Chapter 1

INTRODUCTION

The performance limitation of any wireless network will always be at the physical layer, because, fundamentally, the amount of information that can be transferred between two locations is limited by the availability of spectrum, the laws of electromagnetic propagation, and the principles of information theory.

There are three basic ways in which the efficiency of a wireless network may be improved: (i) deploying access points more densely; (ii) using more spectrum; and (iii) increasing the *spectral efficiency*, that is, the number of bits that can be conveyed per second in each unit of bandwidth. While future wireless systems and standards are likely to use an ever-increasing access point density and use new spectral bands, the need for maximizing the spectral efficiency in a given band is never going to vanish.

The use of multiple antennas, also known as *multiple-input, multiple-output* (MIMO) technology, is the only viable approach for substantial improvement of spectral efficiency. While mostly developed during the last two decades, it is noteworthy that a basic idea behind MIMO is almost a century old: in [1], directional beamforming using an antenna array was suggested to permit more aggressive frequency reuse of scarce spectrum – in this case, very low frequency – for transoceanic communication.

MIMO technology is logically classified into one of three categories, whose development occurred during roughly disjoint epochs: Point-to-Point MIMO, Multiuser MIMO, and Massive MIMO. This book is about Massive MIMO, which arguably will be the ultimate embodiment of MIMO technology. The following sections explain these incarnations of MIMO and their important differences. This treatment is intended to be a quick overview, and subsequent chapters will expand upon the concepts introduced here.

1.1 Point-to-Point MIMO

Point-to-Point MIMO emerged in the late 1990s [2–11] and represents the simplest form of MIMO: a base station equipped with an antenna array serves a terminal equipped with an antenna array; see Figure 1.1. Different terminals are orthogonally multiplexed, for example via a combination of time- and frequency-division multiplexing. In what follows, we summarize some basic facts about Point-to-Point MIMO. More details, along with derivations of all formulas given here, are provided in Section C.3.

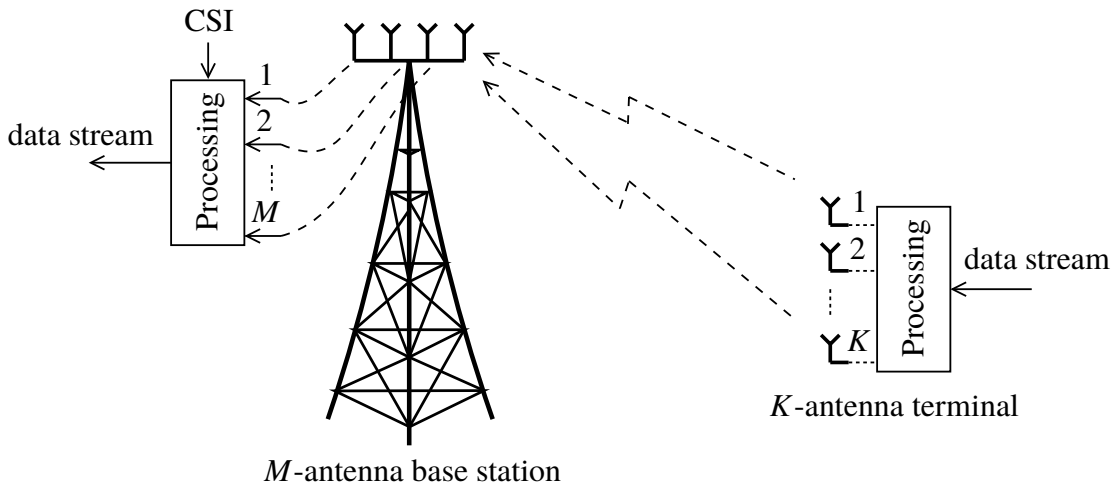
In each channel use, a vector is transmitted and a vector is received. In the presence of additive white Gaussian noise at the receiver, Shannon theory yields the following formulas for the link spectral efficiency (in b/s/Hz):

$$C^{\text{ul}} = \log_2 \left| \boldsymbol{I}_M + \frac{\rho_{\text{ul}}}{K} \boldsymbol{G}\boldsymbol{G}^{\text{H}} \right|, \tag{1.1}$$

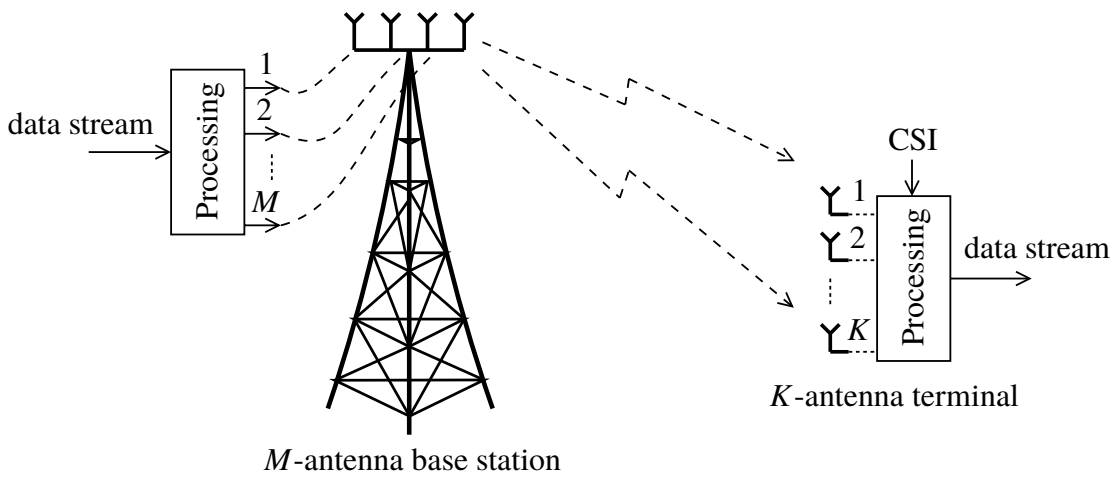
$$\begin{aligned} C^{\text{dl}} &= \log_2 \left| \boldsymbol{I}_K + \frac{\rho_{\text{dl}}}{M} \boldsymbol{G}^{\text{H}}\boldsymbol{G} \right| \\ &\overset{(a)}{=} \log_2 \left| \boldsymbol{I}_M + \frac{\rho_{\text{dl}}}{M} \boldsymbol{G}\boldsymbol{G}^{\text{H}} \right|. \end{aligned} \tag{1.2}$$

In (1.1) and (1.2), $\boldsymbol{G}$ is an $M \times K$ matrix that represents the frequency response of the channel between the base station array and the terminal array; ρ_{ul} and ρ_{dl} are the uplink and downlink signal-to-noise ratios (SNRs), which are proportional to the corresponding total radiated powers; M is the number of base station antennas; and K is the number of terminal antennas. Also, in (a) we used Sylvester's determinant theorem. The normalization by K and M reflects the fact that for constant values of ρ_{ul} and ρ_{dl} total radiated power is independent of the number of antennas. The spectral efficiency values in (1.1) and (1.2) require the receiver to know $\boldsymbol{G}$ but do not require the transmitter to know $\boldsymbol{G}$. Performance can be improved somewhat if the transmitter also acquires channel state information (CSI). However, this requires special effort and is seldom seen in practice – see Section C.3 for the associated capacity formula.

In isotropic (rich) scattering propagation environments, well modeled by independent Rayleigh fading, for sufficiently high SNRs, C^{ul} and C^{dl} scale linearly with $\min(M, K)$ and logarithmically with the SNR. Hence, in theory, the link spectral efficiency can be increased by simultaneously using large arrays at the transmitter and the receiver, that is, making M and K large. In practice, however, three factors seriously limit the usefulness of Point-to-Point MIMO, even with large arrays at both ends of the link. First, the terminal equipment is complicated, requiring independent RF chains per antenna as well as the use of advanced digital processing to separate the data streams. Second, more fundamentally, the propagation environment must support $\min(M, K)$ independent streams. This is often not the case in practice when compact arrays are used. Line-of-sight (LoS) conditions are



(a) Uplink.



(b) Downlink.

Figure 1.1. Point-to-Point MIMO.

particularly stressing. Third, near the cell edge, where normally a majority of the terminals are located and where SNR is typically low because of high path loss, the spectral efficiency scales slowly with $\min(M, K)$. Figure 1.2 illustrates this problem on the downlink for a terminal with $K = 4$ antennas and an SNR of -3 dB.

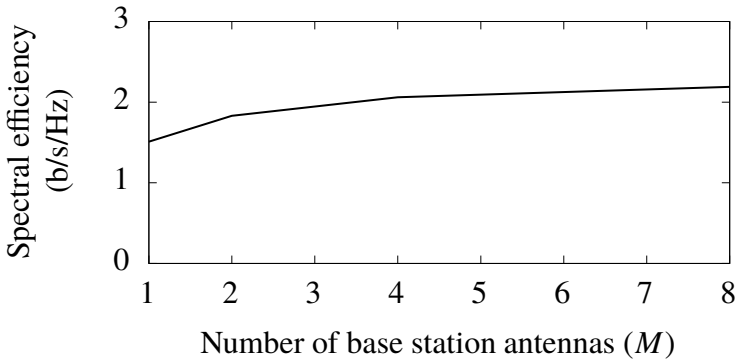


Figure 1.2. Downlink spectral efficiency with Point-to-Point MIMO for a terminal at the cell edge with $K = 4$ antennas, no CSI at the base station, and an SNR of -3 dB.

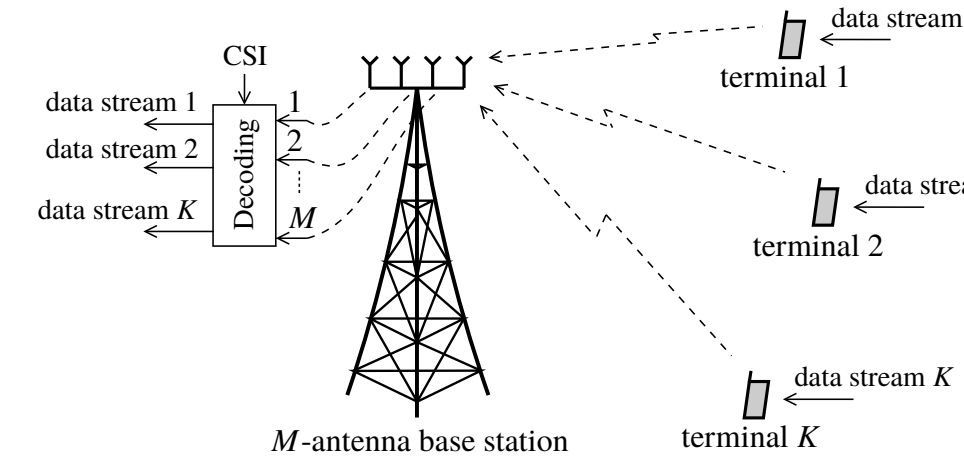
1.2 Multiuser MIMO

The idea of Multiuser MIMO is for a single base station to serve a multiplicity of terminals using the same time-frequency resources; see Figure 1.3. Effectively, the Multiuser MIMO scenario is obtained from the Point-to-Point MIMO setup by breaking up the K-antenna terminal into multiple autonomous terminals. This section summarizes some basic results of Multiuser MIMO. More details, and derivations of all formulas stated here, are given in Section C.4.

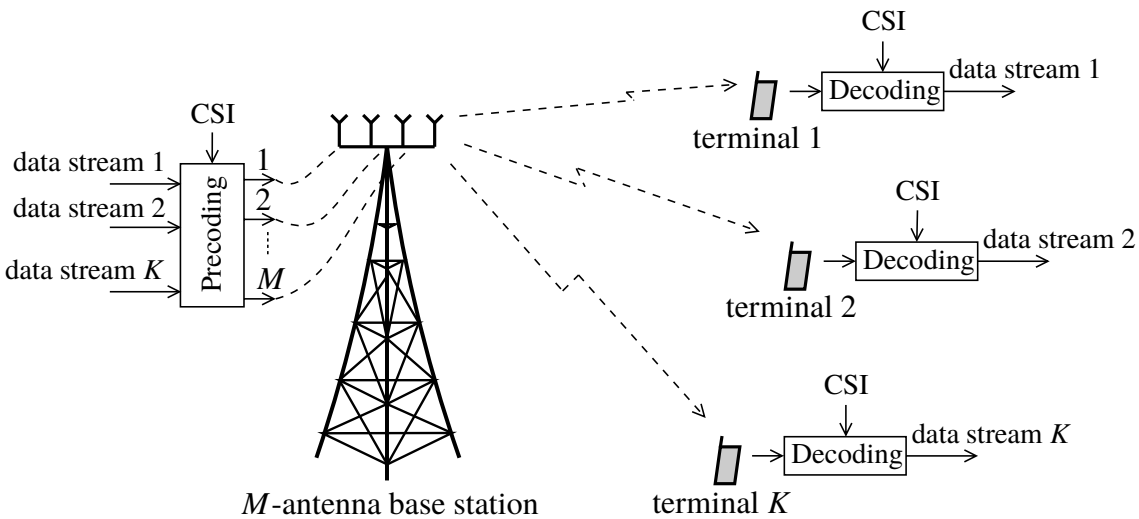
The basic concept of serving several terminals simultaneously using an antenna array at the base station is quite old [12–19]. However, a rigorous information-theoretic understanding of Multiuser MIMO emerged much later [20–23]. The transition in thinking from Point-to-Point MIMO to Multiuser MIMO is explained in some detail in [24].

Our discussion in this section is confined to that particular form of Multiuser MIMO for which there is a comprehensive Shannon theory which provides the ultimate performance of the system and specifies how this performance may be approached arbitrarily closely. It will be convenient to call this *conventional* Multiuser MIMO, even though it is doubtful if such a system has ever been reduced to practice.

Throughout this book, we assume that terminals in Multiuser MIMO have a single antenna. Hence, in the setup in Figure 1.3 the base station serves K terminals. Let $\boldsymbol{G}$ be an $M \times K$ matrix corresponding to the frequency response between the base station array and the K



(a) Uplink.



(b) Downlink.

Figure 1.3. Multiuser MIMO.

terminals. The uplink and downlink sum spectral efficiencies are given by

$$C^{\text{ul}} = \log_2 \left| \boldsymbol{I}_M + \rho_{\text{ul}} \boldsymbol{G}\boldsymbol{G}^{\text{H}} \right|, \tag{1.3}$$

$$C^{\text{dl}} = \max_{\substack{\nu_k \geq 0 \\ \sum_{k=1}^{K} \nu_k \leq 1}} \log_2 \left| \boldsymbol{I}_M + \rho_{\text{dl}} \boldsymbol{G}\boldsymbol{D}_{\boldsymbol{\nu}}\boldsymbol{G}^{\text{H}} \right|, \tag{1.4}$$

where $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_K]^{\text{T}}$, ρ_{ul} is the uplink SNR per terminal, and ρ_{dl} is the downlink SNR. (For given ρ_{ul}, the total uplink power is K times greater than for the Point-to-Point MIMO

model.) The computation of downlink capacity according to (1.4) requires the solution of a convex optimization problem. The possession of CSI is crucial to both (1.3) and (1.4). On uplink, the base station alone must know the channels, and each terminal has to be told its permissible transmission rate separately. On the downlink, both the base station and the terminals must have CSI.

Note that the terminal antennas in the point-to-point case can cooperate, whereas the terminals in the multiuser case cannot. Quite remarkably, however, the inability of the terminals to cooperate in the multiuser system does not compromise the uplink sum spectral efficiency as seen by comparing (1.1) and (1.3). Note also that the downlink capacity (1.4) may exceed the downlink capacity in (1.2) for Point-to-Point MIMO, because (1.4) assumes that the base station knows $\boldsymbol{G}$, where as (1.2) does not.

Multiuser MIMO has two fundamental advantages over Point-to-Point MIMO. First, it is much less sensitive to assumptions about the propagation environment. For example, LoS conditions are stressing for Point-to-Point MIMO, but not for Multiuser MIMO, as explained in Chapter 7. Second, Multiuser MIMO requires only single-antenna terminals. Notwithstanding these virtues, two factors seriously limit the practicality of Multiuser MIMO in its originally conceived form. First, to achieve the spectral efficiencies in (1.3) and (1.4) requires complicated signal processing by both the base station and the terminals. Second, and more seriously, on the downlink both the base station and the terminals must know $\boldsymbol{G}$, which requires substantial resources to be set aside for transmission of pilots in both directions. For these reasons, the original form of Multiuser MIMO is not scalable either with respect to M or to K.

1.3 Massive MIMO

Originally conceived in [25,26], Massive MIMO is a useful and scalable version of Multiuser MIMO. This section introduces the basic Massive MIMO concepts.

Consideration of net spectral efficiency alone according to the rigorous Shannon theory that underlies (1.3) and (1.4) suggests the optimality of a rough parity between M and K in conventional Multiuser MIMO: further growth of M only yields logarithmically increasing throughputs while incurring linearly increasing amounts of time spent on training. Massive MIMO represents a clean break from conventional Multiuser MIMO. Measures are taken such that one operates farther from the Shannon limit, but paradoxically achieves much better performance than any conventional Multiuser MIMO system.

There are three fundamental distinctions between Massive MIMO and conventional Multiuser MIMO. First, only the base station learns $\boldsymbol{G}$. Second, M is typically much

larger than K, although this does not have to be the case. Third, simple linear signal processing is used both on the uplink and on the downlink. These features render Massive MIMO *scalable* with respect to the number of base station antennas, M.

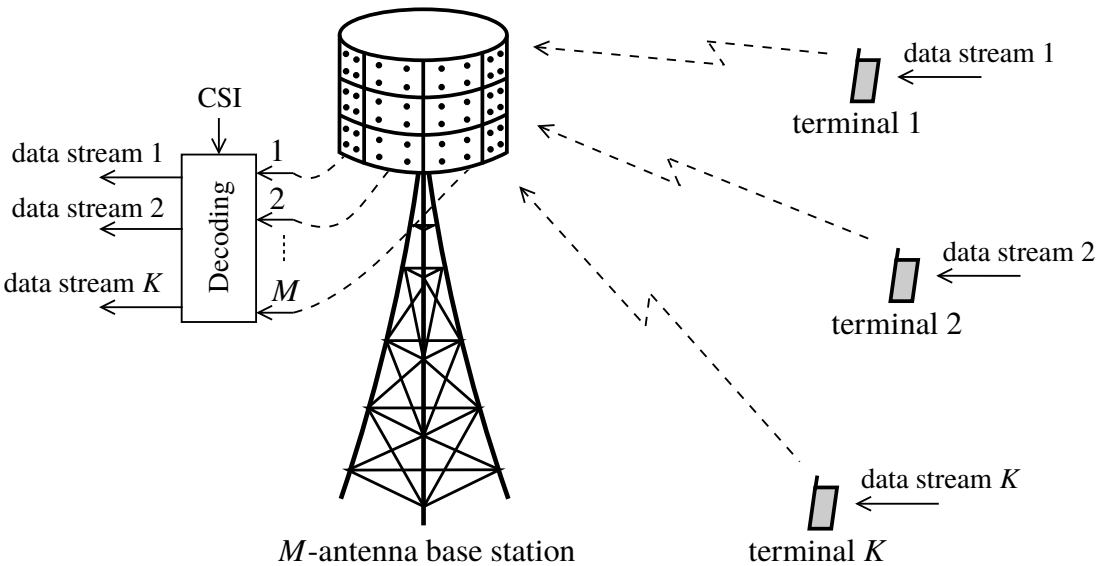
Figure 1.4 illustrates the basic Massive MIMO setup. Each base station is equipped with a large number of antennas, M, and serves a cell with a large number of terminals, K. The terminals typically (and throughout this book) have a single antenna each. Different base stations serve different cells, and with the possible exception of power control and pilot assignment, Massive MIMO uses no cooperation among base stations.

Either in uplink or in downlink transmissions, all terminals occupy the full time-frequency resources concurrently. On uplink, the base station has to recover the individual signals transmitted by the terminals. On the downlink, the base station has to ensure that each terminal receives only the signal intended for it. The base station's multiplexing and de-multiplexing signal processing is made possible by utilizing a large number of antennas and by its possession of CSI.

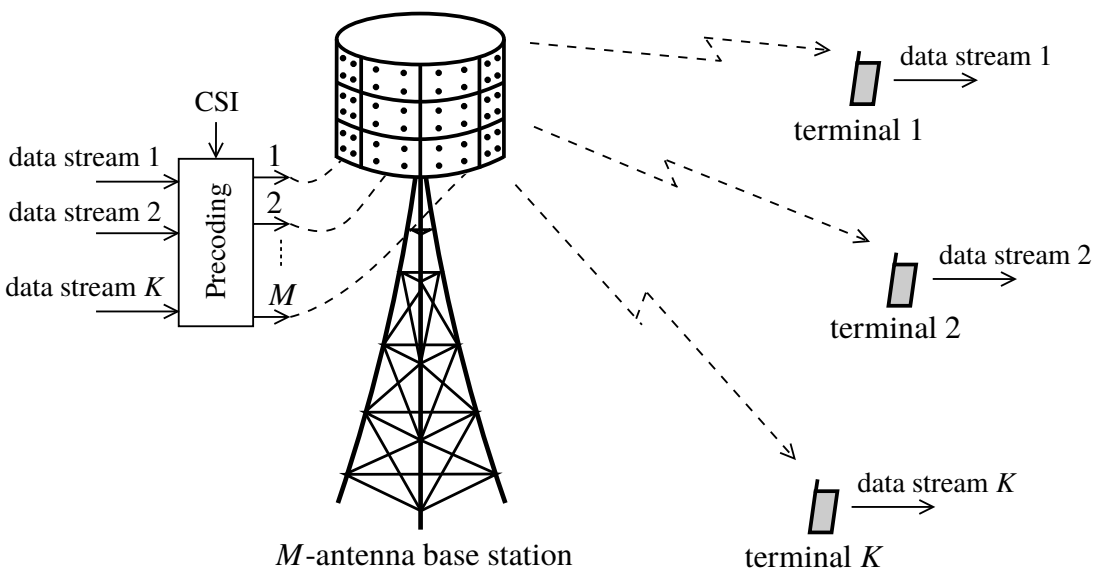
Under LoS propagation conditions, the base station creates, for each terminal, a beam within a narrow angular window centered around the direction to the terminal; see Figure 1.5(a). The more antennas, the narrower are the beams. By contrast, in the presence of local scattering, the signal seen at any given point in space is the superposition of many independently scattered and reflected components that may add up constructively or destructively. When the transmitted waveforms are properly chosen, these components superimpose constructively precisely at the locations of the terminals; see Figure 1.5(b). The more antennas, the more sharply the power focuses onto the terminals. When focusing the power, the use of sufficiently accurate CSI at the base station is essential. In time-division duplex operation (TDD), the base station acquires CSI by measuring pilots transmitted by the terminals, and exploiting reciprocity between the uplink and downlink channel. This requires reciprocity calibration of the transceiver hardware, as discussed in Section 8.7. However, phase-calibrated arrays are not required, since by virtue of the reciprocity a phase offset between any two antennas will affect the uplink and the downlink in the same way.

Increasing the number of antennas, M, always improves performance, in terms of both reduced radiated power and in terms of the number of terminals that can be simultaneously served. In Chapters 3 and 4, we give rigorous lower bounds on Massive MIMO spectral efficiency, and these bounds account for all overhead and imperfections associated with estimating the channels from uplink pilots.

The use of large numbers of antennas at the base station is instrumental not only to obtain high sum spectral efficiencies in a cell, but, more importantly, to provide uniformly good service to many terminals simultaneously. An additional consequence of using large numbers of antennas is that the required signal processing and resource allocation

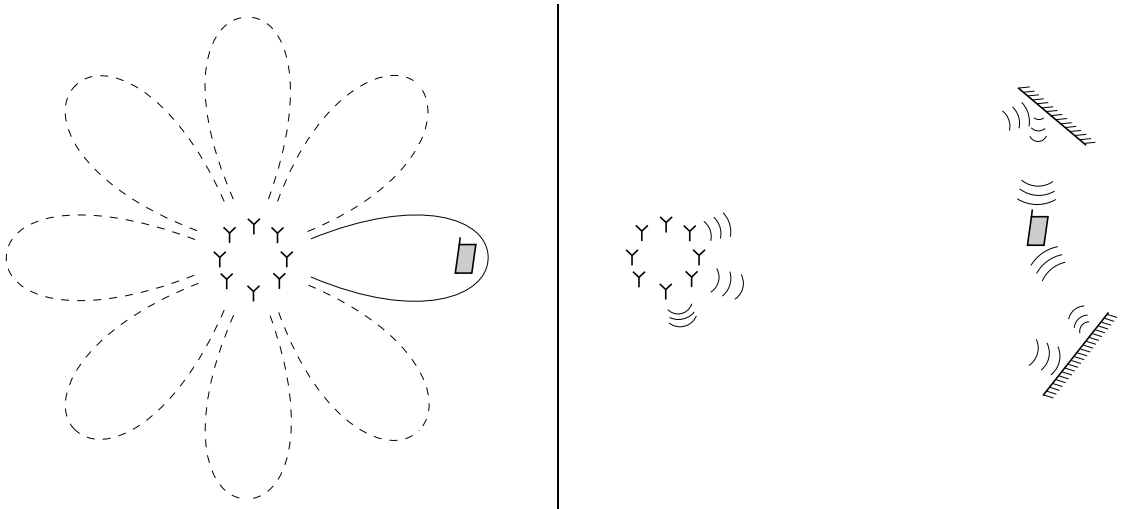


(a) Uplink.



(b) Downlink.

Figure 1.4. Massive MIMO.



(a) LoS propagation conditions.

(b) Propagation environment with scattering.

Figure 1.5. The effect of precoding in different propagation environments.

simplifies, owing to a phenomenon known as *channel hardening*. The significance of channel hardening is that effects of small-scale fading and frequency dependence disappear when M is large. Specifically, consider a terminal with M-dimensional channel response $\boldsymbol{g}$; if beamforming with a beamforming vector $\boldsymbol{a}$ is applied, then the terminal sees a scalar channel with gain $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{g}$. When M is large, by virtue of the law of large numbers, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{g}$ is close to its expected value, $\mathsf{E}\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{g}\right\}$ (a deterministic number). This means that the resulting *effective channel* between each terminal and the base station is a scalar channel with known, frequency-independent gain and additive noise. We show in Chapters 3 and 4 that the capacity of this channel can be rigorously, and without approximations, characterized in terms of an *effective signal-to-interference-plus-noise ratio (SINR)*. Importantly, this characterization does not rely on channel hardening and it is valid for any M and K; however, by virtue of channel hardening, most relevant capacity bounds are tight only when M is reasonably large. This characterization in turn facilitates the use of simple schemes for resource allocation and power control, as further explained in Chapter 5. Furthermore, channel hardening renders channel estimation at the terminals, and the associated transmission of downlink pilots, unnecessary in most cases.

Another benefit of channel hardening in Massive MIMO is that the effective scalar channel seen by each terminal behaves much like an additive white Gaussian noise (AWGN) channel, and hence standard coding and modulation techniques devised for the AWGN channel tend to work well. To illustrate this point, consider the empirical link performance example shown in Figure 1.6. Here, an array with $M = 100$ antennas serves $K = 40$ terminals that transmit simultaneously in the uplink, using QPSK modulation and a rate-1/2 channel code,

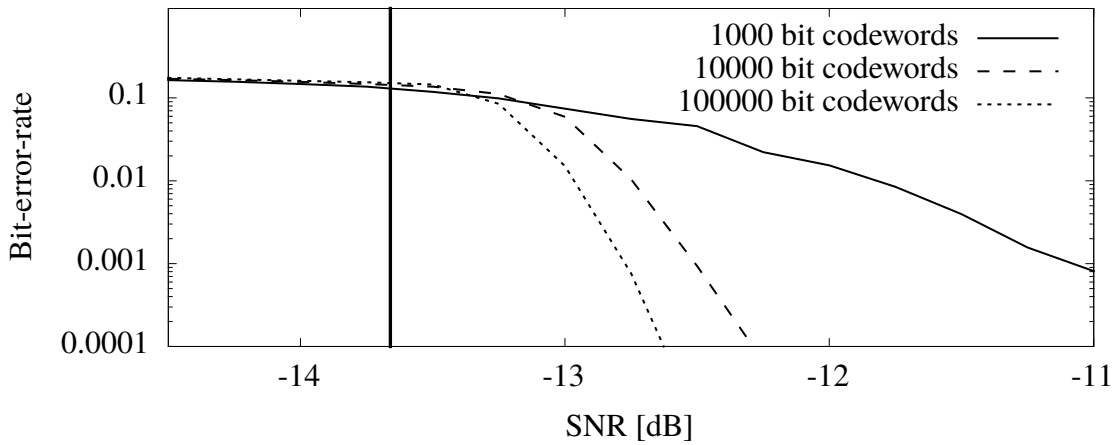


Figure 1.6. Empirical performance of a Massive MIMO uplink with $M = 100$ antennas and $K = 40$ terminals using QPSK modulation with a rate-1/2 low-density parity check code. The vertical solid line represents the SNR threshold obtained from the closed-form lower bound on the spectral efficiency derived in Chapter 3.

with a coherence interval (see Chapter 2 for the exact definition) length of 400 samples. The fading is Rayleigh and independent between the antennas and between the coherence intervals, and there is coding across coherence intervals. All terminals have the same path loss and transmit with the same power, and there is no shadow fading. The channel code is a state-of-the-art low-density parity check code optimized for the AWGN channel [27]. The base station learns the uplink channels through received pilot signals, and each terminal transmits its own orthogonal pilot sequence of length 40. The instantaneous sum spectral efficiency is equal to the number of terminals, K, times the number of bits per symbol, times the code rate: $40 \times 2 \times (1/2) = 40$ b/s/Hz. Assuming, for the sake of argument, that there is only transmission in the uplink, the net sum spectral efficiency is equal to the fraction of the coherence interval spent on payload data transmission multiplied by the instantaneous sum spectral efficiency: $(1 - 40/400) \times 40 = 36$ b/s/Hz. The receiver performs maximum-ratio processing followed by channel decoding. The lower bound on the instantaneous ergodic sum spectral efficiency for this case is $K \log_2(1 + \text{SINR})$ where SINR is the effective SINR given in the upper right corner of Table 3.1 (the same for all terminals). Equating this bound to 40 b/s/Hz, we find that the minimum required SNR is -13.66 dB. For a block length of 100 000 bits, the SNR gap to the bound is about 1 dB. While not shown in Figure 1.6, the smaller the ratio K/M, the closer is the effective channel seen by each terminal to an AWGN channel, and the smaller is the gap to the corresponding capacity bound.