network densification?” *IEEE Commun. Mag.* 54(10): 184–190.
- [20] Annadreddy, V. S. and V. V. Veeravalli. 2011a. “Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on the capacity region”. *IEEE Trans. Inf. Theory.* 55(7): 3032–3050.
- [21] Annadreddy, V. S. and V. V. Veeravalli. 2011b. “Sum capacity of MIMO interference channels in the low interference regime”. *IEEE Trans. Inf. Theory.* 57(5): 2565–2581.
- [22] ApS, M. 2016. *MOSEK optimization suite release 8.0.0.42*. URL: <http://docs.mosek.com/8.0/intro.pdf>.
- [23] Ashikhmin, A. and T. Marzetta. 2012. “Pilot contamination precoding in multi-cell large scale antenna systems”. In: *IEEE International Symposium on Information Theory Proceedings (ISIT)*. 1137–1141.
- [24] Ashraf, I., F. Boccardi, and L. Ho. 2011. “SLEEP mode techniques for small cell deployments”. *IEEE Commun. Mag.* 49(8): 72–79.
- [25] Asplund, H., J.-E. Berg, F. Harrysson, J. Medbo, and M. Riback. 2007. “Propagation characteristics of polarized radio waves in cellular communications”. In: *Proc. IEEE VTC-Fall*. 839–843.
- [26] Auer, G., O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading, E. Katranaras, M. Olsson, D. Sabella, P. Skillermak, and W. Wajda. 2012. *D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown*. INFSO-ICT-247733 EARTH, ver. 2.0.
- [27] Auer, G., V. Giannini, C. Desset, I. Godor, P. Skillermak, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske. 2011. “How much energy is needed to run a wireless network?” *IEEE Wireless Commun.* 18(5): 40–49.

- [28] Baird, C. A. and C. L. Zahm. 1971. “Performance criteria for narrowband array processing”. In: *Proc. IEEE Conf. Decision and Control*. 564–565.
- [29] Baumert, L. D. and M. Hall. 1965. “Hadamard matrices of the Williamson type”. *Math. Comp.* 19(6): 442–447.
- [30] Bazelon, C. and G. McHenry. 2015. “Mobile broadband spectrum: A vital resource for the U.S. economy”. *Tech. rep.* The Brattle Group.
- [31] Bengtsson, E., P. Karlsson, F. Tufvesson, J. Vieira, S. Malkowsky, L. Liu, F. Rusek, and O. Edfors. 2016. “Transmission Schemes for Multiple Antenna Terminals in real Massive MIMO systems”. In: *Proc. IEEE GLOBECOM*.
- [32] Bengtsson, E., F. Tufvesson, and O. Edfors. 2015. “UE antenna properties and their influence on massive MIMO system performance”. In: *Proc. EuCAP*.
- [33] Bengtsson, M. and B. Ottersten. 2001. “Optimal and suboptimal transmit beamforming”. In: *Handbook of Antennas in Wireless Communications*. Ed. by L. C. Godara. CRC Press.
- [34] Bertsekas, D. 1999. *Nonlinear programming*. 2nd edition. Athena Scientific.
- [35] Bethanabhotla, D., O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire. 2016. “Optimal user-cell association for massive MIMO wireless networks”. *IEEE Trans. Wireless Commun.* 15(3): 1835–1850.
- [36] Biglieri, E., J. Proakis, and S. Shamai. 1998. “Fading channels: Information-theoretic and communications aspects”. *IEEE Trans. Inf. Theory*. 44(6): 2619–2691.
- [37] Biguesh, M. and A. Gershman. 2004. “Downlink channel estimation in cellular systems with antenna arrays at base stations using channel probing with feedback”. *EURASIP J. Appl. Signal Process.* (9): 1330–1339.
- [38] Björnson, E. and B. Ottersten. 2010. “A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance”. *IEEE Trans. Signal Process.* 58(3): 1807–1820.

- [39] Björnson, E., M. Bengtsson, and B. Ottersten. 2012. “Pareto characterization of the multicell MIMO performance region with simple receivers”. *IEEE Trans. Signal Process.* 60(8): 4464–4469.
- [40] Björnson, E., M. Bengtsson, and B. Ottersten. 2014a. “Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure”. *IEEE Signal Process. Mag.* 31(4): 142–148.
- [41] Björnson, E., E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski. 2017a. “A random access protocol for pilot allocation in crowded massive MIMO systems”. *IEEE Trans. Wireless Commun.* 16(4): 2220–2234.
- [42] Björnson, E., J. Hoydis, M. Kountouris, and M. Debbah. 2014b. “Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits”. *IEEE Trans. Inf. Theory.* 60(11): 7112–7139.
- [43] Björnson, E., J. Hoydis, and L. Sanguinetti. 2017b. “Massive MIMO has Unlimited Capacity”. *CoRR*. abs/1705.00538. URL: <http://arxiv.org/abs/1705.00538>.
- [44] Björnson, E., J. Hoydis, and L. Sanguinetti. 2017c. “Pilot contamination is not a fundamental asymptotic limitation in massive MIMO”. In: *Proc. IEEE ICC*.
- [45] Björnson, E., N. Jaldén, M. Bengtsson, and B. Ottersten. 2011. “Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission”. *IEEE Trans. Signal Process.* 59(12): 6086–6101.
- [46] Björnson, E. and E. Jorswieck. 2013. “Optimal resource allocation in coordinated multi-cell systems”. *Foundations and Trends in Communications and Information Theory*. 9(2-3): 113–381.
- [47] Björnson, E., E. Jorswieck, M. Debbah, and B. Ottersten. 2014c. “Multi-objective signal processing optimization: The way to balance conflicting metrics in 5G systems”. *IEEE Signal Process. Mag.* 31(6): 14–23.

- [48] Björnson, E. and E. G. Larsson. 2015. “Three practical aspects of massive MIMO: Intermittent user activity, pilot synchronism, and asymmetric deployment”. In: *Proc. IEEE GLOBECOM Workshops*.
- [49] Björnson, E., E. G. Larsson, and M. Debbah. 2016a. “Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?” *IEEE Trans. Wireless Commun.* 15(2): 1293–1308.
- [50] Björnson, E., E. G. Larsson, and T. L. Marzetta. 2016b. “Massive MIMO: Ten myths and one critical question”. *IEEE Commun. Mag.* 54(2): 114–123.
- [51] Björnson, E., M. Kountouris, and M. Debbah. 2013a. “Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination”. In: *Proc. IEEE ICT*.
- [52] Björnson, E., M. Kountouris, M. Bengtsson, and B. Ottersten. 2013b. “Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users”. *IEEE Trans. Signal Process.* 61(13): 3431–3446.
- [53] Björnson, E., M. Matthaiou, and M. Debbah. 2015a. “Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design”. *IEEE Trans. Wireless Commun.* 14(8): 4353–4368.
- [54] Björnson, E., M. Matthaiou, A. Pitarokilis, and E. G. Larsson. 2015b. “Distributed massive MIMO in cellular networks: Impact of imperfect hardware and number of oscillators”. In: *Proc. EUSIPCO*.
- [55] Björnson, E. and B. Ottersten. 2008. “Post-user-selection quantization and estimation of correlated Frobenius and spectral channel norms”. In: *Proc. IEEE PIMRC*.
- [56] Björnson, E., P. Zetterberg, M. Bengtsson, and B. Ottersten. 2013c. “Capacity limits and multiplexing gains of MIMO channels with transceiver impairments”. *IEEE Commun. Lett.* 17(1): 91–94.

- [57] Björnson, E., L. Sanguinetti, and M. Debbah. 2016c. “Massive MIMO with imperfect channel covariance information”. In: *Proc. ASILOMAR*.
- [58] Björnson, E., L. Sanguinetti, J. Hoydis, and M. Debbah. 2014d. “Designing multi-user MIMO for energy efficiency: When is massive MIMO the answer?” In: *Proc. IEEE WCNC*. 242–247.
- [59] Björnson, E., L. Sanguinetti, J. Hoydis, and M. Debbah. 2015c. “Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?” *IEEE Trans. Wireless Commun.* 14(6): 3059–3075.
- [60] Björnson, E., L. Sanguinetti, and M. Kountouris. 2016d. “Deploying dense networks for maximal energy efficiency: Small cells meet massive MIMO”. *IEEE J. Sel. Areas Commun.* 34(4): 832–847.
- [61] Björnson, E., R. Zakhour, D. Gesbert, and B. Ottersten. 2010. “Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical CSI”. *IEEE Trans. Signal Process.* 58(8): 4298–4310.
- [62] Blandino, S., C. Dessel, A. Bourdoux, L. V. der Perre, and S. Pollin. 2017. “Analysis of out-of-band interference from saturated power amplifiers in Massive MIMO”. In: *Proc. IEEE EuCNC*.
- [63] Boche, H. and M. Schubert. 2002. “A general duality theory for uplink and downlink beamforming”. In: *Proc. IEEE VTC-Fall*. 87–91.
- [64] Bock, F. and B. Ebstein. 1964. “Assignment of transmitter powers by linear programming”. *IEEE Trans. Electromagn. Compat.* 6(2): 36–44.
- [65] Boström, J. 2015. “Spectrum for mobile – A Swedish perspective for 2020 and beyond”. *Tech. rep.* Swedish Post and Telecom Authority (PTS). URL: <http://wireless.kth.se/wp-content/uploads/2015/02/KTH-Frekvenser-f%5C%22%7Bo%7Dr-4G-och-5G.pdf>.
- [66] Boyd, S., S.-J. Kim, L. Vandenberghe, and A. Hassibi. 2007. “A tutorial on geometric programming”. *Optimization and Engineering*. 8: 67–127.

- [67] Boyd, S. and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.
- [68] Brennan, D. G. 1959. “Linear diversity combining techniques”. *Proc. IRE*. 43(6): 1975–1102.
- [69] Bussgang, J. J. 1952. “Crosscorrelation functions of amplitude-distorted Gaussian signals”. *Tech. rep.* No. 216. Research Laboratory of Electronics, Massachusetts Institute of Technology.
- [70] Cadambe, V. and S. Jafar. 2008. “Interference alignment and degrees of freedom of the K -user interference channel”. *IEEE Trans. Inf. Theory*. 54(8): 3425–3441.
- [71] Cai, D. W. H., T. Q. S. Quek, C. W. Tan, and S. H. Low. 2012. “Max-min SINR coordinated multipoint downlink transmission—Duality and algorithms”. *IEEE Trans. Signal Process.* 60(10): 5384–5395.
- [72] Caire, G. 2017. “On the Ergodic Rate Lower Bounds with Applications to Massive MIMO”. *CoRR*. abs/1705.03577. URL: <http://arxiv.org/abs/1705.03577>.
- [73] Caire, G., N. Jindal, M. Kobayashi, and N. Ravindran. 2010. “Multiuser MIMO achievable rates with downlink training and channel state feedback”. *IEEE Trans. Inf. Theory*. 56(6): 2845–2866.
- [74] Caire, G. and S. Shamai. 2003. “On the achievable throughput of a multiantenna Gaussian broadcast channel”. *IEEE Trans. Inf. Theory*. 49(7): 1691–1706.
- [75] Carvalho, E. D. and D.T.M. Slock. 1997. “Cramer-Rao bounds for semi-blind, blind and training sequence based channel estimation”. In: *Proc. IEEE SPAWC*. 129–132.
- [76] Carvalho, E. de, E. Björnson, J. H. Sørensen, P. Popovski, and E. G. Larsson. 2017. “Random access protocols for Massive MIMO”. *IEEE Commun. Mag.* 54(5): 216–222.
- [77] Chen, J. and V. Lau. 2014. “Two-Tier Precoding for FDD Multi-Cell Massive MIMO Time-Varying Interference Networks”. *IEEE J. Sel. Areas Commun.* 32(6): 1230–1238.

- [78] Chen, Z. and E. Björnson. 2017. “Channel Hardening and Favorable Propagation in Cell-Free Massive MIMO with Stochastic Geometry”. *CoRR*. abs/1710.00395. URL: <http://arxiv.org/abs/1710.00395>.
- [79] Chen, Z. N. and K.-M. Luk. 2009. *Antennas for base stations in wireless communications*. McGraw-Hill.
- [80] Cheng, H. V., E. Björnson, and E. G. Larsson. 2017. “Optimal pilot and payload power control in single-cell massive MIMO systems”. *IEEE Trans. Signal Process.*
- [81] Chiang, M., P. Hande, T. Lan, and C. Tan. 2008. “Power control in wireless cellular networks”. *Foundations and Trends in Networking*. 2(4): 355–580.
- [82] Chien, T. V., E. Björnson, and E. G. Larsson. 2016. “Joint power allocation and user association optimization for massive MIMO systems”. *IEEE Trans. Wireless Commun.* 15(9): 6384–6399.
- [83] Chizhik, D., J. Ling, P. W. Wolniansky, R. A. Valenzuela, N. Costa, and K. Huber. 2003. “Multiple-input-multiple-output measurements and modeling in Manhattan”. *IEEE J. Sel. Areas Commun.* 21(3): 321–331.
- [84] Choi, J., D. Love, and P. Bidigare. 2014. “Downlink Training Techniques for FDD Massive MIMO Systems: Open-Loop and Closed-Loop Training with Memory”. *IEEE J. Sel. Topics Signal Process.* 8(5): 802–814.
- [85] Chong, C.-C., C.-M. Tan, D. I. Laurenson, S. McLaughlin, M. A. Beach, and A. R. Nix. 2003. “A new statistical wideband spatio-temporal channel model for 5-GHz band WLAN systems”. *IEEE J. Sel. Areas Commun.* 21(2): 139–150.
- [86] Cisco. 2016. “Visual networking index: Global mobile data traffic forecast update, 2015 – 2020”. *Tech. rep.*
- [87] Clerckx, B., G. Kim, and S. Kim. 2008. “Correlated Fading in Broadcast MIMO Channels: Curse or Blessing?” In: *Proc. IEEE GLOBECOM*.
- [88] Clerckx, B. and C. Oestges. 2013. *MIMO wireless networks: Channels, techniques and standards for multi-antenna, multi-user and multi-cell systems*. Academic Press.

- [89] Coldrey, M. 2008. "Modeling and capacity of polarized MIMO channels". In: *Proc. IEEE VTC-Spring*. IEEE. 440–444.
- [90] Common Public Radio Interface (CPRI). 2015. "Interface specification". *Tech. rep.*
- [91] Cooper, M. 2010. "The Myth of Spectrum Scarcity". *Tech. rep.* DYNA llc. URL: <https://ecfsapi.fcc.gov/file/7020396128.pdf>.
- [92] Costa, E. and S. Pupolin. 2002. "M-QAM-OFDM system performance in the presence of a nonlinear amplifier and phase noise". *IEEE Trans. Commun.* 50(3): 462–472.
- [93] Couillet, R. and M. Debbah. 2011. *Random matrix methods for wireless communications*. Cambridge University Press.
- [94] Cover, T. M. and J. A. Thomas. 1991. *Elements of information theory*. Wiley.
- [95] Cui, S., A. Goldsmith, and A. Bahai. 2004. "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks". *IEEE J. Sel. Areas Commun.* 22(6): 1089–1098.
- [96] Cupo, R. L., G. D. Golden, C. C. Martin, K. L. Sherman, N. R. Sollenberger, J. H. Winters, and P. W. Wolniansky. 1997. "A four-element adaptive antenna array for IS-136 PCS base stations". In: *Proc. IEEE VTC*. 1577–1581.
- [97] D. Gesbert, L. Pittman, and M. Kountouris. 2006. "Transmit Correlation-Aided Scheduling in multiuser MIMO networks". In: *Proc. IEEE ICASSP*. Vol. 4. 249–252.
- [98] D. Hammarwall, M. Bengtsson, and B. Ottersten. 2008. "Utilizing the Spatial Information Provided by Channel Norm Feedback in SDMA Systems". *IEEE Trans. Signal Process.* 56(7): 3278–3293.
- [99] Dabeer, O. and U. Madhow. 2010. "Channel estimation with low-precision analog-to-digital conversion". In: *Proc. IEEE ICC*.
- [100] Dahrouj, H. and W. Yu. 2010. "Coordinated beamforming for the multicell multi-antenna wireless system". *IEEE Trans. Wireless Commun.* 9(5): 1748–1759.
- [101] Demir, A., A. Mehrotra, and J. Roychowdhury. 2000. "Phase noise in oscillators: A unifying theory and numerical methods for characterization". *IEEE Trans. Circuits Syst. I*. 47(5): 655–674.

- [102] Dessel, C. and B. Debaillie. 2016. “Massive MIMO for energy-efficient communications”. In: *Proc. EuMC*. 138–141.
- [103] Dessel, C. and L. V. der Perre. 2015. “Validation of low-accuracy quantization in massive MIMO and constellation EVM analysis”. In: *Proc. IEEE EuCNC*.
- [104] Dhillon, H. S., M. Kountouris, and J. G. Andrews. 2013. “Downlink MIMO HetNets: Modeling, ordering results and performance analysis”. *IEEE Trans. Wireless Commun.* 12(10): 5208–5222.
- [105] Doukopoulos, X. G. and G. V. Moustakides. 2008. “Fast and stable subspace tracking”. *IEEE Trans. Signal Process.* 56(4): 4790–4807.
- [106] Duel-Hallen, A., J. Holtzman, and Z. Zvonar. 1995. “Multiuser Detection for CDMA Systems”. *IEEE Personal Commun.* 2(2): 46–58.
- [107] Durisi, G., A. Tarable, C. Camarda, R. Devassy, and G. Montorsi. 2014. “Capacity bounds for MIMO microwave backhaul links affected by phase noise”. *IEEE Trans. Commun.* 62(3): 920–929.
- [108] Erceg, V., P. Soma, D. S. Baum, and S. Catreux. 2004. “Multiple-input multiple-output fixed wireless radio channel measurements and modeling using dual-polarized antennas at 2.5 GHz”. *IEEE Trans. Wireless Commun.* 3(6): 2288–2298.
- [109] Ericsson. 2017. “Ericsson mobility report”. *Tech. rep.* URL: <http://www.ericsson.com/mobility-report>.
- [110] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio frequency (RF) system scenarios (Release 8)*. 2008. 3GPP TS 36.942.
- [111] Farhang, A., N. Marchetti, L. E. Doyle, and B. Farhang-Boroujeny. 2014. “Filter bank multicarrier for massive MIMO”. In: *Proc. IEEE VTC-Fall*.
- [112] *Feasibility study for further advancements for E-UTRA (Release 12)*. 2014. 3GPP TS 36.912.
- [113] Feng, D., C. Jiang, G. Lim, L. J. Cimini, G. Feng, and G. Y. Li. 2013. “A survey of energy-efficient wireless communications”. *IEEE Commun. Surveys Tuts.* 15(1): 167–178.

- [114] Fernandes, F., A. Ashikhmin, and T. Marzetta. 2013. “Inter-cell interference in noncooperative TDD large scale antenna systems”. *IEEE J. Sel. Areas Commun.* 31(2): 192–201.
- [115] Flordelis, J., X. Gao, G. Dahman, F. Rusek, O. Edfors, and F. Tufvesson. 2015. “Spatial separation of closely-spaced users in measured massive multi-user MIMO channels”. In: *Proc. IEEE ICC*. 1441–1446.
- [116] Frefkiel, R. H. 1970. “A high-capacity mobile radiotelephone system modelusing a coordinated small-zone approach”. *IEEE Trans. Veh. Technol.* 19(2): 173–177.
- [117] Friis, H. T. and C. B. Feldman. 1937. “A multiple unit steerable antenna for short-wave reception”. *Proc. IRE.* 25(7): 841–917.
- [118] Friis, H. T. 1946. “A note on a simple transmission formula”. *Proc. IRE.* 34(5): 254–256.
- [119] *Further advancements for E-UTRA physical layer aspects (Release 9)*. 2010. 3GPP TS 36.814.
- [120] Gao, X., O. Edfors, F. Rusek, and F. Tufvesson. 2011. “Linear pre-coding performance in measured very-large MIMO channels”. In: *Proc. IEEE VTC Fall*.
- [121] Gao, X., O. Edfors, F. Rusek, and F. Tufvesson. 2015a. “Massive MIMO performance evaluation based on measured propagation data”. *IEEE Trans. Wireless Commun.* 14(7): 3899–3911.
- [122] Gao, X., O. Edfors, F. Tufvesson, and E. G. Larsson. 2015b. “Massive MIMO in real propagation environments: Do all antennas contribute equally?” *IEEE Trans. Commun.* 63(11): 3917–3928.
- [123] Gao, X., F. Tufvesson, and O. Edfors. 2013. “Massive MIMO channels—Measurements and models”. In: *Proc. ASILOMAR*. 280–284.
- [124] Gauger, M., J. Hoydis, C. Hoek, H. Schlesinger, A. Pascht, and S. t. Brink. 2015. “Channel measurements with different antenna array geometries for massive MIMO systems”. In: *Proc. of 10th Int. ITG Conf. on Systems, Commun. and Coding*. 1–6.

- [125] Gerlach, D. and A. Paulraj. 1994. “Adaptive transmitting antenna arrays with feedback”. *IEEE Signal Process. Lett.* 1(10): 150–152.
- [126] Gesbert, D., S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu. 2010. “Multi-cell MIMO cooperative networks: A new look at interference”. *IEEE J. Sel. Areas Commun.* 28(9): 1380–1408.
- [127] Gesbert, D., M. Kountouris, R. W. Heath, C.-B. Chae, and T. Sälzer. 2007. “Shifting the MIMO paradigm”. *IEEE Signal Process. Mag.* 24(5): 36–46.
- [128] Ghosh, A., J. Zhang, J. G. Andrews, and R. Muhammed. 2010. *Fundamentals of LTE*. Prentice Hall.
- [129] Goldsmith, A., S. A. Jafar, N. Jindal, and S. Vishwanath. 2003. “Capacity limits of MIMO channels”. *IEEE J. Sel. Areas Commun.* 21(5): 684–702.
- [130] Gonthier, G. 2008. “Formal proof—The four-color theorem”. *Notices of the AMS*. 55(11): 1382–1393.
- [131] Gopalakrishnan, B. and N. Jindal. 2011. “An analysis of pilot contamination on multi-user MIMO cellular systems with many antennas”. In: *Proc. IEEE SPAWC*.
- [132] Grant, M. and S. Boyd. 2011. “CVX: Matlab software for disciplined convex programming”. <http://cvxr.com/cvx>.
- [133] Guillaud, M., D. Slock, and R. Knopp. 2005. “A practical method for wireless channel reciprocity exploitation through relative calibration”. In: *Proc. ISSPA*. 403–406.
- [134] Guo, K. and G. Ascheid. 2013. “Performance analysis of multi-cell MMSE based receivers in MU-MIMO systems with very large antenna arrays”. In: *Proc. IEEE WCNC*.
- [135] Guo, K., Y. Guo, G. Fodor, and G. Ascheid. 2014. “Uplink power control with MMSE receiver in multi-cell MU-massive-MIMO systems”. In: *Proc. IEEE ICC*. 5184–5190.
- [136] Gustavsson, U., C. Sánchez-Perez, T. Eriksson, F. Athley, G. Durisi, P. Landin, K. Hausmair, C. Fager, and L. Svensson. 2014. “On the impact of hardware impairments on massive MIMO”. In: *Proc. IEEE GLOBECOM*.

- [137] Haghifatshoar, S. and G. Caire. 2017. “Massive MIMO Pilot Decontamination and Channel Interpolation via Wideband Sparse Channel Estimation”. *IEEE Trans. Wireless Commun.*
- [138] Han, C., T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C. X. Wang, T. A. Le, M. R. Nakhai, J. Zhang, and L. Hanzo. 2011. “Green radio: Radio techniques to enable energy-efficient wireless networks”. *IEEE Commun. Mag.* 49(6): 46–54.
- [139] Harris, P. et al. 2016. “Serving 22 Users in Real-Time with a 128-Antenna Massive MIMO Testbed”. In: *Proc. IEEE International Workshop on Signal Processing Systems (SiPS)*.
- [140] Hasan, Z., H. Boostanimehr, and V. K. Bhargava. 2011. “Green cellular networks: a survey, some research issues and challenges”. *IEEE Commun. Surveys Tuts.* 13(4): 524–540.
- [141] Heath, R. W., N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed. 2016. “An overview of signal processing techniques for millimeter wave MIMO systems”. *IEEE J. Sel. Topics Signal Process.* 10(3): 436–453.
- [142] Hochwald, B. M., T. L. Marzetta, and V. Tarokh. 2004. “Multiple-antenna channel hardening and its implications for rate feedback and scheduling”. *IEEE Trans. Inf. Theory.* 60(9): 1893–1909.
- [143] Hoeg, W. and T. Lauterbach. 2009. *Digital Audio Broadcasting: Principles and Applications of DAB, DAB+ and DMB*. John Wiley & Sons, Ltd.
- [144] Holma, H. and A. Toskala. 2011. *LTE for UMTS: Evolution to LTE-Advanced*. 2nd edition. Wiley.
- [145] Honig, M. L. and W. Xiao. 2001. “Performance of reduced-rank linear interference suppression”. *IEEE Trans. Inf. Theory.* 47(5): 1928–1946.
- [146] Hoorfar, A. and M. Hassani. 2008. “Inequalities on the Lambert W function and hyperpower function”. *J. Inequalities in Pure and Applied Math.* 9(2): 1–5.
- [147] Horlin, F. and A. Bourdoux. 2008. *Digital front-end compensation for emerging wireless systems*. John Wiley & Sons, Ltd.

- [148] Hoydis, J., S. ten Brink, and M. Debbah. 2013a. “Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?” *IEEE J. Sel. Areas Commun.* 31(2): 160–171.
- [149] Hoydis, J., M. Debbah, and M. Kobayashi. 2011. “Asymptotic moments for interference mitigation in correlated fading channels”. In: *Proc. IEEE ISIT*.
- [150] Hoydis, J., C. Hoek, T. Wild, and S. ten Brink. 2012. “Channel measurements for large antenna arrays”. In: *Proc. IEEE ISWCS*.
- [151] Hoydis, J., K. Hosseini, S. t. Brink, and M. Debbah. 2013b. “Making smart use of excess antennas: Massive MIMO, small cells, and TDD”. *Bell Labs Technical Journal*. 18(2): 5–21.
- [152] Hu, D., L. He, and X. Wang. 2016. “Semi-blind pilot decontamination for massive MIMO systems”. *IEEE Trans. Wireless Commun.* 15(1): 525–536.
- [153] Huang, Y., C. Dessel, A. Bourdoux, W. Dehaene, and L. V. der Perre. 2017. “Massive MIMO processing at the semiconductor edge: Exploiting the system and circuit margins for power savings”. In: *Proc. IEEE ICASSP*. 3474–3478.
- [154] Huh, H., G. Caire, H. Papadopoulos, and S. Ramprashad. 2012. “Achieving “Massive MIMO” spectral efficiency with a not-so-large number of antennas”. *IEEE Trans. Wireless Commun.* 11(9): 3226–3239.
- [155] Ingemarsson, C. and O. Gustafsson. 2015. “On fixed-point implementation of symmetric matrix inversion”. In: *Proc. ECCTD*. 1–4.
- [156] Irmer, R., H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel. 2011. “Coordinated multipoint: Concepts, performance, and field trial results”. *IEEE Commun. Mag.* 49(2): 102–111.
- [157] Isheden, C., Z. Chong, E. Jorswieck, and G. Fettweis. 2012. “Framework for link-level energy efficiency optimization with informed transmitter”. *IEEE Trans. Wireless Commun.* 11(8): 2946–2957. DOI: 10.1109/TWC.2012.060412.111829.

- [158] Jacobsson, S., G. Durisi, M. Coldrey, T. Goldstein, and C. Studer. 2016. “Quantized precoding for massive MU-MIMO”. <https://arxiv.org/abs/1610.07564>.
- [159] Jaeckel, S., L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein. 2016. “QuaDRiGa - Quasi deterministic radio channel generator, user manual and documentation”. *Tech. rep.* v1.4.8-571. Fraunhofer Heinrich Hertz Institute.
- [160] Jamsa, T., P. Kyosti, and K. Kusume. 2015. *D1.4: METIS Channel Models*. ICT-317669-METIS.
- [161] Jiang, Z., A. F. Molisch, G. Caire, and Z. Niu. 2015. “Achievable rates of FDD Massive MIMO systems with spatial channel correlation”. *IEEE Trans. Wireless Commun.* 14(5): 2868–2882.
- [162] Jindal, N. 2006. “MIMO broadcast channels with finite-rate feedback”. *IEEE Trans. Inf. Theory*. 52(11): 5045–5060.
- [163] Jindal, N., S. Vishwanath, and A. Goldsmith. 2004. “On the Duality of Gaussian Multiple-Access and Broadcast Channels”. *IEEE Trans. Inf. Theory*. 50(5): 768–783.
- [164] Joham, M., W. Utschick, and J. Nossek. 2005. “Linear transmit processing in MIMO communications systems”. *IEEE Trans. Signal Process.* 53(8): 2700–2712.
- [165] Johnson, C. 2012. *Long Term Evolution IN BULLETS*. CreateSpace Independent Publishing Platform.
- [166] Jorswieck, E. and H. Boche. 2007. “Majorization and matrix-monotone functions in wireless communications”. *Foundations and Trends in Communications and Information Theory*. 3(6): 553–701.
- [167] Jorswieck, E. and E. Larsson. 2008. “The MISO interference channel from a game-theoretic perspective: A combination of selfishness and altruism achieves Pareto optimality”. In: *Proc. IEEE ICASSP*. 5364–5367.
- [168] Jorswieck, E. and H. Boche. 2004. “Optimal transmission strategies and impact of correlation in multiantenna systems with different types of channel state information”. *IEEE Trans. Signal Process.* 52(12): 3440–3453.

- [169] Jose, J., A. Ashikhmin, T. L. Marzetta, and S. Vishwanath. 2011. “Pilot contamination and precoding in multi-cell TDD systems”. *IEEE Trans. Commun.* 10(8): 2640–2651.
- [170] Josse, N. L., C. Laot, and K. Amis. 2008. “Efficient series expansion for matrix inversion with application to MMSE equalization”. *IEEE Commun. Lett.* 12(1): 35–37.
- [171] Joung, H., H.-S. Jo, C. Mun, and J.-G. Yook. 2014. “Capacity loss due to polarization-mismatch and space-correlation on MISO channel”. *IEEE Trans. Wireless Commun.* 13(4): 2124–2136.
- [172] Kahn, L. R. 1954. “Ratio squarer”. *Proc. IRE.* 42(11): 1704.
- [173] Kammoun, A., A. Müller, E. Björnson, and M. Debbah. 2014. “Linear precoding based on polynomial expansion: Large-scale multi-cell MIMO systems”. *IEEE J. Sel. Topics Signal Process.* 8(5): 861–875.
- [174] Kang, D., D. Kim, Y. Cho, J. Kim, B. Park, C. Zhao, and B. Kim. 2011. “1.6 - 2.1 GHz broadband Doherty power amplifiers for LTE handset applications”. In: *Proc. IEEE MTT-S*. 1–4.
- [175] Kay, S. M. 1993. *Fundamentals of statistical signal processing: Estimation theory*. Prentice Hall.
- [176] Kelly, F., A. Maulloo, and D. Tan. 1997. “Rate control for communication networks: Shadow prices, proportional fairness and stability”. *J. Operational Research Society.* 49(3): 237–252.
- [177] Kermoal, J., L. Schumacher, K. I. Pedersen, P. Mogensen, and F. Frederiksen. 2002. “A stochastic MIMO radio channel model with experimental validation”. *IEEE J. Sel. Areas Commun.* 20(6): 1211–1226.
- [178] Khansefid, A. and H. Minn. 2014. “Asymptotically optimal power allocation for massive MIMO uplink”. In: *Proc. IEEE GlobalSIP*. 627–631.
- [179] Khanzadi, M. R., G. Durisi, and T. Eriksson. 2015. “Capacity of SIMO and MISO phase-noise channels with common/separate oscillators”. *IEEE Trans. Commun.* 63(9): 3218–3231.
- [180] Ko, K. and J. Lee. 2012. “Multiuser MIMO user selection based on chordal distance”. *IEEE Trans. Commun.* 60(3): 649–654.

- [181] Korb, M. and T. G. Noll. 2010. “LDPC decoder area, timing, and energy models for early quantitative hardware cost estimates”. In: *Proc. IEEE SOCC*. 169–172.
- [182] Kotecha, J. and A. Sayeed. 2004. “Transmit signal design for optimal estimation of correlated MIMO channels”. *IEEE Trans. Signal Process.* 52(2): 546–557.
- [183] Krishnamoorthy, A. and D. Menon. 2013. “Matrix inversion using Cholesky decomposition”. In: *Proc. Alg. Arch. Arrangements Applicat.* 70–72.
- [184] Krishnan, N., R. D. Yates, and N. B. Mandayam. 2014. “Uplink linear receivers for multi-cell multiuser MIMO with pilot contamination: large system analysis”. *IEEE Trans. Wireless Commun.* 13(8): 4360–4373.
- [185] Kumar, R. and J. Gurugubelli. 2011. “How green the LTE technology can be?” In: *Proc. Wireless VITAE*.
- [186] Lahiri, K., A. Raghunathan, S. Dey, and D. Panigrahi. 2002. “Battery-driven system design: A new frontier in low power design”. In: *Proc. ASP-DAC/VLSI*. 261–267.
- [187] Lakshminaryana, S., J. Hoydis, M. Debbah, and M. Assaad. 2010. “Asymptotic analysis of distributed multi-cell beamforming”. In: *Proc. IEEE PIMRC*. IEEE. 2105–2110.
- [188] Lapidoth, A. 2002. “On phase noise channels at high SNR”. In: *Proc. IEEE ITW*.
- [189] Ledoit, O. and M. Wolf. 2004. “A well-conditioned estimator for large-dimensional covariance matrices”. *J. Multivariate Anal.* 88(2): 365–411.
- [190] Lei, Z. and T. Lim. 1998. “Simplified polynomial-expansion linear detectors for DS-CDMA systems”. *Electronics Lett.* 34(16): 1561–1563.
- [191] Li, L., A. Ashikhmin, and T. Marzetta. 2013a. “Pilot contamination precoding for interference reduction in large scale antenna systems”. In: *Allerton*. 226–232.
- [192] Li, M., S. Jin, and X. Gao. 2013b. “Spatial orthogonality-based pilot reuse for multi-cell massive MIMO transmission”. In: *Proc. WCSP*.

- [193] Li, X., E. Björnson, E. G. Larsson, S. Zhou, and J. Wang. 2017. “Massive MIMO with multi-cell MMSE processing: Exploiting all pilots for interference suppression”. *EURASIP J. Wirel. Commun. Netw.* (117).
- [194] Li, X., E. Björnson, S. Zhou, and J. Wang. 2016. “Massive MIMO with multi-antenna users: When are additional user antennas beneficial?” In: *Proc. IEEE ICT*.
- [195] Liu, Y., T. Wong, and W. Hager. 2007. “Training signal design for estimation of correlated MIMO channels with colored interference”. *IEEE Trans. Signal Process.* 55(4): 1486–1497.
- [196] Löfberg, J. 2004. “YALMIP: A Toolbox for modeling and optimization in MATLAB”. In: *Proc. IEEE CACSD*. 284–289.
- [197] Lopez-Perez, D., M. Ding, H. Claussen, and A. H. Jafari. 2015. “Enhanced intercell interference coordination challenges in heterogeneous networks”. *IEEE Commun. Surveys Tuts.* 17(4): 2078–2101.
- [198] López-Pérez, D., M. Ding, H. Claussen, and A. H. Jafari. 2015. “Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments”. *IEEE Commun. Surveys Tuts.* 17(4): 2078–2101.
- [199] Love, R. and V. Nangia. 2009. “Uplink physical channel structure”. In: *LTE - The UMTS Long Term Evolution: From Theory to Practice*. Ed. by S. Sesia, I. Toufik, and M. Baker. Wiley. Chap. 17. 377–403.
- [200] Lu, W. and M. D. Renzo. 2015. “Stochastic Geometry Modeling of Cellular Networks: Analysis, Simulation and Experimental Validation”. In: *ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*.
- [201] Luo, Z.-Q. and S. Zhang. 2008. “Dynamic spectrum management: Complexity and duality”. *IEEE J. Sel. Topics Signal Process.* 2(1): 57–73.
- [202] Lupas, R. and S. Verdu. 1989. “Linear multiuser detectors for synchronous code-division multiple-access channels”. *IEEE Trans. Inf. Theory*. 35(1): 123–136.

- [203] Ma, J. and L. Ping. 2014. “Data-aided channel estimation in large antenna systems”. *IEEE Trans. Signal Process.* 62(12): 3111–3124.
- [204] MacDonald, V. H. 1979. “The cellular concept”. *Bell System Technical Journal*. 58(1): 15–41.
- [205] Madhow, U. and M. L. Honig. 1994. “MMSE interference suppression for direct-sequence spread-spectrum CDMA”. *IEEE Trans. Commun.* 42(12): 3178–3188.
- [206] Marks, B. R. and G. P. Wright. 1978. “A general inner approximation algorithm for nonconvex mathematical programs”. *Operations Research*. 26(4): 681–683.
- [207] Martinez, A. O., E. De Carvalho, and J. O. Nielsen. 2014. “Towards very large aperture massive MIMO: A measurement based study”. In: *Proc. IEEE GLOBECOM Workshops*. 281–286.
- [208] Marzetta, T. L. 2010. “Noncooperative cellular wireless with unlimited numbers of base station antennas”. *IEEE Trans. Wireless Commun.* 9(11): 3590–3600.
- [209] Marzetta, T. L. 2015. “Massive MIMO: An introduction”. *Bell Labs Technical Journal*. 20: 11–22.
- [210] Marzetta, T. L., E. G. Larsson, H. Yang, and H. Q. Ngo. 2016. *Fundamentals of Massive MIMO*. Cambridge University Press.
- [211] Marzetta, T. L., G. H. Tucci, and S. H. Simon. 2011. “A random matrix-theoretic approach to handling singular covariance estimates”. *IEEE Trans. Inf. Theory*. 57(9): 6256–6271.
- [212] Marzetta, T. and A. Ashikhmin. 2011. “MIMO system having a plurality of service antennas for data transmission and reception and method thereof”. US Patent. 8594215.
- [213] Mathecken, P., T. Riihonen, S. Werner, and R. Wichman. 2011. “Performance analysis of OFDM with Wiener phase noise and frequency selective fading channel”. *IEEE Trans. Commun.* 59(5): 1321–1331.
- [214] Medard, M. 2000. “The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel”. *IEEE Trans. Inf. Theory*. 46(3): 933–946.

- [215] Mehrpouyan, H., A. Nasir, S. Blostein, T. Eriksson, G. Karagianidis, and T. Svensson. 2012. “Joint estimation of channel and oscillator phase noise in MIMO systems”. *IEEE Trans. Signal Process.* 60(9): 4790–4807.
- [216] Meshkati, F., H. V. Poor, S. C. Schwartz, and N. B. Mandayam. 2005. “An energy-efficient approach to power control and receiver design in wireless data networks”. *IEEE Trans. Commun.* 53(11): 1885–1894.
- [217] Mestre, X. 2008. “Improved Estimation of Eigenvalues and Eigenvectors of Covariance Matrices Using Their Sample Estimates”. *IEEE Trans. Inf. Theory.* 54(11): 5113–5129.
- [218] Mezghani, A. and J. A. Nossek. 2007. “On ultra-wideband MIMO systems with 1-bit quantized outputs: Performance analysis and input optimization”. In: *Proc. IEEE ISIT*. 1286–1289.
- [219] Mezghani, A. and J. A. Nossek. 2011. “Power efficiency in communication systems from a circuit perspective”. In: *Proc. IEEE ISCAS*. 1896–1899.
- [220] Mo, J. and R. W. Heath. 2015. “Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information”. *IEEE Trans. Signal Process.* 63(20): 5498–5512.
- [221] Mo, J. and J. Walrand. 2000. “Fair end-to-end window-based congestion control”. *IEEE/ACM Trans. Netw.* 8(5): 556–567.
- [222] Mo, J. and R. W. Heath. 2014. “High SNR capacity of millimeter wave MIMO systems with one-bit quantization”. In: *Proc. IEEE ITA*. IEEE. 1–5.
- [223] Moghadam, N. N., P. Zetterberg, P. Händel, and H. Hjalmarsson. 2012. “Correlation of distortion noise between the branches of MIMO transmit antennas”. In: *Proc. IEEE PIMRC*.
- [224] Mohammed, S. 2014. “Impact of transceiver power consumption on the energy efficiency of zero-forcing detector in massive MIMO systems”. *IEEE Trans. Commun.* 62(11): 3874–3890.
- [225] Molisch, A. F. 2007. *Wireless communications*. John Wiley & Sons.

- [226] Mollén, C., J. Choi, E. G. Larsson, and R. W. Heath. 2017. “Uplink performance of wideband Massive MIMO with one-bit ADCs”. *IEEE Trans. Wireless Commun.* 16(1): 87–100.
- [227] Mollén, C., U. Gustavsson, T. Eriksson, and E. G. Larsson. 2016a. “Out-of-band radiation measure for MIMO arrays with beamformed transmission”. In: *Proc. IEEE ICC*.
- [228] Mollén, C., E. G. Larsson, and T. Eriksson. 2016b. “Waveforms for the massive MIMO Downlink: Amplifier Efficiency, Distortion and Performance”. *IEEE Trans. Commun.* 64(12): 5050–5063.
- [229] Moshavi, S., E. G. Kanterakis, and D. L. Schilling. 1996. “Multi-stage linear receivers for DS-CDMA systems”. *Int. J. Wireless Information Networks*. 3(1): 1–17.
- [230] Motahari, A. S. and A. K. Khandani. 2009. “Capacity bounds for the Gaussian interference channel”. *IEEE Trans. Inf. Theory*. 55(2): 620–643.
- [231] Müller, A., A. Kammoun, E. Björnson, and M. Debbah. 2016. “Linear precoding based on polynomial expansion: Reducing complexity in massive MIMO”. *EURASIP J. Wirel. Commun. Netw.*
- [232] Müller, R., L. Cottatellucci, and M. Vehkaperä. 2014. “Blind pilot decontamination”. *IEEE J. Sel. Topics Signal Process.* 8(5): 773–786.
- [233] Muller, R. and S. Verdú. 2001. “Design and analysis of low-complexity interference mitigation on vector channels”. *IEEE J. Sel. Areas Commun.* 19(8): 1429–1441.
- [234] Al-Naffouri, T. Y., M. Sharif, and B. Hassibi. 2009. “How much does transmit correlation affect the sum-rate scaling of MIMO Gaussian broadcast channels?” *IEEE Trans. Commun.* 57(2): 562–572.
- [235] Nam, J., A. Adhikary, J.-Y. Ahn, and G. Caire. 2014. “Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling”. *IEEE J. Sel. Topics Signal Process.* 8(5): 876–890.
- [236] Nayebi, E., A. Ashikhmin, T. L. Marzetta, and H. Yang. 2015. “Cell-Free Massive MIMO Systems”. In: *Proc. Asilomar*.

- [237] Nayebi, E., A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao. 2017. “Precoding and Power Optimization in Cell-Free Massive MIMO Systems”. *IEEE Trans. Wireless Commun.* 16(7): 4445–4459.
- [238] Neumann, D., M. Joham, and W. Utschick. 2014. “Suppression of pilot-contamination in massive MIMO systems”. In: *Proc. IEEE SPAWC*. 11–15.
- [239] Neumann, D., M. Joham, and W. Utschick. 2017. “On MSE Based Receiver Design for Massive MIMO”. In: *Proc. SCC*.
- [240] Ngo, H. Q., A. E. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta. 2015. “Cell-free massive MIMO: Uniformly great service for everyone”. In: *Proc. IEEE SPAWC*.
- [241] Ngo, H. Q., A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta. 2017. “Cell-Free Massive MIMO Versus Small Cells”. *IEEE Trans. Wireless Commun.* 16(3): 1834–1850.
- [242] Ngo, H. Q. and E. Larsson. 2012. “EVD-based channel estimations for multicell multiuser MIMO with very large antenna arrays”. In: *Proc. IEEE ICASSP*.
- [243] Ngo, H. Q. and E. G. Larsson. 2017. “No Downlink Pilots Are Needed in TDD Massive MIMO”. *IEEE Trans. Wireless Commun.* 16(5): 2921–2935.
- [244] Ngo, H. Q., E. G. Larsson, and T. L. Marzetta. 2013. “Energy and spectral efficiency of very large multiuser MIMO systems”. *IEEE Trans. Commun.* 61(4): 1436–1449.
- [245] Ngo, H. Q., E. G. Larsson, and T. L. Marzetta. 2014a. “Aspects of favorable propagation in massive MIMO”. In: *Proc. EUSIPCO*.
- [246] Ngo, H. Q., M. Matthaiou, and E. G. Larsson. 2012. “Performance analysis of large scale MU-MIMO with optimal linear receivers”. In: *Proc. IEEE Swe-CTW*. 59–64.
- [247] Ngo, H. Q., M. Matthaiou, and E. G. Larsson. 2014b. “Massive MIMO with optimal power and training duration allocation”. *IEEE Commun. Lett.* 3(6): 605–608.

- [248] Nishimori, K., K. Cho, Y. Takatori, and T. Hori. 2001. “Automatic calibration method using transmitting signals of an adaptive array for TDD systems”. *IEEE Trans. Veh. Technol.* 50(6): 1636–1640.
- [249] Oestges, C., B. Clerckx, M. Guillaud, and M. Debbah. 2008. “Dual-polarized wireless communications: From propagation models to system performance evaluation”. *IEEE Trans. Wireless Commun.* 7(10): 4019–4031.
- [250] Palomar, D. and M. Chiang. 2006. “A tutorial on decomposition methods for network utility maximization”. *IEEE J. Sel. Areas Commun.* 24(8): 1439–1451.
- [251] Palomar, D. and Y. Jiang. 2006. “MIMO transceiver design via majorization theory”. *Foundations and Trends in Communications and Information Theory*. 3(4-5): 331–551.
- [252] Park, J. and B. Clerckx. 2014. “Multi-polarized multi-user massive MIMO: Precoder design and performance analysis”. In: *Proc. EUSIPCO*. IEEE. 326–330.
- [253] Park, J. and B. Clerckx. 2015. “Multi-user linear precoding for multi-polarized Massive MIMO system under imperfect CSIT”. *IEEE Trans. Wireless Commun.* 14(5): 2532–2547.
- [254] Paulraj, A. and C. Papadias. 1997. “Space-time processing for wireless communications”. *IEEE Signal Process. Mag.* 14(6): 49–83.
- [255] Paulraj, A., R. Nabar, and D. Gore. 2003. *Introduction to space-time wireless communications*. Cambridge University Press.
- [256] Pedersen, K. I., P. E. Mogensen, and B. H. Fleury. 1997. “Power azimuth spectrum in outdoor environments”. *Electronics Lett.* 33(18): 1583–1584.
- [257] Peterson, H. O., H. H. Beverage, and J. B. Moore. 1931. “Diversity telephone receiving system of R.C.A. communications, Inc.” *Proc. IRE*. 19(4): 562–584.
- [258] Petrovic, D., W. Rave, and G. Fettweis. 2007. “Effects of phase noise on OFDM systems with and without PLL: Characterization and compensation”. *IEEE Trans. Commun.* 55(8): 1607–1616.

- [259] Pi, Z. and F. Khan. 2011. “An introduction to millimeter-wave mobile broadband systems”. *IEEE Commun. Mag.* 49(6): 101–107.
- [260] Pillai, S. U., T. Suel, and S. Cha. 2005. “The Perron-Frobenius theorem: Some of its applications”. *IEEE Signal Process. Mag.* 22(2): 62–75.
- [261] Pinsker, M. S., V. V. Prelov, and E. C. van der Meulen. 1998. “Information Transmission over Channels with Additive-Multiplicative Noise”. In: *Proc. IEEE ISIT*. 239.
- [262] Pitarokilis, A., E. Björnson, and E. G. Larsson. 2016. “Performance of the massive MIMO uplink with OFDM and phase noise”. *IEEE Wireless Commun. Lett.* 20(8): 1595–1598.
- [263] Pitarokilis, A., E. Björnson, and E. G. Larsson. 2017. “On the Effect of Imperfect Timing Synchronization on Pilot Contamination”. In: *Proc. IEEE ICC*.
- [264] Pitarokilis, A., S. K. Mohammed, and E. G. Larsson. 2012. “On the optimality of single-carrier transmission in large-scale antenna systems”. *IEEE Wireless Commun. Lett.* 1(4): 276–279.
- [265] Pitarokilis, A., S. K. Mohammed, and E. G. Larsson. 2015. “Uplink performance of time-reversal MRC in massive MIMO systems subject to phase noise”. *IEEE Trans. Wireless Commun.* 14(2): 711–723.
- [266] Pizzo, A., D. Verenzuela, L. Sanguinetti, and E. Björnson. 2017. “Network Deployment for Maximal Energy Efficiency in Uplink with Zero-Forcing”. In: *Proc. IEEE GLOBECOM*.
- [267] Polyanskiy, Y., H. Poor, and S. Verdú. 2010. “Channel coding rate in the finite blocklength regime”. *IEEE Trans. Inf. Theory*. 56(5): 2307–2359.
- [268] Poon, A. S. Y., R. W. Brodersen, and D. N. C. Tse. 2005. “Degrees of freedom in multiple-antenna channels: A signal space approach”. *IEEE Trans. Inf. Theory*. 51(2): 523–536.
- [269] Qian, L. P., Y. J. Zhang, and J. Huang. 2009. “MAPEL: Achieving global optimality for a non-convex wireless power control problem”. *IEEE Trans. Wireless Commun.* 8(3): 1553–1563.

- [270] Qiao, D., S. Choi, and K. G. Shin. 2002. “Goodput analysis and link adaptation for IEEE 802.11 a Wireless LANs”. *IEEE Trans. Mobile Comp.* 1(4): 278–292.
- [271] Qualcomm. 2012. “Rising to meet the 1000x mobile data challenge”. *Tech. rep.* Qualcomm Incorporated.
- [272] Rangan, S., T. S. Rappaport, and E. Erkip. 2014. “Millimeter-wave cellular wireless networks: Potentials and challenges”. *Proc. IEEE.* 102(3): 366–385.
- [273] Rao, X. and V. Lau. 2014. “Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems”. *IEEE J. Sel. Areas Commun.* 62(12): 3261–3271.
- [274] Rappaport, T. S., R. W. Heath Jr, R. C. Daniels, and J. N. Murdock. 2014. *Millimeter wave wireless communications*. Pearson Education.
- [275] Rappaport, T. S., S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez. 2013. “Millimeter wave mobile communications for 5G cellular: It will work!” *IEEE Access.* 1: 335–349.
- [276] Rashid-Farrokhi, F., L. Tassiulas, and K. J. R. Liu. 1998. “Joint optimal power control and beamforming in wireless networks using antenna arrays”. *IEEE Trans. Commun.* 46(10): 1313–1324.
- [277] Ring, D. H. 1947. “Mobile Telephony - Wide Area Coverage”. *Bell Laboratories Technical Memorandum*.
- [278] Rockafellar, R. 1993. “Lagrange multipliers and optimality”. *SIAM Review.* 35(2): 183–238.
- [279] Rogalin, R., O. Y. Bursalioglu, H. Papadopoulos, G. Caire, A. F. Molisch, A. Michaloliakos, V. Balan, and K. Psounis. 2014. “Scalable synchronization and reciprocity calibration for distributed multiuser MIMO”. *IEEE Trans. Wireless Commun.* 13(4): 1815–1831.
- [280] Roy, R. H. and B. Ottersten. 1991. “Spatial division multiple access wireless communication systems”. US Patent. 5515378.

- [281] Rusek, F., D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson. 2013. “Scaling up MIMO: Opportunities and challenges with very large arrays”. *IEEE Signal Process. Mag.* 30(1): 40–60.
- [282] Sadek, M., A. Tarighat, and A. Sayed. 2007. “A leakage-based precoding scheme for downlink multi-user MIMO channels”. *IEEE Trans. Wireless Commun.* 6(5): 1711–1721.
- [283] Saleh, A. A. and R. A. Valenzuela. 1987. “A statistical model for indoor multipath propagation”. *IEEE J. Sel. Areas Commun.* 5(2): 128–137.
- [284] Salz, J. and J. H. Winters. 1994. “Effect of fading correlation on adaptive arrays in digital mobile radio”. *IEEE Trans. Veh. Technol.* 43(4): 1049–1057.
- [285] Sanguinetti, L., R. Couillet, and M. Debbah. 2016a. “Large system analysis of base station cooperation for power minimization”. *IEEE Trans. Wireless Commun.* 15(8): 5480–5496.
- [286] Sanguinetti, L., A. A. D’Amico, M. Morelli, and M. Debbah. 2016b. “Random access in uplink massive MIMO systems: how to exploit asynchronicity and excess antennas”. In: *Proc. GLOBECOM*.
- [287] Sarkar, T. K., Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma. 2003. “A survey of various propagation models for mobile communication”. *IEEE Antennas Propag. Mag.* 45(3): 51–82.
- [288] Saxena, V., G. Fodor, and E. Karipidis. 2015. “Mitigating pilot contamination by pilot reuse and power control schemes for massive MIMO systems”. In: *Proc. IEEE VTC-Spring*.
- [289] Sayeed, A. 2002. “Deconstructing multiantenna fading channels”. *IEEE Trans. Signal Process.* 50(10): 2563–2579.
- [290] Schenk, T. 2008. *RF imperfections in high-rate wireless systems: Impact and digital compensation*. Springer.
- [291] Schulte, H. J. and W. A. Cornell. 1960. “A high-capacity mobile radiotelephone system model using a coordinated small-zone approach”. *IEEE Trans. Veh. Technol.* 9(1): 49–53.

- [292] Sessler, G. and F. Jondral. 2005. “Low complexity polynomial expansion multiuser detector for CDMA systems”. *IEEE Trans. Veh. Technol.* 54(4): 1379–1391.
- [293] Shafi, M., M. Zhang, A. L. Moustakas, P. J. Smith, A. F. Molisch, F. Tufvesson, and S. H. Simon. 2006. “Polarized MIMO channels in 3-D: Models, measurements and mutual information”. *IEEE J. Sel. Areas Commun.* 24(3): 514–527.
- [294] Shamai, S. and B. M. Zaidel. 2001. “Enhancing the cellular downlink capacity via co-processing at the transmitting end”. In: *Proc. IEEE VTC-Spring*. Vol. 3. 1745–1749.
- [295] Shang, X., B. Chen, G. Kramer, and H. V. Poor. 2011. “Noisy-interference sum-rate capacity of parallel Gaussian interference channels”. *IEEE Trans. Inf. Theory*. 57(1): 210–226.
- [296] Shang, X., G. Kramer, B. Chen, and H. V. Poor. 2009. “A new outer bound and the noisy-interference sum-rate capacity for Gaussian interference channels”. *IEEE Trans. Inf. Theory*. 55(2): 689–699.
- [297] Shannon, C. E. 1948. “A mathematical theory of communication”. *Bell System Technical Journal*. 27: 379–423, 623–656.
- [298] Shannon, C. E. 1949. “Communication in the presence of noise”. *Proc. IRE*. 37(1): 10–21.
- [299] Shariati, N., E. Björnson, M. Bengtsson, and M. Debbah. 2014. “Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics”. *IEEE J. Sel. Topics Signal Process.* 8(5): 815–830.
- [300] Shepard, C., H. Yu, N. Anand, L. Li, T. Marzetta, R. Yang, and L. Zhong. 2012. “Argos: Practical many-antenna base stations”. In: *Proc. ACM MobiCom*.
- [301] Shi, D., G. Foschini, M. Gans, and J. Kahn. 2000. “Fading correlation and its effect on the capacity of multielement antenna systems”. *IEEE Trans. Commun.* 48(3): 502–513.
- [302] Sifaou, H., A. Kammoun, L. Sanguinetti, M. Debbah, and M. S. Alouini. 2017. “Max-min SINR in large-scale single-cell MU-MIMO: Asymptotic analysis and low-complexity transceivers”. *IEEE Trans. Signal Process.* 65(7): 1841–1854.

- [303] Simeone, O., N. Levy, A. Sanderovich, O. Somekh, B. M. Zaidel, H. V. Poor, and S. Shamai. 2012. “Cooperative wireless cellular systems: An information-theoretic view”. *Foundations and Trends in Communications and Information Theory*. 8(1-2): 1–177.
- [304] Somekh, O. and S. Shamai. 2000. “Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading”. *IEEE Trans. Inf. Theory*. 46(4): 1401–1425.
- [305] Sørensen, J. H., E. de Carvalho, C. Stefanovic, and P. Popovski. 2016. “Coded pilot access: A random access solution for massive MIMO systems”. *CoRR*. abs/1605.05862. URL: <http://arxiv.org/abs/1605.05862>.
- [306] Studer, C. and G. Durisi. 2016. “Quantized massive MU-MIMO-OFDM uplink”. *IEEE Trans. Commun.* 64(6): 2387–2399.
- [307] Sturm, J. 1999. “Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones”. *Optimization Methods and Software*. 11-12: 625–653.
- [308] Swales, S. C., M. A. Beach, D. J. Edwards, and J. P. McGeehan. 1990. “The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems”. *IEEE Trans. Veh. Technol.* 39(1): 56–67.
- [309] Tomatis, F. and S. Sesia. 2009. “Synchronization and cell search”. In: *LTE - The UMTS Long Term Evolution: From Theory to Practice*. Ed. by S. Sesia, I. Toufik, and M. Baker. Wiley. Chap. 7. 141–157.
- [310] Tomba, L. 1998. “On the effect of Wiener phase noise in OFDM systems”. *IEEE Trans. Commun.* 46(5): 580–583.
- [311] Tombaz, S., K. W. Sung, and J. Zander. 2012. “Impact of densification on energy efficiency in wireless access networks”. In: *Proc. IEEE GLOBECOM Workshop*. 57–62.
- [312] Tombaz, S., A. Västberg, and J. Zander. 2011. “Energy- and cost-efficient ultra-high-capacity wireless access”. *IEEE Wireless Commun.* 18(5): 18–24.

- [313] Trump, T. and B. Ottersten. 1996. “Estimation of nominal direction of arrival and angular spread using an array of sensors”. *Signal Processing*. 50(1-2): 57–69.
- [314] Tse, D. and P. Viswanath. 2005. *Fundamentals of wireless communications*. Cambridge University Press.
- [315] Tulino, A. M. and S. Verdú. 2004. “Random matrix theory and wireless communications”. *Foundations and Trends in Communications and Information Theory*. 1(1): 1–182.
- [316] Tütüncü, R., K. Toh, and M. Todd. 2003. “Solving semidefinite-quadratic-linear programs using SDPT3”. *Mathematical Programming*. 95(2): 189–217.
- [317] Tuy, H. 2000. “Monotonic optimization: Problems and solution approaches”. *SIAM J. Optim.* 11(2): 464–494.
- [318] Tuy, H., F. Al-Khayyal, and P. Thach. 2005. “Monotonic optimization: Branch and cut methods”. In: *Essays and Surveys in Global Optimization*. Ed. by C. Audet, P. Hansen, and G. Savard. Springer US.
- [319] Upadhyya, K., S. A. Vorobyov, and M. Vehkapera. 2017a. “Downlink Performance of Superimposed Pilots in Massive MIMO systems”. *CoRR*. abs/1606.04476. URL: <http://arxiv.org/abs/1606.04476>.
- [320] Upadhyya, K., S. A. Vorobyov, and M. Vehkapera. 2017b. “Superimposed Pilots Are Superior for Mitigating Pilot Contamination in Massive MIMO”. *IEEE Trans. Signal Process.* 65(11): 2917–2932.
- [321] Va, V., J. Choi, and R. W. Heath. 2017. “The Impact of Beamwidth on Temporal Channel Variation in Vehicular Channels and its Implications”. *IEEE Trans. Veh. Technol.*
- [322] Valkama, M. 2011. “RF impairment compensation for future radio systems”. In: *Multi-Mode/Multi-Band RF Transceivers for Wireless Communications*. Ed. by G. Hueber and R. B. Staszewski. John Wiley & Sons, Inc. 453–496.
- [323] Vaughan, R. G. 1990. “Polarization diversity in mobile communications”. *IEEE Trans. Veh. Technol.* 39(3): 177–186.

- [324] Veen, B. D. V. and K. M. Buckley. 1988. “Beamforming: a versatile approach to spatial filtering”. *IEEE ASSP Mag.* 5(2): 4–24.
- [325] Venkatesan, S., A. Lozano, and R. Valenzuela. 2007. “Network MIMO: Overcoming intercell interference in indoor wireless systems”. In: *Proc. IEEE ACSSC*. 83–87.
- [326] Verdú, S. 1990. “On channel capacity per unit cost”. *IEEE Trans. Inf. Theory*. 36(5): 1019–1030.
- [327] Verenzuela, D., E. Björnson, and M. Matthaiou. 2016. “Hardware design and optimal ADC resolution for uplink massive MIMO systems”. In: *Proc. SAM Workshop*.
- [328] Verenzuela, D., E. Björnson, and L. Sanguinetti. 2017. “Spectral and Energy Efficiency of Superimposed Pilots in Uplink Massive MIMO”. *CoRR*. abs/1709.07722. URL: <http://arxiv.org/abs/1709.07722>.
- [329] Vieira, J., S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. C. Wong, V. Öwall, O. Edfors, and F. Tufvesson. 2014a. “A flexible 100-antenna testbed for massive MIMO”. In: *Proc. IEEE GLOBECOM Workshop*. 287–293.
- [330] Vieira, J., F. Rusek, O. Edfors, S. Malkowsky, L. Liu, and F. Tufvesson. 2017. “Reciprocity Calibration for Massive MIMO: Proposal, Modeling, and Validation”. *IEEE Trans. Wireless Commun.* 16(5): 3042–3056.
- [331] Vieira, J., R. Rusek, and F. Tufvesson. 2014b. “Reciprocity calibration methods for massive MIMO based on antenna coupling”. In: *Proc. IEEE GLOBECOM*.
- [332] Viering, I., H. Hofstetter, and W. Utschick. 2002. “Spatial long-term variations in urban, rural and indoor environments”. In: *COST273 5th Meeting, Lisbon, Portugal*.
- [333] Vinogradova, J., E. Björnson, and E. G. Larsson. 2016a. “Detection and mitigation of jamming attacks in massive MIMO systems using random matrix theory”. In: *Proc. IEEE SPAWC*.
- [334] Vinogradova, J., E. Björnson, and E. G. Larsson. 2016b. “On the separability of signal and interference-plus-noise subspaces in blind pilot decontamination”. In: *Proc. IEEE ICASSP*.

- [335] Viswanath, P. and D. N. C. Tse. 2003. “Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality”. *IEEE Trans. Inf. Theory*. 49(8): 1912–1921.
- [336] Wagner, S., R. Couillet, M. Debbah, and D. Slock. 2012. “Large system analysis of linear precoding in MISO broadcast channels with limited feedback”. *IEEE Trans. Inf. Theory*. 58(7): 4509–4537.
- [337] Wallace, J. W. and M. A. Jensen. 2001. “Measured characteristics of the MIMO wireless channel”. In: *Proc. IEEE VTC-Fall*. Vol. 4. 2038–2042.
- [338] Wallace, J. W. and M. A. Jensen. 2002. “Modeling the indoor MIMO wireless channel”. *IEEE Trans. Antennas Propag.* 50(5): 591–599.
- [339] Wallis, J. S. 1976. “On the existence of Hadamard matrices”. *Journal of Combinatorial Theory*. 21(2): 188–195.
- [340] Wang, H., P. Wang, L. Ping, and X. Lin. 2009. “On the impact of antenna correlation in multi-user MIMO systems with rate constraints”. *IEEE Commun. Lett.* 13(12): 935–937.
- [341] Weeraddana, P., M. Codreanu, M. Latva-aho, A. Ephremides, and C. Fischione. 2012. “Weighted sum-rate maximization in wireless networks: A review”. *Foundations and Trends in Networking*. 6(1-2): 1–163.
- [342] Weingarten, H., Y. Steinberg, and S. Shamai. 2006. “The capacity region of the Gaussian multiple-input multiple-output broadcast channel”. *IEEE Trans. Inf. Theory*. 52(9): 3936–3964.
- [343] Wen, C.-K., S. Jin, K.-K. Wong, C.-J. Wang, and G. Wu. 2015. “Joint channel-and-data estimation for large-MIMO systems with low-precision ADCs”. In: *Proc. IEEE ISIT*. 1237–1241.
- [344] Wenk, M. 2010. *MIMO-OFDM testbed: Challenges, implementations, and measurement results. Series in microelectronics*. Hartung-Gorre.
- [345] Wiesel, A., Y. Eldar, and S. Shamai. 2006. “Linear precoding via conic optimization for fixed MIMO receivers”. *IEEE Trans. Signal Process.* 54(1): 161–176.

- [346] Wiesel, A., Y. Eldar, and S. Shamai. 2008. “Zero-forcing pre-coding and generalized inverses”. *IEEE Trans. Signal Process.* 56(9): 4409–4418.
- [347] WINNER II Channel Models. 2008. “Deliverable 1.1. 2 v. 1.2”. *Tech. rep.*
- [348] Winters, J. H. 1984. “Optimum combining in digital mobile radio with cochannel interference”. *IEEE J. Sel. Areas Commun.* 2(4): 528–539.
- [349] Winters, J. H. 1987. “Optimum combining for indoor radio systems with multiple users”. *IEEE Trans. Commun.* 35(11): 1222–1230.
- [350] Winters, J. H. 1998. “Smart antennas for wireless systems”. *IEEE Personal Commun.* 5(1): 23–27.
- [351] Wu, J., Y. Zhang, M. Zukerman, and E. K. N. Yung. 2015. “Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey”. *IEEE Commun. Surveys Tuts.* 17(2): 803–826.
- [352] Wu, M., B. Yin, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer. 2014. “Large-scale MIMO detection for 3GPP LTE: Algorithms and FPGA implementations”. *IEEE J. Sel. Topics Signal Process.* 8(5): 916–929.
- [353] Wyner, A. D. 1994. “Shannon-theoretic approach to a Gaussian cellular multiple-access channel”. *IEEE Trans. Inf. Theory.* 40(6): 1713–1727.
- [354] Xiao, H., Y. Chen, Y.-N. R. Li, and Z. Lu. 2015. “CSI feedback for massive MIMO system with dual-polarized antennas”. In: *Proc. IEEE PIMRC*. IEEE. 2324–2328.
- [355] Xiao, M., S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C.-L. I, and A. Ghosh. 2017. “Millimeter Wave Communications for Future Mobile Networks”. *IEEE J. Sel. Areas Commun.* 35(9): 1909–1935.
- [356] Xu, J., W. Xu, and F. Gong. 2017. “On Performance of Quantized Transceiver in Multiuser Massive MIMO Downlinks”. *IEEE Wireless Commun. Lett.*

- [357] Yang, H. and T. L. Marzetta. 2013a. “Performance of conjugate and zero-forcing beamforming in large-scale antenna systems”. *IEEE J. Sel. Areas Commun.* 31(2): 172–179.
- [358] Yang, H. and T. L. Marzetta. 2013b. “Total energy efficiency of cellular large scale antenna system multiple access mobile networks”. In: *Proc. IEEE Online GreenComm.* 27–32.
- [359] Yang, H. and T. L. Marzetta. 2014. “A macro cellular wireless network with uniformly high user throughputs”. In: *Proc. IEEE VTC-Fall*.
- [360] Yates, R. 1995. “A framework for uplink power control in cellular radio systems”. *IEEE J. Sel. Areas Commun.* 13(7): 1341–1347.
- [361] Ye, Q., O. Y. Bursalioglu, H. C. Papadopoulos, C. Caramanis, and J. G. Andrews. 2016. “User Association and Interference Management in Massive MIMO HetNets”. *IEEE Trans. Commun.* 64(5): 2049–2064.
- [362] Yin, H., L. Cottatellucci, D. Gesbert, R. R. Müller, and G. He. 2016. “Robust Pilot Decontamination Based on Joint Angle and Power Domain Discrimination”. *IEEE Trans. Signal Process.* 64(11): 2990–3003.
- [363] Yin, H., D. Gesbert, M. Filippou, and Y. Liu. 2013. “A co-ordinated approach to channel estimation in large-scale multiple-antenna systems”. *IEEE J. Sel. Areas Commun.* 31(2): 264–273.
- [364] Young, W. R. 1979. “Advanced mobile phone service: Introduction, background, and objectives”. *Bell System Technical Journal*. 58(1): 1–14.
- [365] Yu, K., M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach. 2004. “Modeling of wide-band MIMO radio channels based on NLoS indoor measurements”. *IEEE Trans. Veh. Technol.* 53(3): 655–665.
- [366] Yu, W. 2006. “Uplink-downlink duality via minimax duality”. *IEEE Trans. Inf. Theory*. 52(2): 361–374.
- [367] Yu, W. and T. Lan. 2007. “Transmitter optimization for the multi-antenna downlink with per-antenna power constraints”. *IEEE Trans. Signal Process.* 55(6): 2646–2660.

- [368] Zakhour, R. and D. Gesbert. 2009. “Coordination on the MISO interference channel using the virtual SINR framework”. In: *Proc. ITG Workshop on Smart Antennas (WSA)*.
- [369] Zander, J. 1992. “Performance of optimum transmitter power control in cellular radio systems”. *IEEE Trans. Veh. Technol.* 41(1): 57–62.
- [370] Zander, J. and M. Frodigh. 1994. “Comment on “Performance of optimum transmitter power control in cellular radio systems””. *IEEE Trans. Veh. Technol.* 43(3): 636.
- [371] Zappone, A. and E. Jorswieck. 2015. “Energy Efficiency in Wireless Networks via Fractional Programming Theory”. *Foundations and Trends in Communications and Information Theory*. 11(3-4): 185–396.
- [372] Zarei, S., W. Gerstacker, R. R. Müller, and R. Schober. 2013. “Low-complexity linear precoding for downlink large-scale MIMO systems”. In: *Proc. IEEE PIMRC*.
- [373] Zetterberg, P. and B. Ottersten. 1995. “The spectrum efficiency of a base station antenna array system for spatially selective transmission”. *IEEE Trans. Veh. Technol.* 44(3): 651–660.
- [374] Zetterberg, P. 2011. “Experimental investigation of TDD reciprocity-based zero-forcing transmit precoding”. *EURASIP J. Adv. Signal Process.* (137541).
- [375] Zhang, H., N. Mehta, A. Molisch, J. Zhang, and H. Dai. 2008. “Asynchronous interference mitigation in cooperative base station systems”. *IEEE Trans. Wireless Commun.* 7(1): 155–165.
- [376] Zhang, J., Y. Wei, E. Björnson, Y. Han, and X. Li. 2017. “Spectral and Energy Efficiency of Cell-Free Massive MIMO Systems with Hardware Impairments”. In: *Proc. WCSP*.
- [377] Zhang, W. 2012. “A general framework for transmission with transceiver distortion and some applications”. *IEEE Trans. Commun.* 60(2): 384–399.
- [378] Zhu, X., Z. Wang, L. Dai, and C. Qian. 2015. “Smart pilot assignment for Massive MIMO”. *IEEE Commun. Lett.* 19(9): 1644–1647.