Massive MIMO Communications

Trinh Van Chien and Emil Björnson

Book Chapter

Part of: 5G Mobile Communications, Ed. Wei Xiang, Kan Zheng, Xuemin (Sherman) Shen, 2017, pp. 77-116. ISBN: 978-3-319-34206-1

DOI: http://dx.doi.org/10.1007/978-3-319-34208-5_4

Massive MIMO Communications

Trinh Van Chien and Emil Björnson

Abstract Every new network generation needs to make a leap in area data throughput, to manage the growing wireless data traffic. The Massive MIMO technology can bring at least ten-fold improvements in area throughput by increasing the spectral efficiency (bit/s/Hz/cell), while using the same bandwidth and density of base stations as in current networks. These extraordinary gains are achieved by equipping the base stations with arrays of a hundred antennas to enable spatial multiplexing of tens of user terminals. This chapter explains the basic motivations and communication theory behind the Massive MIMO technology, and provides implementation-related design guidelines.

1 Introduction

Much higher area data throughput is required in future cellular networks, since the global demand for wireless data traffic is continuously growing. This goal can be achieved without the need for more bandwidth or additional base stations if the spectral efficiency (measured in bit/s/Hz/cell) is improved. This chapter explains why the Massive MIMO (multiple-input multiple-output) communication technology, where multi-antenna base stations spatially multiplex a multitude of user terminals over the entire bandwidth, is well-suited for this purpose. The rationale behind the Massive MIMO concept and its transmission protocol is explained from a historical perspective in Sect. 2. Next, Sect. 3 provides a basic communication theoretic performance analysis. Closed-form spectral efficiency expressions are derived and the key properties and performance limitations of Massive MIMO are highlighted. The chapter

Trinh Van Chien
Linköping University, Department of Electrical Engineering (ISY), SE-581 83 Linköping, Sweden
e-mail: trinh.van.chien@liu.se

Emil Björnson
Linköping University, Department of Electrical Engineering (ISY), SE-581 83 Linköping, Sweden
e-mail: emil.bjornson@liu.se

is concluded by Sect. 4 where implementation-related design guidelines are given, particularly regarding power allocation and the reuse of pilot sequences for efficient channel estimation. Multi-cell simulations are provided to showcase that the Massive MIMO technology can provide ten-fold or even 50-fold improvements in spectral efficiency over contemporary technology, without the need for advanced signal processing or network coordination. Finally, the full mathematical details are provided in Appendix at the end of this chapter.

2 Importance of Improving the Spectral Efficiency

The wireless information traffic has doubled every two and a half years since the beginning of wireless communications, as observed by Martin Cooper at ArrayComm in the nineties. Different technologies and use cases have dominated in different periods, but the exponential increase is currently driven by wireless data traffic in cellular and local area networks. There are no indications that this trend will break anytime soon; in fact, a slightly faster traffic growth is predicted in the well-reputed Cisco Visual Networking Index and Ericsson Mobility Report.

To keep up with the rapid traffic growth, a key goal of the 5G technologies is to improve the area throughput by orders of magnitude; $100\times$ and even $1000\times$ higher throughput are regularly mentioned as 5G design goals. The area throughput of a wireless network is measured in bit/s/km^2 and can be modeled as follows:

$$\text{Area throughput (bit/s/km}^2\text{)} =$$
$$\text{Bandwidth (Hz)} \times \text{Cell density (cells/km}^2\text{)} \times \text{Spectral efficiency (bit/s/Hz/cell).}$$

This simple formula reveals that there are three main components that can be improved to yield higher area throughput: (1) more bandwidth can be allocated for 5G services; (2) the network can be densified by adding more cells with independently operating access points; and (3) the efficiency of the data transmissions (per cell and for a given amount of bandwidth) can be improved.

The improvements in area throughput in previous network generations have greatly resulted from cell densification and allocation of more bandwidth. In urban environments, where contemporary networks are facing the highest traffic demands, cellular networks are nowadays deployed with a few hundred meters inter-site distances and wireless local area networks (WLANs) are available almost everywhere. Further cell densification is certainly possible, but it appears that we are reaching a saturation point. Moreover, the most valuable frequency bands are below 6 GHz because these frequencies can provide good network coverage and service quality, while higher bands might only work well under short-range line-of-sight conditions. In a typical country like Sweden, the cellular and WLAN technologies have in total been allocated more than 1 GHz of bandwidth in the interval below 6 GHz and thus we cannot expect any major bandwidth improvements either.

In contrast, the spectral efficiency (SE) has not seen any major improvements in previous network generations. Hence, it might be a factor that can be greatly improved in the future and possibly become the primary way to achieve high area throughput in 5G networks. In this chapter, we describe the rationale and background of the physical-layer technology Massive multiple-input multiple-output (MIMO), which provides the means to improve the SE of future networks by one or two orders of magnitude.

2.1 Multi-User MIMO Communication

The SE of a single-input single-output (SISO) communication channel, from a single-antenna transmitter to a single-antenna receiver, is upper bounded by the Shannon capacity, which has the form $\log_2(1+\overline{SNR})$ bit/s/Hz for additive white Gaussian noise (AWGN) channels. The SISO capacity is thus a logarithmic function of the signal-to-noise ratio (SNR), denoted here as $\overline{SNR}$. To improve the SE we need to increase the SNR, which corresponds to increasing the power of the transmitted signal. For example, suppose we have a system that operates at 2 bit/s/Hz and we would like to double its SE to 4 bit/s/Hz, then this corresponds to improving the SNR by a factor 5, from 3 to 15. The next doubling of the SE, from 4 to 8 bit/s/Hz, requires another 17 times more power. In other words, the logarithm of the SE expression forces us to increase the transmit power exponentially fast to achieve a linear increase in the SE of the SISO channel. This is clearly a very inefficient and non-scalable way to improve the SE, and the approach also breaks down when there are interfering transmissions in other cells that scale their transmit powers in the same manner. We therefore need to identify another way to improve the SE of cellular networks.

Each base station (BS) in a cellular network serves a multitude of user terminals. Traditionally, the time/frequency resources have been divided into resource blocks and only one of the user terminals was active per block. This terminal can then receive a single data stream with an SE quantified as $\log_2(1+\overline{SNR})$. The efficient way to increase the SE of a cellular network is to have multiple parallel transmissions. If there are $\mathbb{G}$ parallel and independent transmissions, the sum SE becomes $\mathbb{G}\log_2(1+\overline{SNR})$ where $\mathbb{G}$ acts as a multiplicative pre-log factor. Parallel transmissions can be realized by having multiple transmit antennas and multiple receive antennas. There are two distinct cases:

1. Point-to-point MIMO [39], where a BS with multiple antennas communicates with a single user terminal having multiple antennas.
2. Multi-user MIMO [34], where a BS with multiple antennas communicates with multiple user terminals, each having one or multiple antennas.

There are many reasons why multi-user MIMO is the most scalable and attractive solution [17]. Firstly, the wavelength is 5-30 cm in the frequency range of cellular communication (1-6 GHz). This limits the number of antennas that can be deployed

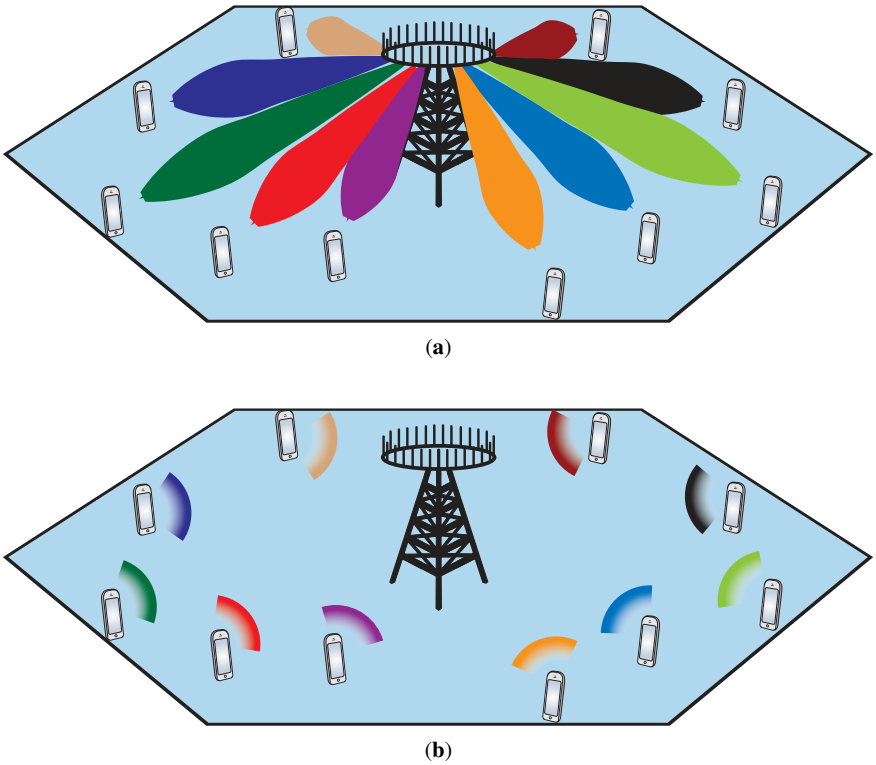
in a compact user terminal for point-to-point MIMO, while one can have almost any number of spatially separated single-antenna terminals in multi-user MIMO. This is an important distinction since the number of simultaneous data streams that can be separated by MIMO processing equals the minimum of the number of transmit and receive antennas. Secondly, the wireless propagation channel to a user terminal is likely to have only a few dominating paths, which limits the ability to convey multiple parallel data streams to a terminal in point-to-point MIMO. The corresponding restriction on multi-user MIMO is that the users need to be, say, a few meters apart to have sufficiently different channel characteristics, which is a very loose restriction that is true in most practical scenarios. Thirdly, advanced signal processing is needed at the terminals in point-to-point MIMO to detect the multiple data streams, while each terminal in multi-user MIMO only needs to detect a single data stream.

The canonical multi-user MIMO system consists of a BS with M antennas that serves K single-antenna terminals; see Fig. 1 for a schematic illustration. The BS multiplexes one data stream per user in the downlink and receives one stream per user in the uplink. Simply speaking, the BS uses its antennas to direct each signal towards its desired receiver in the downlink, and to separate the multiple signals received in the uplink. If the terminal is equipped with multiple antennas, it is often beneficial to use these extra antennas to mitigate interference and improve the SNR rather than sending multiple data streams [6]. For the ease of exposition, this chapter concentrates on single-antenna terminals. In this case, $\min(M,K)$ represents the maximal number of data streams that can be simultaneously transmitted in the cell, while still being separable in the spatial domain. The number $\min(M,K)$ is referred to as the *multiplexing gain* of a multi-user MIMO system.

2.2 Lessons Learned

The research on multi-user MIMO, particularly with multi-antenna BSs, has been going on for decades. Some notable early works are the array processing papers [1, 38, 44, 47], the patent [36] on spatial division multiple access (SDMA), and the seminal information-theoretic works [11, 18, 42, 43, 46] that characterized the achievable multi-user capacity regions, assuming that perfect channel state information (CSI) is available in the system. In this section, we summarize some of the main design insights that have been obtained over the years.

Capacity-achieving transmission schemes for multi-user MIMO are based upon non-linear signal processing; for example, the dirty-paper coding (DPC) scheme that achieves the downlink capacity and the successive interference cancelation (SIC) scheme that achieves the uplink capacity. The intuition behind these schemes is that the inter-user interference needs to be suppressed, by interference-aware transmit processing or iterative interference-aware receive processing, to achieve the optimal performance. These non-linear schemes naturally require extensive computations and accurate CSI, because otherwise the attempts to subtract interference cause more harm than good.



(**a**)

(**b**)

Fig. 1 Illustration of the downlink and uplink transmission in a multi-user MIMO system, where the BS is equipped with M antennas and serves K user terminals simultaneously. This illustration focuses on line-of-sight propagation where the downlink signals can be viewed as angular beams, but multi-user MIMO works equally well in non-line-of-sight conditions. (**a**) Downlink in multi-user MIMO. (**b**) Uplink in multi-user MIMO

How large are the gains of optimal non-linear processing (e.g., DPC and SIC) over simplified linear processing schemes where each user terminal is treated separately? To investigate this, let us provide a numerical example where $K = 10$ user terminals are simultaneously served by a BS with M antennas. For simplicity, each user is assumed to have an average SNR of 5 dB, there is perfect CSI available everywhere, and the channels are modeled as uncorrelated Rayleigh fading (this is defined in detail in Sect. 3). Figure 2 shows the average sum SE, as a function of M, achieved by sum capacity-achieving non-linear processing and a simplified linear processing scheme called zero-forcing (ZF), which attempts to suppress all interference. The results are representative for both uplink and downlink transmissions.

This simulation shows that the non-linear processing greatly outperforms linear ZF when $M \approx K$. The operating point $M = K$ makes particular sense from a multiplexing perspective since the multiplexing gain $\min(M,K)$ does not improve if we let M increase for a fixed K. Nevertheless, Fig. 2 shows that there are other rea-

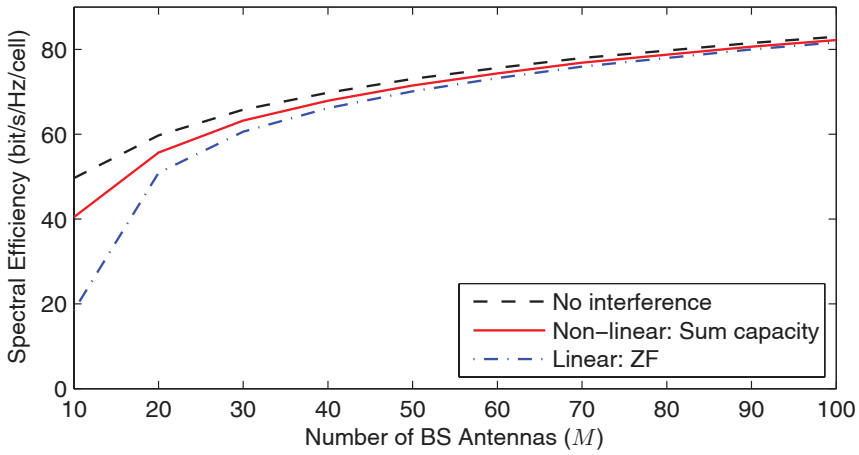


Fig. 2 Average spectral efficiency in a multi-user MIMO system with $K = 10$ users and varying number of BS antennas. Each user has an average SNR of 5 dB and the channels are Rayleigh fading. The sum capacity is compared with the performance of linear ZF processing and the upper bound when neglecting all interference. The results are representative for both uplink and downlink

sons to consider $M > K$; the capacity increases and the performance with linear ZF processing approaches the capacity. Already at $M = 20$ (i.e., $M/K = 2$) there is only a small gap between optimal non-linear processing and linear ZF. In fact, both schemes also approach the upper curve in Fig. 2 which represents the upper bound where the interference between the users is neglected. This shows that we can basically serve all the K users as if each one of them was alone in the cell.

First lesson learned: Linear processing, such as ZF, provides a sum spectral efficiency close to the sum capacity when $M \gg K$.

The performance analysis and optimization of linear processing schemes have received much attention from academic researchers. While non-linear schemes are hard to implement but relatively easy to analyze and optimize, linear processing schemes have proved to have the opposite characteristics. In particular, computing the optimal downlink linear precoding is an NP-hard problem in many cases [27], which requires monotonic optimization tools to solve; see for example [9]. Nevertheless, the suboptimal ZF curve in Fig. 2 was generated without any complicated optimization, thus showing that the optimal linear processing obtained in [9] can only bring noticable gains over simple ZF for $M \approx K$, which is the regime where we have learnt not to operate.

As mentioned earlier, the BS needs CSI in multi-user MIMO systems to separate the signals associated with the different users. Perfect CSI can typically not be achieved in practice, since the channels are changing over time and frequency, and thus must be estimated using limited resources. The channel estimation of a frequency-selective channel can be handled by splitting the frequency resources into

multiple independent frequency-flat subchannels that can be estimated separately. A known pilot sequence is transmitted over each such subchannel and the received signal is used to estimate the channel response. In order to explore all spatial channel dimensions, this sequence must at least have the same length as the number of transmit antennas [8]. This means that a pilot sequence sent by the BS needs to have the length M, while the combined pilot sequence sent by the single-antenna user terminals needs to have the length K.

There are two ways of implementing the downlink and uplink transmission over a given frequency band. In frequency division duplex (FDD) mode the bandwidth is split into two separate parts: one for the uplink and one for the downlink. Pilot sequences are needed in both the downlink and the uplink due to the frequency-selective fading, giving an average pilot length of $(M+K)/2$ per subchannel. There is an alternative time-division duplex (TDD) mode where the whole bandwidth is used for both downlink and uplink transmission, but separated in time. If the system switches between downlink and uplink faster than the channels are changing, then it is sufficient to learn the channels in only one of the directions. This leads to an average pilot length of $\min(M,K)$ per subchannel, if we send pilots only in the most efficient direction. In the preferable operating regime of $M \gg K$, we note that TDD systems should send pilots only in the uplink and the pilot length becomes $\min(M,K) = K$. We conclude that TDD is the preferable mode since it not only requires shorter pilots than FDD, but is also highly scalable since the pilot length is independent of the number of BS antennas.

We give a concrete numerical example in Fig. 3 for downlink transmission with $K = 10$ users, an SNR of 5 dB, and uncorrelated Rayleigh fading channels. Two linear precoding schemes are considered; (a) maximum ratio (MR) and (b) zero-forcing (ZF). These schemes are later defined mathematically in Sect. 3. This simulation compares the SE obtained when having perfect CSI with the performance when having CSI estimated with pilot sequences of length τ_p. The SE is shown as a function of the number of BS antennas, M, and we compare TDD mode using $\tau_p = K = 10$ with FDD mode using either $\tau_p = 10$, $\tau_p = M$, or $\tau_p = \min(M,50)$, where the latter models an arbitrarily chosen maximum pilot length of 50 (e.g., motivated by pilot overhead constraints).

In TDD mode there is a visible performance loss in Fig. 3 as compared to having perfect CSI. The loss with MR precoding is very small, which shows that it is robust to estimation errors. The performance loss is larger for ZF precoding, since estimation errors make it harder to suppress interference, but we notice that ZF anyway provide higher performance than MR for all considered M. We notice that the performance losses are substantially constant irrespective of the number of BS antennas, thus TDD systems always benefit from adding more antennas. In contrast, FDD systems only benefits from adding more antennas if the pilot sequences are also made longer, as in the case $\tau_p = M$. With $\tau_p = 10$ there is no benefit from having more than 10 antennas, while the performance saturates at 50 antennas when $\tau_p = \min(M,50)$. In summary, TDD operation is fully scalable with respect to the number of BS antennas, while FDD operation can only handle more antennas by also increasing the pilot overhead. It is practically feasible to deploy FDD systems

with many antennas, particularly for slowly varying channels where we can accept a large pilot overhead, but TDD is always the better choice in this respect.

Second lesson learned: The channel estimation is simplified when operating in TDD mode, since the pilot sequences only need to be of length K irrespective of the number of BS antennas M.

Note that the uplink works in the same way in the TDD and FDD modes, while the distinct benefit of TDD in terms of scalability appears in the downlink.

2.2.1 Favorable Propagation

Recall from Fig. 2 that by adding more BS antennas, both the sum capacity-achieving non-linear processing and the simplified linear ZF processing approached the case without interference. This is not a coincidence but a fundamental property that is referred to as *favorable propagation*.

Let $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{C}^M$ represent the channel responses between a BS and two different user terminals. If these vectors are non-zero and orthogonal in the sense that

$$\mathbf{h}_1^H \mathbf{h}_2 = 0, \tag{1}$$

where $(\cdot)^H$ denotes the conjugate transpose, then the BS can completely separate the signals s_1, s_2 transmitted by the users when it observes $\mathbf{y} = \mathbf{h}_1 s_1 + \mathbf{h}_2 s_2$. By simply computing the inner product between $\mathbf{y}$ and $\mathbf{h}_1$, the BS obtains

$$\mathbf{h}_1^H \mathbf{y} = \mathbf{h}_1^H \mathbf{h}_1 s_1 + \mathbf{h}_1^H \mathbf{h}_2 s_2 = \|\mathbf{h}_1\|^2 s_1 \tag{2}$$

where the inter-user interference disappeared due to (1). The same thing can be done for the second user: $\mathbf{h}_2^H \mathbf{y} = \|\mathbf{h}_2\|^2 s_2$. Note that the BS needs perfect knowledge of $\mathbf{h}_1$ and $\mathbf{h}_2$ to compute these inner products. The channel orthogonality in (1) is called favorable propagation, since the two users can communicate with the BS without affecting each other.

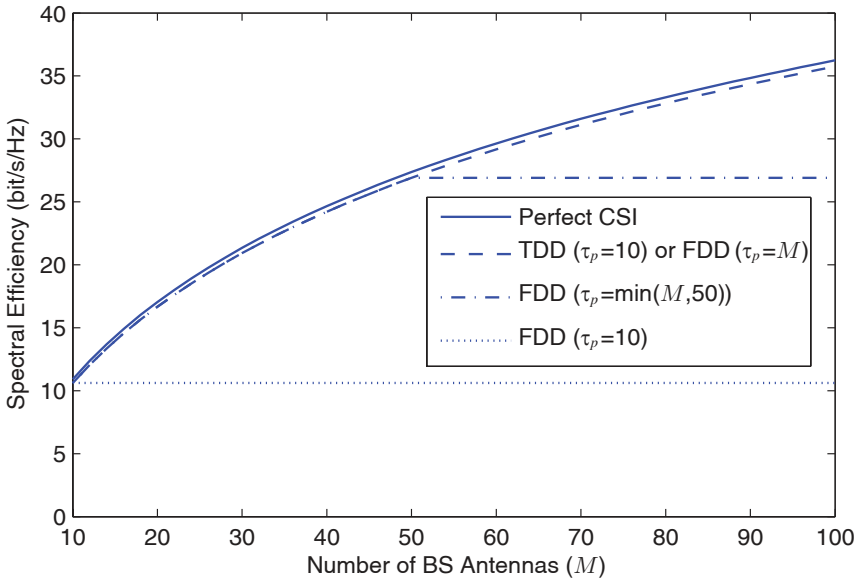
Is there a chance that practical channels offer favorable propagation? Probably not according to the strict definition that $\mathbf{h}_1^H \mathbf{h}_2 = 0$, but an approximate form of favorable propagation is achieved in non-line-of-sight scenarios with rich scattering:

Lemma 1. *Suppose that $\mathbf{h}_1 \in \mathbb{C}^M$ and $\mathbf{h}_2 \in \mathbb{C}^M$ have independent random entries with zero mean, identical distribution, and bounded fourth-order moments, then*

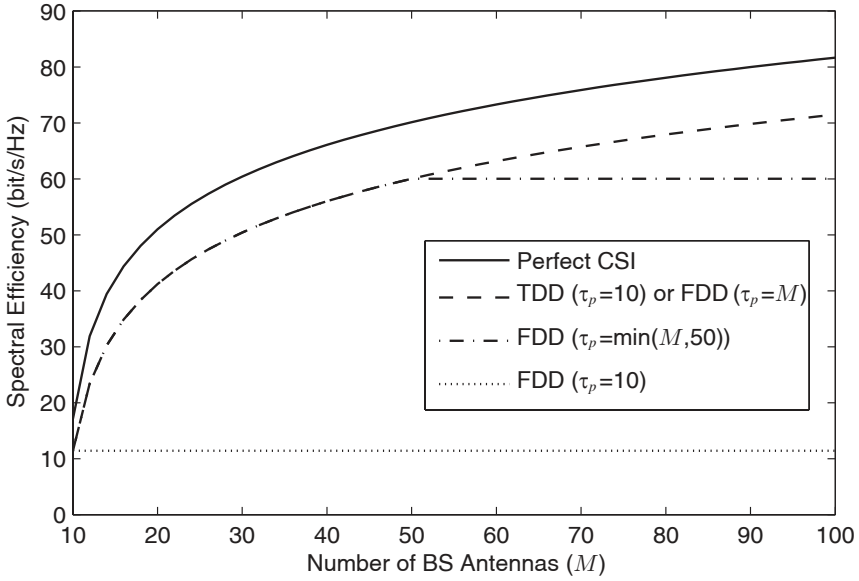
$$\frac{\mathbf{h}_1^H \mathbf{h}_2}{M} \to 0 \tag{3}$$

almost surely as $M \to \infty$.

Proof. This is a consequence of the law of large numbers. A direct proof is provided along with Theorem 3.7 in [14]. ■



(**a**)



(**b**)

Fig. 3 Average downlink spectral efficiency, as a function of the number of BS antennas, with different processing schemes and different types of CSI available at the BS. (**a**) Downlink simulation with maximum ratio precoding. (**b**) Downlink simulation with zero-forcing precoding

This lemma shows that the inner product between $\mathbf{h}_1$ and $\mathbf{h}_2$, if normalized with the number of BS antennas, goes asymptotically to zero as M increases. We refer to this as *asymptotic favorable propagation* and note that this phenomenon explains the behaviors in Fig. 2; the difference between having no inter-user interference and suppressing the interference by ZF becomes smaller and smaller as the number of antennas increases, because the loss in desired signal gain when using ZF reduces when the user channels become more orthogonal.

One special case in which Lemma 1 holds is $\mathbf{h}_1, \mathbf{h}_2 \sim \mathbb{C}\mathcal{N}(\mathbf{0}, \mathbf{I}_M)$, where $\mathbb{C}\mathcal{N}(\cdot,\cdot)$ denotes a multi-variate circularly symmetric complex Gaussian distribution and $\mathbf{I}_M$ is the $M \times M$ identity matrix. This is known as *uncorrelated Rayleigh fading* and in this case one can even prove that the variance of the inner product in (3) is $1/M$ and thus decreases linearly with the number of antennas [31]. Many academic works on Massive MIMO systems consider Rayleigh fading channels, due to the analytic tractability of Gaussian distributions. Nevertheless, Lemma 1 shows that asymptotic favorable propagation holds for other random channel distributions as well. This mathematical result can be extended to also include correlation between the elements in a channel vector. One can also derive similar analytic results for line-of-sight propagation [31] and behaviors that resemble asymptotic favorable propagation have been observed also in the real-world multi-user MIMO channel measurements presented in [16, 20].

Third lesson learned: Most wireless channels seem to provide asymptotic favorable propagation.

This lesson is yet another reason to design multi-user MIMO systems with $M \gg K$. It is, however, important to note that $(\mathbf{h}_1^H \mathbf{h}_2)/M \to 0$, as $M \to \infty$, does not imply that $\mathbf{h}_1^H \mathbf{h}_2 \to 0$. Strict favorable propagation is unlikely to appear in practical or theoretical channels. In fact, the inner product $\mathbf{h}_1^H \mathbf{h}_2$ grows roughly as $\sqrt{M}$ for Rayleigh fading channels. The key point is that this correlation has a negligible impact, since the SE depends on $(\mathbf{h}_1^H \mathbf{h}_2)/M$ which goes to zero roughly as $1/\sqrt{M}$. Moreover, the main suppression of inter-user interference appears already at relatively small number of antennas due to the square root.

2.3 Massive MIMO Concept

The Massive MIMO concept was proposed in the seminal paper [28] and described in the patent [29], both of which have received numerous scientific awards. Massive MIMO takes multi-user MIMO communications to a new level by designing a highly scalable communication protocol that utilizes the three lessons described in Sect. 2.2. The basic information and communication theoretic limits of this 5G technology were established in early works such as [3, 19, 21, 23, 30]. In this chapter we define Massive MIMO as follows:

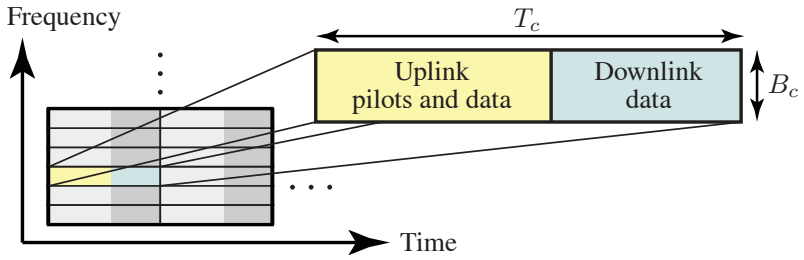


Fig. 4 Illustration of the basic Massive MIMO transmission protocol, where the time-frequency resources are divided into coherence intervals, each containing $\tau_c = B_c T_c$ transmission symbols. Each coherence interval contains uplink pilot sequences and can be used for both uplink and downlink payload data transmission based on TDD operation

Massive MIMO is a multi-user MIMO system with M antennas and K users per BS. The system is characterized by $M \gg K$ and operates in TDD mode using linear uplink and downlink processing.

This definition does not manifest any particular ratio between M and K, or any particular orders of magnitude that these parameters should have. One attractive example is a system with M in the range of 100 to 200 antennas, serving between $K = 1$ and $K = 40$ users depending on the data traffic variations. The first public real-time implementation of Massive MIMO is the LuMaMi testbed described in [41], which features $M = 100$ and $K = 10$. We stress that other definitions of Massive MIMO are available in other works and can both be more restrictive (e.g., require certain dimensionality of M and K) and looser (e.g., also include FDD mode), but in this chapter we only consider the definition above.

The BS antenna array typically consists of M dipole antennas, each having an effective size $\lambda/2 \times \lambda/2$, where λ is the wavelength. This means that an array area of 1 m^2 can fit 100 antennas at a 1.5 GHz carrier frequency and 400 antennas at 3 GHz. Each antenna is attached to a separate transceiver chain, so that the system can access the individual received signal at each antenna and select the individual signals to be transmitted from each antenna. The array can have any geometry; linear, rectangular, cylindrical, and distributed arrays are described in [25]. It is important to note that no model of the array geometry is exploited in the Massive MIMO processing, thus the antennas can be deployed arbitrarily without any geometrical array calibration.

The basic Massive MIMO transmission protocol is illustrated in Fig. 4. The time-frequency resources are divided into blocks of size B_c Hz and T_c s, with the purpose of making each user channel approximately frequency-flat and static within a block. Hence, the bandwidth B_c is selected to be smaller or equal to the anticipated channel coherence bandwidth among the users, while T_c is smaller or equal to the anticipated channel coherence time of the users. For this particular reason, each block is referred to as a *coherence interval*. The number of transmission symbols that fit into

a coherence interval is given by $\tau_c = B_c T_c$, due to the Nyquist-Shannon sampling theorem. The dimensionality of the coherence interval depends greatly on the anticipated system application. For example, a coherence interval of $\tau_c = 200$ symbols can be obtained with $B_c = 200$ kHz and $T_c = 1$ ms, which supports highway user velocities in urban environments at 2 GHz carrier frequencies. Much larger coherence intervals (e.g., τ_c at the order of 10^3 or 10^4) can be obtained by limiting the application to scenarios with low user mobility and short delay spread.

Each coherence interval is operated in TDD mode and can contain both downlink and uplink payload transmissions. To enable channel estimation at the BS, τ_p of the symbols in each coherence interval are allocated for uplink transmission of pilot sequences (where $\tau_p \geq K$), while the remaining $\tau_c - \tau_p$ symbols can be allocated arbitrarily between uplink and downlink payload data transmissions.

We let γ^{UL} and γ^{DL} denote the fractions of uplink and downlink payload transmission, respectively. This means that the uplink contains $\gamma^{\mathrm{UL}}(\tau_c - \tau_p)$ data symbols and the downlink contains $\gamma^{\mathrm{DL}}(\tau_c - \tau_p)$ data symbols per coherence interval. Naturally, these fractions satisfies $\gamma^{\mathrm{UL}} + \gamma^{\mathrm{DL}} = 1$ and $\gamma^{\mathrm{UL}}, \gamma^{\mathrm{DL}} \geq 0$. Notice that no downlink pilots are assumed in this protocol, since the effective precoded channels converge to their mean values when the BS has many antennas (due to the law of large numbers). It is certainly possible to also send a small amount of downlink pilots, particularly for estimating the small fading variations of the effective precoded channels, but the additional gains from doing this appears to be small in many relevant Massive MIMO cases [33].

Based on this definition of Massive MIMO, the next sections analyze how large SEs that the transmission protocol can offer in 5G cellular networks.

3 Performance Analysis

In this section, we describe the uplink detection and downlink precoding of a Massive MIMO network, and analyze the achievable system performance. We consider a basic Massive MIMO network comprising L cells, each consisting of a BS with M antennas and K single-antenna user terminals.

The channel response between the lth BS and user k in the ith cell is denoted by $\mathbf{h}_{i,k}^l = [h_{i,k,1}^l \ldots h_{i,k,M}^l]^T \in \mathbb{C}^M$, where $(\cdot)^T$ denotes the transpose. These channel vectors are ergodic random variables that are assumed to take new independent realizations in each coherence interval; recall the Massive MIMO protocol described in Sect. 2.3. To show that the general concept of Massive MIMO is applicable in any propagation environment, we keep the performance analysis general by only defining the basic statistical channel properties: the mean value and variance of each channel coefficient $h_{i,k,m}^l$ (note that m stands for the mth antenna at BS l, for $m = 1, \ldots, M$). We let

$$\bar{\mathbf{h}}_{i,k}^l = \mathbb{E}\{\mathbf{h}_{i,k}^l\} = [\bar{h}_{i,k,1}^l \ldots \bar{h}_{i,k,M}^l]^T \tag{4}$$

denote the vector of mean values. The variance of the mth coefficient of $\mathbf{h}_{i,k}^{l}$ is denoted by

$$\beta_{i,k}^{l} = \mathbb{V}\{h_{i,k,m}^{l}\}, \tag{5}$$

which is independent of the antenna index m (assuming that the large-scale fading is stationary over the BS array). We also assume that each BS and user can keep perfect track of these long-term statistical properties, and that the user channels are statistically independent.

Using these channel properties, we now analyze the uplink and the downlink.

3.1 Uplink with Linear Detection

For each uplink symbol, the received baseband signal $\mathbf{y}_l \in \mathbb{C}^M$ at the lth BS is modeled as

$$\mathbf{y}_l = \sum_{i=1}^{L} \sum_{k=1}^{K} \mathbf{h}_{i,k}^{l} \sqrt{p_{i,k}} x_{i,k} + \mathbf{n}_l, \tag{6}$$

where $x_{i,k}$ is the normalized transmission symbol (with $\mathbb{E}\{|x_{i,k}|^2\} = 1$) and $p_{i,k}$ is the transmit power of user k in cell i. The receiver hardware at the BS is contaminated by additive white noise, as modeled by the vector $\mathbf{n}_l \in \mathbb{C}^M$ which is zero-mean circularly symmetric complex Gaussian distributed with variance σ_{UL}^2; that is, $\mathbf{n}_l \sim \mathbb{C}\mathcal{N}(\mathbf{0}, \sigma_{\mathrm{UL}}^2 \mathbf{I}_M)$.

The matrix notations $\mathbf{H}_i^l = [\mathbf{h}_{i,1}^l \ldots \mathbf{h}_{i,K}^l] \in \mathbb{C}^{M \times K}$, $\mathbf{P}_i = \mathrm{diag}(p_{i,1}, \ldots, p_{i,K}) \in \mathbb{C}^{K \times K}$, and $\mathbf{x}_i = [x_{i,1} \ldots x_{i,K}]^T \in \mathbb{C}^K$ can be used to write the multi-cell multi-user MIMO system model from (6) in a compact matrix form:

$$\mathbf{y}_l = \sum_{i=1}^{L} \mathbf{H}_i^l \mathbf{P}_i^{1/2} \mathbf{x}_i + \mathbf{n}_l. \tag{7}$$

The channels $\mathbf{h}_{i,k}^l$ need to be estimated at BS l to perform good detection and this is done in the uplink by letting each user transmit a sequence of τ_p pilot symbols; see Fig. 4. We let $\tau_p = fK$ for some positive integer f (e.g., $1, 2, \ldots$) which is called the *pilot reuse factor*. This allows for linear independence between a total of τ_p different pilot sequences. This is, by design, sufficient to allocate independent pilot sequence to the K users in each cell and to also divide the L cells into f disjoint cell groups having fully independent pilot sequences. The benefit of having multiple cell groups is reduced interference during the pilot transmission and the corresponding gains in estimation quality are quantified below.

The uplink received signal $\mathbf{Y}_l^{\mathrm{pilot}} \in \mathbb{C}^{M \times \tau_p}$ at the lth BS during pilot transmission is

$$\mathbf{Y}_l^{\mathrm{pilot}} = \sum_{i=1}^{L} \mathbf{H}_i^l \mathbf{P}_i^{1/2} \mathbf{\Phi}_i^H + \mathbf{N}_l \tag{8}$$

and collects the received signal from (7) over the τ_p pilot symbols. Here, $\boldsymbol{\Phi}_i = [\boldsymbol{\phi}_{i,1} \ldots \boldsymbol{\phi}_{i,K}] \in \mathbb{C}^{\tau_p \times K}$ denotes the pilot matrix used by the K users in the ith cell, where $\boldsymbol{\phi}_{i,k} \in \mathbb{C}^{\tau_p}$ is the pilot sequence used by the kth user in that cell. The pilot matrix satisfies $\boldsymbol{\Phi}_i^H \boldsymbol{\Phi}_i = \tau_p \mathbf{I}_K$. Moreover, $\boldsymbol{\Phi}_l^H \boldsymbol{\Phi}_i = \tau_p \mathbf{I}_K$ if cell l and cell j belong to the same cell group (i.e., use the same set of pilots), while $\boldsymbol{\Phi}_l^H \boldsymbol{\Phi}_i = \mathbf{0}$ if the two cells belong to different cell groups. For notational convenience, we let $\mathscr{P}_l \subset \{1, \ldots, L\}$ denote the set of cell indices that belong to the same cell group as cell l, including l itself. Some particular examples are given later in Fig. 7.

By using the channel mean and variances, defined in the beginning of Sect. 3, we can use the linear minimum mean square error (LMMSE) estimator to separately acquire each element of $\mathbf{h}_{i,k}^l$ from the received pilot signal (8), which was proposed in [37] as a low-complexity estimation scheme. The channel estimate $\hat{\mathbf{h}}_{i,k}^l$ related to the true channel response $\mathbf{h}_{i,k}^l$ is given by the following lemma.

Lemma 2. *Suppose that BS l estimates each channel coefficient separately from its received signal* (8) *using an LMMSE estimator. BS l can then estimate the channel to the kth user in the jth cell as*

$$\hat{\mathbf{h}}_{j,k}^l = \bar{\mathbf{h}}_{j,k}^l + \frac{\sqrt{p_{j,k}}\beta_{j,k}^l}{\sum_{i \in \mathscr{P}_j} p_{i,k}\tau_p \beta_{i,k}^l + \sigma_{\mathrm{UL}}^2} \left(\mathbf{Y}_l^{\mathrm{pilot}} \boldsymbol{\phi}_{j,k} - \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}}\tau_p \bar{\mathbf{h}}_{i,k}^l \right). \tag{9}$$

Each element of the uncorrelated estimation error $\mathbf{e}_{j,k}^l = \mathbf{h}_{j,k}^l - \hat{\mathbf{h}}_{j,k}^l$ has zero mean and the variance

$$\mathrm{MSE}_{j,k}^l = \beta_{j,k}^l \left(1 - \frac{p_{j,k}\tau_p \beta_{j,k}^l}{\sum_{i \in \mathscr{P}_j} p_{i,k}\tau_p \beta_{i,k}^l + \sigma_{\mathrm{UL}}^2} \right). \tag{10}$$

Proof. The proof is available in Appendix at the end of this chapter. ■

It is worth emphasizing that the estimation error variance in (10) is independent of M, thus the estimation quality per channel coefficient is not affected by adding more antennas at the BS. Note that Lemma 2 holds for any correlation between the channel coefficients, since each coefficient is estimated separately. If the channel coefficients are correlated, with a known correlation structure and distribution, the estimation quality would improve with the number of antennas if the estimator is modified appropriately [8]. We also stress that the estimation error is only affected by noise and interference from the users in the same cell group that are allocated the same pilot sequence. In addition, we notice that the estimate in (9) can be computed using elementary linear algebra operations, with low computational complexity.

Using the channel estimates derived in Lemma 2, in this chapter, we analyze the performance of a Massive MIMO network with non-cooperative BSs. During uplink payload data transmission this means that the BS in cell l only utilizes its own received signal $\mathbf{y}_l$ in (6) and only targets to detect the signals sent by its own K users. Signals coming from users in other cells are perceived as inter-cell interference and eventually treated as additional noise. The BS in cell l discriminates the

signal transmitted by its kth user from the interference by multiplying the received signal in (6) with a linear detection vector $\mathbf{v}_{l,k} \in \mathbb{C}^M$ as follows:

$$\begin{aligned}\mathbf{v}_{l,k}^H \mathbf{y}_l &= \sum_{i=1}^{L}\sum_{t=1}^{K} \mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l \sqrt{p_{i,t}} x_{i,t} + \mathbf{v}_{l,k}^H \mathbf{n}_l \\ &= \underbrace{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l \sqrt{p_{l,k}} x_{l,k}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{t=1 \\ t\neq k}}^{K} \mathbf{v}_{l,k}^H \mathbf{h}_{l,t}^l \sqrt{p_{l,t}} x_{l,t}}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{i=1 \\ i\neq l}}^{L}\sum_{t=1}^{K} \mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l \sqrt{p_{i,t}} x_{i,t}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{v}_{l,k}^H \mathbf{n}_l}_{\text{Residual noise}}\end{aligned} \tag{11}$$

where $x_{i,t}$ is the transmitted data symbol from user t in cell i. As seen from (11), the processed received signal is the superposition of four parts: the desired signal, intra-cell interference, inter-cell interference, and residual noise. Since the linear detection vector $\mathbf{v}_{l,k}$ appears in all these terms, it can be used to amplify the desired signal, suppress the interference, and/or suppress the noise. More precisely, by gathering the detection vectors at BS l in matrix form as $\mathbf{V}_l = [\mathbf{v}_{l,1} \ldots \mathbf{v}_{l,K}] \in \mathbb{C}^{M\times K}$, there are two main schemes being considered in the Massive MIMO literature: maximum ratio (MR) and zero-forcing (ZF). These are given by

$$\mathbf{V}_l = \begin{cases} \hat{\mathbf{H}}_l^l, & \text{for MR,} \\ \hat{\mathbf{H}}_l^l \left((\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l \right)^{-1}, & \text{for ZF.} \end{cases} \tag{12}$$

MR detection exploits the M observations in $\mathbf{y}_l$ to maximize the ratio between the average signal gain in (11) and the norm of the detection vector:

$$\mathbb{E}\left\{ \frac{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l}{\|\mathbf{v}_{l,k}\|} \right\} = \frac{\mathbf{v}_{l,k}^H \hat{\mathbf{h}}_{l,k}^l}{\|\mathbf{v}_{l,k}\|} \leq \|\hat{\mathbf{h}}_{l,k}^l\| \tag{13}$$

where the expectation is computed with respect to the zero-mean channel estimation error. The inequality in (13) is satisfied with equality by $\mathbf{v}_{l,k} = \hat{\mathbf{h}}_{l,k}^l$ (leading to MR detection with $\mathbf{V}_l = \hat{\mathbf{H}}_l^l$). In contrast, the ZF detection matrix utilizes the M observations over the antennas to minimize the average intra-cell interference, while retaining the desired signals:

$$\mathbb{E}\{\mathbf{V}_l^H \mathbf{H}_l^l \mathbf{P}_l^{1/2} \mathbf{x}_l\} = \mathbf{V}_l^H \hat{\mathbf{H}}_l^l \mathbf{P}_l^{1/2} \mathbf{x}_l = \left((\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l \right)^{-1} \left((\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l \right) \mathbf{P}_l^{1/2} \mathbf{x}_l = \mathbf{P}_l^{1/2} \mathbf{x}_l \tag{14}$$

where the expectation is computed with respect to the zero-mean channel estimation error and the second equality follows from the ZF detection matrix definition. The average processed signal becomes $\mathbf{P}_l^{1/2} \mathbf{x}_l = [\sqrt{p_{l,1}} x_{l,1} \ldots \sqrt{p_{l,K}} x_{l,K}]^T$, which contains no intra-cell interference. Note that the inverse of the $K \times K$ matrix $(\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l$

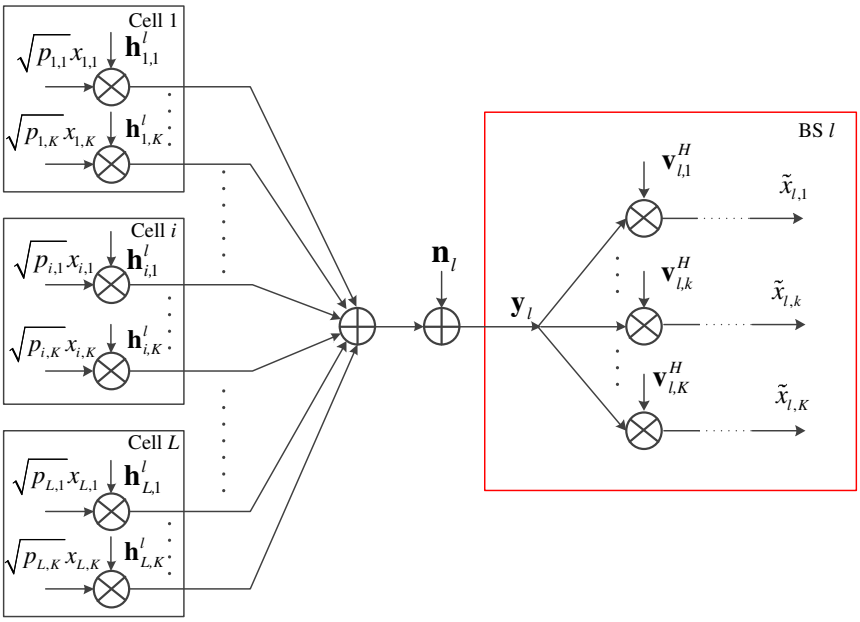


Fig. 5 Block diagram of the uplink transmission with linear detection in a multi-cell multi-user MIMO network, where BS l receives a linear combination of the signals transmitted from all K users in all L cells

only exists if $M \geq K$. There are also multi-cell variants of ZF detection that can be used to cancel out inter-cell interference; see for example [2] and [7].

A block diagram of the uplink transmission with linear detection is provided in Fig. 5. The purpose of the detection is to make the detected signal $\tilde{x}_{l,k}$ at BS l equal to the true signal $x_{l,k}$, at least up to a scaling factor. Due to noise and estimation errors, there is always a mismatch between the signals which is why the communication link has a limited capacity. If the true signal $x_{l,k}$ originates from a discrete constellation set $\mathscr{X}$ (e.g., a quadrature amplitude modulation (QAM)), $\tilde{x}_{l,k}$ is selected based on $\mathbf{v}_{l,k}^H \mathbf{y}_l$ by finding the minimum distance over all the candidates $x \in \mathscr{X}$:

$$\tilde{x}_{l,k} = \min_{x \in \mathscr{X}} \left| \mathbf{v}_{l,k}^H \mathbf{y}_l - \mathbf{v}_{l,k}^H \hat{\mathbf{h}}_{l,k}^l \sqrt{p_{l,k}} x \right|^2. \tag{15}$$

This expression can be utilized to compute bit error rates and similar uncoded performance metrics. Since modern communication systems apply channel coding over relatively long data blocks, to protect against errors, the ergodic channel capacity is a more appropriate performance metric in 5G networks. It merits to note that the ergodic capacitites of the individual communication links are hard to characterize exactly, particularly under imperfect channel knowledge, but tractable lower bounds are obtained by the following theorem.

Theorem 1. *In the uplink, a lower bound on the ergodic capacity of an arbitrary user k in cell l is*

$$R_{l,k}^{\mathrm{UL}} = \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2\left(1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}\right), \tag{16}$$

where the signal-to-interference-and-noise ratio (SINR) is

$$\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \frac{p_{l,k} \left| \mathbb{E}\left\{ \mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l \right\} \right|^2}{\sum\limits_{i=1}^{L} \sum\limits_{t=1}^{K} p_{i,t} \mathbb{E}\left\{ |\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l|^2 \right\} - p_{l,k} \left| \mathbb{E}\left\{ \mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l \right\} \right|^2 + \sigma_{\mathrm{UL}}^2 \mathbb{E}\left\{ \|\mathbf{v}_{l,k}\|^2 \right\}}. \tag{17}$$

Proof. The proof is available in Appendix at the end of this chapter. ■

Theorem 1 demonstrates that the achievable SE of an arbitrary user k in cell l in a Massive MIMO network can be described by an SINR term $\mathrm{SINR}_{l,k}^{\mathrm{UL}}$ that contains expectations with respect to the small-scale channel fading. The numerator contains the gain of the desired signal, while the denominator contains three different terms. The first term is the average power of all the signals, including both multi-user interference and the desired signal, while the second term subtracts the part of the desired signal power that is usable for decoding. The third term is the effective noise power. The pre-log factor $(1 - \frac{\tau_p}{\tau_c})$ compensates for the fact that τ_p/τ_c of the transmission symbols contain pilots instead of payload data. The SE is also multiplied by γ^{UL}, which was defined earlier as the fraction of uplink data. Clearly, MR detection aims at maximizing the numerator of $\mathrm{SINR}_{l,k}^{\mathrm{UL}}$, while ZF detection tries to minimize the intra-cell interference.

The expectations in Theorem 1 can be computed numerically for any channel distribution and any detection scheme. In the case of MR detection, the desired signal gain $|\mathbb{E}\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\}|^2$ grows as M^2 for most channel distributions, while the noise term $\sigma_{\mathrm{UL}}^2 \mathbb{E}\{\|\mathbf{v}_{l,k}\|^2\}$ only grows as M and thus becomes less significant the more antennas are deployed at the BS. This property is known as the *array gain* from coherent detection. The behavior of the multi-user interference terms greatly depends on the channel distribution, but typically these terms will also have the slower scaling of M [31], except for users that interfered with each other during pilot transmission (i.e., appeared in each other's expressions (10) for the estimation error variance). The latter is a phenomenon called *pilot contamination* and is further discussed later in this chapter.

To demonstrate these properties in detail, we now consider the special case in which the channel between BS l and user k in cell i is uncorrelated Rayleigh fading:

$$\mathbf{h}_{i,k}^l \sim \mathbb{C}\mathcal{N}\left(\mathbf{0}, \beta_{i,k}^l \mathbf{I}_M\right). \tag{18}$$

Hence, $\bar{\mathbf{h}}_{i,k}^l = \mathbb{E}\{\mathbf{h}_{i,k}^l\} = \mathbf{0}$, which means that there is no line-of-sight channel component. This special case is relevant in rich-scattering environments where the channel does not have any statistically dominating directivity.

Subsequently, the LMMSE estimate in Lemma 2 simplifies to

$$\hat{\mathbf{h}}_{j,k}^{l} = \frac{\sqrt{p_{j,k}}\beta_{j,k}^{l}}{\sum\limits_{i\in\mathscr{P}_j} p_{i,k}\tau_p\beta_{i,k}^{l} + \sigma_{\mathrm{UL}}^2}\mathbf{Y}_{l}^{\mathrm{pilot}}\boldsymbol{\phi}_{j,k} \tag{19}$$

and becomes circularly-symmetric complex Gaussian distributed:

$$\hat{\mathbf{h}}_{j,k}^{l} \sim \mathbb{C}\mathscr{N}\left(\mathbf{0}, (\beta_{j,k}^{l} - \mathrm{MSE}_{j,k}^{l})\mathbf{I}_M\right). \tag{20}$$

There is an important relationship between the two estimated channels $\hat{\mathbf{h}}_{l,k}^{l}$ and $\hat{\mathbf{h}}_{i,k}^{l}$ for cell indices i and l such as $i \in \mathscr{P}_l$ expressed by

$$\hat{\mathbf{h}}_{i,k}^{l} = \frac{\sqrt{p_{i,k}}\beta_{i,k}^{l}}{\sqrt{p_{l,k}}\beta_{l,k}^{l}}\hat{\mathbf{h}}_{l,k}^{l}. \tag{21}$$

This equation shows that BS l cannot tell apart the channels of users that send the same pilot sequence; the estimates are the same up to a scaling factor. This fact is the cause of pilot contamination and will have a key impact on the performance, as shown later.

Moreover, the LMMSE estimator in (19) is also the MMSE estimator in the special case of Rayleigh fading, since the channels are Gaussian distributed [24]. By using these key properties, the ergodic SE in Theorem 1 can be computed in closed form for MR and ZF detection, as shown by the following corollary.

Corollary 1. *In the uplink, if all channels are uncorrelated Rayleigh fading, the lower bound on the ergodic capacity of user k in cell l stated in Theorem 1 becomes*

$$R_{l,k}^{\mathrm{UL}} = \gamma^{\mathrm{UL}}\left(1 - \frac{\tau_p}{\tau_c}\right)\log_2\left(1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}\right), \tag{22}$$

where the SINR is

$$\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \frac{Gp_{l,k}\beta_{l,k}^{l}\frac{p_{l,k}\tau_p\beta_{l,k}^{l}}{\sum_{i'\in\mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^{l}+\sigma_{\mathrm{UL}}^2}}{G\sum\limits_{i\in\mathscr{P}_l\setminus\{l\}} p_{i,k}\beta_{i,k}^{l}\frac{p_{i,k}\tau_p\beta_{i,k}^{l}}{\sum_{i'\in\mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^{l}+\sigma_{\mathrm{UL}}^2} + \sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} p_{i,t}z_{i,t}^{l} + \sigma_{\mathrm{UL}}^2} \tag{23}$$

and the parameters G and $z_{i,t}^{l}$ depend on the choice of detection scheme. MR gives $G = M$ and $z_{i,t}^{l} = \beta_{i,t}^{l}$, while ZF gives $G = M - K$ and

$$z_{i,t}^{l} = \begin{cases} \mathrm{MSE}_{i,t}^{l}, & \textit{for } i \in \mathscr{P}_l, \\ \beta_{i,t}^{l}, & \textit{otherwise}. \end{cases}$$

Proof. The proof is available in Appendix at the end of this chapter. ■

The closed-form achievable SE expressions in Corollary 1 provide many insights on the advantages of spatial multi-user multiplexing and the effects of channel estimation. Firstly, the desired signal term in the numerator of (23) scales with the number of BS antennas, proportionally to M and $M-K$ with MR and ZF, respectively. This array gain is multiplied with the average received signal power per antenna, $p_{l,k}\beta_{l,k}^l$, and the relative channel estimation quality

$$\frac{p_{l,k}\tau_p\beta_{l,k}^l}{\sum_{i'\in\mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^l+\sigma_{\mathrm{UL}}^2}, \tag{24}$$

which is a number between 0 and 1 (where 1 is perfect CSI and 0 is no CSI).

Secondly, we notice that the first term of the denominator in (23) has a similar structure as the desired signal and represents the coherent pilot contamination—interference that is amplified along with the desired signals due the BS's inability to tell apart users that use the same pilot sequence. The pilot contamination degrades the SINR by adding additional interference that scales as M or $M-K$, depending on the detection scheme. However, since pilot contamination only arises at BS l from the interfering user in cell i in $\mathscr{P}_l$, the network can suppress pilot contamination by increasing the pilot reuse factor f and by designing the cell groups appropriately. To understand how to suppress pilot contamination, we have a look at the ratio between the pilot contamination term and the signal term in (23):

$$\frac{G\sum\limits_{i\in\mathscr{P}_l\setminus\{l\}} p_{i,k}\beta_{i,k}^l \frac{p_{i,k}\tau_p\beta_{i,k}^l}{\sum_{i'\in\mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^l+\sigma_{\mathrm{UL}}^2}}{Gp_{l,k}\beta_{l,k}^l \frac{p_{l,k}\tau_p\beta_{l,k}^l}{\sum_{i'\in\mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^l+\sigma_{\mathrm{UL}}^2}} = \sum_{i\in\mathscr{P}_l\setminus\{l\}} \left(\frac{p_{i,k}\beta_{i,k}^l}{p_{l,k}\beta_{l,k}^l}\right)^2. \tag{25}$$

This ratio represents the relative strength of the pilot contamination and (25) should preferably be small. The pilot contamination caused by UE k in cell i is small whenever $\beta_{i,k}^l/\beta_{l,k}^l$ is small, which occurs when either $\beta_{l,k}^l$ is large (i.e., the desired user is close to its serving BS) or $\beta_{i,k}^l$ is small (i.e., the interfering cell is far away). The cell groups should be designed based on these properties, and this issue is further discussed in Sect. 4.2.

Thirdly, the performance in Corollary 1 is also affected by classical noise and interference. Since MR focuses only on maximizing the SNR, the interference term $\sum_{i=1}^L\sum_{t=1}^K p_{i,t}\beta_{i,t}^l$ is simply the average signal power received at any antenna of BS l. In contrast, ZF pays attention to the intra-cell interference and takes no notice of the noise. The interference suppression replaces the full channel variance $\beta_{i,t}^l$ in the aforementioned interference summation with the estimation error variance $\mathrm{MSE}_{i,t}^l$ for cells $i\in\mathscr{P}_l$. Due to the imperfect CSI (i.e., $\mathrm{MSE}_{i,t}^l>0$) not all intra-cell interference can be removed by ZF. However, the pilot contamination also has the positive effect that not only intra-cell interference is suppressed, but also the inter-cell interference coming from other users in the same cell group (which use the same pilots as in cell l). The fact that the interference and noise terms are independent

of M, while the desired signal scales with M, is a consequence of the asymptotic favorable propagation that was described in Sect. 2.2.1.

If we limit the scope to a single-cell network, achievable SE expressions can be obtained directly from Corollary 1 by simply setting $\mathscr{P}_l = \{l\}$ and removing the interference from all other cells $j \in \{1,\ldots,L\} \setminus \{l\}$. For simplicity of exposition, we leave out the cell index l in this special case.

Corollary 2. *In the single-cell uplink, if all channels are uncorrelated Rayleigh fading, a lower bound on the ergodic SE of an arbitrary user k is given by*

$$R_k^{\mathrm{UL}} = \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \frac{G p_k^2 \tau_p \beta_k^2}{\left(p_k \tau_p \beta_k + \sigma_{\mathrm{UL}}^2\right) \left(\sum\limits_{t=1}^{K} p_t z_t + \sigma_{\mathrm{UL}}^2\right)}\right). \tag{26}$$

Here, the parameters G and z_t depend on the detection scheme. MR gives $G = M$ and $z_t = \beta_t$, while ZF gives $G = M - K$ and $z_t = \frac{\beta_t \sigma_{\mathrm{UL}}^2}{p_t \tau_p \beta_t + \sigma_{\mathrm{UL}}^2}$.

This corollary shows that the spatial multi-user multiplexing capability is even greater in isolated single-cell networks. The most notable difference compared to a multi-cell network is the lack of inter-cell interference, both during data and pilot transmission. In other words, the interference only originates from users within the own cell, while pilot contamination vanishes thanks to the orthogonality of all pilot sequences in the cell. The SE per cell is therefore higher in single-cell networks than in the multi-cell networks—at least if the cell geometry is the same and we only neglect inter-cell interference. The motivation of having multiple cells is, of course, to cover a larger area and thereby achieve much higher total SE. The scenarios when the interference suppression of ZF is beneficial as compared to MR can be identified from Corollary 2 as the cases when

$$\frac{M-K}{\sum\limits_{t=1}^{K} \frac{p_t \beta_t \sigma_{\mathrm{UL}}^2}{p_t \tau_p \beta_t + \sigma_{\mathrm{UL}}^2} + \sigma_{\mathrm{UL}}^2} > \frac{M}{\sum\limits_{t=1}^{K} p_t \beta_t + \sigma_{\mathrm{UL}}^2}. \tag{27}$$

To summarize, we have derived uplink SE expressions for Massive MIMO networks, for general channel distributions in Theorem 1 and for Rayleigh fading in Corollary 1. In the latter case, the expressions are in closed form and can thus be computed and analyzed directly, without having to simulate any channel fading realizations. These expressions are used in Sect. 4 to illustrate the anticipated performance of Massive MIMO networks.

3.2 Downlink with Linear Precoding

Next, we consider the downlink of a Massive MIMO network where the BSs are transmitting signals to their users. For an arbitrary BS l, we let $\mathbf{x}_l \in \mathbb{C}^M$ denote the transmitted signal vector intended for its K users. We consider linear precoding where this vector is computed as

$$\mathbf{x}_l = \sum_{t=1}^{K} \sqrt{\rho_{l,t}} \mathbf{w}_{l,t} s_{l,t}, \tag{28}$$

where the payload symbol $s_{l,t}$ intended for user t in cell l has unit transmit power $\mathbb{E}\{|s_{l,t}|^2\} = 1$ and $\rho_{l,t}$ represents the transmit power allocated to this particular user. Moreover, $\mathbf{w}_{l,t} \in \mathbb{C}^M$, for $t = 1, \ldots, K$, are the corresponding linear precoding vectors that determine the spatial directivity of the signal sent to each user.

The received signal $y_{l,k} \in \mathbb{C}$ at user k in cell l is modeled as

$$y_{l,k} = \sum_{i=1}^{L} (\mathbf{h}_{l,k}^{i})^H \mathbf{x}_i + n_{l,k}, \tag{29}$$

where $n_{l,k} \sim \mathbb{C}\mathcal{N}(0, \sigma_{\mathrm{DL}}^2)$ is the additive white noise with variance σ_{DL}^2. Notice that $\mathbf{h}_{l,k}^{i}$ is the same channel response as in the uplink, due to the reciprocity of physical propagation channels (within a coherence interval).[1] A block diagram of the downlink transmission is provided in Fig. 6. Since there are no downlink pilots in the Massive MIMO protocol described in Sect. 2.3, the users are assumed to only know the channel statistics. The lack of instantaneous CSI would greatly reduce the performance of small MIMO systems, but Massive MIMO works well without it since the effective precoded channels quickly approach their mean as more antennas are added. Hence, coherent downlink reception is possible using only statistical CSI. This leads to a low-complexity communication solution where all the intelligence is placed at the BS. Since the ergodic capacity is hard to characterize in this case, the following theorem derives a lower bound on the capacity between user k in cell l and its serving BS.

Theorem 2. *In the downlink, a lower bound on the ergodic rate an arbitrary user k in cell l is*

$$R_{l,k}^{\mathrm{DL}} = \gamma^{\mathrm{DL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \mathrm{SINR}_{l,k}^{\mathrm{DL}}\right), \tag{30}$$

where the SINR is

[1] In fact, the reciprocal channel is $(\mathbf{h}_{l,k}^{i})^T$, using the regular transpose instead of the conjugate transpose as in (29), but since the only difference is a complex conjugation we can characterize the performance using (29) without loss of generality. The reason to use the conjugate transpose is that the notation becomes easier and the relation to the uplink is clearer.

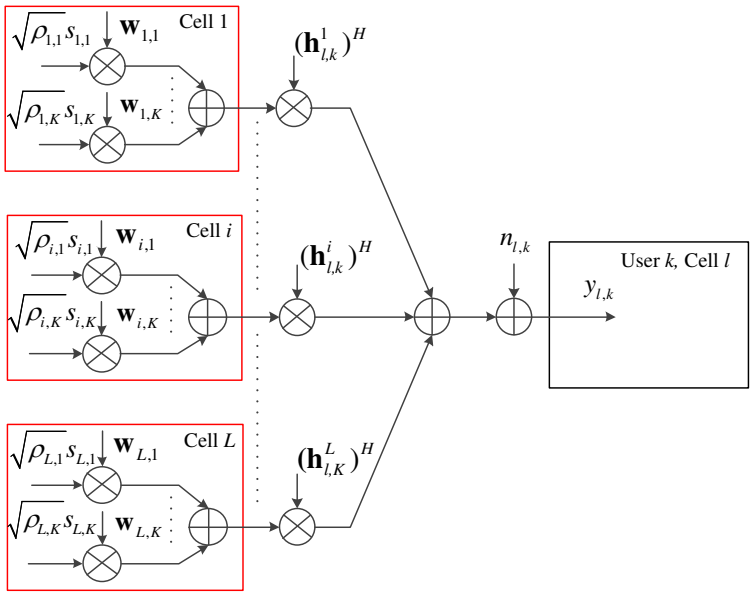


Fig. 6 Block diagram of the downlink transmission with linear precoding in a multi-cell MIMO system, where BSs equipped with M antennas are transmitting signals that reach user k in cell l

$$\mathrm{SINR}_{l,k}^{\mathrm{DL}} = \frac{\rho_{l,k} \left| \mathbb{E}\left\{ (\mathbf{h}_{l,k}^{l})^H \mathbf{w}_{l,k} \right\} \right|^2}{\sum\limits_{i=1}^{L} \sum\limits_{t=1}^{K} \rho_{i,t} \mathbb{E}\left\{ |(\mathbf{h}_{l,k}^{i})^H \mathbf{w}_{i,t}|^2 \right\} - \rho_{l,k} \left| \mathbb{E}\left\{ (\mathbf{h}_{l,k}^{l})^H \mathbf{w}_{l,k}^{l} \right\} \right|^2 + \sigma_{\mathrm{DL}}^2}. \quad (31)$$

Proof. The proof is available in Appendix at the end of this chapter. ■

The downlink achievable SE provided in Theorem 2 holds for any channel distributions and choice of precoding vectors. Since the uplink and downlink channels are reciprocal, it would make sense if the uplink and downlink performance were also somehow connected. The downlink achievable SE in Theorem 2 indeed bears much similarity with the corresponding uplink expression in Theorem 1. The desired signal terms are the same, except for the potentially different transmit power parameters and the fact that the detection vector is replaced by the corresponding precoding vector. The interference terms have a similar structure, but the indices are swapped between the channel vector and the processing vector. This is because the uplink interference arrives through different channels for different users while all the downlink interference from a particular cell comes through the same channel from the BS. These observations lead to the following *uplink-downlink duality* [7, 10]:

Lemma 3. *Suppose that the downlink precoding vectors are selected as*

$$\mathbf{w}_{l,k} = \frac{\mathbf{v}_{l,k}}{\sqrt{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}} \tag{32}$$

based on the uplink detection vectors $\mathbf{v}_{l,k}$*, for all* l *and* k*. For any given uplink powers* $p_{i,t}$ *(for* $i = 1,\ldots,L$ *and* $t = 1,\ldots,K$*), there exist a corresponding set of downlink powers* $\rho_{i,t}$ *(for* $i = 1,\ldots,L$ *and* $t = 1,\ldots,K$*) such that*

$$\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \mathrm{SINR}_{l,k}^{\mathrm{DL}} \tag{33}$$

for all l *and* k*, and*

$$\frac{\sum_{i=1}^{L}\sum_{t=1}^{K} p_{i,t}}{\sigma_{\mathrm{UL}}^2} = \frac{\sum_{i=1}^{L}\sum_{t=1}^{K} \rho_{i,t}}{\sigma_{\mathrm{DL}}^2}. \tag{34}$$

Proof. The proof is available in Appendix at the end of this chapter. ■

This lemma shows that the same performance can be achieved in both the uplink and the downlink, if the downlink power is allocated in a particular way based on the uplink powers and the precoding vectors are selected based on the detection vectors as in (32). The downlink powers are computed according to (72), which is given in Appendix at the end of this chapter since the important thing for now is that there exist a collection of downlink powers that give exactly the same performance in both directions. If $\sigma_{\mathrm{UL}}^2 = \sigma_{\mathrm{DL}}^2$, then the same total transmit power is used in both directions of the Massive MIMO network; however, the power will generally be distributed differently over the users.

Motivated by the uplink-downlink duality, it makes sense to consider MR and ZF precoding as the main downlink precoding schemes. These are defined as

$$\mathbf{w}_{l,k} = \begin{cases} \dfrac{\hat{\mathbf{h}}_{l,k}^{l}}{\sqrt{\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^2\right\}}}, & \text{for MR,} \\ \dfrac{\hat{\mathbf{H}}_{l}^{l}\mathbf{r}_{l,k}}{\sqrt{\mathbb{E}\left\{\|\hat{\mathbf{H}}_{l}^{l}\mathbf{r}_{l,k}\|^2\right\}}}, & \text{for ZF,} \end{cases} \tag{35}$$

where $\mathbf{r}_{l,k}$ denotes the kth column of $((\hat{\mathbf{H}}_{l}^{l})^{H}\hat{\mathbf{H}}_{l}^{l})^{-1}$.

Similar to the uplink performance analysis, we now compute the downlink SE in closed form for uncorrelated Rayleigh fading channels, as defined in (18). Because of the channel reciprocity, the channel estimates obtained at the BSs in the uplink can also be used in the downlink. In particular, the channel estimates $\hat{\mathbf{h}}_{i,k}^{i}$ and $\hat{\mathbf{h}}_{l,k}^{i}$ for cell indices i and l with $l \in \mathscr{P}_i$ are still related as

$$\hat{\mathbf{h}}_{l,k}^{i} = \frac{\sqrt{p_{l,k}}\beta_{l,k}^{i}}{\sqrt{p_{i,k}}\beta_{i,k}^{i}}\hat{\mathbf{h}}_{i,k}^{i}, \tag{36}$$

thus showing that pilot contamination exists also in the downlink; that is, BS i cannot precode signals toward its user k without also precode the signal towards user k in cell $i \in \mathscr{P}_l$. The next corollary specializes Theorem 2 for Rayleigh fading channels.

Corollary 3. *In the downlink, if all channels are uncorrelated Rayleigh fading, the lower bound on the ergodic capacity of user k stated in Theorem 2 becomes*

$$R_{l,k}^{\mathrm{DL}} = \gamma^{\mathrm{DL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \mathrm{SINR}_{l,k}^{\mathrm{DL}}\right), \tag{37}$$

where the SINR is

$$\mathrm{SINR}_{l,k}^{\mathrm{DL}} = \frac{G\rho_{l,k}\beta_{l,k}^{l} \frac{p_{l,k}\tau_p\beta_{l,k}^{l}}{\sum_{i' \in \mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^{l} + \sigma_{\mathrm{UL}}^2}}{G \sum\limits_{i \in \mathscr{P}_l \setminus \{l\}} \rho_{i,k}\beta_{l,k}^{i} \frac{p_{l,k}\tau_p\beta_{l,k}^{i}}{\sum_{i' \in \mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^{i} + \sigma_{\mathrm{UL}}^2} + \sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} \rho_{i,t} z_{l,k}^{i} + \sigma_{\mathrm{DL}}^2}. \tag{38}$$

The parameters G and $z_{l,k}^{i}$ are specified by the precoding scheme. MR precoding gives $G = M$ and $z_{l,k}^{i} = \beta_{l,k}^{i}$, while ZF precoding gives $G = M - K$ and

$$z_{l,k}^{i} = \begin{cases} \mathrm{MSE}_{l,k}^{i}, & \text{for } i \in \mathscr{P}_l, \\ \beta_{l,k}^{i}, & \text{otherwise}. \end{cases}$$

Proof. The proof is available in Appendix at the end of this chapter. ■

For Rayleigh fading channels, Corollary 3 shows that the array gain, pilot contamination, and all other attributes of MR and ZF precoding are very similar to the uplink counterparts. Hence, the same kind of observations can be made from Corollary 3 as previously done for Corollary 1.

In the single-cell scenario, the SE expression in Corollary 3 simplifies to the following result.

Corollary 4. *In the single-cell downlink, if all channels are uncorrelated Rayleigh fading, a lower bound on the ergodic SE of an arbitrary user k is given by*

$$R_{k}^{\mathrm{DL}} = \gamma^{\mathrm{DL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \frac{G\rho_k p_k \tau_p \beta_k^2}{\left(p_k\tau_p\beta_k + \sigma_{\mathrm{UL}}^2\right)\left(z_k \sum_{t=1}^{K} \rho_t + \sigma_{\mathrm{DL}}^2\right)}\right). \tag{39}$$

The parameters G and z_k depend on the precoding scheme. MR gives $G = M$ and $z_k = \beta_k$, while ZF obtains $G = M - K$ and $z_k = \frac{\beta_k \sigma_{\mathrm{UL}}^2}{p_k\tau_p\beta_k + \sigma_{\mathrm{UL}}^2}$.

We conclude the analytical part of this chapter by recalling that the uplink and downlink spectral efficiencies with Massive MIMO can be easily computed from Theorem 1 and Theorem 2 for any channel distributions and processing schemes. In the uncorrelated Rayleigh fading case there are even closed-form expressions. The same SINR performance can be achieved in the uplink and downlink, based on what is known as uplink-downlink duality. The intuition is that the downlink precoding and uplink detection vectors should be the same, but that the power allocation needs to be adapted differently in the two cases.

4 Design Guidelines and Anticipated Spectral Efficiency Gains

In this section, we provide some basic design guidelines for Massive MIMO networks and showcase the SEs that the technology can deliver to 5G networks according to the theory developed in Sect. 3. For illustrative purposes, we consider a classic cellular network topology with hexagonal cells, where each cell can be illustrated as in Fig. 1. In other words, the BS is deployed in the center of the cell, while the K users are distributed over the cell area. When many cells of this type are placed next to each other, the cellular network has the shape showed in Fig. 7. While conventional cellular networks use sectorization to split each cell into, say, three static sectors, this is not assumed here. This is because the spatial transceiver processing at the BS in Massive MIMO basically creates K virtual sectors, adapted dynamically to the positions of the current set of users.

4.1 Power Allocation

The average transmit power of user k in cell j is denoted by $p_{j,k}$ in the uplink and by $\rho_{j,k}$ in the downlink. These are important design parameters that determine the SEs of the users; see Theorem 1 (for the uplink) and Theorem 2 (for the downlink). Since inter-user interference is an important factor in any multi-user MIMO system, each transmit power coefficient affects not only the strength of the desired signal at the desired user, but also the amount of interference caused to all the other users in the network (although the interference is most severe within a cell and between neighboring cells). The selection of these transmit power coefficients is referred to as *power allocation* and needs to be addressed properly.

A key property of Massive MIMO is that the small-scale fading in time and the frequency-selective fading variations are negligible, since they essentially average out over the many antennas at each BS. For example, the SE expressions for Rayleigh fading channels in Corollaries 1–4 only depend on the channel variances $\beta_{i,k}^{l}$ and not on the instantaneous realizations of the corresponding channel vectors $\mathbf{h}_{i,k}^{l}$. Therefore, there is no need to change the power allocation between each coherence interval, but only over the longer time frame where the channel variances change, due to modifications in the large-scale propagation behaviors (e.g., caused by user mobility). This is a substantial increase of the time frame in which power allocation decisions are to be made, from milliseconds to seconds. This fact makes it possible to optimize and coordinate the power allocation across cells, in ways that have not been possible in the past due computational or delay limitations.

A structured approach to power allocation is to find the transmit powers that jointly maximize the network utility functions $U^{\mathrm{UL}}(\{R_{l,k}^{\mathrm{UL}}\})$ and $U^{\mathrm{DL}}(\{R_{l,k}^{\mathrm{DL}}\})$ in the uplink and downlink, respectively. These utilities are increasing functions of the users' SEs, where $\{R_{l,k}^{\mathrm{UL}}\}$ and $\{R_{l,k}^{\mathrm{DL}}\}$ denote the sets of all SEs. Some particular examples of network utility functions are [5]

$$U(\{R_{l,k}\}) = \begin{cases} \sum_{l=1}^{L}\sum_{k=1}^{K} R_{l,k}, & \text{Sum utility,} \\ \prod_{l=1}^{L}\prod_{k=1}^{K} R_{l,k}, & \text{Proportional fairness,} \\ \min_{l\in\{1,\ldots,L\},k\in\{1,\ldots,K\}} R_{l,k}, & \text{Max-min fairness,} \end{cases} \tag{40}$$

where we have omitted the uplink/downlink superscripts since the same type of utility function can be utilized in both cases. These utilities are often maximized with respect to a given power budget per user (in the uplink) and per BS (in the downlink). For brevity, we will not provide any further mathematical details, but briefly outline what is known around power allocation for Massive MIMO.

Maximization of the sum utility (SU) provides high SEs to users with good average channel conditions, at the expense of low SE for users with bad average channel conditions. In contrast, max-min fairness (MMF) enforces that each user should get equal SE, which effectively means that users with good channels reduce their SEs to cause less interference to the users with bad channels. Proportional fairness (PF) can be shown to lie in between these extremes. The SU achieves the highest sum SE, since this is really what is optimized by this utility function, while MMF trades some of the sum SE to obtain a uniform user experience. The choice of network utility function is a matter of subjective taste, since there is no objectively optimal utility function [5]. Nevertheless, there seems to be a trend towards more fairness-emphasizing utilities in the Massive MIMO literature [7, 32, 45], motivated by the fact that contemporary networks are designed to provide high peak rates, while the cell edge performance is modest and needs to be improved in 5G. In the uplink, another important aspect to consider in the power allocation is the fact that a BS cannot simultaneously receive desired user signals of very different power levels, since then the weak signals will then drown in the quantization noise caused by the analog-to-digital conversion. Hence, even if the channel attenuation might differ by 50 dB within a cell, these variations need to be brought down to, say, 10 dB by the uplink power allocation.

From a numerical optimization perspective, the downlink power allocation problem (for fixed uplink power allocation) has the same mathematical structure as the seemingly different scenario of single-antenna multi-cell communications with perfect CSI. The downlink utility optimization can therefore be solved using the techniques described in [4, 27, 35] and references therein. In general, the PF and MMF utilities give rise to convex optimization problems that can be solved efficiently with guaranteed convergence to the global optimum. These algorithms can also be implemented in a distributed fashion [4]. The SU problem is, in contrast, provable non-convex and hard to solve [27], which means that the optimal solution cannot be found under any practical constraints on complexity.

The uplink power allocation is more complicated than power allocation in the downlink; for example, because the SE expression in Corollary 1 contains both $p_{j,k}$ and $p_{j,k}^2$ (while the downlink SE expressions only contain the linear term $\rho_{j,k}$). Nevertheless, there are several efficient algorithms that maximize the MMF utility [13, 45, 12], and the approach in [12] can also maximize the SU and PF utilities with MR and ZF detection. The work [26] provides an alternative methodology to maximize an approximation of the SU metric for other detection methods.

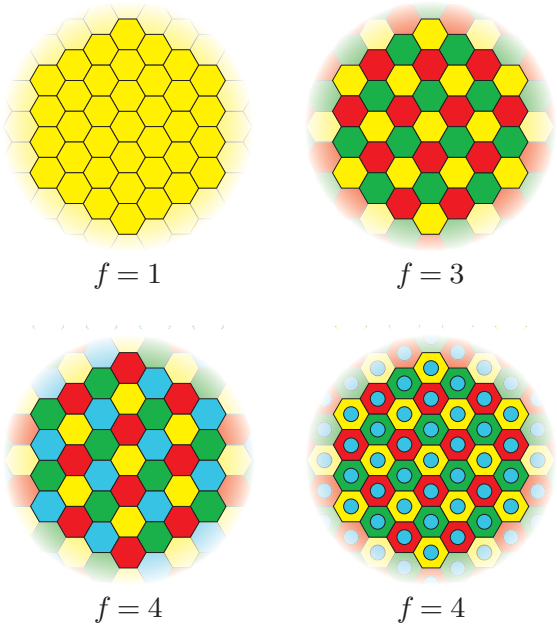


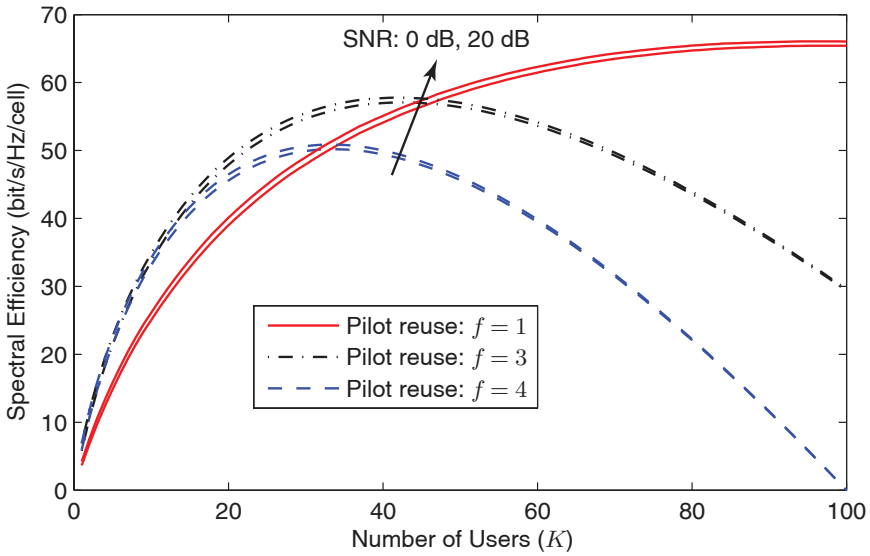
Fig. 7 Illustration of potential symmetric reuse patterns created by three different pilot reuse factors, f, in a cellular network with hexagonal cells. In the lower right case, each cell is divided into two sub-cells with different sets of pilots. If j is the index of a particular cell, then $\mathscr{P}_j$ is the index set of all cells having the same color. Only the cells with the same color use the same pilot sequences, and thereby degrade each other's CSI estimation quality and cause pilot contaminated interference

In summary, power allocation is used in Massive MIMO to distribute the sum SE over the individual users. There are plenty of algorithms that can be used to optimize the power allocation, depending on the utility function that is used in the system.

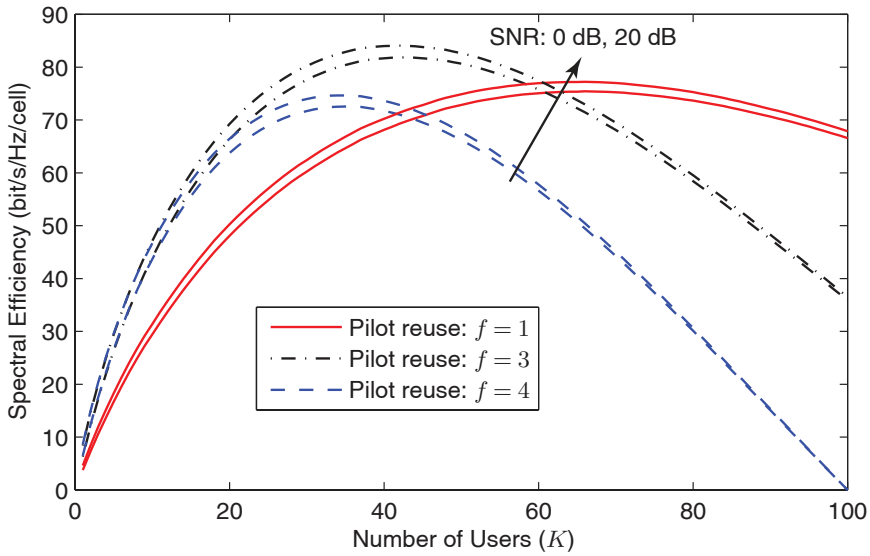
4.2 Non-Universal Pilot Reuse

An important insight from the theoretical analysis in Sect. 3 is that the SE of a particular cell j is influenced by the pilot signaling carried out in other cells. The degradations in CSI estimation quality and pilot contaminated interference are caused only by the interfering cells in $\mathscr{P}_j$ that use the same pilot sequences as cell j. Since the channel attenuation of the interference increases with distance, one would like these interfering cells to be as far away from cell j as possible—and the same is desirable for all cells in $\mathscr{P}_j$.

Recall that the pilot reuse factor $f = \tau_p/K$ was assumed to be an integer in Sect. 3, which leads to a division of the L cells into f disjoint cell groups. The case $f = 1$ is known as universal pilot reuse and $f > 1$ is called non-universal pi-



a. Maximum ratio detection



b. Zero-forcing detection

Fig. 8 Average spectral efficiency, as a function of the number of users, with different processing schemes and pilot reuse factors. Two different SNR levels are considered: $\delta/\sigma_{\text{BS}}^2 = 0$ dB or 20 dB

lot reuse. Since the hexagonal cell topology has a multiple of six cells in each tier of interfering cells, the smallest pilot reuse factors that give rise to symmetric pilot reuse patterns are $f = 1$, $f = 3$, and $f = 4$ [15]. Examples of such reuse patterns are given in Fig. 7, where cells with different colors use different subsets of the pilot sequences. The cells with the same color use exactly the same subset of pilots and therefore cause pilot contamination to each other, while there is no contamination between cells with different colors. If the center cell in Fig. 7 has index j, then $\mathscr{P}_j$ is the set of all cells having the same color. By increasing the pilot reuse factor, there are more colors and therefore fewer interfering cells in each group. We note that with a pilot reuse factor of $f = 4$, one can either divide the cells into four different disjoint groups (as in the lower left example in Fig. 7) or divide each cell into two subcells: cell edge and cell center (as in the lower right example in Fig. 7). The latter is known as *fractional pilot reuse* and can be used to have less frequent pilot reuse at the cell edges than in the cell centers [2], because it is users at the cell edges that are most sensitive to pilot contamination.

To give a concrete example, consider a Massive MIMO scenario with $M = 200$ BS antennas and a coherence interval of $\tau_c = 400$ symbols. The users are assumed to be uniformly distributed in the cell, except for the 10% cell center, and the channels are modeled as uncorrelated Rayleigh fading with a distant-dependent channel attenuation with pathloss exponent 3.7. We consider the pilot reuse factors $f \in \{1, 3, 4\}$, but not the fractional pilot reuse case. Recall from the uplink-downlink duality in Lemma 3 that the same SE is achievable in the uplink and the downlink, thus it is sufficient to study the uplink. We assume a simple power allocation policy

$$p_{j,k} = \frac{\delta}{\beta_{j,k}^{j}} \quad j = 1, \ldots, L, \quad k = 1, \ldots, K, \tag{41}$$

where $\delta \geq 0$ is a design parameter that determines the SNR achieved at each BS antenna: $p_{j,k}\beta_{j,k}^{j}/\sigma_{\text{BS}}^2 = \delta/\sigma_{\text{BS}}^2$.[2] This is called statistical channel inversion power allocation.

Fig. 8 shows the average SE for different number of users, for both MR and ZF detection. The first observation from Fig. 8 is that the two SNR levels, $\delta/\sigma_{\text{BS}}^2 = 0$ dB and 20 dB, give essentially the same performance. This shows that Massive MIMO works equally well at high and low SNRs, since the array gain makes the SE interference-limited and not noise-limited. Next, we notice that different pilot reuse factors are desirable at different user loads (i.e., number of users K). A pilot reuse of $f = 3$ is desired at low load, while $f = 1$ is needed to reduce the prelog factor $(1 - fK/\tau_c)$ when K is large. By selecting f properly, one can always operate on the top curve in Fig. 8 and then Massive MIMO can provide a high SE over a wide range of different number of users. In fact, the technology provides a relatively stable SE for any $K > 10$. This removes the need for intricate scheduling in Massive MIMO networks, because all active users can basically be served simultaneously in every coherence interval (or at least up to $\tau_c/2$ users, to leave half of the coherence

[2] This explicit uplink power allocation policy is very similar to what the uplink MMF utility would give [4], but has the benefit of not requiring any numerical optimization.

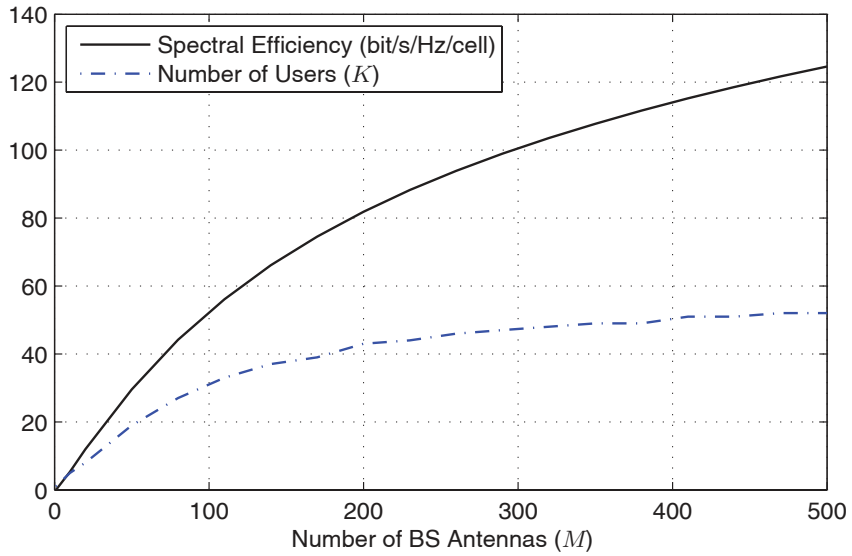


Fig. 9 Average spectral efficiency, as a function of the number of BS antennas, with ZF processing, a pilot reuse factor $f = 3$, and an SNR of 0 dB. The number of users is optimized for each antenna number to yield the highest SE, and the corresponding number of users is also shown

interval for data, which is a number that is typically more than a hundred [7]); the high sum SE is then shared between all the users.

Another observation is that the difference in SE between ZF and MR is relatively small in Fig. 8; ZF only gives a performance gain of between 3% and 45%, depending on the number of users. This should be compared with the single-cell simulation in Fig. 3, where ZF provided more than twice the SEs as MR. The reason for the more modest performance gap is that also ZF suffers from interference in the multi-cell case, since the pilot contamination and many inter-cell interferers make it impossible to cancel all interference.

In summary, the pilot reuse factor is an important design parameter in Massive MIMO networks and the best choice depends on the user load. As shown in [4], it also depends on the propagation environment and the number of BS antennas.

4.3 How High Spectral Efficiency can Massive MIMO Deliver?

We conclude this chapter by showcasing the SE that the Massive MIMO technology can deliver in the uplink and downlink of 5G networks—which is the same due to the uplink-downlink duality. We continue the previous simulation example from Fig. 8, but focus on ZF processing with pilot reuse $f = 3$ and a power allocation policy that gives an SNR of 0 dB to everyone. Note that these design choices are motivated by the previous simulation results.

Fig. 9 shows the SE as a function of the number of BS antennas M. The number of active users is optimized for each M to get the highest SE, and the optimal user numbers are also shown in the figure. A reasonable performance baseline is the IMT-Advanced requirements for 4G networks, provided in [22]. This document specifies spectral efficiencies in the range of 2-3 bit/s/Hz/cell, depending on the simulation scenario. In comparison, the Massive MIMO network considered in Fig. 9 achieves 52 bit/s/Hz/cell using $M = 100$ antennas, which is a $17\times$ to $26\times$ improvement over IMT-Advanced. With $M = 400$ antennas the Massive MIMO system achieves 114 bit/s/Hz/cell, which is an incredible $38\times$ to $57\times$ improvement over IMT-Advanced. These improvements are between one and two orders of magnitude!

It is important to notice that the number of active users increase alongside the SE in Fig. 9. If one divides the top curve with the bottom curve, this gives the SE per user. Interestingly, the SE per user lies in the modest range of 1 to 2.5 bit/s/Hz. Such spectral efficiencies can be achieved in practice using conventional modulation schemes, such as 16-QAM with appropriate channel coding.

In conclusion, Massive MIMO can theoretically provide ten-fold or even 50-fold improvements in SE over IMT-Advanced. These huge improvements are mainly achieved by serving many users simultaneously, while the SE per user is in the conventional range. Huge gains are achieved already at $M = 100$ or fewer BS antennas. These are indeed very encouraging results indicating that the Massive MIMO technology is key to not only improve the SE, but can also be the driving force towards achieving orders of magnitude higher area throughput in 5G technologies.

Acknowledgements The authors of this chapter have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641985 (5Gwireless). The authors are also supported by ELLIIT and CENIIT. We would also like to thank Daniel Verenzuela and Meysam Sadeghi for helping out with the proof-reading.

Appendix

Proof of Lemma 2

Recall that $\boldsymbol{\phi}_{j,k} \in \mathbb{C}^{\tau_p}$ is the pilot sequence used by the kth user in the jth cell, where $\boldsymbol{\Phi}_j = [\boldsymbol{\phi}_{j,1} \ldots \boldsymbol{\phi}_{j,K}]$. Since the desired channel $\mathbf{h}_{j,k}^l$ only appears as $\mathbf{h}_{j,k}^l \boldsymbol{\phi}_{j,k}^H$ in (8), a sufficient statistic for estimating this channel is given by

$$\begin{aligned} \mathbf{Y}_l^{\text{pilot}} \boldsymbol{\phi}_{j,k} &= \sum_{i=1}^{L} \mathbf{H}_l^l \mathbf{P}_i \boldsymbol{\Phi}_i^H \boldsymbol{\phi}_{j,k} + \mathbf{N}_l \boldsymbol{\phi}_{j,k} \\ &= \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}} \tau_p \mathbf{h}_{i,k}^l + \tilde{\mathbf{n}}_{l,j,k} \end{aligned} \tag{42}$$

where $\tilde{\mathbf{n}}_{l,j,k} = \mathbf{N}_l \boldsymbol{\phi}_{j,k} = [\tilde{n}_{l,j,k,1} \ldots \tilde{n}_{l,j,k,M}]^T \sim \mathbb{C}\mathscr{N}(\mathbf{0}, \tau_p \sigma_{\text{UL}}^2 \mathbf{I}_M)$. The second equality follows from the assumed orthogonality of the pilot sequences.

Based on (42), we compute a separate LMMSE estimate of each element of $\mathbf{h}_{j,k}^{l}$. If $y_{l,j,k,m} \in \mathbb{C}$ denotes the mth row of the vector in (42), then

$$y_{l,j,k,m} = \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}} \tau_p h_{i,k,m}^{l} + \tilde{n}_{l,j,k,m}. \tag{43}$$

By the definition of LMMSE estimation [24], the LMMSE estimate of $h_{j,k,m}^{l}$ is given by

$$\hat{h}_{j,k,m}^{l} = \mathbb{E}\left\{h_{j,k,m}^{l}\right\} + \frac{\mathrm{Cov}\left\{h_{j,k,m}^{l}, y_{l,j,k,m}\right\}}{\mathbb{V}\left\{y_{l,j,k,m}\right\}} \left(y_{l,j,k,m} - \mathbb{E}\left\{y_{l,j,k,m}\right\}\right) \tag{44}$$

where we recall that $\mathbb{E}\left\{h_{j,k,m}^{l}\right\} = \bar{h}_{j,k,m}^{l}$ by definition and $\mathrm{Cov}\{\cdot,\cdot\}$ stands for covariance. Moreover, we have

$$\mathbb{E}\left\{y_{l,j,k,m}\right\} = \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}} \tau_p \bar{h}_{i,k,m}^{l}, \tag{45}$$

$$\mathrm{Cov}\left\{h_{j,k,m}^{l}, y_{l,j,k,m}\right\} = \sqrt{p_{j,k}} \tau_p \beta_{j,k}^{l}, \tag{46}$$

$$\mathbb{V}\left\{y_{l,j,k,m}\right\} = \sum_{i \in \mathscr{P}_j} p_{i,k} \tau_p^2 \beta_{i,k}^{l} + \tau_p \sigma_{\mathrm{UL}}^2. \tag{47}$$

The estimation expression in (9) is obtained by substituting (45)–(47) into (44) and writing the result in vector form. The variance of the estimate is then given by

$$\mathbb{V}\left\{\hat{h}_{j,k,m}^{l}\right\} = \frac{\left|\mathrm{Cov}\left\{h_{j,k,m}^{l}, y_{l,j,k,m}\right\}\right|^2}{\mathbb{V}\left\{y_{l,j,k,m}\right\}} = \frac{p_{j,k} \tau_p (\beta_{j,k}^{l})^2}{\sum_{i \in \mathscr{P}_j} p_{i,k} \tau_p \beta_{i,k}^{l} + \sigma_{\mathrm{UL}}^2}, \tag{48}$$

while the estimation error variance in (10) is obtained as $\beta_{j,k}^{l} - \mathbb{V}\left\{\hat{h}_{j,k,m}^{l}\right\}$ since the LMMSE estimate and its error are uncorrelated [24].

Proof of Theorem 1

The ergodic capacity $C_{l,k}^{\mathrm{UL}}$ with linear detection and pilot-based channel estimation is the supremum of the mutual information between the input signal $x_{l,k}$ and the output signal $\mathbf{v}_{l,k}^{H}\mathbf{y}_l$ in (11). The supremum is taken over the distribution of the unit-variance input signal $x_{l,k}$, thus a lower bound is obtained by assuming that $x_{l,k} \sim \mathbb{C}\mathscr{N}(0,1)$. Let $\widehat{\mathscr{H}}$ denote the channel estimates available as side-information at the receiver. We then have that

$$\begin{aligned} C_{l,k}^{\mathrm{UL}} &\geq \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) I(x_{l,k}; \mathbf{v}_{l,k}^{H}\mathbf{y}_l, \widehat{\mathscr{H}}) \\ &= \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \left(h(x_{l,k}) - h(x_{l,k} | \mathbf{v}_{l,k}^{H}\mathbf{y}_l, \widehat{\mathscr{H}})\right) \\ &= \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \left(\log_2(\pi e) - h(x_{l,k} | \mathbf{v}_{l,k}^{H}\mathbf{y}_l, \widehat{\mathscr{H}})\right) \end{aligned} \tag{49}$$

where $I(\cdot;\cdot)$ denotes the mutual information under the suboptimally assumed Gaussian signal distribution and $h(\cdot)$ is the differential entropy function. The first equality follows from the definition of mutual information and the second equality uses the entropy expression for complex Gaussian

random variables. The factor $\gamma^{\mathrm{UL}}(1-\frac{\tau_p}{\tau_c})$ is the fraction of transmission symbols used for uplink data in each coherence interval. It remains to characterize $h(x_{l,k}|\mathbf{v}_{l,k}^H\mathbf{y}_l,\widehat{\mathscr{H}})$, which is done by finding a tractable upper bound on this term:

$$\begin{aligned} h(x_{l,k}|\mathbf{v}_{l,k}^H\mathbf{y}_l,\widehat{\mathscr{H}}) &= h(x_{l,k}-\alpha\mathbf{v}_{l,k}^H\mathbf{y}_l|\mathbf{v}_{l,k}^H\mathbf{y}_l,\widehat{\mathscr{H}}) \\ &\leq h(x_{l,k}-\alpha\mathbf{v}_{l,k}^H\mathbf{y}_l) \\ &\leq \log_2\left(\pi e\mathbb{E}\{|x_{l,k}-\alpha\mathbf{v}_{l,k}^H\mathbf{y}_l|^2\}\right) \end{aligned} \tag{50}$$

where the equality follows from the fact that subtracting a known variable $\alpha\mathbf{v}_{l,k}^H\mathbf{y}_l$, for some deterministic scalar α, does not change the entropy. The first inequality follows from dropping the knowledge of $\mathbf{v}_{l,k}^H\mathbf{y}_l$ and $\widehat{\mathscr{H}}$ which increases the entropy, while the second inequality follows from exploiting the fact that the highest entropy is obtained when $x_{l,k}-\alpha\mathbf{v}_{l,k}^H\mathbf{y}_l$ is a zero-mean complex Gaussian random variable with the same second-order moment as the original variable has.

The last step of the proof is to select α to get the tightest upper bound in (50), which corresponds to the minimization problem

$$\min_{\alpha}\ \mathbb{E}\left\{|x_{l,k}-\alpha\mathbf{v}_{l,k}^H\mathbf{y}_l|^2\right\} = \frac{1}{1+\mathrm{SINR}_{l,k}^{\mathrm{UL}}}, \tag{51}$$

which is solved by first computing the expectation with respect to the independent signals $x_{i,t}$, for all i and t, then finding the optimal α by equating the first derivative of the expression (with respect to α) to zero, and substituting it back into the expression. From (49)–(51) we now have

$$\begin{aligned} C_{l,k}^{\mathrm{UL}} &\geq \gamma^{\mathrm{UL}}\left(1-\frac{\tau_p}{\tau_c}\right)\left(\log_2(\pi e)-\log_2\left(\pi e\frac{1}{1+\mathrm{SINR}_{l,k}^{\mathrm{UL}}}\right)\right) \\ &= \gamma^{\mathrm{UL}}\left(1-\frac{\tau_p}{\tau_c}\right)\log_2\left(1+\mathrm{SINR}_{l,k}^{\mathrm{UL}}\right), \end{aligned} \tag{52}$$

which is the result stated in the theorem.

Proof of Corollary 1

Before computing the SINR expression in (23) for Rayleigh fading channels, we recall that

$$\mathbb{V}\left\{h_{i,t,m}^l\right\} = \beta_{i,t}^l, \tag{53}$$

$$\mathbb{V}\left\{\hat{h}_{i,t,m}^l\right\} = \frac{p_{i,t}\tau_p(\beta_{i,t}^l)^2}{\sum_{i'\in\mathscr{P}_l}p_{i',t}\tau_p\beta_{i',t}^l+\sigma_{\mathrm{UL}}^2}, \tag{54}$$

$$\mathrm{MSE}_{i,t}^l = \beta_{i,t}^l\left(1-\frac{p_{i,t}\tau_p\beta_{i,t}^l}{\sum_{i'\in\mathscr{P}_l}p_{i',t}\tau_p\beta_{i',t}^l+\sigma_{\mathrm{UL}}^2}\right), \tag{55}$$

for the channel between an arbitrary user t in cell i ($i=1,\ldots,L$) and BS l. Note that m is used as an arbitrary antenna index since the channel variance is the same for all antennas. The corollary is first proved in the case of MR detection, where $\mathbf{v}_{l,k}=\hat{\mathbf{h}}_{l,k}^l$, in which case $\mathrm{SINR}_{l,k}^{\mathrm{UL}}$ in Theorem 1 becomes

$$\mathrm{SINR}_{l,k}^{\mathrm{MR,UL}} = \frac{p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{l,k}^{l}\right\}\right|^{2}}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{i,t}^{l}|^{2}\right\} - p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{l,k}^{l}\right\}\right|^{2} + \sigma_{\mathrm{UL}}^{2}\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{2}\right\}}. \quad (56)$$

It remains to compute the expectations in the numerator and denominator of (56). Since $\mathbf{h}_{l,k}^{l} = \hat{\mathbf{h}}_{l,k}^{l} + \mathbf{e}_{l,k}^{l}$, as stated in Lemma 2, the numerator is computed as

$$p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{l,k}^{l}\right\}\right|^{2} = M^{2}p_{l,k}\left(\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\right)^{2}. \quad (57)$$

When computing the denominator, we decompose its first term into three parts based on the pilot reuse; the first two parts contain the cells that use the same pilot sequences as cell l (i.e., all cells in $\mathscr{P}_l$) and the third part contains the remaining cells. We then observe that

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{i,t}^{l}|^{2}\right\} - p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{l,k}^{l}\right\}\right|^{2} + \sigma_{\mathrm{UL}}^{2}\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{2}\right\} \\
&= \sum_{i\in\mathscr{P}_l} p_{i,k}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{i,k}^{l}|^{2}\right\} + \sum_{i\in\mathscr{P}_l}\sum_{\substack{t=1\\t\neq k}}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{i,t}^{l}|^{2}\right\} \\
&\quad + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{i,t}^{l}|^{2}\right\} - p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^{H}\mathbf{h}_{l,k}^{l}\right\}\right|^{2} + \sigma_{\mathrm{UL}}^{2}\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{2}\right\} \quad (58) \\
&= M^{2}\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sum_{i\in\mathscr{P}_l\setminus\{l\}} p_{i,k}\mathbb{V}\left\{\hat{h}_{i,k,m}^{l}\right\} + M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\} \\
&\quad + M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\} + M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sigma_{\mathrm{UL}}^{2}.
\end{aligned}$$

The first term in the second expression of (58) demonstrates the effect of pilot contamination and is computed by using (21) and the independence between the MMSE estimate and its estimation error. Besides, we handle the expectation $\mathbb{E}\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{4}\}$ by virtue of Lemma 2.9 in [40], since $\hat{\mathbf{h}}_{l,k}^{l}(\hat{\mathbf{h}}_{l,k}^{l})^{H}$ is an $M\times M$ central complex Wishart matrix with one degree of freedom:

$$\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{4}\right\} = \mathbb{E}\left\{\mathrm{tr}^{2}\left(\hat{\mathbf{h}}_{l,k}^{l}(\hat{\mathbf{h}}_{l,k}^{l})^{H}\right)\right\} = M(M+1)\left(\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\right)^{2}, \quad (59)$$

where $\mathrm{tr}(\cdot)$ stands for the trace of a matrix. In contrast, the second term of the middle expression of (58) is computed by the fact that the remaining users in $\mathscr{P}_l$ use pilot sequences that are orthogonal to the pilot sequence of user k. The third term in (58) is computed based on the independence between the channel estimates in cell l and the channels in other cells not belong to $\mathscr{P}_l$, while the last term follows from the fact that $\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{2}\right\} = M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}$.

Substituting (57) and (58) into (56), the SINR expression with MR detection becomes

$$\mathrm{SINR}_{l,k}^{\mathrm{MR,UL}} = \frac{Mp_{l,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}}{M\sum\limits_{i\in\mathscr{P}_l\setminus\{l\}} p_{i,k}\mathbb{V}\left\{\hat{h}_{i,k,m}^{l}\right\} + \sum\limits_{i\in\mathscr{P}_l}\sum\limits_{t=1}^{K} p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\} + \sigma_{\mathrm{UL}}^{2}} \quad (60)$$

which equals the expression in the corollary by further substituting (53)–(55) into (60).

In case of ZF detection, the channel inversion structure yields the property

$$\mathbb{E}\left\{\mathbf{v}_{l,k}^{H}\mathbf{h}_{l,k}^{l}\right\} = 1. \quad (61)$$

Additionally, the noise term in (17) is computed as

$$\sigma_{\mathrm{UL}}^2 \mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\} = \sigma_{\mathrm{UL}}^2 \mathbb{E}\left\{\mathrm{tr}\left[\left((\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l\right)^{-1}\right]_{k,k}\right\} = \frac{\sigma_{\mathrm{UL}}^2}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\}} \tag{62}$$

by utilizing the fact that $(\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l$ is a $K \times K$ central complex Wishart matrix with M degrees of freedom and applying Lemma 2.10 in [40] to compute the trace of the inverse. Note that $[\cdot]_{k,k}$ is used here to denote the kth diagonal element of a matrix.

Substituting (61) and (62) into (17), we achieve the SINR expression

$$\mathrm{SINR}_{l,k}^{\mathrm{ZF,UL}} = \frac{p_{l,k}}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l|^2\right\} - p_{l,k} + \frac{\sigma_{\mathrm{UL}}^2}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\}}}. \tag{63}$$

To compute the remaining expectations, we utilize the pilot reuse patterns together with the ZF properties to decompose the expectation term in (63) into three terms:

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l|^2\right\} \\
&= \sum_{i\in\mathscr{P}_l} p_{i,k}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \hat{\mathbf{h}}_{i,k}^l|^2\right\} + \sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \mathbf{e}_{i,t}^l|^2\right\} + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l|^2\right\} \\
&= \sum_{i\in\mathscr{P}_l} \frac{p_{i,k}^2(\beta_{i,k}^l)^2}{p_{l,k}(\beta_{l,k}^l)^2} + \sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K} \frac{p_{i,t}\ \mathrm{MSE}_{\ i,t}^{\ l}}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\}} + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} \frac{p_{i,t}\mathbb{V}\left\{h_{i,t,m}^l\right\}}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\}}.
\end{aligned} \tag{64}$$

In the last equality of (64), the first term is obtained by utilizing the relationship between user channels for cells in $\mathscr{P}_l$ as stated in (21). The second and third terms follow directly from the independence between the ZF detection vector, the estimation errors for channels in $\mathscr{P}_l$ and the complete channels for cells not in $\mathscr{P}_l$. Moreover, Lemma 2.10 in [40] is again used to compute the expectation of the inverse of the central complex Wishart matrix $(\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l$. Substituting (64) back into (63) and utilizing the properties in (53)–(55), the final SINR expression for ZF is obtained.

Proof of Theorem 2

Substituting (28) into (29), the received signal at user k in cell l is

$$\begin{aligned}
y_{l,k} &= \sum_{i=1}^{L}(\mathbf{h}_{l,k}^i)^H \sum_{t=1}^{K}\sqrt{\rho_{i,t}}\mathbf{w}_{i,t}s_{i,t} + n_{l,k} \\
&= \underbrace{\sqrt{\rho_{l,k}}(\mathbf{h}_{l,k}^l)^H \mathbf{w}_{l,k}s_{l,k}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{t=1\\t\neq k}}^{K}\sqrt{\rho_{l,t}}(\mathbf{h}_{l,k}^l)^H \mathbf{w}_{l,t}s_{l,t}}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{i=1\\i\neq l}}^{L}\sum_{t=1}^{K}\sqrt{\rho_{i,t}}(\mathbf{h}_{l,k}^i)^H \mathbf{w}_{i,t}s_{i,t}}_{\text{Inter-cell interference}} + \underbrace{n_{l,k}}_{\text{Noise}}.
\end{aligned} \tag{65}$$

The last row of (65) shows that $s_{l,k}$ is the desired signal that we want to detect, under additive noise, intra-cell, and inter-cell interference. Similar to Theorem 1, if $C_{l,k}^{\mathrm{DL}}$ is the ergodic capacity with linear precoding, then we compute a lower bound on the mutual information between $s_{l,k}$ and $y_{l,k}$ by considering a potentially suboptimal Gaussian signal distribution, $s_{l,k} \sim \mathscr{CN}(0,1)$, and

bounding the corresponding conditional mutual information $I(s_{l,k};y_{l,k})$ as follows:

$$\begin{aligned} C_{l,k}^{\mathrm{DL}} &\geq \gamma^{\mathrm{DL}} \left(1-\frac{\tau_p}{\tau_c}\right) I(s_{l,k};y_{l,k}) \\ &\geq \gamma^{\mathrm{DL}} \left(1-\frac{\tau_p}{\tau_c}\right) \left(\log_2(\pi e) - h(s_{l,k}|y_{l,k})\right) \\ &\geq \log_2\left(1+\mathrm{SINR}_{l,k}^{\mathrm{DL}}\right) \end{aligned} \tag{66}$$

where the inequalities follow from the same procedures as in (50)–(52). The lower bound on the ergodic capacity $R_{l,k}^{\mathrm{DL}}$ in (30) is then obtained. Note that in contrast to the proof of Theorem 1, the receiver does not have any side-information with channel estimates in the downlink.

Proof of Lemma 3

Let $\xi_{l,1},\ldots,\xi_{l,K}$ be the uplink SINRs of the K users in cell l that are achieved by Theorem 1 for the given detection vectors and uplink power coefficients, such that the equations $\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \xi_{l,k}$ hold for $l=1,\ldots,L$ and $k=1,\ldots,K$. From this condition we get

$$\xi_{l,k} \frac{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}{\left|\mathbb{E}\left\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\right\}\right|^2} = \frac{p_{l,k}}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} p_{i,t} \frac{\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}} - p_{l,k} \frac{\left|\mathbb{E}\left\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\right\}\right|^2}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}} + \sigma_{\mathrm{UL}}^2}, \tag{67}$$

by multiplying each side of the equation $\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \xi_{l,k}$ with $\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\} / |\mathbb{E}\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\}|^2$.

The goal is to prove that also $\mathrm{SINR}_{l,k}^{\mathrm{DL}} = \xi_{l,k}$ holds if the downlink precoding vectors in (32) are used and the downlink transmit power coefficients are selected appropriately. According to the definition of the downlink precoding vectors, the equation $\mathrm{SINR}_{l,k}^{\mathrm{DL}} = \xi_{l,k}$ can be written as

$$\xi_{l,k} \frac{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2} = \frac{\rho_{l,k}}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} \rho_{i,t} \frac{\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H \mathbf{v}_{i,t}|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{i,t}\|^2\right\}} - \rho_k \frac{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}} + \sigma_{\mathrm{DL}}^2}, \tag{68}$$

by multiplying each side of the equation with $\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\} / |\mathbb{E}\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\}|^2$.

Let us define a diagonal matrix $\mathbf{D}$ and a matrix $\mathbf{\Psi}$, both of size $KL \times KL$. Let $\mathbf{D}_l \in \mathbb{C}^{K\times K}$ be the lth diagonal block of $\mathbf{D}$ and let $\mathbf{\Psi}_{l,i} \in \mathbb{C}^{K\times K}$ be the (l,i)th block of $\mathbf{\Psi}$. The elements of these blocks are defined as

$$[\mathbf{D}_l]_{k,k} = \frac{\xi_{l,k}\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2}, \tag{69}$$

$$\left[\mathbf{\Psi}_{l,i}\right]_{k,t} = \begin{cases} \frac{\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H \mathbf{v}_{i,t}|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{i,t}\|^2\right\}} - \frac{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}, & \text{for } l=i \text{ and } t=k, \\ \frac{\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H \mathbf{v}_{i,t}|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{i,t}\|^2\right\}}, & \text{otherwise.} \end{cases} \tag{70}$$

Using this notation, the KL equations in (67) and (68) respectively become

$$\frac{\mathbf{p}}{\sigma^2_{\mathrm{UL}}} = \left(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}^T\right)^{-1}\mathbf{D}\mathbf{1}_{KL}, \tag{71}$$

$$\frac{\boldsymbol{\rho}}{\sigma^2_{\mathrm{DL}}} = \left(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}\right)^{-1}\mathbf{D}\mathbf{1}_{KL}, \tag{72}$$

where $\mathbf{p} = [\mathbf{p}_1^T \ \ldots \ \mathbf{p}_L^T]^T$ and $\mathbf{p}_i = [p_{i,1} \ \ldots \ p_{i,K}]^T$ contain the uplink transmit powers, $\boldsymbol{\rho} = [\boldsymbol{\rho}_1^T \ \ldots \ \boldsymbol{\rho}_L^T]^T$ and $\boldsymbol{\rho}_i = [\rho_{i,1} \ \ldots \ \rho_{i,K}]^T$ contain the downlink transmit powers, $\mathbf{1}_{KL}$ is a $KL \times 1$ vector with only ones, and $\mathbf{I}_{KL}$ is the $KL \times KL$ identity matrix. These equations give the uplink and downlink transmit powers that provide the SINRs $\xi_{l,1}, \ldots, \xi_{l,K}$ in cell l, but only if the inverses $(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}^T)^{-1}$ and $(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi})^{-1}$ exist.

Since $\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}^T$ and $\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}$ have the same eigenvalues, either both or none of the inverses exist. Recall that we have selected $\xi_{l,k}$ (for $l = 1, \ldots, L$, $k = 1, \ldots, K$) as the SINRs that were actually achieved in the uplink, thus the inverses must exist and (72) gives the downlink transmit powers that achieves the same SINRs in the downlink as in the uplink. It is also straightforward to show that

$$\frac{\mathbf{1}_{KL}^T\mathbf{p}}{\sigma^2_{\mathrm{UL}}} = \frac{\boldsymbol{\rho}^T\mathbf{1}_{KL}}{\sigma^2_{\mathrm{DL}}}, \tag{73}$$

which corresponds to the relationship between the total transmit power in the uplink and downlink stated in the lemma.

Proof of Corollary 3

The proof follows along the same lines as the proof of Corollary 1, because the same expectations are involved, thus the variances summarized in (53)–(55) are still useful. We briefly summarize the proof of Corollary 3 as follows.

We need to compute all the expectations in (31). MR precoding gives the desired signal power

$$\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H\mathbf{w}_{l,k}\right\}\right|^2 = \frac{\rho_{l,k}}{M\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\}}\left|\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^l\|^2\right\}\right|^2 = M\rho_{l,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\} \tag{74}$$

and the denominator is computed as

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H\mathbf{w}_{i,t}|^2\right\} - \rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H\mathbf{w}_{l,k}^l\right\}\right|^2 + \sigma^2_{\mathrm{DL}} \\
&= \sum_{i\in\mathscr{P}_l}\rho_{i,k}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H\mathbf{w}_{i,k}|^2\right\} + \sum_{i\in\mathscr{P}_l}\sum_{\substack{t=1\\t\neq k}}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H\mathbf{w}_{i,t}|^2\right\} \\
&\quad + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H\mathbf{w}_{i,t}|^2\right\} - \rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H\mathbf{w}_{l,k}^l\right\}\right|^2 + \sigma^2_{\mathrm{DL}} \\
&= M\sum_{i\in\mathscr{P}_l\setminus\{l\}}\rho_{i,k}\mathbb{V}\left\{\hat{h}_{i,k,m}^i\right\} + \sum_{i=1}^{L}\sum_{t=1}^{K}\rho_{i,t}\mathbb{V}\left\{h_{l,k,m}^i\right\} + \sigma^2_{\mathrm{DL}}.
\end{aligned} \tag{75}$$

Substituting (74) and (75) into (31), yields the SINR expression stated for MR in the corollary.

Next, we consider ZF precoding for which we notice that

$$\mathbb{E}\left\{\left\|\hat{\mathbf{H}}_i^i\mathbf{r}_{i,t}\right\|^2\right\} = \mathbb{E}\left\{\left[\left((\hat{\mathbf{H}}_i^i)^H\hat{\mathbf{H}}_i^i\right)^{-1}\right]_{t,t}\right\} = \frac{1}{(M-K)\mathbb{V}\left\{\hat{h}_{i,t,m}^i\right\}} \tag{76}$$

by utilizing the fact that $(\hat{\mathbf{H}}_i^i)^H\hat{\mathbf{H}}_i^i$ is a central complex Wishart matrix with M degrees of freedom. Hence, the ZF precoding vector becomes

$$\mathbf{w}_{i,t} = \sqrt{(M-K)\mathbb{V}\left\{\hat{h}_{i,t,m}^i\right\}}\hat{\mathbf{H}}_i^i\mathbf{r}_{i,t}. \tag{77}$$

Using this precoding vector, we compute the numerator and denominator of (31) as follows:

$$\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H\mathbf{w}_{l,k}^l\right\}\right|^2 = (M-K)\rho_{l,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^l\right\}, \tag{78}$$

$$\begin{aligned}&\sum_{i=1}^{L}\sum_{t=1}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H\mathbf{w}_{i,t}|^2\right\} - \rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H\mathbf{w}_{l,k}\right\}\right|^2 + \sigma_{\mathrm{DL}}^2\\ &= (M-K)\sum_{i\in\mathscr{P}_l\setminus\{l\}}\rho_{i,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^i\right\} + \sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K}\rho_{i,t}\mathrm{MSE}_{l,k}^i + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K}\rho_{i,t}\mathbb{V}\{h_{i,t,m}\} + \sigma_{\mathrm{DL}}^2.\end{aligned} \tag{79}$$

Substituting (78) and (79) into (31), yields the SINR expression stated for ZF in the corollary.

References

1. Anderson, S., Millnert, M., Viberg, M., Wahlberg, B.: An adaptive array for mobile communication systems. IEEE Trans. Veh. Technol. **40**(1), 230–236 (1991)
2. Atzeni, I., Arnau, J., Debbah, M.: Fractional pilot reuse in massive MIMO systems. In: Proc. IEEE ICC (2015)
3. Björnson, E., Hoydis, J., Kountouris, M., Debbah, M.: Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits. IEEE Trans. Inf. Theory **60**(11), 7112–7139 (2014)
4. Björnson, E., Jorswieck, E.: Optimal resource allocation in coordinated multi-cell systems. Foundations and Trends in Communications and Information Theory **9**(2-3), 113–381 (2013)
5. Björnson, E., Jorswieck, E., Debbah, M., Ottersten, B.: Multi-objective signal processing optimization: The way to balance conflicting metrics in 5G systems. IEEE Signal Process. Mag. **31**(6), 14–23 (2014)
6. Björnson, E., Kountouris, M., Bengtsson, M., Ottersten, B.: Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users. IEEE Trans. Signal Process. **61**(13), 3431–3446 (2013)
7. Björnson, E., Larsson, E., Debbah, M.: Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated? IEEE Trans. Wireless Commun. **15**(2), 1293–1308 (2016)
8. Björnson, E., Ottersten, B.: A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance. IEEE Trans. Signal Process. **58**(3), 1807–1820 (2010)
9. Björnson, E., Zheng, G., Bengtsson, M., Ottersten, B.: Robust monotonic optimization framework for multicell MISO systems. IEEE Trans. Signal Process. **60**(5), 2508–2523 (2012)
10. Boche, H., Schubert, M.: A general duality theory for uplink and downlink beamforming. In: Proc. IEEE VTC-Fall, pp. 87–91 (2002)
11. Caire, G., Shamai, S.: On the achievable throughput of a multiantenna Gaussian broadcast channel. IEEE Trans. Inf. Theory **49**(7), 1691–1706 (2003)
12. Cheng, H.V., Björnson, E., Larsson, E.G.: Optimal pilot and payload power control in single-cell massive MIMO systems. IEEE Trans. Signal Process. Submitted
13. Cheng, H.V., Björnson, E., Larsson, E.G.: Uplink pilot and data power control for single cell massive mimo systems with MRC. In: Proc. IEEE ISWCS (2015)

14. Couillet, R., Debbah, M.: Random Matrix Methods for Wireless Communications. Cambridge University Press (2011)
15. Cox, D.: Cochannel interference considerations in frequency reuse small-coverage-area radio systems. IEEE Trans. Commun. **30**(1), 135–142 (1982)
16. Gao, X., Edfors, O., Rusek, F., Tufvesson, F.: Linear pre-coding performance in measured very-large MIMO channels. In: Proc. IEEE VTC Fall (2011)
17. Gesbert, D., Kountouris, M., Heath, R., Chae, C.B., Sälzer, T.: Shifting the MIMO paradigm. IEEE Signal Process. Mag. **24**(5), 36–46 (2007)
18. Goldsmith, A., Jafar, S., Jindal, N., Vishwanath, S.: Capacity limits of MIMO channels. IEEE J. Sel. Areas Commun. **21**(5), 684–702 (2003)
19. Hoydis, J., ten Brink, S., Debbah, M.: Massive MIMO in the UL/DL of cellular networks: How many antennas do we need? IEEE J. Sel. Areas Commun. **31**(2), 160–171 (2013)
20. Hoydis, J., Hoek, C., Wild, T., ten Brink, S.: Channel measurements for large antenna arrays. In: Int. Symp. Wireless Commun. Systems (ISWCS) (2012)
21. Huh, H., Caire, G., Papadopoulos, H., Ramprashad, S.: Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas. IEEE Trans. Wireless Commun. **11**(9), 3226–3239 (2012)
22. ITU: Requirements related to technical performance for imt-advanced radio interface(s). Tech. rep., ITU-R M.2134
23. Jose, J., Ashikhmin, A., Marzetta, T.L., Vishwanath, S.: Pilot contamination and precoding in multi-cell TDD systems. IEEE Trans. Commun. **10**(8), 2640–2651 (2011)
24. Kay, S.: Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice Hall (1993)
25. Larsson, E.G., Tufvesson, F., Edfors, O., Marzetta, T.L.: Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**(2), 186–195 (2014)
26. Li, X., Björnson, E., Larsson, E.G., Zhou, S., Wang, J.: Massive MIMO with multi-cell MMSE processing: Exploiting all pilots for interference suppression. IEEE Trans. Wireless Commun. Submitted, Available: http://arxiv.org/abs/1505.03682
27. Liu, Y.F., Dai, Y.H., Luo, Z.Q.: Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms. IEEE Trans. Signal Process. **59**(3), 1142–1157 (2011)
28. Marzetta, T.L.: Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans. Wireless Commun. **9**(11), 3590–3600 (2010)
29. Marzetta, T.L., Ashikhmin, A.: MIMO system having a plurality of service antennas for data transmission and reception and method thereof. US Patent (2011). 8594215
30. Ngo, H., Larsson, E., Marzetta, T.: Energy and spectral efficiency of very large multiuser MIMO systems. IEEE Trans. Commun. **61**(4), 1436–1449 (2013)
31. Ngo, H., Larsson, E., Marzetta, T.: Aspects of favorable propagation in massive MIMO. In: Proc. EUSIPCO (2014)
32. Ngo, H.Q., Ashikhmin, A.E., Yang, H., Larsson, E.G., Marzetta, T.L.: Cell-free massive MIMO: Uniformly great service for everyone. In: Proc. IEEE SPAWC (2015)
33. Ngo, H.Q., Larsson, E.G., Marzetta, T.L.: Massive MU-MIMO downlink TDD systems with linear precoding and downlink pilots. IEEE Trans. Commun. **61**(4), 1436–1449 (2013)
34. Paulraj, A., Papadias, C.: Space-time processing for wireless communications. IEEE Signal Process. Mag. **14**(6), 49–83 (1997)
35. Qian, L., Zhang, Y., Huang, J.: MAPEL: Achieving global optimality for a non-convex wireless power control problem. IEEE Trans. Wireless Commun. **8**(3), 1553–1563 (2009)
36. Roy, R., Ottersten, B.: Spatial division multiple access wireless communication systems. US Patent (1991). 5515378
37. Shariati, N., Björnson, E., Bengtsson, M., Debbah, M.: Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics. IEEE J. Sel. Topics Signal Process. **8**(5), 815–830 (2014)
38. Swales, S.C., Beach, M.A., Edwards, D.J., McGeehan, J.P.: The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems. IEEE Trans. Veh. Technol. **39**(1), 56–67 (1990)

39. Telatar, E.: Capacity of multi-antenna Gaussian channels. European Trans. Telecom. **10**(6), 585–595 (1999)
40. Tulino, A.M., Verdú, S.: Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory **1**(1), 1–182 (2004)
41. Vieira, J., Malkowsky, S., Nieman, K., Miers, Z., Kundargi, N., Liu, L., Wong, I.C., Öwall, V., Edfors, O., Tufvesson, F.: A flexible 100-antenna testbed for massive MIMO. In: Proc. IEEE Globecom Workshop - Massive MIMO: From Theory to Practice (2014)
42. Viswanath, P., Tse, D.: Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. IEEE Trans. Inf. Theory **49**(8), 1912–1921 (2003)
43. Weingarten, H., Steinberg, Y., Shamai, S.: The capacity region of the Gaussian multiple-input multiple-output broadcast channel. IEEE Trans. Inf. Theory **52**(9), 3936–3964 (2006)
44. Winters, J.: Optimum combining for indoor radio systems with multiple users. IEEE Trans. Commun. **35**(11), 1222–1230 (1987)
45. Yang, H., Marzetta, T.L.: A macro cellular wireless network with uniformly high user throughputs. In: Proc. IEEE VTC-Fall (2014)
46. Yu, W.: Uplink-downlink duality via minimax duality. IEEE Trans. Inf. Theory **52**(2), 361–374 (2006)
47. Zetterberg, P., Ottersten, B.: The spectrum efficiency of a base station antenna array system for spatially selective transmission. IEEE Trans. Veh. Technol. **44**(3), 651–660 (1995)