Chapter 2

MODELS AND PRELIMINARIES

This chapter introduces the basic signal and channel models to be used throughout the book. We use standard complex baseband representations of all signals and noise, with the implicit assumption that all signals are eventually modulated onto a carrier with frequency f_c and wavelength $\lambda = c/f_c$, where c is the speed of light. Also, unless stated explicitly, all Gaussian random variables are complex-valued and circularly symmetric; see Appendix A for a treatment of such variables.

2.1 Single-Antenna Transmitter and Single-Antenna Receiver

The wireless channel takes an input signal $x(t)$, emitted by a transmit antenna, and yields an output signal $y(t)$, observed at a receive antenna. The relation between $x(t)$ and $y(t)$ is linear, owing to the linearity of Maxwell's equations. However, this relation generally is time-varying, since the transmitter, receiver, and other objects in the propagation environment may move relative to one another.

2.1.1 Coherence Time

The time during which the channel can be reasonably well viewed as time-invariant is called the *coherence time* and denoted by T_c (measured in seconds). To relate T_c to the characteristics of the physical propagation environment, we consider a simple two-path propagation model where a transmit antenna emits a signal $x(t)$ that reaches the receiver both directly via a LoS path, and via a single specular reflection; see Figure 2.1(a). If both paths have unit strength, and the bandwidth of $x(t)$ is small enough that time-delays can be

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

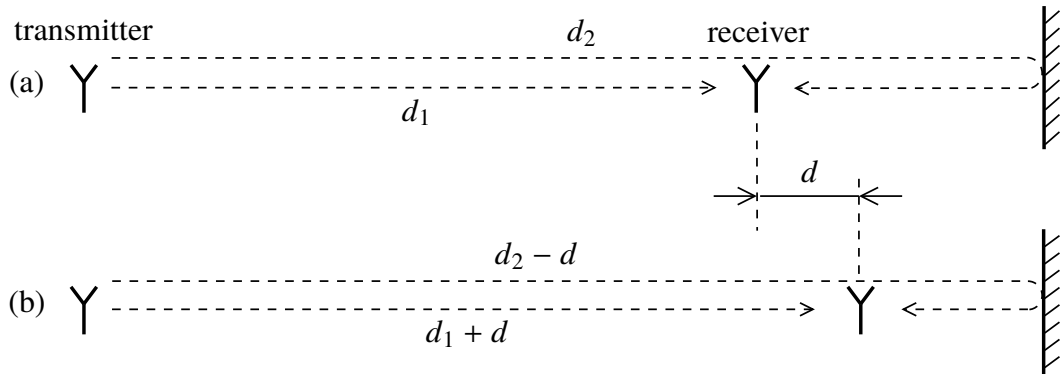


Figure 2.1. Illustration of the two-path propagation model used to motivate the definitions of coherence time and coherence bandwidth.

approximated as phase shifts, then by the superposition principle the received signal is

$$\begin{aligned} y(t) &= \left(e^{-i2\pi f_c \frac{d_1}{c}} + e^{-i2\pi f_c \frac{d_2}{c}}\right) x(t) \\ &= \left(e^{-i2\pi \frac{d_1}{\lambda}} + e^{-i2\pi \frac{d_2}{\lambda}}\right) x(t), \end{aligned} \tag{2.1}$$

where d_1 and d_2 are the propagation path lengths defined in Figure 2.1(a).

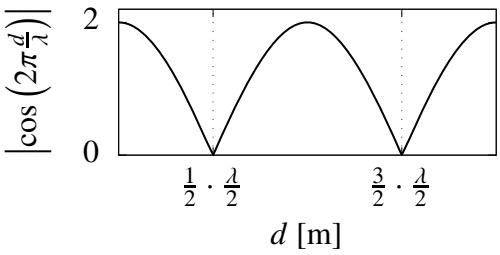
Suppose, for the sake of argument, that when the receiver is located as shown in Figure 2.1(a), d_1/λ and d_2/λ are integers. Then the two paths add up constructively and $y(t) = 2x(t)$. Next, if the receiver is displaced d meters to the right, so that we have the situation in Figure 2.1(b), the received signal will instead be

$$\begin{aligned} y(t) &= \left(e^{-i2\pi \frac{d}{\lambda}} + e^{-i2\pi \frac{-d}{\lambda}}\right) x(t) \\ &= 2\cos\left(2\pi \frac{d}{\lambda}\right) x(t). \end{aligned} \tag{2.2}$$

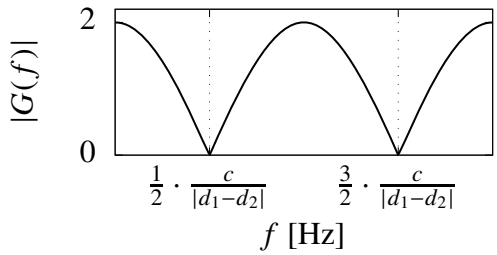
The two paths add up destructively if the cosine in (2.2) is equal to zero. As shown in Figure 2.2(a), this occurs periodically for displacements d that are spaced $\lambda/2$ meters apart. The channel may be considered time-invariant as long as the receiver does not move farther than this distance, $\lambda/2$. This means that if the receiver moves with velocity v meters/second, then the coherence time, T_c, is

$$T_c = \frac{\lambda}{2v} \quad \text{seconds.} \tag{2.3}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press



(a) The coherence time, T_c, is the time it takes to move the distance between two consecutive locations at which the two paths add up destructively, that is, $\lambda/2$ meters.



(b) The coherence bandwidth, B_c, is the frequency separation between two nulls of the frequency response $G(f)$, that is, $c/|d_1 - d_2|$ Hz.

Figure 2.2. Definitions of coherence time and coherence bandwidth for the two-path model in Figure 2.1.

A real propagation environment is considerably more involved than the two-path model of Figure 2.1. It can entail a direct path and a multiplicity of indirect paths via scattering centers of different amplitudes. The overall response is generally complex-valued. Nevertheless, the coherence time as specified by (2.3) is typically a good approximation.

2.1.2 Coherence Bandwidth

Consider now the transmission of a waveform whose time-duration is shorter than the coherence time, T_c. The relation between $x(t)$ and $y(t)$ is then approximately time-invariant, and defined by the channel impulse response $g(t)$ (where $y(t) = \int_{-\infty}^{\infty} d\tau\ g(\tau)x(t-\tau)$) or, equivalently, by the channel frequency response

$$G(f) = \int_{-\infty}^{\infty} dt\ g(t)e^{-i2\pi ft}. \tag{2.4}$$

Generally, the magnitude of the channel frequency response, $|G(f)|$, varies with f. The length of a frequency interval over which $|G(f)|$ is approximately constant is called the *coherence bandwidth* and denoted by B_c (measured in Hz). Consider again the two-path propagation model in Figure 2.1(a), and assume that d_1 and d_2 are fixed and chosen such that d_1/λ and d_2/λ are integers. If a sinusoidal signal, $x(t) = e^{i2\pi ft}$, is transmitted, then the received signal is

$$y(t) = \left(e^{-i2\pi(f_c+f)\frac{d_1}{c}} + e^{-i2\pi(f_c+f)\frac{d_2}{c}}\right)e^{i2\pi ft}. \tag{2.5}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

Hence, the frequency response of the channel is

$$\begin{aligned} G(f) &= e^{-i2\pi(f_c+f)\frac{d_1}{c}} + e^{-i2\pi(f_c+f)\frac{d_2}{c}} \\ &= e^{-i2\pi f\frac{d_1}{c}} + e^{-i2\pi f\frac{d_2}{c}}. \end{aligned} \tag{2.6}$$

The magnitude of the frequency response is

$$\begin{aligned} |G(f)| &= \left| e^{-i2\pi f\frac{d_1}{c}} + e^{-i2\pi f\frac{d_2}{c}} \right| \\ &= 2\left|\cos\left(\pi f\frac{d_1-d_2}{c}\right)\right|, \end{aligned} \tag{2.7}$$

independently of f_c. $|G(f)|$ has zero-crossings at frequencies periodically spaced $c/|d_1 - d_2|$ Hz apart; see Figure 2.2(b). Analogously to the definition of coherence time in (2.3), we define the coherence bandwidth B_c to be the spacing between two nulls of $|G(f)|$, that is

$$B_c = \frac{c}{|d_1 - d_2|} \qquad \text{Hz.} \tag{2.8}$$

While the two-path model represents a simplified description of reality, in practice we expect $|G(f)|$ to be substantially constant over a frequency interval whose length is given by (2.8), where $|d_1 - d_2|$ is the maximum difference in length between different propagation paths from the transmitter to the receiver. As a first-order approximation, $|d_1 - d_2|/c$ is equal to the delay spread of the channel, and $g(t)$ is time-limited to $|d_1 - d_2|/c$ seconds.

2.1.3 Coherence Interval

A time-frequency space of duration T_c seconds and bandwidth B_c Hz is called a *coherence interval*. This is the largest possible time-frequency space within which the effect of the channel reduces to a multiplication by a complex-valued scalar gain g. The magnitude $|g|$ represents the scaling of the waveform envelope and $\arg(g)$ represents the shift in its phase.

According to the sampling theorem, any T-second segment of a waveform $x(t)$ whose energy is substantially contained in a B Hz wide frequency interval can be described in terms of BT (complex-valued) samples taken at intervals of $1/B$ seconds. This means that B_cT_c (complex-valued) samples are required to define a waveform that fits into one coherence interval. We therefore say that a coherence interval has the length

$$\tau_c = B_cT_c \quad \text{samples.} \tag{2.9}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

	Indoors $\lvert d_1 - d_2 \rvert = 30$ meters	Outdoors $\lvert d_1 - d_2 \rvert = 1\,000$ meters
Pedestrian $v = 1.5$ m/s (5.4 km/h)	$B_c = 10$ MHz $T_c = 50$ ms $\tau_c = 500\,000$	$B_c = 300$ kHz $T_c = 50$ ms $\tau_c = 15\,000$
Vehicular $v = 30$ m/s (108 km/h)	N/A	$B_c = 300$ kHz $T_c = 2.5$ ms $\tau_c = 750$

Table 2.1. First-order estimates of the coherence time T_c, coherence bandwidth B_c, and sample length of the coherence interval, τ_c, for some different propagation scenarios, at a carrier frequency of 2 GHz (λ =15 cm).

Consider a waveform $x(t)$, occupying one coherence interval, transmitted over a channel having the same coherence interval. The output of the noisy channel, sampled at rate B_c, takes the form

$$y_n = g x_n + w_n, \quad n = 0, \ldots, \tau_c - 1, \tag{2.10}$$

where x_n is the input, y_n is the output, g represents the channel gain, and $\{w_n\}$ denote samples of additive receiver noise. Throughout the book, we assume that the noise is a stationary random process having flat bandlimited spectral support, $[-B_c, B_c]$. The noise autocorrelation function is proportional to $\mathrm{sinc}(B_c t)$, so noise samples $\{w_n\}$ taken at intervals of $1/B_c$ seconds are uncorrelated.

Some typical values of B_c, T_c, and τ_c, computed using (2.3), (2.8), and (2.9), are shown in Table 2.1 for a carrier frequency of $f_c = 2$ GHz. The values of $|d_1 - d_2|$ depend on the exact characteristics of the propagation environment, and the values in the table are only first-order estimates. An important observation is that outdoors in high mobility, τ_c is only on the order of a few hundred samples.

2.1.4 Interpretation of T_c and B_c in Terms of Nyquist Sampling Rate

It will be convenient in our subsequent analyses to pretend that the channel is static during each coherence interval, as in (2.10). In reality, however, the amplitude and phase of g smoothly evolve between consecutive samples and therefore have some dependence on n. In a practical system, estimates of g acquired from pilots may require interpolation for which

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

the sampling theorem provides a rational basis.

The sampling theorem applies rigorously to a function that is strictly bandlimited to $[-B, B]$ and is sampled at intervals of $1/B$ for all time. Our definitions of coherence time and coherence bandwidth are equivalent to specifications of Nyquist sampling intervals for functions that are substantially bandlimited. Thus, coherence time (2.3) is associated with motion over half of one cycle of a sinewave, while coherence bandwidth (2.8) is equivalent to the reciprocal of the channel delay spread. In practice, one may not be dealing with strictly bandlimited functions (in particular, if there is strong near-field propagation, or reverberation), but there is still ample precedent for invoking the sampling theorem. The concept of the coherence interval (2.9), while exceedingly useful, is somewhat nebulous, and in actual systems the nominal interval may have to be shortened to provide an adequate design margin, especially in case a terminal is served over only a few consecutive coherence intervals, or if residual carrier frequency offsets remain, or to accommodate special applications that require high-accuracy interpolation. In the case studies of Chapter 6, we adopt a factor of two design margin in specifying the coherence interval.

2.1.5 TDD Coherence Interval Structure

As pointed out in Section 1.4, TDD operation is ideal for Massive MIMO because the training burden is independent of the number of base station antennas. Throughout the book, we assume half-duplex TDD so that only one end of the link is transmitting at any one time, either the base station or the terminals. As a consequence, the coherence interval naturally divides into uplink and downlink subintervals, not necessarily of equal duration. Figure 2.3 illustrates two possible configuration, where Figure 2.3(a) includes provision for downlink as well as uplink pilots, and Figure 2.3(b) has only uplink pilots. Not shown are guard intervals between uplink and downlink transmissions.

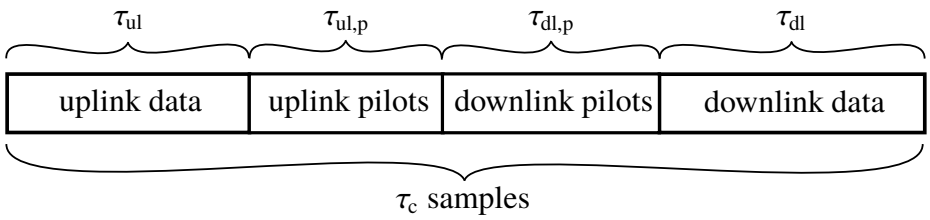
Let τ_{ul} be the number of samples per coherence interval spent on transmission of uplink payload data, $\tau_{\mathrm{ul,p}}$ the number of samples per coherence interval spent on uplink pilots, τ_{dl} the number of samples used for transmission of downlink payload data, and $\tau_{\mathrm{dl,p}}$ the number of samples allocated for downlink pilots. For the Figure 2.3(a) structure,

$$\tau_{\mathrm{ul}} + \tau_{\mathrm{ul,p}} + \tau_{\mathrm{dl,p}} + \tau_{\mathrm{dl}} = \tau_{\mathrm{c}}. \tag{2.11}$$

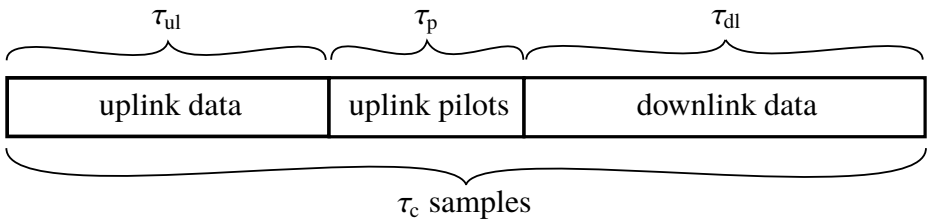
We show later that uplink pilots alone are sufficient to make TDD Massive MIMO work, and for the remainder of the book we assume the coherence interval structure of Figure 2.3(b). For the sake of simplicity, we drop the subscript $(\cdot)_{\mathrm{ul}}$ from the parameter $\tau_{\mathrm{ul,p}}$, and the structural constraint becomes

$$\tau_{\mathrm{ul}} + \tau_{\mathrm{p}} + \tau_{\mathrm{dl}} = \tau_{\mathrm{c}}. \tag{2.12}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press



(a) With downlink pilots.



(b) Without downlink pilots.

Figure 2.3. Allocation of the samples in a coherence interval.

2.1.6 The Coherence Interval in the Context of OFDM Modulation

Orthogonal frequency-division multiplexing (OFDM) is a popular modulation scheme that is fundamentally simple and has numerous attractive properties. The use of OFDM also facilitates a natural interpretation of the coherence interval concept. However, nothing said in this book is specific to OFDM. All spectral efficiency results to be given in subsequent chapters are valid regardless of the particular modulation scheme that is eventually used in an implementation.

OFDM uses the (fast) discrete Fourier transform to decompose a frequency-selective channel into many parallel channels called *subcarriers*; see Figure 2.4. By virtue of the *cyclic prefix* (see below), the effect of the channel on each subcarrier is purely multiplicative, and each subcarrier sees a flat-fading channel. While there are other versions of OFDM, here we treat only the variant that uses a cyclic prefix.

Transmission entails a sequence of OFDM symbols, each symbol consisting of a useful part of length T_u seconds, preceded by a cyclic prefix (also known as guard interval) of T_{cp} seconds. In total, each OFDM symbol is $T_s = T_{cp} + T_u$ seconds long; see Figure 2.5. The useful part carries N_s samples which are obtained by discrete Fourier-transformation of N_s information symbols, and the cyclic prefix replicates the T_{cp} last seconds of the useful part.

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

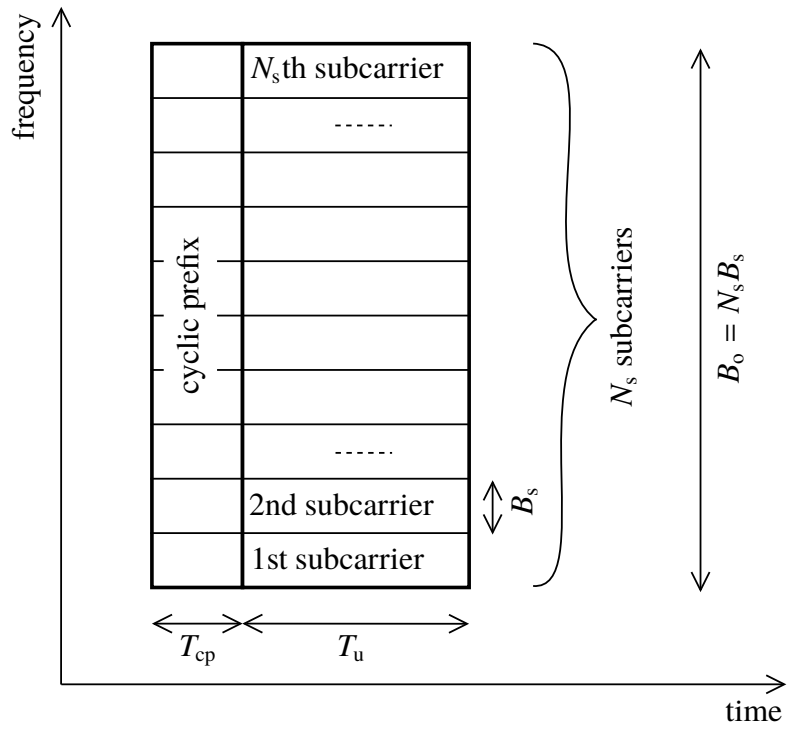


Figure 2.4. Time-frequency domain view of an OFDM symbol, including its cyclic prefix.

The effect of prepending the cyclic prefix is that the linear convolution that represents the effect of the channel impulse response is transformed into a circular convolution, which is equivalent to multiplication in the frequency domain. OFDM renders the original wideband delay-spread channel into many parallel narrowband flat-fading channels. The number of subcarriers is N_s and the frequency separation between neighboring subcarriers is $B_s = 1/T_u$. Hence, the total bandwidth occupied by an OFDM symbol is

$$B_o = N_s B_s = \frac{N_s}{T_u}. \tag{2.13}$$

In order for consecutive OFDM symbols not to interfere, T_{cp} must be at least as large as the channel delay spread.

In practice, several consecutive OFDM symbols are grouped together into one *slot*. We denote the number of OFDM symbols in a slot by N_{slot}, and the duration of one slot by

$$T_{slot} = N_{slot} T_s. \tag{2.14}$$

We assume that $T_{slot} \leq T_c$, so that the channel is time-invariant during one slot. However,

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

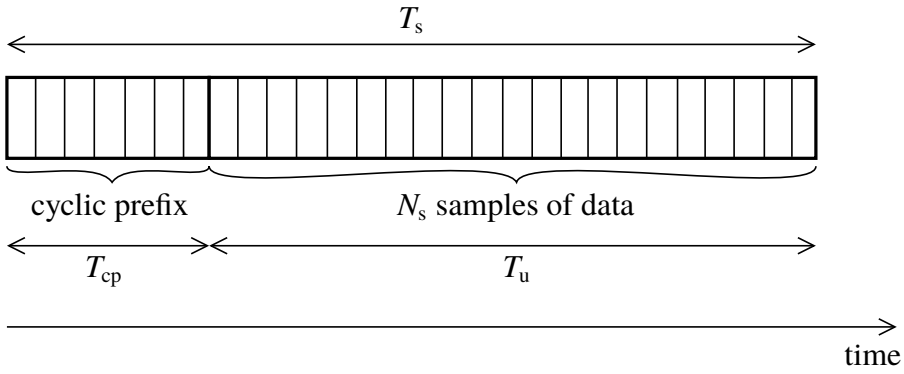


Figure 2.5. Structure of an OFDM symbol, in the time domain.

T_{slot} is not necessarily equal to T_c. In fact, it may be expedient to use a slot that is shorter than the coherence time, say if reduced latency were a consideration.

Normally, the total OFDM symbol bandwidth, B_o, is much greater than the channel coherence bandwidth, B_c, while the subcarrier bandwidth B_s is smaller than B_c. We denote the number of consecutive subcarriers in the frequency domain that fit into one coherence bandwidth by N_{smooth}, assumed to be an integer here. Then,

$$B_c = N_{smooth} B_s. \tag{2.15}$$

The number N_{smooth} represents the number of subcarriers over which the channel frequency response is smooth (approximately constant).

If $T_c = T_{slot}$, a coherence interval consists of N_{smooth} neighboring subcarriers in the frequency domain and N_{slot} consecutive OFDM symbols in the time domain. Each slot consists of

$$\frac{N_s}{N_{smooth}} = \frac{B_o}{B_c} \tag{2.16}$$

coherence intervals that are located adjacent to one another in the frequency domain; see Figure 2.6. The length of a coherence interval, measured in samples is

$$\begin{aligned} B_c T_c &= B_c T_{slot} \\ &= N_{smooth} B_s N_{slot} T_s \\ &= \frac{N_{smooth}}{T_u} N_{slot} T_s \\ &= \frac{T_s}{T_u} N_{smooth} N_{slot}. \end{aligned} \tag{2.17}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

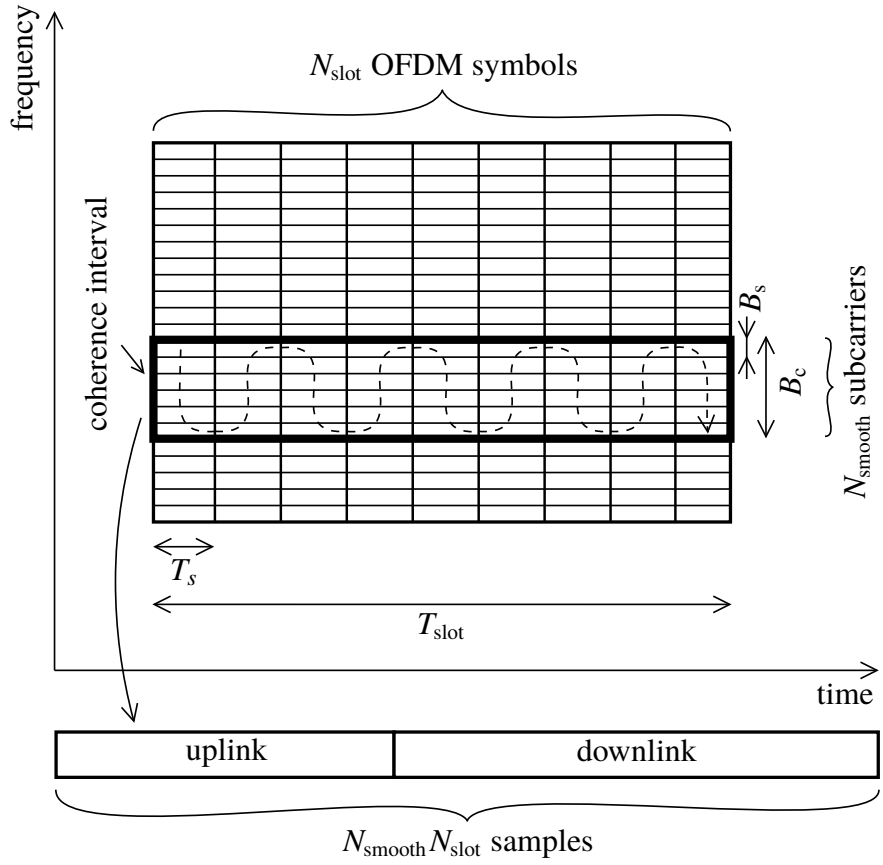


Figure 2.6. A slot comprises N_{slot} consecutive OFDM symbols, each of which contains N_{s} subcarriers. If $T_{\text{c}} = T_{\text{slot}}$, a coherence interval spans N_{slot} OFDM symbols and N_{smooth} subcarriers and each slot contains $N_{\text{s}}/N_{\text{smooth}}$ coherence intervals. The lower part of the figure shows a possible mapping between the time-frequency domain and the samples in a coherence interval.

A fraction $T_{\text{u}}/T_{\text{s}}$ of each coherence interval is useful, and the rest is spent on the cyclic prefix. Thus, the number of useful samples per coherence interval is

$$\frac{T_{\text{u}}}{T_{\text{s}}} B_{\text{c}} T_{\text{c}} = N_{\text{smooth}} N_{\text{slot}}, \tag{2.18}$$

which is equal to τ_{c} as defined in (2.9), up to a factor that reflects the loss of useful samples associated with the cyclic prefix. All samples in a coherence interval are affected by a scaling with a channel gain g as in (2.10). Figure 2.6 also shows a possible mapping between the time-frequency domain and the $N_{\text{smooth}} N_{\text{slot}}$ samples in a coherence interval.

Table 2.2 shows parameters of a sample OFDM system, where the channel delay spread is

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

OFDM symbol duration	T_s	$\frac{1}{14}$ ms
OFDM symbol duration, useful part	T_u	$\frac{1}{15}$ ms
Cyclic prefix duration	T_{cp}	$\frac{1}{14\times15}$ ms
Subcarrier spacing	B_s	15 kHz
Coherence bandwidth	B_c	210 kHz
Number of subcarriers within coherence bandwidth	N_{smooth}	14
Slot duration	T_{slot}	2 ms
Number of OFDM symbols within one slot	N_{slot}	28
Number of useful samples per coherence interval, if $T_c = T_{slot}$	$N_{smooth}N_{slot}$	392

Table 2.2. Parameters of a sample OFDM system.

assumed equal to the duration of the cyclic prefix, T_{cp}.

2.1.7 Small-Scale and Large-Scale Fading

Within a coherence interval, as illustrated in Figure 2.6 (for OFDM), the complex-valued gain between any pair of antennas is substantially constant, and is denoted by the symbol g. It is useful to factor g as follows:

$$g = \sqrt{\beta}h. \tag{2.19}$$

The positive real number, β, called the *large-scale fading* coefficient, embodies range-dependent path loss and shadow fading, it is virtually independent of frequency, and is strongly correlated over many wavelengths of space. The complex-valued number h, representing *small-scale fading*, models range dependent phase shift and constructive and destructive interference among different propagation paths.

In all ensuing analyses, we will assume that the small-scale fading is Rayleigh; that is, $h \sim \mathrm{CN}(0,1)$. The assumption of Rayleigh fading permits the use of Bayesian analysis and it makes ergodic capacity a legitimate performance criterion. Rayleigh fading is also straightforward to justify with simple physical models. For example, in isotropic scattering, h represents the combined effect of many independent propagation paths so by the superposition principle and the central limit theorem, h will be approximately circularly symmetric Gaussian. While all quantitative performance analyses in this book rest on the Rayleigh fading assumption, in Chapter 7 we will argue that under the extreme opposite

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

LoS propagation regime, Massive MIMO is still fully functional. We will assume that large-scale fading coefficients are known a priori to anyone who needs to know them, but that small-scale fading is a priori known to nobody.

2.1.8 Normalized Signal Model and SNR

Henceforth, we will work with the following normalized model for the received signal:

$$y = \sqrt{\rho} g x + w, \tag{2.20}$$

where w is receiver noise and ρ is a dimensionless constant that scales the transmitted signal. Throughout, we adopt the convention that each transmitted signal x has zero mean and satisfies a unit power constraint, $\mathsf{E}\{x\} = 0$ and $\mathsf{E}\left\{|x|^2\right\} \leq 1$. We also assume that the noise is circularly symmetric Gaussian with unit variance, denoted $w \sim \mathrm{CN}(0, 1)$, and is independent of x. This gives ρ the interpretation of a *signal-to-noise ratio* (SNR) in the following sense: if the median of β equals unity, and the transmitter expends its maximum permitted power, then ρ is the median SNR measured at the receiver.

We denote the SNR associated with the uplink and downlink by ρ_{ul} respectively ρ_{dl}. Hence, on the uplink,

$$y = \sqrt{\rho_{\mathrm{ul}}} g x + w, \tag{2.21}$$

and on the downlink,

$$y = \sqrt{\rho_{\mathrm{dl}}} g x + w, \tag{2.22}$$

where in both cases, x is the transmitted signal, y is the received signal, and w represents noise. The uplink and downlink SNRs are different in general, owing to differences in the transmit powers and the noise figures at the base station and the terminal.

2.2 Multiple Base Station Antennas and Multiple Terminals

We now consider cellular Massive MIMO whereby base stations, each equipped with an array of M antennas, serve simultaneously a multiplicity of terminals in their designated areas via spatial multiplexing. We introduce the signal models and assumptions that underlie the spectral efficiency analysis in the ensuing chapters.

We confine the discussion entirely to the case when the terminals have a single antenna each. The case of multiple-antenna terminals can be handled, for example, by treating each terminal antenna as a separate terminal.

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

2.2.1 Single-Cell System

We first consider the case of a single base station that simultaneously serves K terminals. We call the area where the terminals are located a *cell*, and refer to the corresponding scenario as *single-cell*, emphasizing the fact that there is no inter-cell interference to account for.

Let g_k^m be the channel gain between the kth terminal and the mth base station antenna; see Figure 2.7. We will assume that the base station antennas are configured in a compact array, so that the paths between a given terminal and all base station antennas are affected by the same large-scale fading, but by different small-scale fading. Hence,

$$g_k^m = \sqrt{\beta_k} h_k^m, \qquad k = 1, \ldots, K, \qquad m = 1, \ldots, M, \tag{2.23}$$

where β_k is a large-scale fading coefficient that depends on k but not on m, and h_k^m represents the effect of small-scale fading. We let $\boldsymbol{G}$ be a matrix that comprises the channel gains between all terminals and all base station antennas,

$$\boldsymbol{G} = \begin{bmatrix} g_1^1 & \cdots & g_K^1 \\ \vdots & \ddots & \vdots \\ g_1^M & \cdots & g_K^M \end{bmatrix}. \tag{2.24}$$

Throughout all performance analyses, we will assume that the small-scale fading is Rayleigh and *independent* between the antennas and the terminals, so that $\{h_k^m\}$ are i.i.d. $\mathrm{CN}(0, 1)$ random variables. We discuss this assumption further in Chapter 7.

Uplink

Consider the uplink. If the terminals simultaneously transmit the K signals $x_1, \ldots, x_K$, then the mth base station antenna receives the signal,

$$y_m = \sqrt{\rho_{\mathrm{ul}}} \sum_{k=1}^{K} g_k^m x_k + w_m, \tag{2.25}$$

where w_m is receiver noise. As before, we assume that $w_m \sim \mathrm{CN}(0, 1)$. Additionally, we will assume that the noise is uncorrelated across the antennas; that is, $\{w_m\}$ are independent. The transmit powers of the terminals are individually constrained,

$$\mathsf{E}\left\{|x_k|^2\right\} \leq 1, \tag{2.26}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

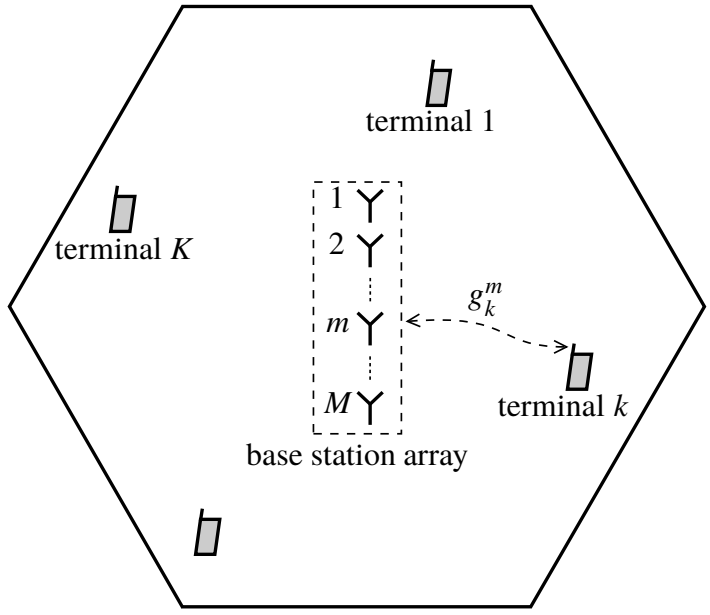


Figure 2.7. Single-cell propagation model.

as in Section 2.1.8, and the transmitted signals have zero mean: $\mathsf{E}\{x_k\} = 0$. Collectively, according to (2.25) the M antennas receive a vector $\boldsymbol{y} = [y_1, \ldots, y_M]^{\mathrm{T}}$,

$$\begin{aligned} \boldsymbol{y} &= \sqrt{\rho_{\mathrm{ul}}} \sum_{k=1}^{K} \boldsymbol{g}_k x_k + \boldsymbol{w} \\ &= \sqrt{\rho_{\mathrm{ul}}} \boldsymbol{G}\boldsymbol{x} + \boldsymbol{w}, \end{aligned} \tag{2.27}$$

where $\boldsymbol{g}_k$ is the kth column of $\boldsymbol{G}$, $\boldsymbol{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$ and $\boldsymbol{w} = [w_1, \ldots, w_M]^{\mathrm{T}}$.

Downlink

On the downlink, the M base station antennas transmit the M-vector $\boldsymbol{x}$, and via reciprocity, the kth terminal receives

$$y_k = \sqrt{\rho_{\mathrm{dl}}} \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{x} + w_k, \tag{2.28}$$

where w_k is noise. In vector form,

$$\boldsymbol{y} = \sqrt{\rho_{\mathrm{dl}}} \boldsymbol{G}^{\mathrm{T}} \boldsymbol{x} + \boldsymbol{w}, \tag{2.29}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

where $\boldsymbol{y} = [y_1, \ldots, y_K]^{\mathrm{T}}$ and $\boldsymbol{w} \triangleq [w_1, \ldots, w_K]^{\mathrm{T}}$. As before, we assume that the noise samples $\{w_k\}$ are i.i.d. CN(0, 1). Analogously to the single-antenna case in Section 2.1.8, we assume that $\boldsymbol{x}$ is normalized such that

$$\mathsf{E}\left\{\|\boldsymbol{x}\|^2\right\} \leq 1. \tag{2.30}$$

This normalization corresponds to enforcing a long-term constraint on the sum of the radiated power from all antennas. While this assumption is analytically convenient, it is not the only possibility. For example, one could alternatively consider power constraints for each antenna individually.

Signal-to-Noise Ratio

As in the single-antenna case, the quantities ρ_{ul} and ρ_{dl} have interpretations in terms of SNR. In the current context, on the uplink, if the median of β_k is unity for a given terminal and the terminal transmits with its maximum permitted power, then ρ_{ul} is the median SNR for that terminal, measured at any of the base station antennas. On the downlink, if the total permitted power were radiated through only one transmit antenna, say the first one, such that $\mathsf{E}\left\{|x_1|^2\right\} = 1$ and $x_2 = \cdots = x_M = 0$, and if additionally the median of β_k were equal to unity, then the median SNR measured at the kth terminal would be ρ_{dl}.

2.2.2 Multi-Cell System

Next we consider a *multi-cell* scenario. Here multiple base stations coexist, with some geographical separation, and each base station serves terminals in its associated cell. The antennas at each base station work coherently together, but different base stations do not cooperate. Generally, the carrier frequency used in a particular cell is reused in other cells, and inter-cell interference then results.

Throughout, we assume that there are K terminals in each cell. This assumption is made only for simplicity and, in reality, there may, of course, be a varying number of terminals in each cell. We will also assume synchronized operation, such that, at any given point in time, either all base stations simultaneously transmit or all terminals simultaneously transmit. This assumption is not strictly necessary, and it does not necessarily result in optimal system performance, but it is convenient to make for purposes of analysis.

Uplink

Consider first the uplink. The signal received at the mth base station antenna in the lth cell, denoted by y_{lm}, is a superposition of signals transmitted from the K terminals in the

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

lth cell, and the $K(L-1)$ terminals in all interfering cells $l' = 1, \ldots, l-1, l+1, \ldots, L$. Mathematically,

$$y_{lm} = \sqrt{\rho_{\text{ul}}} \sum_{k=1}^{K} g_{lk}^{lm} x_{lk} + \sqrt{\rho_{\text{ul}}} \sum_{\substack{l'=1 \\ l' \neq l}}^{L} \sum_{k=1}^{K} g_{l'k}^{lm} x_{l'k} + w_{lm}, \tag{2.31}$$

where $x_{l'k}$ is the signal transmitted by the kth terminal in the l'th cell and $g_{l'k}^{lm}$ is the channel gain from the kth terminal in the l'th cell to the mth base station antenna in the lth cell; see Figure 2.8. The last term in (2.31), w_{lm}, represents additive receiver noise, which we assume is CN(0, 1) and independent among different m and l. In vector form,

$$\boldsymbol{y}_l = \sqrt{\rho_{\text{ul}}} \boldsymbol{G}_l^l \boldsymbol{x}_l + \sqrt{\rho_{\text{ul}}} \sum_{\substack{l'=1 \\ l' \neq l}}^{L} \boldsymbol{G}_{l'}^l \boldsymbol{x}_{l'} + \boldsymbol{w}_l, \tag{2.32}$$

where $\boldsymbol{y}_l = [y_{l1}, \ldots, y_{lM}]^{\text{T}}$, $\boldsymbol{w}_l = [w_{l1}, \ldots, w_{lM}]^{\text{T}}$,

$$\boldsymbol{G}_{l'}^l = \begin{bmatrix} g_{l'1}^{l1} & \cdots & g_{l'K}^{l1} \\ \vdots & \ddots & \vdots \\ g_{l'1}^{lM} & \cdots & g_{l'K}^{lM} \end{bmatrix} \tag{2.33}$$

is an $M \times K$ matrix that contains all channel gains from the terminals in the l'th cell to the base station array in the lth cell, and the K-vector $\boldsymbol{x}_{l'} = [x_{l'1}, \ldots, x_{l'K}]^{\text{T}}$ contains the signals transmitted by the terminals in the l'th cell. In (2.33), analogous to (2.23),

$$g_{l'k}^{lm} = \sqrt{\beta_{l'k}^l} h_{l'k}^{lm}, \tag{2.34}$$

where $\beta_{l'k}^l$ models the large-scale fading associated with propagation from the kth terminal in the l'th cell to the base station array in the lth cell, and $h_{l'k}^{lm}$ models small-scale fading.

Downlink

Next, consider the downlink. Under the assumption of reciprocity, the kth terminal in the lth cell receives

$$y_{lk} = \sqrt{\rho_{\text{dl}}} \boldsymbol{g}_{lk}^{l\text{T}} \boldsymbol{x}_l + \sqrt{\rho_{\text{dl}}} \sum_{\substack{l'=1 \\ l' \neq l}}^{L} \boldsymbol{g}_{lk}^{l'\text{T}} \boldsymbol{x}_{l'} + w_{lk}, \tag{2.35}$$

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

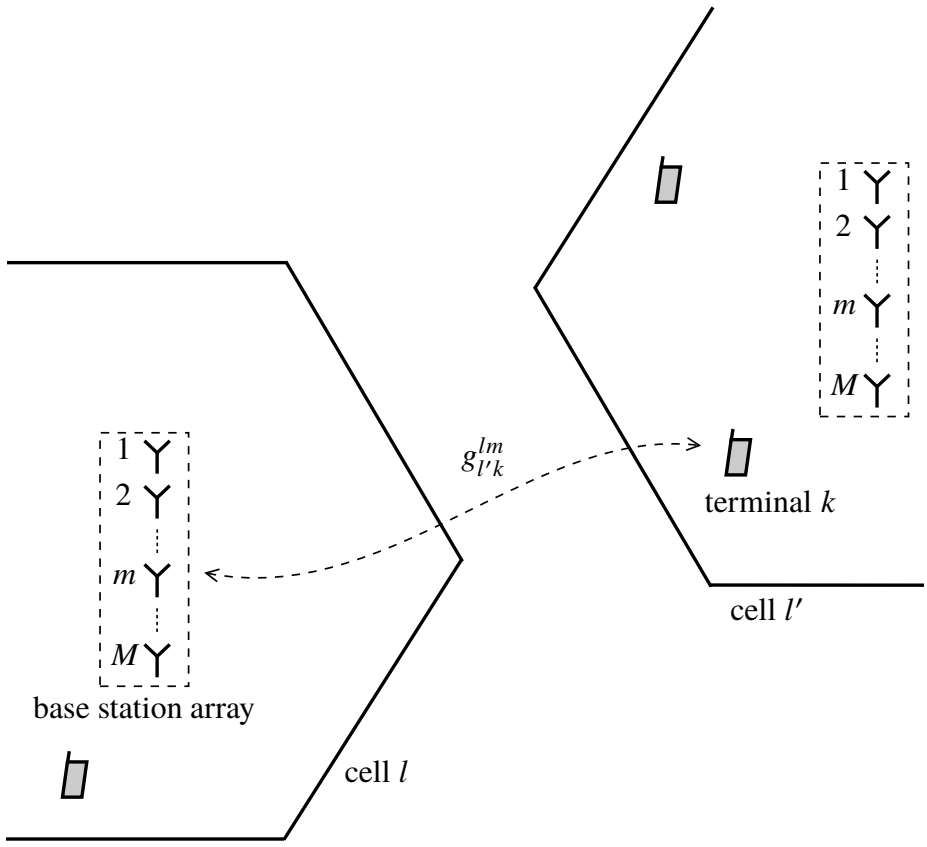


Figure 2.8. Multi-cell propagation model.

where $\boldsymbol{g}_{lk}^{l'}$ is the kth column of $\boldsymbol{G}_l^{l'}$, $\boldsymbol{x}_{l'}$ represents the M-vector transmitted by the array in the l'th cell, and w_{lk} is noise with distribution CN(0, 1). Collectively, the K terminals in the lth cell will receive the K-vector,

$$\boldsymbol{y}_l = \sqrt{\rho_{\mathrm{dl}}}\boldsymbol{G}_l^{l\mathrm{T}}\boldsymbol{x}_l + \sqrt{\rho_{\mathrm{dl}}}\sum_{\substack{l'=1\\ l'\neq l}}^{L}\boldsymbol{G}_l^{l'\mathrm{T}}\boldsymbol{x}_{l'} + \boldsymbol{w}_l, \tag{2.36}$$

where $\boldsymbol{y}_l = [y_{l1}, \ldots, y_{lK}]^{\mathrm{T}}$ and $\boldsymbol{w}_l = [w_{l1}, \ldots, w_{lK}]^{\mathrm{T}}$ is a K-vector of noise with i.i.d. CN(0, 1) elements.

Similarly to the single-cell case, we will assume that the small-scale fading is Rayleigh and independent between all antennas and all terminals, so that $\{h_{l'k}^{lm}\}$ are i.i.d. CN(0, 1) random variables.

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

2.3 Capacity Bounds as Performance Metric

In Massive MIMO, after appropriate signal processing, the effective channel associated with each of the terminals is a *scalar point-to-point channel*. Each time this channel is used, it takes a (complex-valued) scalar input symbol x and delivers a (complex-valued) output signal y. The action of the channel is characterized by the conditional probability distribution of y given x.

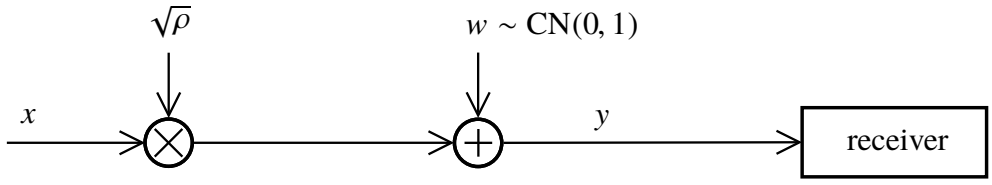
To communicate a message over a point-to-point scalar channel, the transmitter maps the message onto a sequence of symbols $\{x_n\}$, and the receiver recovers the message from the sequence of samples $\{y_n\}$. The effective number of bits conveyed per transmitted symbol, denoted by R, is called the *rate* and is measured in bits per channel use (bpcu). Recall that a waveform contained in a time-frequency space of bandwidth B Hz and time-duration T seconds can be described by BT samples; see Section 2.1.3. Hence, transmitting a waveform with bandwidth B Hz and time-duration T seconds is equivalent to transmitting BT symbols $\{x_n\}$. Therefore, the rate R is usually termed *spectral efficiency* and measured in bits per second per Hertz (b/s/Hz).

According to Shannon's noisy channel coding theorem, there exists a quantity C (unit: bpcu) called the channel *capacity*, which determines a rate R at which error-free communication is possible, asymptotically, when coding over many transmitted symbols. More precisely, the noisy channel coding theorem states that for any given probability of error ϵ, and any given "gap from capacity" ζ, there exist a block length N, and a coding scheme that achieves the rate $R = C - \zeta$ with a probability of a decoding error less than ϵ. Generally, achieving rates R that are close to C requires that N be large, and in the limit when ζ is forced towards zero, the required value of N tends to infinity.

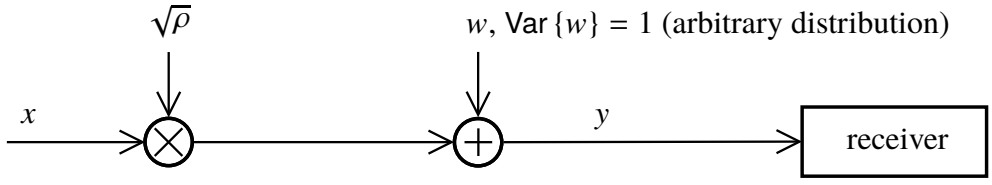
For several channels, exact expressions for the capacity are known. In many cases, however, only bounds on capacity, also known as *achievable rates*, are available. Throughout this book, motivated by the channel coding theorem, we will use such capacity bounds as the primary performance metric. As illustrated in Section 1.3, these capacity bounds can typically be approached closely in practice by using state-of-the-art channel coding techniques.

In what follows, we present some key results on capacity and capacity bounds for point-to-point scalar channels that will be needed for the performance analysis in Chapters 3 and 4. Section C.2 contains comprehensive derivations of these results.

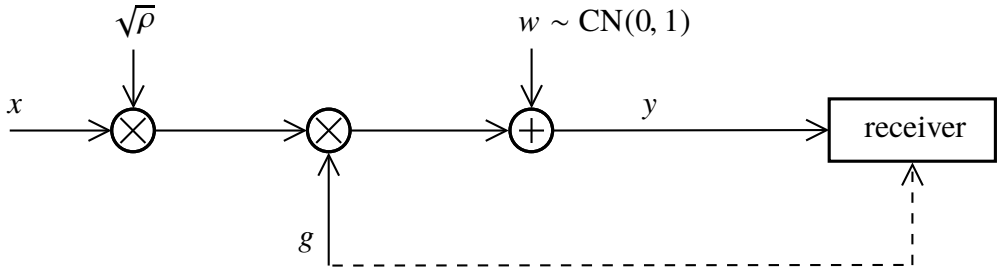
https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press



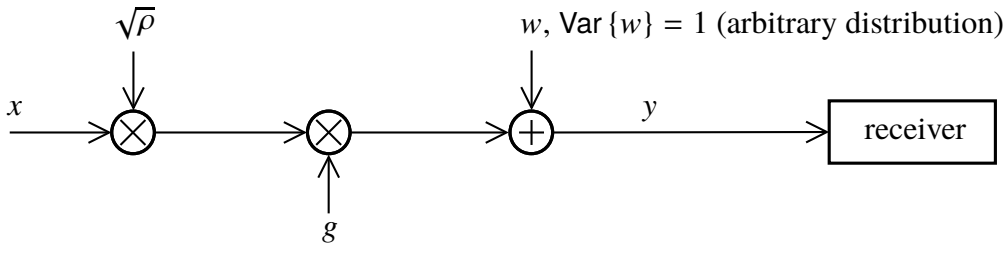
(a) Deterministic channel with additive Gaussian noise.



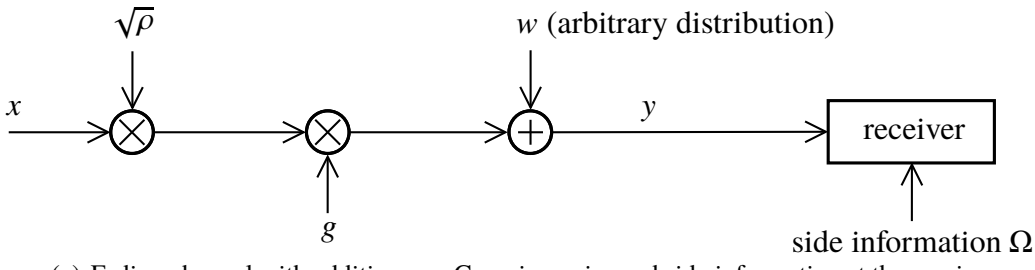
(b) Deterministic channel with additive non-Gaussian noise.



(c) Fading channel with additive Gaussian noise and perfect CSI at the receiver.



(d) Fading channel with additive non-Gaussian noise and no CSI at the receiver.



(e) Fading channel with additive non-Gaussian noise and side information at the receiver.

Figure 2.9. Scalar point-to-point channels.

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

2.3.1 Deterministic Channel with Additive Gaussian Noise

The most fundamental example of a scalar point-to-point channel is the deterministic channel with additive Gaussian noise; see Figure 2.9(a). Here,

$$y = \sqrt{\rho}x + w, \tag{2.37}$$

where w is noise that is independent of x and has distribution $\mathrm{CN}(0, 1)$, and ρ is a constant. A new independent realization of w is drawn for every transmitted symbol x. The transmitted symbol x satisfies the power constraint $\mathsf{E}\left\{|x|^2\right\} \leq 1$. Hence, ρ has the meaning of SNR, as in Section 2.1.8. The capacity of this channel is

$$C = \log_2(1 + \rho), \tag{2.38}$$

and this capacity is achieved when the input symbols x are Gaussian distributed.

2.3.2 Deterministic Channel with Additive Non-Gaussian Noise

The next case of interest is when (2.37) applies, $\mathsf{E}\{w\} = 0$ and $\mathsf{Var}\{w\} = 1$, but w is not necessarily Gaussian; see Figure 2.9(b). Assuming that x and w are uncorrelated, but not necessarily independent,

$$\mathsf{E}\{x^* w\} = 0, \tag{2.39}$$

the capacity is lower bounded as follows:

$$C \geq \log_2(1 + \rho). \tag{2.40}$$

In contrast to the Gaussian noise case in Section 2.3.1, the optimal distribution of the input symbol x is generally not Gaussian.

2.3.3 Fading Channel with Additive Gaussian Noise and Perfect CSI at the Receiver

We next introduce fading. The first model of interest is that of a fading channel with Gaussian noise and a gain g that is perfectly known to the receiver but unknown to the transmitter. Here,

$$y = \sqrt{\rho}gx + w, \tag{2.41}$$

where ρ, x, and w have the same meaning as in Section 2.3.1, x and w are independent, and, in addition, g is a random variable that represents the fading channel gain and which

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

is independent of x and w; see Figure 2.9(c). New independent realizations of g and w are drawn for each transmitted symbol x. The distribution of g can be arbitrary. The capacity is

$$C = \mathsf{E}\left\{\log_2\left(1 + \rho|g|^2\right)\right\}. \tag{2.42}$$

Operationally, the capacity in (2.42) only has a meaning if there is coding across all sources of randomness in the channel – including both the noise and channel gain.[1] To stress this fact, C is often called the *ergodic* capacity.

2.3.4 Fading Channel with Additive Non-Gaussian Noise and no CSI at the Receiver

We next extend the previous model to the case of an unknown channel gain and non-Gaussian noise; see Figure 2.9(d). Equation (2.41) applies; $\mathsf{E}\left\{|x|^2\right\} \leq 1$, $\mathsf{E}\{w\} = 0$, and $\mathsf{Var}\{w\} = 1$, however, w is not necessarily Gaussian. The signal x and the noise w are uncorrelated, but not necessarily independent. Neither the transmitter, nor the receiver knows g. Also, g and x are independent; however, no assumption is made on the statistical relation between g and w.

To obtain a simple capacity bound, we rewrite the expression for y as follows:

$$y = \sqrt{\rho}\mathsf{E}\{g\}x + \sqrt{\rho}\left(g - \mathsf{E}\{g\}\right)x + w. \tag{2.43}$$

The receiver's lack of knowledge about g is captured by the second term, $\sqrt{\rho}\left(g - \mathsf{E}\{g\}\right)x$. A direct calculation shows that the second and third terms of (2.43) are mutually uncorrelated, and uncorrelated with x. Considering the two last terms of (2.43) as *effective noise*, the channel in (2.43) is, with appropriate normalization, equivalent to the model we treated in Section 2.3.2. Using (2.40) results in

$$C \geq \log_2\left(1 + \frac{\rho\,|\mathsf{E}\{g\}|^2}{\rho\mathsf{Var}\{g\} + 1}\right). \tag{2.44}$$

The bound in (2.44) is mainly useful when g fluctuates only slightly around its expected value $\mathsf{E}\{g\}$, so that $\mathsf{Var}\{g\}$ is small. This will be the case in many of the derivations of capacity bounds for Massive MIMO in Chapters 3 and 4. The train of reasoning can be summarized as follows: (i) Owing to the linear beamforming, each terminal sees a scalar channel with unknown gain g, and additive uncorrelated effective noise that comprises receiver noise and interference. (ii) The effect of the lack of knowledge of g, captured

[1] By contrast, in case each codeword sees only one realization of g, the capacity expression in (2.42) is irrelevant. Instead, a quantity called *outage capacity* must be considered; see, for example, [29].

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

in the deviation $g - \mathsf{E}\{g\}$, is treated as additional uncorrelated effective noise. By virtue of channel hardening, while g is random and unknown, it fluctuates only slightly around $\mathsf{E}\{g\}$ so this additional effective noise is small. (iii) The variances of all effective noise terms depend only on second- and fourth-order moments of Gaussian random variables, and hence can be computed in closed form.

2.3.5 Fading Channel with Non-Gaussian Noise and Side Information

The final, and most general, case of interest is that of a fading channel with non-Gaussian noise, where the receiver has access to *side information* quantified via a random variable Ω; see Figure 2.9(e). The received signal is given by (2.41) where $\mathsf{E}\left\{|x|^2\right\} \leq 1$, and w has an arbitrary distribution. The side information Ω may be correlated with g. Hence, while the receiver has no direct access to g, its knowledge of Ω may convey implicit information about g. We assume that x is independent of g and of Ω, that w has zero mean, and that x and w are uncorrelated (but not necessarily independent), conditioned on Ω, in the precise sense that

$$\begin{aligned}\mathsf{E}\{w|\Omega\} &= \mathsf{E}\{x^* w|\Omega\} \\ &= \mathsf{E}\{g^* x^* w|\Omega\} \\ &= 0. \end{aligned} \tag{2.45}$$

Then the capacity is bounded as follows:

$$C \geq \mathsf{E}\left\{\log_2\left(1 + \frac{\rho\,|\mathsf{E}\{g|\Omega\}|^2}{\rho \mathsf{Var}\{g|\Omega\} + \mathsf{Var}\{w|\Omega\}}\right)\right\}, \tag{2.46}$$

where the outer expectation is with respect to Ω.

Three special cases of (2.46) are noteworthy – in these special cases, we assume additionally that $\mathsf{Var}\{w\} = 1$:

1. In the absence of side information Ω, we revert to the case in Section 2.3.4. The bound (2.46) then reduces to (2.44), as expected.

2. If the receiver knows g so that $\Omega = g$, and w is independent of g, then (2.46) specializes to

$$C \geq \mathsf{E}\left\{\log_2\left(1 + \rho\,|g|^2\right)\right\}. \tag{2.47}$$

   The right-hand side of (2.47) is equal to (2.42). Hence, the bound (2.47) is tight in the case of Gaussian noise.

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

3. In the absence of fading, g is deterministic; say $g = 1$ for simplicity. Then (2.46) becomes

$$C \geq \mathsf{E}\left\{\log_2\left(1 + \frac{\rho}{\mathsf{Var}\{w|\Omega\}}\right)\right\}. \tag{2.48}$$

By applying Jensen's inequality (see Section C.1), specifically (C.3), to (2.48), we find that

$$\begin{aligned} C &\geq \log_2\left(1 + \frac{\rho}{\mathsf{Var}\{w\}}\right) \\ &= \log_2(1+\rho). \end{aligned} \tag{2.49}$$

The bound in (2.49) coincides with the bound derived in Section 2.3.2, as expected. Moreover, comparing with (2.38), we see that the bound is tight in the special case of Gaussian noise.

2.4 Summary of Key Points

The key points of this chapter are the following:

- A *coherence interval* is a time-frequency space during which the channel is substantially time-invariant and frequency-flat, so that its effect is well modeled as multiplication by a complex-valued scalar gain. The duration of a coherence interval is the channel coherence time T_c, and the bandwidth of a coherence interval is the channel coherence bandwidth B_c. Each coherence interval contains $\tau_c = B_c T_c$ complex-valued samples, and is split into an uplink and a downlink part. Table 2.1 shows some typical values of B_c, T_c, and τ_c.

  If OFDM modulation is used, two important quantities are the number of subcarriers over which the channel frequency response is approximately constant, N_{smooth}, and the number of OFDM symbols in a slot, N_{slot}. If the slot duration is equal to the channel coherence time, then each coherence interval spans N_{smooth} subcarriers and N_{slot} consecutive OFDM symbols. For this case, Figure 2.6 shows a possible mapping from OFDM symbols and subcarriers onto the samples in a coherence interval.

- In a system with L cells, K terminals per cell, and where each cell is served by an M-element base station array, propagation is modeled as follows. On the uplink, if

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

the K terminals in the l'th cell collectively transmit the K-vector $\boldsymbol{x}_{l'}$, the base station in the lth cell observes the M-vector

$$\boldsymbol{y}_l = \sqrt{\rho_{\text{ul}}}\boldsymbol{G}_l^l \boldsymbol{x}_l + \sqrt{\rho_{\text{ul}}} \sum_{\substack{l'=1 \\ l' \neq l}}^{L} \boldsymbol{G}_{l'}^l \boldsymbol{x}_{l'} + \boldsymbol{w}_l, \tag{2.50}$$

where $\boldsymbol{w}_l$ is a vector of receiver noise and the (m, k)th element of $\boldsymbol{G}_{l'}^l$, denoted by $g_{l'k}^{lm}$, contains the channel gain between the kth terminal in the l'th cell and the mth base station antenna in the lth cell. On the downlink, if the l'th base station transmits the M-vector $\boldsymbol{x}_{l'}$ then the terminals in the lth cell receive

$$\boldsymbol{y}_l = \sqrt{\rho_{\text{dl}}}\boldsymbol{G}_l^{l\text{T}} \boldsymbol{x}_l + \sqrt{\rho_{\text{dl}}} \sum_{\substack{l'=1 \\ l' \neq l}}^{L} \boldsymbol{G}_l^{l'\text{T}} \boldsymbol{x}_{l'} + \boldsymbol{w}_l, \tag{2.51}$$

where again, $\boldsymbol{w}_l$ is receiver noise. In (2.50) and (2.51), ρ_{ul} and ρ_{dl} have the operational interpretation of SNR.

Each channel coefficient can be decomposed as

$$g_{l'k}^{lm} = \sqrt{\beta_{l'k}^l} h_{l'k}^{lm}, \tag{2.52}$$

where $\beta_{l'k}^l$ represents the attenuation due to large-scale fading (that includes path loss and shadow fading), and $h_{l'k}^{lm}$ represents the effect of the small-scale fading.

- In Massive MIMO, we show in Chapters 3 and 4 that the effective channel to each terminal is a scalar, point-to-point channel. Ergodic capacity (bounds) for the different scalar point-to-point channels illustrated in Figure 2.9 are summarized in Table 2.3.

https://doi.org/10.1017/CBO9781316799895.003 Published online by Cambridge University Press

Deterministic channel with additive Gaussian noise; see Section 2.3.1 and Figure 2.9(a)	$C = \log_2(1+\rho)$, if x and w are independent
Deterministic channel with additive non-Gaussian noise; see Section 2.3.2 and Figure 2.9(b)	$C \geq \log_2(1+\rho)$, if $\mathsf{E}\{x^* w\} = 0$
Fading channel with additive Gaussian noise and perfect CSI at the receiver; see Section 2.3.3 and Figure 2.9(c)	$C = \mathsf{E}\left\{\log_2\left(1+\rho\,\lvert g\rvert^2\right)\right\}$, if x, w and g are mutually independent
Fading channel with additive non-Gaussian noise and no CSI at the receiver; see Section 2.3.4 and Figure 2.9(d)	$C \geq \log_2\left(1+\frac{\rho\,\lvert\mathsf{E}\{g\}\rvert^2}{\rho\mathsf{Var}\{g\}+1}\right)$, if $\mathsf{E}\{x^* w\} = 0$
Fading channel with additive non-Gaussian noise and side information; see Section 2.3.5 and Figure 2.9(e)	$C \geq \mathsf{E}\left\{\log_2\left(1+\frac{\rho\,\lvert\mathsf{E}\{g\lvert\Omega\}\rvert^2}{\rho\mathsf{Var}\{g\lvert\Omega\}+\mathsf{E}\{\lvert w\rvert^2\lvert\Omega\}}\right)\right\}$, if $\mathsf{E}\{w\lvert\Omega\} = \mathsf{E}\{x^* w\lvert\Omega\} = \mathsf{E}\{g^* x^* w\lvert\Omega\} = 0$

Table 2.3. Capacity bounds for the scalar point-to-point channels. In all cases, $\mathsf{E}\{w\} = 0$ and x is independent of g and Ω, but no other assumptions (other than those explicitly stated) on statistical independence are made.