

Capacity of wireless channels

In the previous two chapters, we studied *specific* techniques for communication over wireless channels. In particular, Chapter 3 is centered on the point-to-point communication scenario and there the focus is on diversity as a way to mitigate the adverse effect of fading. Chapter 4 looks at cellular wireless networks as a whole and introduces several multiple access and interference management techniques.

The present chapter takes a more fundamental look at the problem of communication over wireless fading channels. We ask: what is the *optimal* performance achievable on a given channel and what are the techniques to achieve such optimal performance? We focus on the point-to-point scenario in this chapter and defer the multiuser case until Chapter 6. The material covered in this chapter lays down the theoretical basis of the modern development in wireless communication to be covered in the rest of the book.

The framework for studying performance limits in communication is *information theory*. The basic measure of performance is the *capacity* of a channel: the maximum rate of communication for which arbitrarily small error probability can be achieved. Section 5.1 starts with the important example of the AWGN (additive white Gaussian noise) channel and introduces the notion of capacity through a heuristic argument. The AWGN channel is then used as a building block to study the capacity of wireless fading channels. Unlike the AWGN channel, there is no single definition of capacity for fading channels that is applicable in all scenarios. Several notions of capacity are developed, and together they form a systematic study of performance limits of fading channels. The various capacity measures allow us to see clearly the different types of resources available in fading channels: power, diversity and degrees of freedom. We will see how the diversity techniques studied in Chapter 3 fit into this big picture. More importantly, the capacity results suggest an alternative technique, *opportunistic communication*, which will be explored further in the later chapters.

5.1 AWGN channel capacity

Information theory was invented by Claude Shannon in 1948 to characterize the limits of reliable communication. Before Shannon, it was widely believed that the only way to achieve reliable communication over a noisy channel, i.e., to make the error probability as small as desired, was to reduce the data rate (by, say, repetition coding). Shannon showed the surprising result that this belief is incorrect: by more intelligent coding of the information, one can in fact communicate at a strictly *positive* rate but at the same time with as small an error probability as desired. However, there is a maximal rate, called the *capacity* of the channel, for which this can be done: if one attempts to communicate at rates above the channel capacity, then it is *impossible* to drive the error probability to zero.

In this section, the focus is on the familiar (real) AWGN channel:

$$y[m] = x[m] + w[m], \quad (5.1)$$

where $x[m]$ and $y[m]$ are real input and output at time m respectively and $w[m]$ is $\mathcal{N}(0, \sigma^2)$ noise, independent over time. The importance of this channel is two-fold:

- It is a building block of all of the wireless channels studied in this book.
- It serves as a motivating example of what capacity means operationally and gives some sense as to why arbitrarily reliable communication is possible at a strictly positive data rate.

5.1.1 Repetition coding

Using uncoded BPSK symbols $x[m] = \pm\sqrt{P}$, the error probability is $Q\left(\sqrt{P/\sigma^2}\right)$. To reduce the error probability, one can repeat the same symbol N times to transmit the one bit of information. This is a repetition code of block length N , with codewords $\mathbf{x}_A = \sqrt{P}[1, \dots, 1]^t$ and $\mathbf{x}_B = \sqrt{P}[-1, \dots, -1]^t$. The codewords meet a power constraint of P joules/symbol. If \mathbf{x}_A is transmitted, the received vector is

$$\mathbf{y} = \mathbf{x}_A + \mathbf{w}, \quad (5.2)$$

where $\mathbf{w} = (w[1], \dots, w[N])^t$. Error occurs when \mathbf{y} is closer to \mathbf{x}_B than to \mathbf{x}_A , and the error probability is given by

$$Q\left(\frac{\|\mathbf{x}_A - \mathbf{x}_B\|}{2\sigma}\right) = Q\left(\sqrt{\frac{NP}{\sigma^2}}\right), \quad (5.3)$$

which decays exponentially with the block length N . The good news is that communication can now be done with arbitrary reliability by choosing a large

enough N . The bad news is that the data rate is only $1/N$ bits per symbol time and with increasing N the data rate goes to zero.

The reliably communicated data rate with repetition coding can be marginally improved by using multilevel PAM (generalizing the two-level BPSK scheme from earlier). By repeating an M -level PAM symbol, the levels equally spaced between $\pm\sqrt{P}$, the rate is $\log M/N$ bits per symbol time¹ and the error probability for the inner levels is equal to

$$Q\left(\frac{\sqrt{NP}}{(M-1)\sigma}\right). \quad (5.4)$$

As long as the number of levels M grows at a rate less than \sqrt{N} , reliable communication is guaranteed at large block lengths. But the data rate is bounded by $(\log \sqrt{N})/N$ and this still goes to zero as the block length increases. Is that the price one must pay to achieve reliable communication?

5.1.2 Packing spheres

Geometrically, repetition coding puts all the codewords (the M levels) in just one dimension (Figure 5.1 provides an illustration; here, all the codewords are on the same line). On the other hand, the signal space has a large number of dimensions N . We have already seen in Chapter 3 that this is a very inefficient way of packing codewords. To communicate more efficiently, the codewords should be spread in all the N dimensions.

We can get an estimate on the maximum number of codewords that can be packed in for the given power constraint P , by appealing to the classic sphere-packing picture (Figure 5.2). By the law of large numbers, the N -dimensional received vector $\mathbf{y} = \mathbf{x} + \mathbf{w}$ will, with high probability, lie within

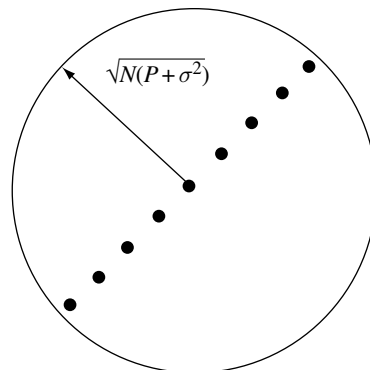
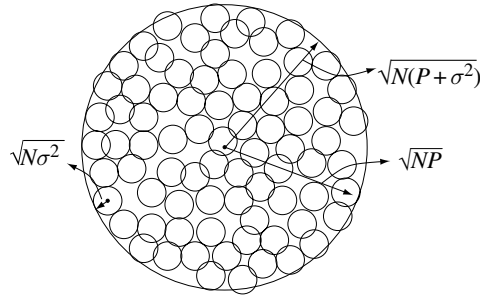


Figure 5.1 Repetition coding packs points inefficiently in the high-dimensional signal space.

¹ In this chapter, all logarithms are taken to be to the base 2 unless specified otherwise.

Figure 5.2 The number of noise spheres that can be packed into the y -sphere yields the maximum number of codewords that can be reliably distinguished.



a y -sphere of radius $\sqrt{N(P + \sigma^2)}$; so without loss of generality we need only focus on what happens inside this y -sphere. On the other hand

$$\frac{1}{N} \sum_{m=1}^N w^2[m] \rightarrow \sigma^2 \quad (5.5)$$

as $N \rightarrow \infty$, by the law of large numbers again. So, for N large, the received vector \mathbf{y} lies, with high probability, near the surface of a *noise sphere* of radius $\sqrt{N}\sigma$ around the transmitted codeword (this is sometimes called the *sphere hardening* effect). Reliable communication occurs as long as the noise spheres around the codewords do not overlap. The maximum number of codewords that can be packed with non-overlapping noise spheres is the ratio of the volume of the y -sphere to the volume of a noise sphere:²

$$\frac{\left(\sqrt{N(P + \sigma^2)}\right)^N}{\left(\sqrt{N\sigma^2}\right)^N}. \quad (5.6)$$

This implies that the maximum number of bits per symbol that can be reliably communicated is

$$\frac{1}{N} \log \left(\frac{\left(\sqrt{N(P + \sigma^2)}\right)^N}{\left(\sqrt{N\sigma^2}\right)^N} \right) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right). \quad (5.7)$$

This is indeed the capacity of the AWGN channel. (The argument might sound very heuristic. Appendix B.5 takes a more careful look.)

The sphere-packing argument only yields the maximum number of codewords that can be packed while ensuring reliable communication. How to construct codes to achieve the promised rate is another story. In fact, in Shannon's argument, he never explicitly constructed codes. What he showed is that if

² The volume of an N -dimensional sphere of radius r is proportional to r^N and an exact expression is evaluated in Exercise B.10.

one picks the codewords randomly and independently, with the components of each codeword i.i.d. $\mathcal{N}(0, P)$, then with very high probability the randomly chosen code will do the job at any rate $R < C$. This is the so-called *i.i.d. Gaussian code*. A sketch of this random coding argument can be found in Appendix B.5.

From an engineering standpoint, the essential problem is to identify easily encodable and decodable codes that have performance close to the capacity. The study of this problem is a separate field in itself and Discussion 5.1 briefly chronicles the success story: codes that operate very close to capacity have been found and can be implemented in a relatively straightforward way using current technology. In the rest of the book, these codes are referred to as “capacity-achieving AWGN codes”.

Discussion 5.1 Capacity-achieving AWGN channel codes

Consider a code for communication over the real AWGN channel in (5.1). The ML decoder chooses the nearest codeword to the received vector as the most likely transmitted codeword. The closer two codewords are to each other, the higher the probability of confusing one for the other: this yields a geometric design criterion for the set of codewords, i.e., place the codewords as far apart from each other as possible. While such a set of maximally spaced codewords are likely to perform very well, this in itself does not constitute an *engineering* solution to the problem of code construction: what is required is an arrangement that is “easy” to describe and “simple” to decode. In other words, the computational complexity of encoding and decoding should be practical.

Many of the early solutions centered around the theme of ensuring efficient ML decoding. The search of codes that have this property leads to a rich class of codes with nice algebraic properties, but their performance is quite far from capacity. A significant breakthrough occurred when the stringent ML decoding was relaxed to an *approximate* one. An iterative decoding algorithm with near ML performance has led to *turbo* and *low density parity check* codes.

A large ensemble of linear parity check codes can be considered in conjunction with the iterative decoding algorithm. Codes with good performance can be found offline and they have been verified to perform very close to capacity. To get a feel for their performance, we consider some sample performance numbers. The capacity of the AWGN channel at 0 dB SNR is 0.5 bits per symbol. The error probability of a carefully designed LDPC code in these operating conditions (rate 0.5 bits per symbol, and the signal-to-noise ratio is equal to 0.1 dB) with a block length of 8000 bits is approximately 10^{-4} . With a larger block length, much smaller error probabilities have been achieved. These modern developments are well surveyed in [100].

The capacity of the AWGN channel is probably the most well-known result of information theory, but it is in fact only a special case of Shannon's general theory applied to a specific channel. This general theory is outlined in Appendix B. All the capacity results used in the book can be derived from this general framework. To focus more on the implications of the results in the main text, the derivation of these results is relegated to Appendix B. In the main text, the capacities of the channels looked at are justified by either

Summary 5.1 Reliable rate of communication and capacity

- Reliable communication at rate R bits/symbol means that one can design codes at that rate with arbitrarily small error probability.
- To get reliable communication, one *must* code over a long block; this is to exploit the law of large numbers to average out the randomness of the noise.
- Repetition coding over a long block can achieve reliable communication, but the corresponding data rate goes to zero with increasing block length.
- Repetition coding does not pack the codewords in the available degrees of freedom in an efficient manner. One can pack a number of codewords that is exponential in the block length and still communicate reliably. This means the data rate can be strictly positive even as reliability is increased arbitrarily by increasing the block length.
- The maximum data rate at which reliable communication is possible is called the capacity C of the channel.
- The capacity of the (real) AWGN channel with power constraint P and noise variance σ^2 is:

$$C_{\text{awgn}} = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right), \quad (5.8)$$

and the engineering problem of constructing codes close to this performance has been successfully addressed.

Figure 5.3 summarizes the three communication schemes discussed.

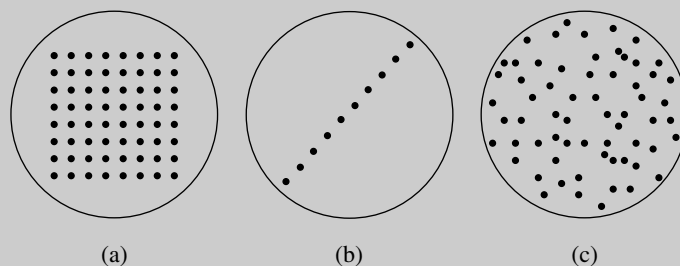


Figure 5.3 The three communication schemes when viewed in N -dimensional space: (a) uncoded signaling: error probability is poor since large noise in any dimension is enough to confuse the receiver; (b) repetition code: codewords are now separated in all dimensions, but there are only a few codewords packed in a single dimension; (c) capacity-achieving code: codewords are separated in all dimensions and there are many of them spread out in the space.

transforming the channels back to the AWGN channel, or by using the type of heuristic sphere-packing arguments we have just seen.

5.2 Resources of the AWGN channel

The AWGN capacity formula (5.8) can be used to identify the roles of the key resources of *power* and *bandwidth*.

5.2.1 Continuous-time AWGN channel

Consider a continuous-time AWGN channel with bandwidth W Hz, power constraint \bar{P} watts, and additive white Gaussian noise with power spectral density $N_0/2$. Following the passband–baseband conversion and sampling at rate $1/W$ (as described in Chapter 2), this can be represented by a discrete-time complex baseband channel:

$$y[m] = x[m] + w[m], \quad (5.9)$$

where $w[m]$ is $\mathcal{CN}(0, N_0)$ and is i.i.d. over time. Note that since the noise is independent in the I and Q components, each use of the complex channel can be thought of as two independent uses of a real AWGN channel. The noise variance and the power constraint per real symbol are $N_0/2$ and $\bar{P}/(2W)$ respectively. Hence, the capacity of the channel is

$$\frac{1}{2} \log \left(1 + \frac{\bar{P}}{N_0 W} \right) \text{ bits per real dimension}, \quad (5.10)$$

or

$$\log \left(1 + \frac{\bar{P}}{N_0 W} \right) \text{ bits per complex dimension}. \quad (5.11)$$

This is the capacity in bits per complex dimension or degree of freedom. Since there are W complex samples per second, the capacity of the continuous-time AWGN channel is

$$C_{\text{awgn}}(\bar{P}, W) = W \log \left(1 + \frac{\bar{P}}{N_0 W} \right) \text{ bits/s}. \quad (5.12)$$

Note that $\text{SNR} := \bar{P}/(N_0 W)$ is the SNR per (complex) degree of freedom. Hence, AWGN capacity can be rewritten as

$$C_{\text{awgn}} = \log(1 + \text{SNR}) \text{ bits/s/Hz}. \quad (5.13)$$

This formula measures the maximum achievable *spectral efficiency* through the AWGN channel as a function of the SNR.

5.2.2 Power and bandwidth

Let us ponder the significance of the capacity formula (5.12) to a communication engineer. One way of using this formula is as a benchmark for evaluating the performance of channel codes. For a system engineer, however, the main significance of this formula is that it provides a high-level way of thinking about how the performance of a communication system depends on the basic *resources* available in the channel, without going into the details of specific modulation and coding schemes used. It will also help identify the bottleneck that limits performance.

The basic resources of the AWGN channel are the received power \bar{P} and the bandwidth W . Let us first see how the capacity depends on the received power. To this end, a key observation is that the function

$$f(\text{SNR}) := \log(1 + \text{SNR}) \quad (5.14)$$

is *concave*, i.e., $f''(x) \leq 0$ for all $x \geq 0$ (Figure 5.4). This means that increasing the power \bar{P} suffers from a law of diminishing marginal returns: the higher the SNR, the smaller the effect on capacity. In particular, let us look at the low and the high SNR regimes. Observe that

$$\log_2(1 + x) \approx x \log_2 e \quad \text{when } x \approx 0, \quad (5.15)$$

$$\log_2(1 + x) \approx \log_2 x \quad \text{when } x \gg 1. \quad (5.16)$$

Thus, when the SNR is low, the capacity increases linearly with the received power \bar{P} : every 3 dB increase in (or, doubling) the power doubles the capacity. When the SNR is high, the capacity increases logarithmically with \bar{P} : every 3 dB increase in the power yields only one additional bit per dimension. This phenomenon should not come as a surprise. We have already seen in

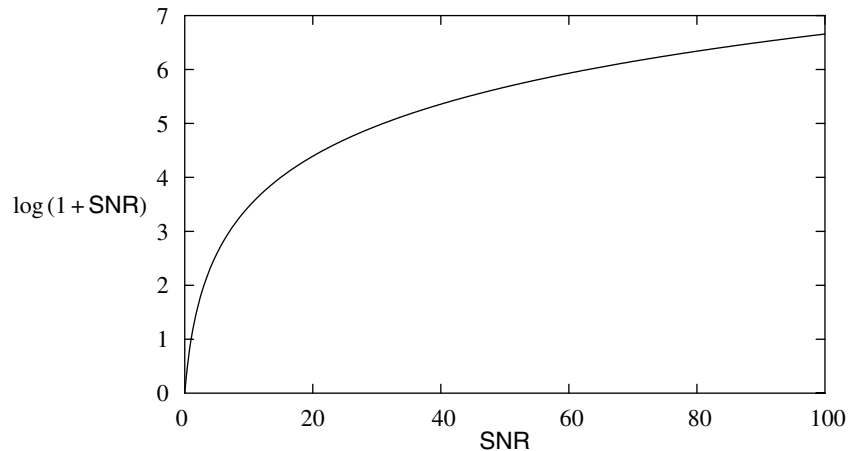


Figure 5.4 Spectral efficiency $\log(1 + \text{SNR})$ of the AWGN channel.

Chapter 3 that packing many bits per dimension is very power-inefficient. The capacity result says that this phenomenon not only holds for specific schemes but is in fact fundamental to *all* communication schemes. In fact, for a fixed error probability, the data rate of uncoded QAM also increases logarithmically with the SNR (Exercise 5.7).

The dependency of the capacity on the bandwidth W is somewhat more complicated. From the formula, the capacity depends on the bandwidth in two ways. First, it increases the degrees of freedom available for communication. This can be seen in the linear dependency on W for a fixed $\text{SNR} = \bar{P}/(N_0W)$. On the other hand, for a given received power \bar{P} , the SNR per dimension decreases with the bandwidth as the energy is spread more thinly across the degrees of freedom. In fact, it can be directly calculated that the capacity is an increasing, concave function of the bandwidth W (Figure 5.5). When the bandwidth is small, the SNR per degree of freedom is high, and then the capacity is insensitive to small changes in SNR. Increasing W yields a rapid increase in capacity because the increase in degrees of freedom more than compensates for the decrease in SNR. The system is in the *bandwidth-limited* regime. When the bandwidth is large such that the SNR per degree of freedom is small,

$$W \log \left(1 + \frac{\bar{P}}{N_0W} \right) \approx W \left(\frac{\bar{P}}{N_0W} \right) \log_2 e = \frac{\bar{P}}{N_0} \log_2 e. \quad (5.17)$$

In this regime, the capacity is proportional to the *total* received power across the entire band. It is insensitive to the bandwidth, and increasing the bandwidth has a small impact on capacity. On the other hand, the capacity is now linear in the received power and increasing power has a significant effect. This is the *power-limited* regime.

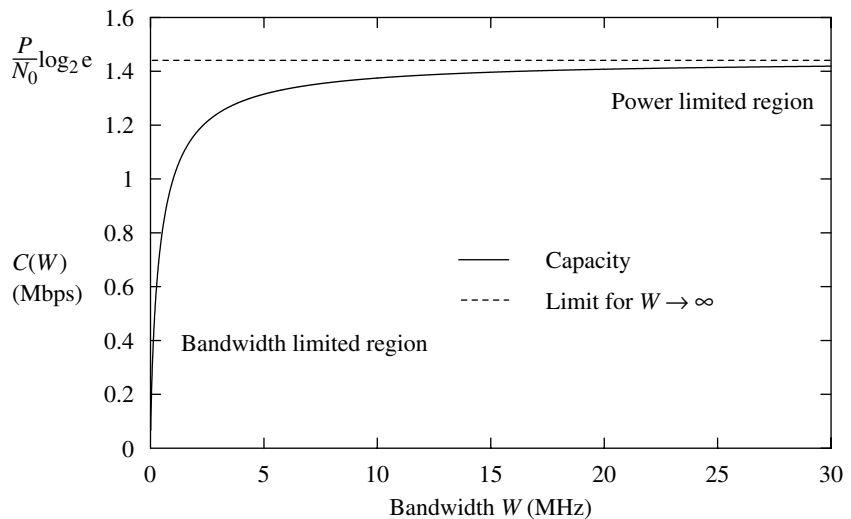


Figure 5.5 Capacity as a function of the bandwidth W . Here $\bar{P}/N_0 = 10^6$.

As W increases, the capacity increases monotonically (why must it?) and reaches the asymptotic limit

$$C_\infty = \frac{\bar{P}}{N_0} \log_2 e \text{ bits/s} \quad (5.18)$$

This is the infinite bandwidth limit, i.e., the capacity of the AWGN channel with only a power constraint but no limitation on bandwidth. It is seen that even if there is no bandwidth constraint, the capacity is finite.

In some communication applications, the main objective is to minimize the required energy per bit \mathcal{E}_b rather than to maximize the spectral efficiency. At a given power level \bar{P} , the minimum required energy per bit \mathcal{E}_b is $\bar{P}/C_{\text{awgn}}(\bar{P}, W)$. To minimize this, we should be operating in the most power-efficient regime, i.e., $\bar{P} \rightarrow 0$. Hence, the minimum \mathcal{E}_b/N_0 is given by

$$\left(\frac{\mathcal{E}_b}{N_0}\right)_{\min} = \lim_{\bar{P} \rightarrow 0} \frac{\bar{P}}{C_{\text{awgn}}(\bar{P}, W)N_0} = \frac{1}{\log_2 e} = -1.59 \text{ dB}. \quad (5.19)$$

To achieve this, the SNR per degree of freedom goes to zero. The price to pay for the energy efficiency is *delay*: if the bandwidth W is fixed, the communication rate (in bits/s) goes to zero. This essentially mimics the infinite bandwidth regime by spreading the total energy over a long time interval, instead of spreading the total power over a large bandwidth.

It was already mentioned that the success story of designing capacity-achieving AWGN codes is a relatively recent one. In the infinite bandwidth regime, however, it has long been known that *orthogonal codes*³ achieve the capacity (or, equivalently, achieve the minimum \mathcal{E}_b/N_0 of -1.59 dB). This is explored in Exercises 5.8 and 5.9.

Example 5.2 Bandwidth reuse in cellular systems

The capacity formula for the AWGN channel can be used to conduct a simple comparison of the two orthogonal cellular systems discussed in Chapter 4: the narrowband system with frequency reuse versus the wideband system with universal reuse. In both systems, users within a cell are orthogonal and do not interfere with each other. The main parameter of interest is the reuse ratio ρ ($\rho \leq 1$). If W denotes the bandwidth per user within a cell, then each user transmission occurs over a bandwidth of ρW . The parameter $\rho = 1$ yields the full reuse of the wideband OFDM system and $\rho < 1$ yields the narrowband system.

³ One example of orthogonal coding is the Hadamard sequences used in the IS-95 system (Section 4.3.1). Pulse position modulation (PPM), where the position of the on-off pulse (with large duty cycle) conveys the information, is another example.

Here we consider the uplink of this cellular system; the study of the downlink in orthogonal systems is similar. A user at a distance r is heard at the base-station with an attenuation of a factor $r^{-\alpha}$ in power; in free space the decay rate α is equal to 2 and the decay rate is 4 in the model of a single reflected path off the ground plane, cf. Section 2.1.5.

The uplink user transmissions in a neighboring cell that reuses the same frequency band are averaged and this constitutes the interference (this averaging is an important feature of the wideband OFDM system; in the narrowband system in Chapter 4, there is no interference averaging but that effect is ignored here). Let us denote by f_ρ the amount of total out-of-cell interference at a base-station as a *fraction* of the received signal power of a user at the edge of the cell. Since the amount of interference depends on the number of neighboring cells that reuse the same frequency band, the fraction f_ρ depends on the reuse ratio and also on the topology of the cellular system.

For example, in a one-dimensional linear array of base-stations (Figure 5.6), a reuse ratio of ρ corresponds to one in every $1/\rho$ cells using the same frequency band. Thus the fraction f_ρ decays roughly as ρ^α . On the other hand, in a two-dimensional hexagonal array of base-stations, a reuse ratio of ρ corresponds to the nearest reusing base-station roughly a distance of $\sqrt{1/\rho}$ away: this means that the fraction f_ρ decays roughly as $\rho^{\alpha/2}$. The exact fraction f_ρ takes into account geographical features of the cellular system (such as shadowing) and the geographic averaging of the interfering uplink transmissions; it is usually arrived at using numerical simulations (Table 6.2 in [140] has one such enumeration for a full reuse system). In a simple model where the interference is considered to come from the center of the cell reusing the same frequency band, f_ρ can be taken to be $2(\rho/2)^\alpha$ for the linear cellular system and $6(\rho/4)^{\alpha/2}$ for the hexagonal planar cellular system (see Exercises 5.2 and 5.3).

The received SINR at the base-station for a cell edge user is

$$\text{SINR} = \frac{\text{SNR}}{\rho + f_\rho \text{SNR}}, \quad (5.20)$$

where the SNR for the cell edge user is

$$\text{SNR} := \frac{P}{N_0 W d^\alpha}, \quad (5.21)$$

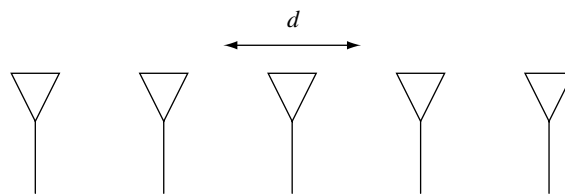


Figure 5.6 A linear cellular system with base-stations along a line (representing a highway).

with d the distance of the user to the base-station and P the uplink transmit power. The operating value of the parameter SNR is decided by the *coverage* of a cell: a user at the edge of a cell has to have a minimum SNR to be able to communicate reliably (at least a fixed minimum rate) with the nearest base-station. Each base-station comes with a capital installation cost and recurring operation costs and to minimize the number of base-stations, the cell size d is usually made as large as possible; depending on the uplink transmit power capability, coverage decides the cell size d .

Using the AWGN capacity formula (cf. (5.14)), the rate of reliable communication for a user at the edge of the cell, as a function of the reuse ratio ρ , is

$$R_\rho = \rho W \log_2(1 + \text{SINR}) = \rho W \log_2 \left(1 + \frac{\text{SNR}}{\rho + f_\rho \text{SNR}} \right) \text{ bits/s.} \quad (5.22)$$

The rate depends on the reuse ratio through the available degrees of freedom and the amount of out-of-cell interference. A large ρ increases the available bandwidth per cell but also increases the amount of out-of-cell interference. The formula (5.22) allows us to study the optimal reuse factor. At low SNR, the system is not degree of freedom limited *and* the interference is small relative to the noise; thus the rate is insensitive to the reuse factor and this can be verified directly from (5.22). On the other hand, at large SNR the interference grows as well and the SINR peaks at $1/f_\rho$. (A general rule of thumb in practice is to set SNR such that the interference is of the same order as the background noise; this will guarantee that the operating SINR is close to the largest value.) The largest rate is

$$\rho W \log_2 \left(1 + \frac{1}{f_\rho} \right). \quad (5.23)$$

This rate goes to zero for small values of ρ ; thus sparse reuse is not favored. It can be verified that universal reuse yields the largest rate in (5.23) for the hexagonal cellular system (Exercise 5.3). For the linear cellular model, the corresponding optimal reuse is $\rho = 1/2$, i.e., reusing the frequency every other cell (Exercise 5.5). The reduction in interference due to less reuse is more dramatic in the linear cellular system when compared to the hexagonal cellular system. This difference is highlighted in the optimal reuse ratios for the two systems at high SNR: universal reuse is preferred for the hexagonal cellular system while a reuse ratio of $1/2$ is preferred for the linear cellular system.

This comparison also holds for a range of SNR between the small and the large values: Figures 5.7 and 5.8 plot the rates in (5.22) for different reuse ratios for the linear and hexagonal cellular systems respectively. Here the power decay rate α is fixed to 3 and the rates are plotted as a function of the SNR for a user at the edge of the cell, cf. (5.21). In the

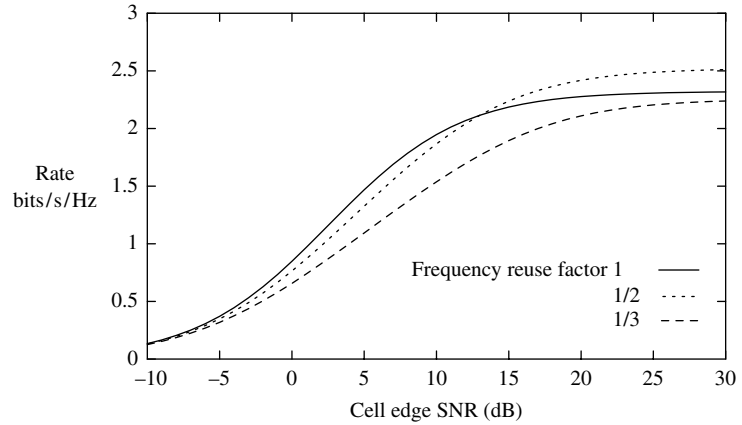


Figure 5.7 Rates in bits/s/Hz as a function of the SNR for a user at the edge of the cell for universal reuse and reuse ratios of 1/2 and 1/3 for the linear cellular system. The power decay rate α is set to 3.

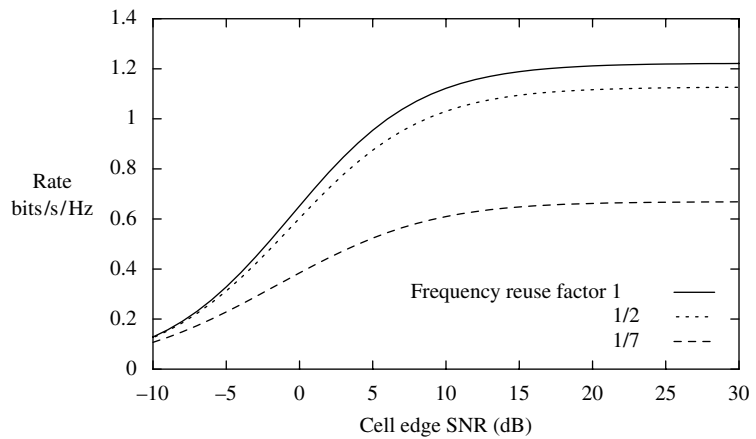


Figure 5.8 Rates in bits/s/Hz as a function of the SNR for a user at the edge of the cell for universal reuse, reuse ratios 1/2 and 1/7 for the hexagonal cellular system. The power decay rate α is set to 3.

hexagonal cellular system, universal reuse is clearly preferred at all ranges of SNR. On the other hand, in a linear cellular system, universal reuse and a reuse of 1/2 have comparable performance and if the operating SNR value is larger than a threshold (10 dB in Figure 5.7), then it pays to reuse, i.e., $R_{1/2} > R_1$. Otherwise, universal reuse is optimal. If this SNR threshold is within the rule of thumb setting mentioned earlier (i.e., the gain in rate is worth operating at this SNR), then reuse is preferred. This Preference has to be traded off with the size of the cell dictated by (5.21) due to a transmit power constraint on the mobile device.

5.3 Linear time-invariant Gaussian channels

We give three examples of channels which are closely related to the simple AWGN channel and whose capacities can be easily computed. Moreover, optimal codes for these channels can be constructed directly from an optimal code for the basic AWGN channel. These channels are *time-invariant*, known to both the transmitter and the receiver, and they form a bridge to the fading channels which will be studied in the next section.

5.3.1 Single input multiple output (SIMO) channel

Consider a SIMO channel with one transmit antenna and L receive antennas:

$$y_\ell[m] = h_\ell x[m] + w_\ell[m] \quad \ell = 1, \dots, L, \quad (5.24)$$

where h_ℓ is the *fixed* complex channel gain from the transmit antenna to the ℓ th receive antenna, and $w_\ell[m]$ is $\mathcal{CN}(0, N_0)$ is additive Gaussian noise independent across antennas. A sufficient statistic for detecting $x[m]$ from $\mathbf{y}[m] := [y_1[m], \dots, y_L[m]]^t$ is

$$\tilde{y}[m] := \mathbf{h}^* \mathbf{y}[m] = \|\mathbf{h}\|^2 x[m] + \mathbf{h}^* \mathbf{w}[m], \quad (5.25)$$

where $\mathbf{h} := [h_1, \dots, h_L]^t$ and $\mathbf{w}[m] := [w_1[m], \dots, w_L[m]]^t$. This is an AWGN channel with received SNR $P\|\mathbf{h}\|^2/N_0$ if P is the average energy per transmit symbol. The capacity of this channel is therefore

$$C = \log \left(1 + \frac{P\|\mathbf{h}\|^2}{N_0} \right) \text{ bits/s/Hz.} \quad (5.26)$$

Multiple receive antennas increase the effective SNR and provide a *power gain*. For example, for $L = 2$ and $|h_1| = |h_2| = 1$, dual receive antennas provide a 3 dB power gain over a single antenna system. The linear combining (5.25) maximizes the output SNR and is sometimes called *receive beamforming*.

5.3.2 Multiple input single output (MISO) channel

Consider a MISO channel with L transmit antennas and a single receive antenna:

$$y[m] = \mathbf{h}^* \mathbf{x}[m] + w[m], \quad (5.27)$$

where $\mathbf{h} = [h_1, \dots, h_L]^t$ and h_ℓ is the (fixed) channel gain from transmit antenna ℓ to the receive antenna. There is a total power constraint of P across the transmit antennas.

In the SIMO channel above, the sufficient statistic is the projection of the L -dimensional received signal onto \mathbf{h} : the projections in orthogonal directions contain noise that is not helpful to the detection of the transmit signal. A natural reciprocal transmission strategy for the MISO channel would send information only in the direction of the channel vector \mathbf{h} ; information sent in any orthogonal direction will be nulled out by the channel anyway. Therefore, by setting

$$\mathbf{x}[m] = \frac{\mathbf{h}}{\|\mathbf{h}\|} \tilde{x}[m], \quad (5.28)$$

the MISO channel is reduced to the scalar AWGN channel:

$$y[m] = \|\mathbf{h}\| \tilde{x}[m] + w[m], \quad (5.29)$$

with a power constraint P on the scalar input. The capacity of this scalar channel is

$$\log \left(1 + \frac{P \|\mathbf{h}\|^2}{N_0} \right) \text{ bits/s/Hz.} \quad (5.30)$$

Can one do better than this scheme? Any reliable code for the MISO channel can be used as a reliable code for the scalar AWGN channel $y[m] = x[m] + w[m]$: if $\{\mathbf{X}_i\}$ are the transmitted $L \times N$ (space-time) code matrices for the MISO channel, then the received $1 \times N$ vectors $\{\mathbf{h}^* \mathbf{X}_i\}$ form a code for the scalar AWGN channel. Hence, the rate achievable by a reliable code for the MISO channel must be at most the capacity of a scalar AWGN channel with the same received SNR. Exercise 5.11 shows that the received SNR $P \|\mathbf{h}\|^2 / N_0$ of the transmission strategy above is in fact the *largest* possible SNR given the transmit power constraint of P . Any other scheme has a lower received SNR and hence its reliable rate must be less than (5.30), the rate achieved by the proposed transmission strategy. We conclude that the capacity of the MISO channel is indeed

$$C = \log \left(1 + \frac{P \|\mathbf{h}\|^2}{N_0} \right) \text{ bits/s/Hz.} \quad (5.31)$$

Intuitively, the transmission strategy maximizes the received SNR by having the received signals from the various transmit antennas add up in-phase (coherently) and by allocating more power to the transmit antenna with the better gain. This strategy, “aligning the transmit signal in the direction of the transmit antenna array pattern”, is called *transmit beamforming*. Through beamforming, the MISO channel is converted into a scalar AWGN channel and thus any code which is optimal for the AWGN channel can be used directly.

In both the SIMO and the MISO examples the benefit from having multiple antennas is a *power* gain. To get a gain in degrees of freedom, one has to use both multiple transmit and multiple receive antennas (MIMO). We will study this in depth in Chapter 7.

5.3.3 Frequency-selective channel

Transformation to a parallel channel

Consider a *time-invariant* L -tap frequency-selective AWGN channel:

$$y[m] = \sum_{\ell=0}^{L-1} h_{\ell} x[m - \ell] + w[m], \quad (5.32)$$

with an average power constraint P on each input symbol. In Section 3.4.4, we saw that the frequency-selective channel can be converted into N_c independent sub-carriers by adding a cyclic prefix of length $L - 1$ to a data vector of length N_c , cf. (3.137). Suppose this operation is repeated over *blocks* of data symbols (of length N_c each, along with the corresponding cyclic prefix of length $L - 1$); see Figure 5.9. Then communication over the i th OFDM block can be written as

$$\tilde{y}_n[i] = \tilde{h}_n \tilde{d}_n[i] + \tilde{w}_n[i] \quad n = 0, 1, \dots, N_c - 1. \quad (5.33)$$

Here,

$$\tilde{\mathbf{d}}[i] := [\tilde{d}_0[i], \dots, \tilde{d}_{N_c-1}[i]]', \quad (5.34)$$

$$\tilde{\mathbf{w}}[i] := [\tilde{w}_0[i], \dots, \tilde{w}_{N_c-1}[i]]', \quad (5.35)$$

$$\tilde{\mathbf{y}}[i] := [\tilde{y}_0[i], \dots, \tilde{y}_{N_c-1}[i]]' \quad (5.36)$$

are the DFTs of the input, the noise and the output of the i th OFDM block respectively. $\tilde{\mathbf{h}}$ is the DFT of the channel scaled by $\sqrt{N_c}$ (cf. (3.138)). Since the overhead in the cyclic prefix relative to the block length N_c can be made arbitrarily small by choosing N_c large, the capacity of the original frequency-selective channel is the same as the capacity of this transformed channel as $N_c \rightarrow \infty$.

The transformed channel (5.33) can be viewed as a collection of sub-channels, one for each sub-carrier n . Each of the sub-channels is an AWGN channel. The

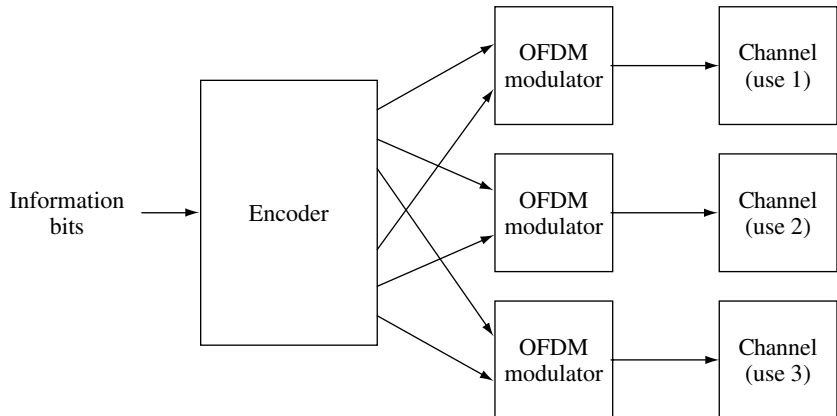


Figure 5.9 A coded OFDM system. Information bits are coded and then sent over the frequency-selective channel via OFDM modulation. Each channel use corresponds to an OFDM block. Coding can be done across different OFDM blocks as well as over different sub-carriers.

transformed noise $\tilde{\mathbf{w}}[i]$ is distributed as $\mathcal{CN}(0, N_0 \mathbf{I})$, so the noise is $\mathcal{CN}(0, N_0)$ in each of the sub-channels and, moreover, the noise is independent across sub-channels. The power constraint on the input symbols in time translates to one on the data symbols on the sub-channels (Parseval theorem for DFTs):

$$\mathbb{E} [\|\tilde{\mathbf{d}}[i]\|^2] \leq N_c P. \quad (5.37)$$

In information theory jargon, a channel which consists of a set of non-interfering sub-channels, each of which is corrupted by independent noise, is called a *parallel channel*. Thus, the transformed channel here is a parallel AWGN channel, with a total power constraint across the sub-channels. A natural strategy for reliable communication over a parallel AWGN channel is illustrated in Figure 5.10. We allocate power to each sub-channel, P_n to the n th sub-channel, such that the total power constraint is met. Then, a separate capacity-achieving AWGN code is used to communicate over each of the sub-channels. The maximum rate of reliable communication using this scheme is

$$\sum_{n=0}^{N_c-1} \log \left(1 + \frac{P_n |\tilde{h}_n|^2}{N_0} \right) \text{ bits/OFDM symbol}. \quad (5.38)$$

Further, the power allocation can be chosen appropriately, so as to maximize the rate in (5.38). The “optimal power allocation”, thus, is the solution to the optimization problem:

$$C_{N_c} := \max_{P_0, \dots, P_{N_c-1}} \sum_{n=0}^{N_c-1} \log \left(1 + \frac{P_n |\tilde{h}_n|^2}{N_0} \right), \quad (5.39)$$

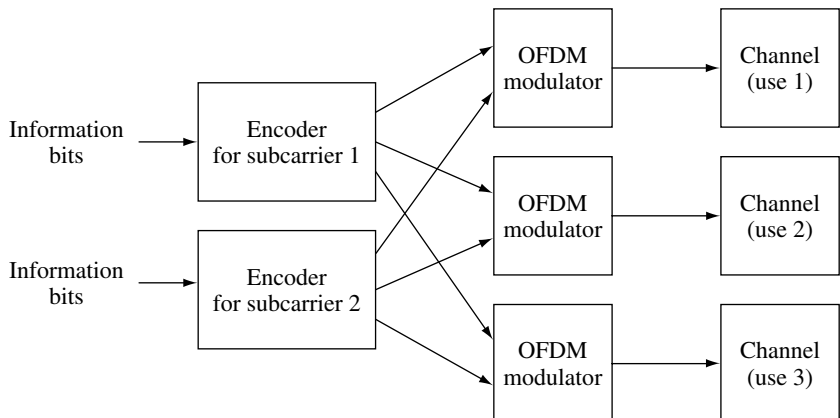


Figure 5.10 Coding independently over each of the sub-carriers. This architecture, with appropriate power and rate allocations, achieves the capacity of the frequency-selective channel.

subject to

$$\boxed{\sum_{n=0}^{N_c-1} P_n = N_c P, \quad P_n \geq 0, \quad n = 0, \dots, N_c - 1.} \quad (5.40)$$

Waterfilling power allocation

The optimal power allocation can be explicitly found. The objective function in (5.39) is jointly concave in the powers and this optimization problem can be solved by Lagrangian methods. Consider the Lagrangian

$$\mathcal{L}(\lambda, P_0, \dots, P_{N_c-1}) := \sum_{n=0}^{N_c-1} \log \left(1 + \frac{P_n |\tilde{h}_n|^2}{N_0} \right) - \lambda \sum_{n=0}^{N_c-1} P_n, \quad (5.41)$$

where λ is the Lagrange multiplier. The Kuhn–Tucker condition for the optimality of a power allocation is

$$\frac{\partial \mathcal{L}}{\partial P_n} \begin{cases} = 0 & \text{if } P_n > 0 \\ \leq 0 & \text{if } P_n = 0. \end{cases} \quad (5.42)$$

Define $x^+ := \max(x, 0)$. The power allocation

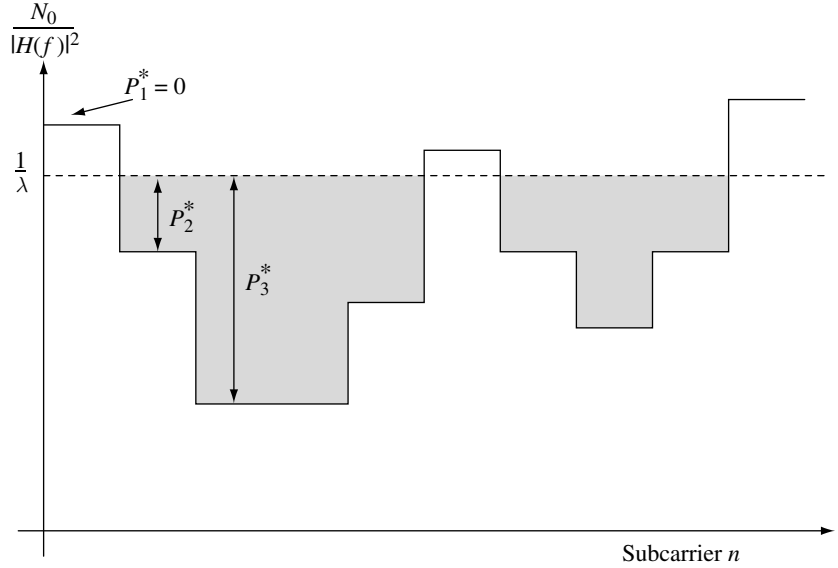
$$P_n^* = \left(\frac{1}{\lambda} - \frac{N_0}{|\tilde{h}_n|^2} \right)^+, \quad (5.43)$$

satisfies the conditions in (5.42) and is therefore optimal, with the Lagrange multiplier λ chosen such that the power constraint is met:

$$\frac{1}{N_c} \sum_{n=0}^{N_c-1} \left(\frac{1}{\lambda} - \frac{N_0}{|\tilde{h}_n|^2} \right)^+ = P. \quad (5.44)$$

Figure 5.11 gives a pictorial view of the optimal power allocation strategy for the OFDM system. Think of the values $N_0/|\tilde{h}_n|^2$ plotted as a function of the sub-carrier index $n = 0, \dots, N_c - 1$, as tracing out the bottom of a vessel. If P units of water per sub-carrier are filled into the vessel, the depth of the water at sub-carrier n is the power allocated to that sub-carrier, and $1/\lambda$ is the height of the water surface. Thus, this optimal strategy is called *waterfilling* or *waterpouring*. Note that there are some sub-carriers where the bottom of the vessel is above the water and no power is allocated to them. In these sub-carriers, the channel is too poor for it to be worthwhile to transmit information. In general, the transmitter allocates more power to the stronger sub-carriers, taking advantage of the better channel conditions, and less or even no power to the weaker ones.

Figure 5.11 Waterfilling power allocation over the N_c sub-carriers.



Observe that

$$\tilde{h}_n = \sum_{\ell=0}^{L-1} h_\ell \exp\left(-\frac{j2\pi\ell n}{N_c}\right), \quad (5.45)$$

is the discrete-time Fourier transform $H(f)$ evaluated at $f = nW/N_c$, where (cf. (2.20))

$$H(f) := \sum_{\ell=0}^{L-1} h_\ell \exp\left(-\frac{j2\pi\ell f}{W}\right), \quad f \in [0, W]. \quad (5.46)$$

As the number of sub-carriers N_c grows, the frequency width W/N_c of the sub-carriers goes to zero and they represent a finer and finer sampling of the continuous spectrum. So, the optimal power allocation converges to

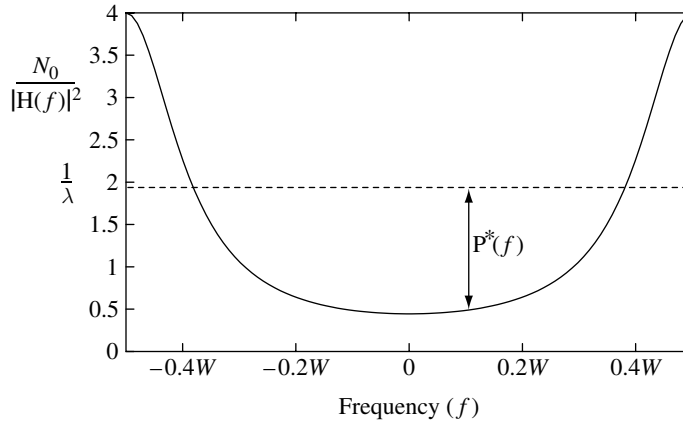
$$P^*(f) = \left(\frac{1}{\lambda} - \frac{N_0}{|H(f)|^2}\right)^+, \quad (5.47)$$

where the constant λ satisfies (cf. (5.44))

$$\int_0^W P^*(f) df = P. \quad (5.48)$$

The power allocation can be interpreted as waterfilling over frequency (see Figure 5.12). With N_c sub-carriers, the largest reliable communication rate

Figure 5.12 Waterfilling power allocation over the frequency spectrum of the two-tap channel (high-pass filter): $h[0] = 1$ and $h[1] = 0.5$.



with independent coding is C_{N_c} bits per OFDM symbol or C_{N_c}/N_c bits/s/Hz (C_{N_c} given in (5.39)). So as $N_c \rightarrow \infty$, the WC_{N_c}/N_c converges to

$$C = \int_0^W \log \left(1 + \frac{P^*(f)|H(f)|^2}{N_0} \right) df \text{ bits/s.} \quad (5.49)$$

Does coding across sub-carriers help?

So far we have considered a very simple scheme: coding independently over each of the sub-carriers. By coding *jointly* across the sub-carriers, presumably better performance can be achieved. Indeed, over a finite block length, coding jointly over the sub-carriers yields a smaller error probability than can be achieved by coding separately over the sub-carriers at the same rate. However, somewhat surprisingly, the *capacity* of the parallel channel is equal to the largest reliable rate of communication with independent coding within each sub-carrier. In other words, if the block length is very large then coding jointly over the sub-carriers cannot increase the rate of reliable communication any more than what can be achieved simply by allocating power and rate over the sub-carriers but not coding across the sub-carriers. So indeed (5.49) is the capacity of the time-invariant frequency-selective channel.

To get some insight into why coding *across* the sub-carriers with large block length does not improve capacity, we turn to a geometric view. Consider a code, with block length $N_c N$ symbols, coding over all N_c of the sub-carriers with N symbols from each sub-carrier. In high dimensions, i.e., $N \gg 1$, the $N_c N$ -dimensional *received* vector after passing through the parallel channel (5.33) lives in an ellipsoid, with different axes stretched and shrunk by the different channel gains \tilde{h}_n . The volume of the ellipsoid is proportional to

$$\prod_{n=0}^{N_c-1} \left(|\tilde{h}_n|^2 P_n + N_0 \right)^N, \quad (5.50)$$

see Exercise 5.12. The volume of the noise sphere is, as in Section 5.1.2, proportional to $N_0^{N_c N}$. The maximum number of distinguishable codewords that can be packed in the ellipsoid is therefore

$$\prod_{n=0}^{N_c-1} \left(1 + \frac{P_n |\tilde{h}_n|^2}{N_0} \right)^N. \quad (5.51)$$

The maximum reliable rate of communication is

$$\frac{1}{N} \log \prod_{n=0}^{N_c-1} \left(1 + \frac{P_n |\tilde{h}_n|^2}{N_0} \right)^N = \sum_{n=0}^{N_c-1} \log \left(1 + \frac{P_n |\tilde{h}_n|^2}{N_0} \right) \text{ bits/OFDM symbol}. \quad (5.52)$$

This is precisely the rate (5.38) achieved by separate coding and this suggests that coding across sub-carriers can do no better. While this sphere-packing argument is heuristic, Appendix B.6 gives a rigorous derivation from information theoretic first principles.

Even though coding across sub-carriers cannot improve the reliable *rate* of communication, it can still improve the *error probability* for a given data rate. Thus, coding across sub-carriers can still be useful in practice, particularly when the block length for each sub-carrier is small, in which case the coding effectively increases the overall block length.

In this section we have used parallel channels to model a frequency-selective channel, but parallel channels will be seen to be very useful in modeling many other wireless communication scenarios as well.

5.4 Capacity of fading channels

The basic capacity results developed in the last few sections are now applied to analyze the limits to communication over wireless fading channels.

Consider the complex baseband representation of a flat fading channel:

$$y[m] = h[m]x[m] + w[m], \quad (5.53)$$

where $\{h[m]\}$ is the fading process and $\{w[m]\}$ is i.i.d. $\mathcal{CN}(0, N_0)$ noise. As before, the symbol rate is W Hz, there is a power constraint of P joules/symbol, and $\mathbb{E}[|h[m]|^2] = 1$ is assumed for normalization. Hence $\text{SNR} := P/N_0$ is the average received SNR.

In Section 3.1.2, we analyzed the performance of uncoded transmission for this channel. What is the ultimate performance limit when information can be coded over a sequence of symbols? To answer this question, we make the simplifying assumption that the receiver can perfectly track the fading process, i.e., coherent reception. As we discussed in Chapter 2, the coherence time of typical wireless channels is of the order of hundreds of symbols and

so the channel varies slowly relative to the symbol rate and can be estimated by say a pilot signal. For now, the *transmitter* is not assumed to have any knowledge of the channel realization other than the statistical characterization. The situation when the transmitter has access to the channel realizations will be studied in Section 5.4.6.

5.4.1 Slow fading channel

Let us first look at the situation when the channel gain is random but remains constant for all time, i.e., $h[m] = h$ for all m . This models the *slow fading* situation where the delay requirement is short compared to the channel coherence time (cf. Table 2.2). This is also called the *quasi-static* scenario.

Conditional on a realization of the channel h , this is an AWGN channel with received signal-to-noise ratio $|h|^2\text{SNR}$. The maximum rate of reliable communication supported by this channel is $\log(1 + |h|^2\text{SNR})$ bits/s/Hz. This quantity is a function of the random channel gain h and is therefore random (Figure 5.13). Now suppose the transmitter encodes data at a rate R bits/s/Hz. If the channel realization h is such that $\log(1 + |h|^2\text{SNR}) < R$, then whatever the code used by the transmitter, the decoding error probability cannot be made arbitrarily small. The system is said to be *in outage*, and the outage probability is

$$p_{\text{out}}(R) := \mathbb{P}\{\log(1 + |h|^2\text{SNR}) < R\}. \quad (5.54)$$

Thus, the best the transmitter can do is to encode the data assuming that the channel gain is strong enough to support the desired rate R . Reliable communication can be achieved whenever that happens, and outage occurs otherwise.

A more suggestive interpretation is to think of the channel as allowing $\log(1 + |h|^2\text{SNR})$ bits/s/Hz of information through when the fading gain is h .

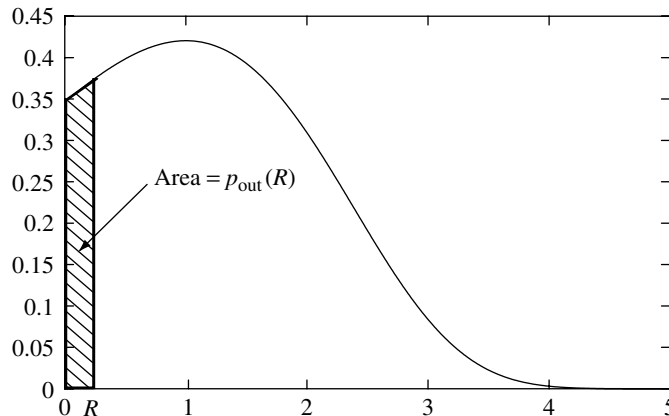


Figure 5.13 Density of $\log(1 + |h|^2\text{SNR})$, for Rayleigh fading and $\text{SNR} = 0$ dB. For any target rate R , there is a non-zero outage probability.

Reliable decoding is possible as long as this amount of information exceeds the target rate.

For Rayleigh fading (i.e., h is $\mathcal{CN}(0, 1)$), the outage probability is

$$p_{\text{out}}(R) = 1 - \exp\left(\frac{-(2^R - 1)}{\text{SNR}}\right). \quad (5.55)$$

At high SNR,

$$p_{\text{out}}(R) \approx \frac{(2^R - 1)}{\text{SNR}}, \quad (5.56)$$

and the outage probability decays as $1/\text{SNR}$. Recall that when we discussed uncoded transmission in Section 3.1.2, the detection error probability also decays like $1/\text{SNR}$. Thus, we see that coding *cannot* significantly improve the error probability in a slow fading scenario. The reason is that while coding can average out the Gaussian white noise, it cannot average out the channel fade, which affects *all* the coded symbols. Thus, deep fade, which is the typical error event in the uncoded case, is also the typical error event in the coded case.

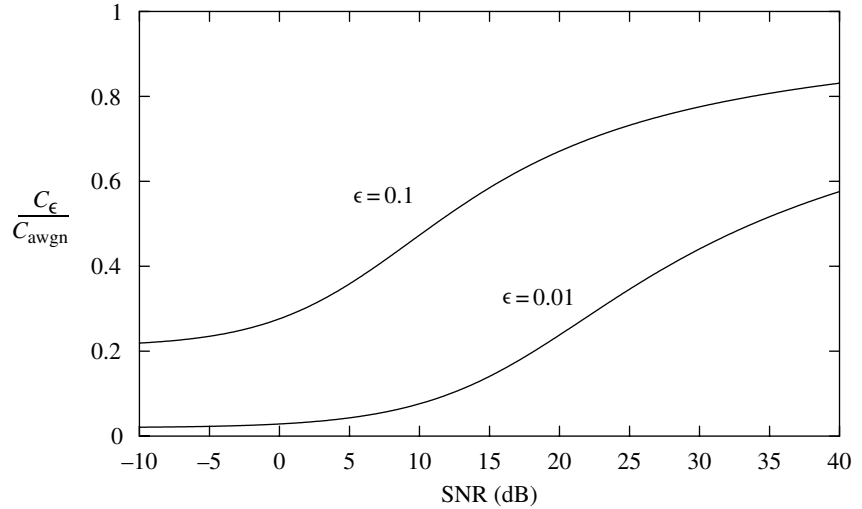
There is a conceptual difference between the AWGN channel and the slow fading channel. In the former, one can send data at a positive rate (in fact, any rate less than C) while making the error probability as small as desired. This cannot be done for the slow fading channel as long as the probability that the channel is in deep fade is non-zero. Thus, the capacity of the slow fading channel in the strict sense is zero. An alternative performance measure is the ϵ -outage capacity C_ϵ . This is the largest rate of transmission R such that the outage probability $p_{\text{out}}(R)$ is less than ϵ . Solving $p_{\text{out}}(R) = \epsilon$ in (5.54) yields

$$C_\epsilon = \log(1 + F^{-1}(1 - \epsilon) \text{SNR}) \text{ bits/s/Hz}, \quad (5.57)$$

where F is the complementary cumulative distribution function of $|h|^2$, i.e., $F(x) := \mathbb{P}\{|h|^2 > x\}$.

In Section 3.1.2, we looked at uncoded transmission and there it was natural to focus only on the high SNR regime; at low SNR, the error probability of uncoded transmission is very poor. On the other hand, for coded systems, it makes sense to consider both the high and the low SNR regimes. For example, the CDMA system in Chapter 4 operates at very low SINR and uses very low-rate orthogonal coding. A natural question is: in which regime does fading have a more significant impact on outage performance? One can answer this question in two ways. Eqn (5.57) says that, to achieve the same rate as the AWGN channel, an extra $10 \log(1/F^{-1}(1 - \epsilon))$ dB of power is needed. This is true regardless of the operating SNR of the environment. Thus the *fade margin* is the same at all SNRs. If we look at the outage capacity at a *given* SNR, however, the impact of fading depends very much on the operating regime. To get a sense, Figure 5.14 plots the ϵ -outage capacity as

Figure 5.14 ϵ -outage capacity as a fraction of AWGN capacity under Rayleigh fading, for $\epsilon = 0.1$ and $\epsilon = 0.01$.



a function of SNR for the Rayleigh fading channel. To assess the impact of fading, the ϵ -outage capacity is plotted as a fraction of the AWGN capacity at the same SNR. It is clear that the impact is much more significant in the low SNR regime. Indeed, at high SNR,

$$C_{\epsilon} \approx \log \text{SNR} + \log(F^{-1}(1 - \epsilon)) \quad (5.58)$$

$$\approx C_{\text{awgn}} - \log\left(\frac{1}{F^{-1}(1 - \epsilon)}\right), \quad (5.59)$$

a constant *difference* irrespective of the SNR. Thus, the relative loss gets smaller at high SNR. At low SNR, on the other hand,

$$C_{\epsilon} \approx F^{-1}(1 - \epsilon)\text{SNR} \log_2 e \quad (5.60)$$

$$\approx F^{-1}(1 - \epsilon)C_{\text{awgn}}. \quad (5.61)$$

For reasonably small outage probabilities, the outage capacity is only a small fraction of the AWGN capacity at low SNR. For Rayleigh fading, $F^{-1}(1 - \epsilon) \approx \epsilon$ for small ϵ and the impact of fading is very significant. At an outage probability of 0.01, the outage capacity is only 1% of the AWGN capacity! Diversity has a significant effect at high SNR (as already seen in Chapter 3), but can be more important at low SNR. Intuitively, the impact of the randomness of the channel is in the received SNR, and the reliable rate supported by the AWGN channel is much more sensitive to the received SNR at low SNR than at high SNR. Exercise 5.10 elaborates on this point.

5.4.2 Receive diversity

Let us increase the diversity of the channel by having L receive antennas instead of one. For given channel gains $\mathbf{h} := [h_1, \dots, h_L]^t$, the capacity was

calculated in Section 5.3.1 to be $\log(1 + \|\mathbf{h}\|^2 \text{SNR})$. Outage occurs whenever this is below the target rate R :

$$p_{\text{out}}^{\text{rx}}(R) := \mathbb{P}\{\log(1 + \|\mathbf{h}\|^2 \text{SNR}) < R\}. \quad (5.62)$$

This can be rewritten as

$$p_{\text{out}}(R) = \mathbb{P}\left\{\|\mathbf{h}\|^2 < \frac{2^R - 1}{\text{SNR}}\right\}. \quad (5.63)$$

Under independent Rayleigh fading, $\|\mathbf{h}\|^2$ is a sum of the squares of $2L$ independent Gaussian random variables and is distributed as Chi-square with $2L$ degrees of freedom. Its density is

$$f(x) = \frac{1}{(L-1)!} x^{L-1} e^{-x}, \quad x \geq 0. \quad (5.64)$$

Approximating e^{-x} by 1 for x small, we have (cf. (3.44)),

$$\mathbb{P}\{\|\mathbf{h}\|^2 < \delta\} \approx \frac{1}{L!} \delta^L, \quad (5.65)$$

for δ small. Hence at high SNR the outage probability is given by

$$p_{\text{out}}(R) \approx \frac{(2^R - 1)^L}{L! \text{SNR}^L}. \quad (5.66)$$

Comparing with (5.55), we see a diversity gain of L : the outage probability now decays like $1/\text{SNR}^L$. This parallels the performance of uncoded transmission discussed in Section 3.3.1: thus, coding cannot increase the diversity gain.

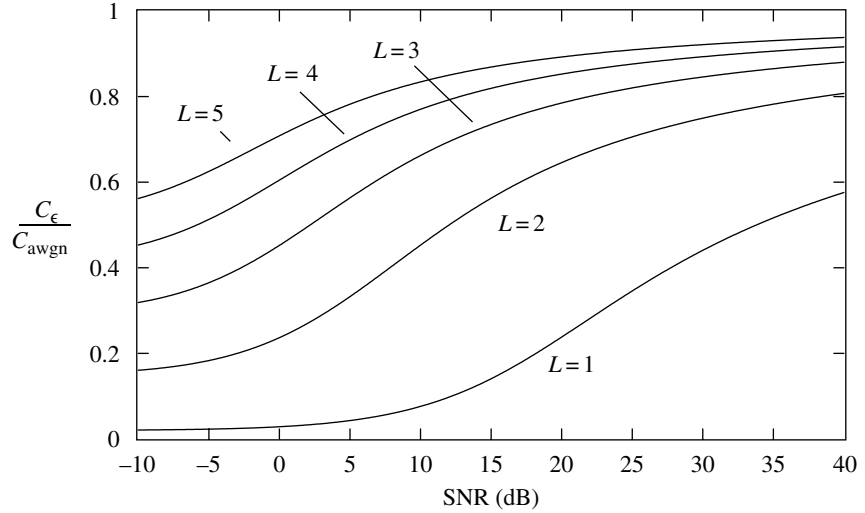
The impact of receive diversity on the ϵ -outage capacity is plotted in Figure 5.15. The ϵ -outage capacity is given by (5.57) with F now the cumulative distribution function of $\|\mathbf{h}\|^2$. Receive antennas yield a diversity gain and an L -fold power gain. To emphasize the impact of the diversity gain, let us normalize the outage capacity C_ϵ by $C_{\text{awgn}} = \log(1 + L\text{SNR})$. The dramatic salutary effect of diversity on outage capacity can now be seen. At low SNR and small ϵ , (5.61) and (5.65) yield

$$C_\epsilon \approx F^{-1}(1 - \epsilon) \text{SNR} \log_2 e \quad (5.67)$$

$$\approx (L!)^{\frac{1}{L}} (\epsilon)^{\frac{1}{L}} \text{SNR} \log_2 e \text{ bits/s/Hz} \quad (5.68)$$

and the loss with respect to the AWGN capacity is by a factor of $\epsilon^{1/L}$ rather than by ϵ when there is no diversity. At $\epsilon = 0.01$ and $L = 2$, the outage capacity is increased to 14% of the AWGN capacity (as opposed to 1% for $L = 1$).

Figure 5.15 ϵ -outage capacity with L -fold receive diversity, as a fraction of the AWGN capacity $\log(1 + L\text{SNR})$, for $\epsilon = 0.01$ and different L .



5.4.3 Transmit diversity

Now suppose there are L transmit antennas but only one receive antenna, with a total power constraint of P . From Section 5.3.2, the capacity of the channel conditioned on the channel gains $\mathbf{h} = [h_1, \dots, h_L]^t$ is $\log(1 + \|\mathbf{h}\|^2 \text{SNR})$. Following the approach taken in the SISO and the SIMO cases, one is tempted to say that the outage probability for a fixed rate R is

$$p_{\text{out}}^{\text{full-csi}}(R) = \mathbb{P}\{\log(1 + \|\mathbf{h}\|^2 \text{SNR}) < R\}, \quad (5.69)$$

which would have been exactly the same as the corresponding SIMO system with 1 transmit and L receive antennas. However, this outage performance is achievable only if the transmitter knows the phases and magnitudes of the gains \mathbf{h} so that it can perform transmit beamforming, i.e., allocate more power to the stronger antennas and arrange the signals from the different antennas to align in phase at the receiver. When the transmitter does not know the channel gains \mathbf{h} , it has to use a fixed transmission strategy that does not depend on \mathbf{h} . (This subtlety does not arise in either the SISO or the SIMO case because the transmitter need not know the channel realization to achieve the capacity for those channels.) How much performance loss does not knowing the channel entail?

Alamouti scheme revisited

For concreteness, let us focus on $L = 2$ (dual transmit antennas). In this situation, we can use the Alamouti scheme, which extracts transmit diversity without transmitter channel knowledge (introduced in Section 3.3.2). Recall from (3.76) that, under this scheme, both the transmitted symbols u_1, u_2 over a block of 2 symbol times see an equivalent scalar fading channel with gain $\|\mathbf{h}\|$

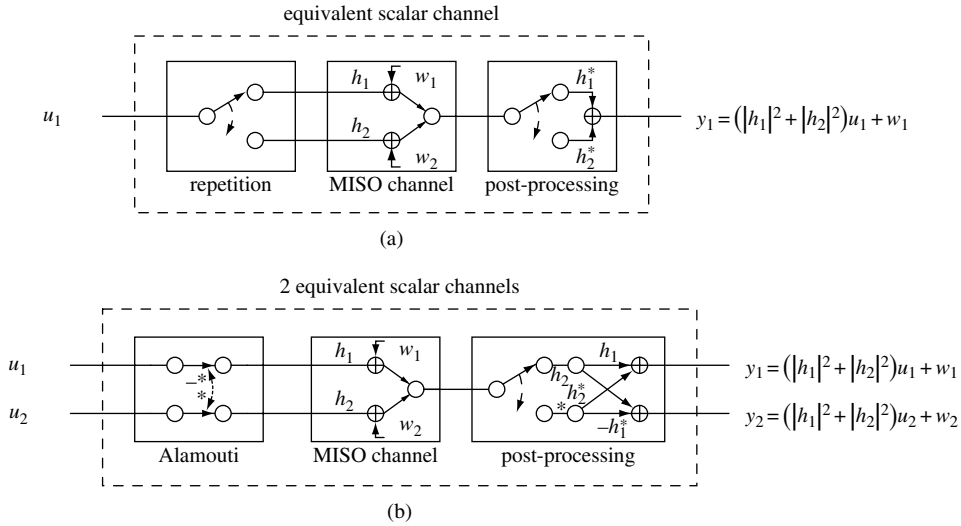


Figure 5.16 A space-time coding scheme combined with the MISO channel can be viewed as an equivalent scalar channel: (a) repetition coding; (b) the Alamouti scheme. The outage probability of the scheme is the outage probability of the equivalent channel.

and additive noise $\mathcal{CN}(0, N_0)$ (Figure 5.16(b)). The energy in the symbols u_1 and u_2 is $P/2$. Conditioned on h_1, h_2 , the capacity of the equivalent scalar channel is

$$\log\left(1 + \|\mathbf{h}\|^2 \frac{\text{SNR}}{2}\right) \text{ bits/s/Hz.} \quad (5.70)$$

Thus, if we now consider successive blocks and use an AWGN capacity-achieving code of rate R over each of the streams $\{u_1[m]\}$ and $\{u_2[m]\}$ separately, then the outage probability of each stream is

$$p_{\text{out}}^{\text{Ala}}(R) = \mathbb{P}\left\{\log\left(1 + \|\mathbf{h}\|^2 \frac{\text{SNR}}{2}\right) < R\right\}. \quad (5.71)$$

Compared to (5.69) when the transmitter knows the channel, the Alamouti scheme performs strictly worse: the loss is 3 dB in the received SNR. This can be explained in terms of the efficiency with which energy is transferred to the receiver. In the Alamouti scheme, the symbols sent at the two transmit antennas in each time are *independent* since they come from two separately coded streams. Each of them has power $P/2$. Hence, the total SNR at the receive antenna at any given time is

$$\frac{(|h_1|^2 + |h_2|^2) \text{SNR}}{2}. \quad (5.72)$$

In contrast, when the transmitter knows the channel, the symbols transmitted at the two antennas are *completely correlated* in such a way that the signals add up in phase at the receive antenna and the SNR is now

$$(|h_1|^2 + |h_2|^2) \text{SNR},$$

a 3-dB power gain over the independent case.⁴ Intuitively, there is a power loss because, without channel knowledge, the transmitter is sending signals that have energy in *all* directions instead of focusing the energy in a specific direction. In fact, the Alamouti scheme radiates energy in a perfectly *isotropic* manner: the signal transmitted from the two antennas has the same energy when projected in any direction (Exercise 5.14).

A scheme radiates energy isotropically whenever the signals transmitted from the antennas are uncorrelated and have equal power (Exercise 5.14). Although the Alamouti scheme does not perform as well as transmit beamforming, it is optimal in one important sense: it has the *best* outage probability among all schemes that radiate energy isotropically. Indeed, any such scheme must have a received SNR equal to (5.72) and hence its outage performance must be no better than that of a scalar slow fading AWGN channel with that received SNR. But this is precisely the performance achieved by the Alamouti scheme.

Can one do even better by radiating energy in a non-isotropic manner (but in a way that does not depend on the random channel gains)? In other words, can one improve the outage probability by correlating the signals from the transmit antennas and/or allocating unequal powers on the antennas? The answer depends of course on the distribution of the gains h_1, h_2 . If h_1, h_2 are i.i.d. Rayleigh, Exercise 5.15 shows, using symmetry considerations, that correlation never improves the outage performance, but it is not necessarily optimal to use all the transmit antennas. Exercise 5.16 shows that uniform power allocation across antennas is always optimal, but the number of antennas used depends on the operating SNR. For reasonable values of target outage probabilities, it is optimal to use all the antennas. This implies that in most cases of interest, *the Alamouti scheme has the optimal outage performance for the i.i.d. Rayleigh fading channel.*

What about for $L > 2$ transmit antennas? An information theoretic argument in Appendix B.8 shows (in a more general framework) that

$$p_{\text{out}}(R) = \mathbb{P} \left\{ \log \left(1 + \|\mathbf{h}\|^2 \frac{\text{SNR}}{L} \right) < R \right\} \quad (5.73)$$

is achievable. This is the natural generalization of (5.71) and corresponds again to isotropic transmission of energy from the antennas. Again, Exercises 5.15 and 5.16 show that this strategy is optimal for the i.i.d. Rayleigh fading channel and for most target outage probabilities of interest. However, there is no natural generalization of the Alamouti scheme for a larger number of transmit antennas (cf. Exercise 3.17). We will return to the problem of outage-optimal code design for $L > 2$ in Chapter 9.

⁴ The addition of two in-phase signals of equal power yields a sum signal that has double the amplitude and four times the power of each of the signals. In contrast, the addition of two independent signals of equal power only *doubles* the power.

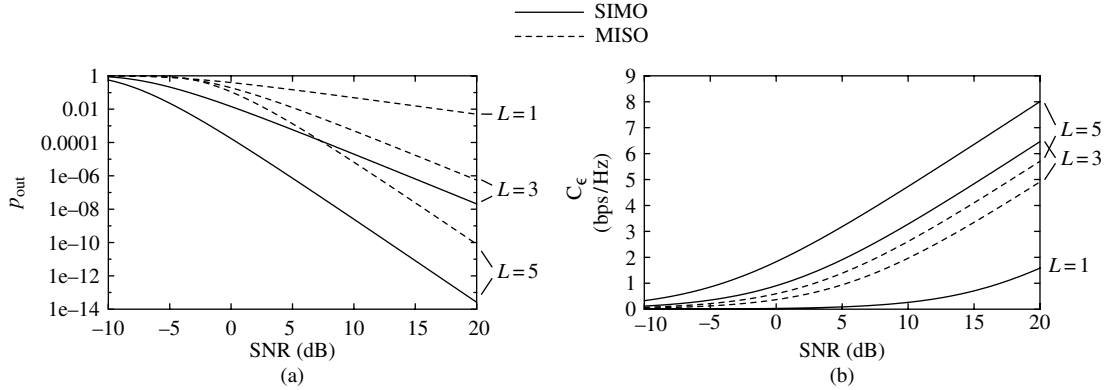


Figure 5.17 Comparison of outage performance between SIMO and MISO channels for different L : (a) outage probability as a function of SNR, for fixed $R = 1$; (b) outage capacity as a function of SNR, for a fixed outage probability of 10^{-2} .

The outage performances of the SIMO and the MISO channels with i.i.d. Rayleigh gains are plotted in Figure 5.17 for different numbers of transmit antennas. The difference in outage performance clearly outlines the asymmetry between receive and transmit antennas caused by the transmitter lacking knowledge of the channel.

Suboptimal schemes: repetition coding

In the above, the Alamouti scheme is viewed as an *inner code* that converts the MISO channel into a scalar channel. The outage performance (5.71) is achieved when the Alamouti scheme is used in conjunction with an *outer code* that is capacity-achieving for the scalar AWGN channel. Other space-time schemes can be similarly used as inner codes and their outage probability analyzed and compared to the channel outage performance.

Here we consider the simplest example, the repetition scheme: the same symbol is transmitted over the L different antennas over L symbol periods, using only one antenna at a time to transmit. The receiver does maximal ratio combining to demodulate each symbol. As a result, each symbol sees an equivalent scalar fading channel with gain $\|\mathbf{h}\|$ and noise variance N_0 (Figure 5.16(a)). Since only one symbol is transmitted every L symbol periods, a rate of LR bits/symbol is required on this scalar channel to achieve a target rate of R bits/symbol on the original channel. The outage probability of this scheme, when combined with an outer capacity-achieving code, is therefore:

$$p_{\text{out}}^{\text{rep}}(R) = \mathbb{P} \left\{ \frac{1}{L} \log(1 + \|\mathbf{h}\|^2 \text{SNR}) < R \right\}. \quad (5.74)$$

Compared to the outage probability (5.73) of the *channel*, this scheme is suboptimal: the SNR has to be increased by a factor of

$$\frac{L(2^R - 1)}{2^{LR} - 1}, \quad (5.75)$$

to achieve the same outage probability for the same target rate R . Equivalently, the reciprocal of this ratio can be interpreted as the maximum achievable *coding gain* over the simple repetition scheme. For a fixed R , the performance loss increases with L : the repetition scheme becomes increasingly inefficient in using the degrees of freedom of the channel. For a fixed L , the performance loss increases with the target rate R . On the other hand, for R small, $2^R - 1 \approx R \ln 2$ and $2^{RL} - 1 \approx RL \ln 2$, so

$$\frac{L(2^R - 1)}{2^{LR} - 1} \approx \frac{LR \ln 2}{LR \ln 2} = 1, \quad (5.76)$$

and there is hardly any loss in performance. Thus, while the repetition scheme is very suboptimal in the high SNR regime where the target rate can be high, it is nearly optimal in the low SNR regime. This is not surprising: the system is degree-of-freedom limited in the high SNR regime and the inefficiency of the repetition scheme is felt more there.

Summary 5.2 Transmit and receive diversity

With receive diversity, the outage probability is

$$p_{\text{out}}^{\text{rx}}(R) := \mathbb{P}\{\log(1 + \|\mathbf{h}\|^2 \text{SNR}) < R\}. \quad (5.77)$$

With transmit diversity and isotropic transmission, the outage probability is

$$p_{\text{out}}^{\text{tx}}(R) := \mathbb{P}\left\{\log\left(1 + \|\mathbf{h}\|^2 \frac{\text{SNR}}{L}\right) < R\right\}, \quad (5.78)$$

a loss of a factor of L in the received SNR because the transmitter has no knowledge of the channel direction and is unable to beamform in the specific channel direction.

With two transmit antennas, capacity-achieving AWGN codes in conjunction with the Alamouti scheme achieve the outage probability.

5.4.4 Time and frequency diversity

Outage performance of parallel channels

Another way to increase channel diversity is to exploit the time-variation of the channel: in addition to coding over symbols within one coherence period, one can code over symbols from L such periods. Note that this is a generalization of the schemes considered in Section 3.2, which take *one* symbol from each coherence period. When coding can be performed over

many symbols from each period, as well as between symbols from different periods, what is the performance limit?

One can model this situation using the idea of *parallel channels* introduced in Section 5.3.3: each of the sub-channels, $\ell = 1, \dots, L$, represents a coherence period of duration T_c symbols:

$$y_\ell[m] = h_\ell x_\ell[m] + w_\ell[m], \quad m = 1, \dots, T_c. \quad (5.79)$$

Here h_ℓ is the (non-varying) channel gain during the ℓ th coherence period. It is assumed that the coherence time T_c is large such that one can code over many symbols in each of the sub-channels. An average transmit power constraint of P on the original channel translates into a total power constraint of LP on the parallel channel.

For a given realization of the channel, we have already seen in Section 5.3.3 that the optimal power allocation across the sub-channels is *waterfilling*. However, since the transmitter does not know what the channel gains are, a reasonable strategy is to allocate equal power P to each of the sub-channels. In Section 5.3.3, it was mentioned that the maximum rate of reliable communication given the fading gains h_ℓ is

$$\sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}) \text{ bits/s/Hz}, \quad (5.80)$$

where $\text{SNR} = P/N_0$. Hence, if the target rate is R bits/s/Hz per sub-channel, then outage occurs when

$$\sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}) < LR. \quad (5.81)$$

Can one design a code to communicate reliably whenever

$$\sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}) > LR? \quad (5.82)$$

If so, an L -fold diversity is achieved for i.i.d. Rayleigh fading: outage occurs only if each of the terms in the sum $\sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR})$ is small.

The term $\log(1 + |h_\ell|^2 \text{SNR})$ is the capacity of an AWGN channel with received SNR equal to $|h_\ell|^2 \text{SNR}$. Hence, a seemingly straightforward strategy, already used in Section 5.3.3, would be to use a capacity-achieving AWGN code with rate

$$\log(1 + |h_\ell|^2 \text{SNR})$$

for the ℓ th coherence period, yielding an average rate of

$$\frac{1}{L} \sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}) \text{ bits/s/Hz}$$

and meeting the target rate whenever condition (5.82) holds. The caveat is that this strategy requires the transmitter to know in advance the channel state during each of the coherence periods so that it can adapt the rate it allocates to each period. This knowledge is not available. *However, it turns out that such transmitter adaptation is unnecessary: information theory guarantees that one can design a single code that communicates reliably at rate R whenever the condition (5.82) is met.* Hence, the outage probability of the time diversity channel is precisely

$$p_{\text{out}}(R) = \mathbb{P} \left\{ \frac{1}{L} \sum_{\ell=1}^L \log(1 + |h_{\ell}|^2 \text{SNR}) < R \right\}. \quad (5.83)$$

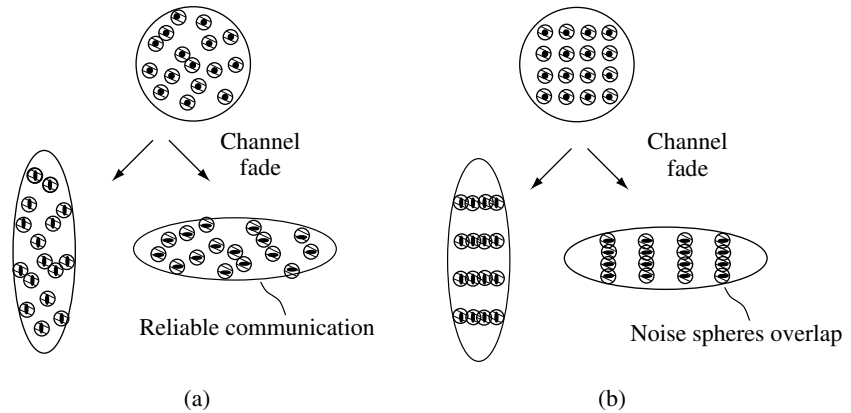
Even though this outage performance can be achieved with or without transmitter knowledge of the channel, the coding strategy is vastly different. With transmitter knowledge of the channel, dynamic rate allocation and separate coding for each sub-channel suffices. Without transmitter knowledge, separate coding would mean using a fixed-rate code for each sub-channel and poor diversity results: errors occur whenever one of the sub-channels is bad. Indeed, coding *across* the different coherence periods is now necessary: if the channel is in deep fade during one of the coherence periods, the information bits can still be protected if the channel is strong in other periods.

A geometric view

Figure 5.18 gives a geometric view of our discussion so far. Consider a code with rate R , coding over all the sub-channels and over one coherence time-interval; the block length is LT_c symbols. The codewords lie in an LT_c -dimensional sphere. The received LT_c -dimensional signal lives in an ellipsoid, with (L groups of) different axes stretched and shrunk by the different sub-channel gains (cf. Section 5.3.3). The ellipsoid is a function of the sub-channel gains, and hence random. The no-outage condition (5.82) has a geometric interpretation: it says that the volume of the ellipsoid is large enough to contain $2^{LT_c R}$ noise spheres, one for each codeword. (This was already seen in the sphere-packing argument in Section 5.3.3.) An outage-optimal code is one that communicates reliably whenever the random ellipsoid is at least this large. The subtlety here is that the *same* code must work for all such ellipsoids. Since the shrinking can occur in any of the L groups of dimensions, a robust code needs to have the property that the codewords are simultaneously well-separated in *each* of the sub-channels (Figure 5.18(a)). A set of independent codes, one for each sub-channel, is not robust: errors will be made when even only one of the sub-channels fades (Figure 5.18(b)).

We have already seen, in the simple context of Section 3.2, codes for the parallel channel which are designed to be well-separated in all the sub-channels. For example, the repetition code and the rotation code in Figure 3.8 have the property that the codewords are separated in both the sub-channels

Figure 5.18 Effect of the fading gains on codes for the parallel channel. Here there are $L = 2$ sub-channels and each axis represents T_c dimensions within a sub-channel. (a) Coding across the sub-channels. The code works as long as the volume of the ellipsoid is big enough. This requires good codeword separation in both the sub-channels. (b) Separate, non-adaptive code for each sub-channel. Shrinking of one of the axes is enough to cause confusion between the codewords.



(here $T_c = 1$ symbol and $L = 2$ sub-channels). More generally, the code design criterion of maximizing the product distance for all pairs of codewords naturally favors codes that satisfy this property. Coding over long blocks affords a larger coding gain; information theory guarantees the existence of codes with large enough coding gain to achieve the outage probability in (5.83).

To achieve the outage probability, one wants to design a code that communicates reliably over *every* parallel channel that is not in outage (i.e., parallel channels that satisfy (5.82)). In information theory jargon, a code that communicates reliably for a class of channels is said to be *universal* for that class. In this language, we are looking for universal codes for parallel channels that are not in outage. In the slow fading scalar channel without diversity ($L = 1$), this problem is the same as the code design problem for a *specific* channel. This is because all scalar channels are ordered by their received SNR; hence a code that works for the channel that is just strong enough to support the target rate will automatically work for all better channels. For parallel channels, each channel is described by a vector of channel gains and there is no natural ordering of channels; the universal code design problem is now non-trivial. In Chapter 9, a universal code design criterion will be developed to construct universal codes that come close to achieving the outage probability.

Extensions

In the above development, a uniform power allocation across the sub-channels is assumed. Instead, if we choose to allocate power P_ℓ to sub-channel ℓ , then the outage probability (5.83) generalizes to

$$p_{\text{out}}(R) = \mathbb{P} \left\{ \sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}_\ell) < LR \right\}, \quad (5.84)$$

where $\text{SNR}_\ell = P_\ell/N_0$. Exercise 5.17 shows that for the i.i.d. Rayleigh fading model, a non-uniform power allocation that does not depend on the channel gains cannot improve the outage performance.

The parallel channel is used to model time diversity, but it can model frequency diversity as well. By using the usual OFDM transformation, a slow frequency-selective fading channel can be converted into a set of parallel sub-channels, one for each sub-carrier. This allows us to characterize the outage capacity of such channels as well (Exercise 5.22).

We summarize the key idea in this section using more suggestive language.

Summary 5.3 Outage for parallel channels

Outage probability for a parallel channel with L sub-channels and the ℓ th channel having random gain h_ℓ :

$$p_{\text{out}}(R) = \mathbb{P} \left\{ \frac{1}{L} \sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}) < R \right\}, \quad (5.85)$$

where R is in bits/s/Hz per sub-channel.

The ℓ th sub-channel allows $\log(1 + |h_\ell|^2 \text{SNR})$ bits of information per symbol through. Reliable decoding can be achieved as long as the *total* amount of information allowed through exceeds the target rate.

5.4.5 Fast fading channel

In the slow fading scenario, the channel remains constant over the transmission duration of the codeword. If the codeword length spans several coherence periods, then time diversity is achieved and the outage probability improves. When the codeword length spans *many* coherence periods, we are in the so-called *fast fading* regime. How does one characterize the performance limit of such a fast fading channel?

Capacity derivation

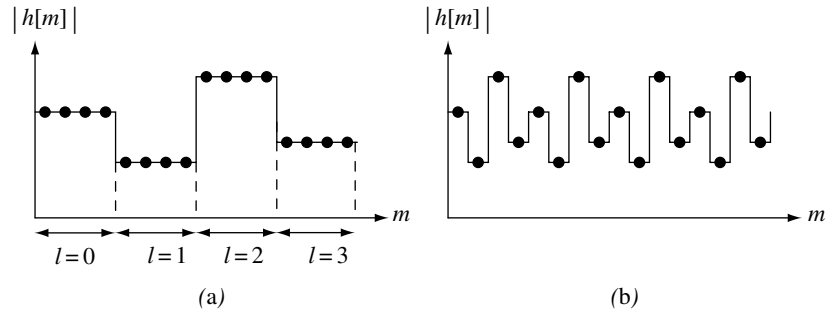
Let us first consider a very simple model of a fast fading channel:

$$y[m] = h[m]x[m] + w[m], \quad (5.86)$$

where $h[m] = h_\ell$ remains constant over the ℓ th coherence period of T_c symbols and is i.i.d. across different coherence periods. This is the so-called *block fading* model; see Figure 5.19(a). Suppose coding is done over L such coherence periods. If $T_c \gg 1$, we can effectively model this as L parallel sub-channels that fade independently. The outage probability from (5.83) is

$$p_{\text{out}}(R) = \mathbb{P} \left\{ \frac{1}{L} \sum_{\ell=1}^L \log(1 + |h_\ell|^2 \text{SNR}) < R \right\}. \quad (5.87)$$

Figure 5.19 (a) Typical trajectory of the channel strength as a function of symbol time under a block fading model. (b) Typical trajectory of the channel strength after interleaving. One can equally think of these plots as rates of flow of information allowed through the channel over time.



For finite L , the quantity

$$\frac{1}{L} \sum_{\ell=1}^L \log(1 + |h_{\ell}|^2 \text{SNR})$$

is random and there is a non-zero probability that it will drop below any target rate R . Thus, there is no meaningful notion of capacity in the sense of maximum rate of arbitrarily reliable communication and we have to resort to the notion of outage. However, as $L \rightarrow \infty$, the law of large numbers says that

$$\frac{1}{L} \sum_{\ell=1}^L \log(1 + |h_{\ell}|^2 \text{SNR}) \rightarrow \mathbb{E}[\log(1 + |h|^2 \text{SNR})]. \quad (5.88)$$

Now we can average over many independent fades of the channel by coding over a large number of coherence time intervals and a reliable rate of communication of $\mathbb{E}[\log(1 + |h|^2 \text{SNR})]$ can indeed be achieved. In this situation, it is now meaningful to assign a positive capacity to the fast fading channel:

$$C = \mathbb{E}[\log(1 + |h|^2 \text{SNR})] \text{ bits/s/Hz} \quad (5.89)$$

Impact of interleaving

In the above, we considered codes with block lengths LT_c symbols, where L is the number of coherence periods and T_c is the number of symbols in each coherence block. To approach the capacity of the fast fading channel, L has to be large. Since T_c is typically also a large number, the overall block length may become prohibitively large for implementation. In practice, shorter codes are used but they are interleaved so that the symbols of each codeword are spaced far apart in time and lie in different coherence periods. (Such interleaving is used for example in the IS-95 CDMA system, as illustrated in Figure 4.4.) Does interleaving impart a performance loss in terms of capacity?

Going back to the channel model (5.86), ideal interleaving can be modeled by assuming the $h[m]$ are now i.i.d., i.e., successive interleaved symbols go through independent fades. (See Figure 5.19(b).) In Appendix B.7.1, it is

shown that for a large block length N and a given realization of the fading gains $h[1], \dots, h[N]$, the maximum achievable rate through this interleaved channel is

$$\frac{1}{N} \sum_{m=1}^N \log(1 + |h[m]|^2 \text{SNR}) \text{ bits/s/Hz.} \quad (5.90)$$

By the law of large numbers,

$$\frac{1}{N} \sum_{m=1}^N \log(1 + |h[m]|^2 \text{SNR}) \rightarrow \mathbb{E}[\log(1 + |h|^2 \text{SNR})] \quad (5.91)$$

as $N \rightarrow \infty$, for almost all realizations of the random channel gains. Thus, even with interleaving, the capacity (5.89) of the fast fading channel can be achieved. The important benefit of interleaving is that this capacity can now be achieved with a much shorter block length.

A closer examination of the above argument reveals why the capacity under interleaving (with $\{h[m]\}$ i.i.d.) and the capacity of the original block fading model (with $\{h[m]\}$ block-wise constant) are the same: the convergence in (5.91) holds for both fading processes, allowing the same long-term average rate through the channel. If one thinks of $\log(1 + |h[m]|^2 \text{SNR})$ as the rate of information flow allowed through the channel at time m , the only difference is that in the block fading model, the rate of information flow is constant over each coherence period, while in the interleaved model, the rate varies from symbol to symbol. See Figure 5.19 again.

This observation suggests that the capacity result (5.89) holds for a much broader class of fading processes. Only the convergence in (5.91) is needed. This says that the time average should converge to the same limit for almost all realizations of the fading process, a concept called *ergodicity*, and it holds in many models. For example, it holds for the Gaussian fading model mentioned in Section 2.4. What matters from the point of view of capacity is only the long-term time average rate of flow allowed, and not on how fast that rate fluctuates over time.

Discussion

In the earlier parts of the chapter, we focused exclusively on deriving the capacities of *time-invariant* channels, particularly the AWGN channel. We have just shown that *time-varying* fading channels also have a well-defined capacity. However, the operational significance of capacity in the two cases is quite different. In the AWGN channel, information flows at a *constant* rate of $\log(1 + \text{SNR})$ through the channel, and reliable communication can take place as long as the coding block length is large enough to average out the white Gaussian noise. The resulting coding/decoding delay is typically much smaller than the delay requirement of applications and this is not a big concern. In the fading channel, on the other hand, information flows

at a *variable* rate of $\log(1 + |h[m]|^2 \text{SNR})$ due to variations of the channel strength; the coding block length now needs to be large enough to average out *both* the Gaussian noise *and* the fluctuations of the channel. To average out the latter, the coded symbols must span many coherence time periods, and this coding/decoding delay can be quite significant. Interleaving reduces the block length but not the coding/decoding delay: one still needs to wait many coherence periods before the bits get decoded. For applications that have a tight delay constraint relative to the channel coherence time, this notion of capacity is not meaningful, and one will suffer from outage.

The capacity expression (5.89) has the following interpretation. Consider a family of codes, one for each possible fading state h , and the code for state h achieves the capacity $\log(1 + |h|^2 \text{SNR})$ bits/s/Hz of the AWGN channel at the corresponding received SNR level. From these codes, we can build a variable-rate coding scheme that adaptively selects a code of appropriate rate depending on what the current channel condition is. This scheme would then have an average throughput of $\mathbb{E}[\log(1 + |h|^2 \text{SNR})]$ bits/s/Hz. For this variable-rate scheme to work, however, the transmitter needs to know the current channel state. The significance of the fast fading capacity result (5.89) is that one can communicate reliably at this rate even when the transmitter is blind and cannot track the channel.⁵

The nature of the information theoretic result that guarantees a code which achieves the capacity of the fast fading channel is similar to what we have already seen in the outage performance of the slow fading channel (cf. (5.83)). In fact, information theory guarantees that a fixed code with the rate in (5.89) is *universal* for the class of ergodic fading processes (i.e., (5.91) is satisfied with the same limiting value). This class of processes includes the AWGN channel (where the channel is fixed for all time) and, at the other extreme, the interleaved fast fading channel (where the channel varies i.i.d. over time). This suggests that capacity-achieving AWGN channel codes (cf. Discussion 5.1) could be suitable for the fast fading channel as well. While this is still an active research area, LDPC codes have been adapted successfully to the fast Rayleigh fading channel.

Performance comparison

Let us explore a few implications of the capacity result (5.89) by comparing it with that for the AWGN channel. The capacity of the fading channel is always less than that of the AWGN channel with the same SNR. This follows directly from Jensen's inequality, which says that if f is a strictly concave function and u is any random variable, then $\mathbb{E}[f(u)] \leq f(\mathbb{E}[u])$, with equality if and only if u is deterministic (Exercise B.2). Intuitively, the gain from

⁵ Note however that if the transmitter can really track the channel, one can do even better than this rate. We will see this next in Section 5.4.6.

the times when the channel strength is above the average cannot compensate for the loss from the times when the channel strength is below the average. This again follows from the law of diminishing marginal return on capacity from increasing the received power.

At low SNR, the capacity of the fading channel is

$$C = \mathbb{E}[\log(1 + |h|^2 \text{SNR})] \approx \mathbb{E}[|h|^2 \text{SNR}] \log_2 e = \text{SNR} \log_2 e \approx C_{\text{awgn}}, \quad (5.92)$$

where C_{awgn} is the capacity of the AWGN channel and is measured in bits per symbol. Hence at low SNR the “Jensen’s loss” becomes negligible; this is because the capacity is approximately linear in the received SNR in this regime. At high SNR,

$$C \approx \mathbb{E}[\log(|h|^2 \text{SNR})] = \log \text{SNR} + \mathbb{E}[\log |h|^2] \approx C_{\text{awgn}} + \mathbb{E}[\log |h|^2], \quad (5.93)$$

i.e., a constant difference with the AWGN capacity at high SNR. This difference is -0.83 bits/s/Hz for the Rayleigh fading channel. Equivalently, 2.5 dB more power is needed in the fading case to achieve the same capacity as in the AWGN case. Figure 5.20 compares the capacity of the Rayleigh fading channel with the AWGN capacity as a function of the SNR. The difference is not that large for the entire plotted range of SNR.

5.4.6 Transmitter side information

So far we have assumed that only the receiver can track the channel. But let us now consider the case when the *transmitter* can track the channel as well. There are several ways in which such channel information can be obtained at the transmitter. In a TDD (time-division duplex) system, the transmitter

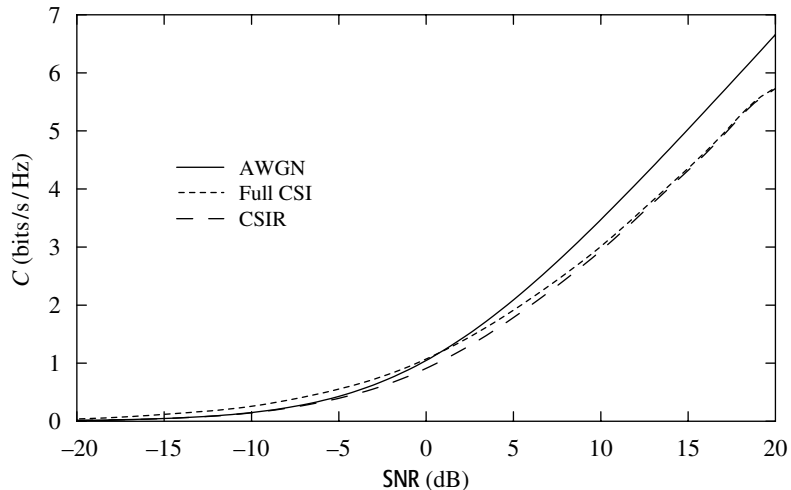


Figure 5.20 Plot of AWGN capacity, fading channel capacity with receiver tracking the channel only (CSIR) and capacity with both transmitter and the receiver tracking the channel (full CSI). (A discussion of the latter is in Section 5.4.6.)

can exploit channel reciprocity and make channel measurements based on the signal received along the opposite link. In an FDD (frequency-division duplex) system, there is no reciprocity and the transmitter will have to rely on feedback information from the receiver. For example, power control in the CDMA system implicitly conveys some channel state information through the feedback in the uplink.

Slow fading: channel inversion

When we discussed the slow fading channel in Section 5.4.1, it was seen that with no channel knowledge at the transmitter, outage occurs whenever the channel cannot support the target data rate R . With transmitter knowledge, one option is now to control the transmit power such that the rate R can be delivered no matter what the fading state is. This is the *channel inversion* strategy: the received SNR is kept constant irrespective of the channel gain. (This strategy is reminiscent of the power control used in CDMA systems, discussed in Section 4.3.) With exact channel inversion, there is zero outage probability. The price to pay is that huge power has to be consumed to invert the channel when it is very bad. Moreover, many systems are also peak-power constrained and cannot invert the channel beyond a certain point. Systems like IS-95 use a combination of channel inversion and diversity to achieve a target rate with reasonable power consumption (Exercise 5.24).

Fast fading: waterfilling

In the slow fading scenario, we are interested in achieving a target data rate within a coherence time period of the channel. In the fast fading case, one is now concerned with the rate averaged over *many* coherence time periods. With transmitter channel knowledge, what is the capacity of the fast fading channel? Let us again consider the simple block fading model (cf. (5.86)):

$$y[m] = h[m]x[m] + w[m], \quad (5.94)$$

where $h[m] = h_\ell$ remains constant over the ℓ th coherence period of T_c ($T_c \gg 1$) symbols and is i.i.d. across different coherence periods. The channel over L such coherence periods can be modeled as a *parallel channel* with L sub-channels that fade independently. For a given realization of the channel gains h_1, \dots, h_L , the capacity (in bits/symbol) of this parallel channel is (cf. (5.39), (5.40) in Section 5.3.3)

$$\max_{P_1, \dots, P_L} \frac{1}{L} \sum_{\ell=1}^L \log \left(1 + \frac{P_\ell |h_\ell|^2}{N_0} \right) \quad (5.95)$$

subject to

$$\frac{1}{L} \sum_{\ell=1}^L P_\ell = P, \quad (5.96)$$

where P is the average power constraint. It was seen (cf. (5.43)) that the optimal power allocation is *waterfilling*:

$$P_\ell^* = \left(\frac{1}{\lambda} - \frac{N_0}{|h_\ell|^2} \right)^+, \quad (5.97)$$

where λ satisfies

$$\frac{1}{L} \sum_{\ell=1}^L \left(\frac{1}{\lambda} - \frac{N_0}{|h_\ell|^2} \right)^+ = P. \quad (5.98)$$

In the context of the frequency-selective channel, waterfilling is done over the OFDM sub-carriers; here, waterfilling is done over time. In both cases, the basic problem is that of power allocation over a parallel channel.

The optimal power P_ℓ allocated to the ℓ th coherence period depends on the channel gain in that coherence period and λ , which in turn depends on all the other channel gains through the constraint (5.98). So it seems that implementing this scheme would require knowledge of the future channel states. Fortunately, as $L \rightarrow \infty$, this non-causality requirement goes away. By the law of large numbers, (5.98) converges to

$$\mathbb{E} \left[\left(\frac{1}{\lambda} - \frac{N_0}{|h|^2} \right)^+ \right] = P \quad (5.99)$$

for almost all realizations of the fading process $\{h[m]\}$. Here, the expectation is taken with respect to the stationary distribution of the channel state. The parameter λ now converges to a constant, depending only on the channel *statistics* but not on the specific *realization* of the fading process. Hence, the optimal power at any time depends only on the channel gain h at that time:

$$P^*(h) = \left(\frac{1}{\lambda} - \frac{N_0}{|h|^2} \right)^+. \quad (5.100)$$

The capacity of the fast fading channel with transmitter channel knowledge is

$$C = \mathbb{E} \left[\log \left(1 + \frac{P^*(h)|h|^2}{N_0} \right) \right] \text{ bits/s/Hz.} \quad (5.101)$$

Equations (5.101), (5.100) and (5.99) together allow us to compute the capacity.

We have derived the capacity assuming the block fading model. The generalization to any ergodic fading process can be done exactly as in the case with no transmitter channel knowledge.

Discussion

Figure 5.21 gives a pictorial view of the waterfilling power allocation strategy. In general, the transmitter allocates more power when the channel is good, taking advantage of the better channel condition, and less or even no power when the channel is poor. This is precisely the opposite of the channel inversion strategy. Note that only the magnitude of the channel gain is needed to implement the waterfilling scheme. In particular, phase information is not required (in contrast to transmit beamforming, for example).

The derivation of the waterfilling capacity suggests a natural variable-rate coding scheme (see Figure 5.22). This scheme consists of a set of codes of different rates, one for each channel state h . When the channel is in state h , the code for that state is used. This can be done since both the transmitter and the receiver can track the channel. A transmit power of $P^*(h)$ is used when

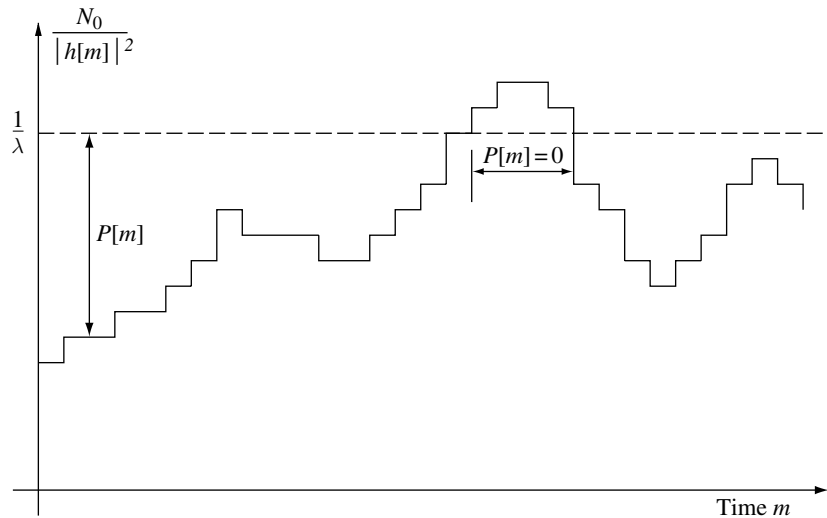
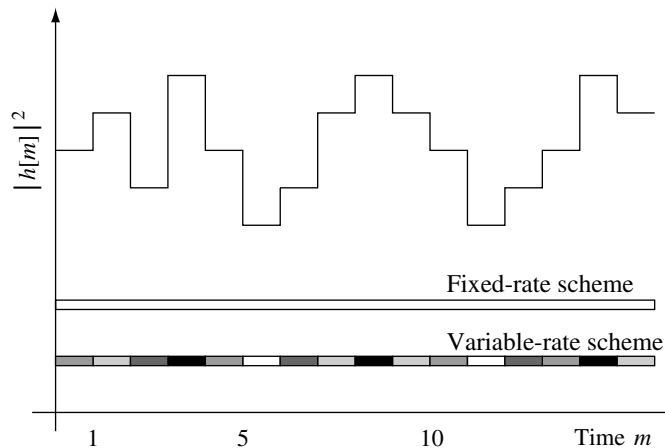


Figure 5.21 Pictorial representation of the waterfilling strategy.

Figure 5.22 Comparison of the fixed-rate and variable-rate schemes. In the fixed-rate scheme, there is only one code spanning many coherence periods. In the variable-rate scheme, different codes (distinguished by different shades) are used depending on the channel quality at that time. For example, the code in white is a low-rate code used only when the channel is weak.



the channel gain is h . The rate of that code is therefore $\log(1 + P^*(h)|h|^2/N_0)$ bits/s/Hz. No coding across channel states is necessary. This is in contrast to the case without transmitter channel knowledge, where a single fixed-rate code with the coded symbols spanning across different coherence time periods is needed (Figure 5.22). Thus, knowledge of the channel state at the transmitter not only allows dynamic power allocation but simplifies the code design problem as one can now use codes designed for the AWGN channel.

Waterfilling performance

Figure 5.20 compares the waterfilling capacity and the capacity with channel knowledge only at the receiver, under Rayleigh fading. Figure 5.23 focuses on the low SNR regime. In the literature the former is also called the capacity with full channel side information (CSI) and the latter is called the capacity with channel side information at the receiver (CSIR). Several observations can be made:

- At low SNR, the capacity with full CSI is significantly larger than the CSIR capacity.
- At high SNR, the difference between the two goes to zero.
- Over a wide range of SNR, the gain of waterfilling over the CSIR capacity is very small.

The first two observations are in fact generic to a wide class of fading models, and can be explained by the fact that the benefit of dynamic power allocation is a *received power gain*: by spending more power when the channel is good, the received power gets boosted up. At high SNR, however, the capacity is insensitive to the received power per degree of freedom and varying the amount of transmit power as a function of the channel state yields a minimal gain (Figure 5.24(a)). At low SNR, the capacity is quite sensitive to the received power (linear, in fact) and so the boost in received power from optimal transmit power allocation provides significant gain. Thus, dynamic

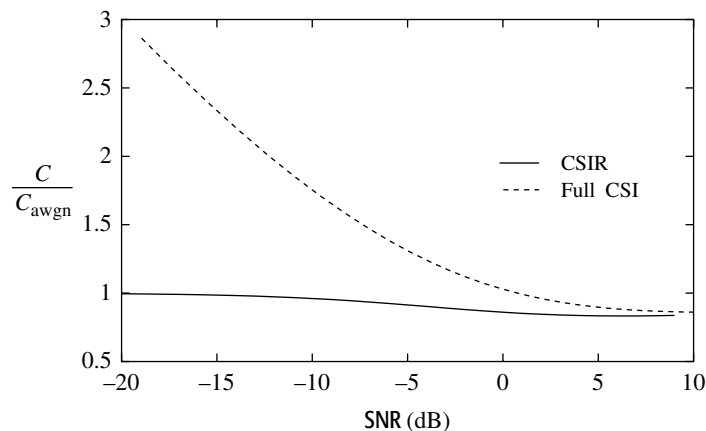


Figure 5.23 Plot of capacities with and without CSI at the transmitter, as a fraction of the AWGN capacity.

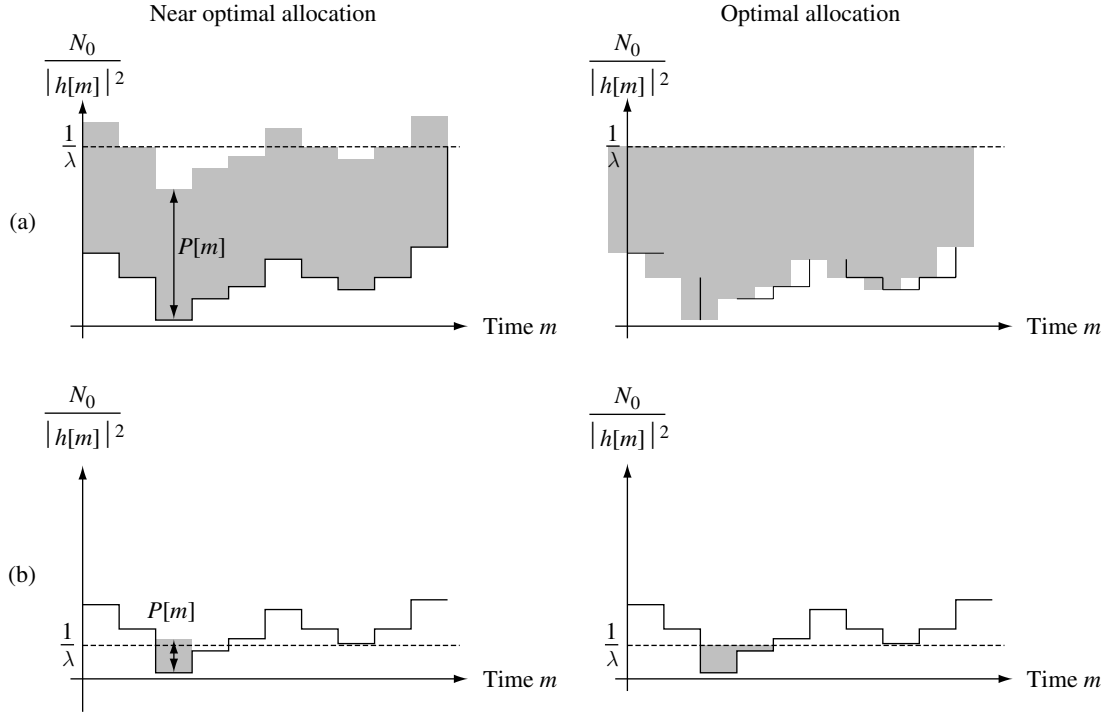


Figure 5.24 (a) High SNR: allocating equal powers at all times is almost optimal. (b) Low SNR: allocating all the power when the channel is strongest is almost optimal.

power allocation is more important in the power-limited (low SNR) regime than in the bandwidth-limited (high SNR) regime.

Let us look more carefully at the low SNR regime. Consider first the case when the channel gain $|h|^2$ has a peak value G_{\max} . At low SNR, the waterfilling strategy transmits information only when the channel is very good, near G_{\max} : when there is very little water, the water ends up at the bottom of the vessel (Figure 5.24(b)). Hence at low SNR

$$\begin{aligned} C &\approx \mathbb{P}\{|h|^2 \approx G_{\max}\} \log\left(1 + G_{\max} \cdot \frac{\text{SNR}}{\mathbb{P}\{|h|^2 \approx G_{\max}\}}\right) \\ &\approx G_{\max} \cdot \text{SNR} \log_2 e \text{ bits/s/Hz.} \end{aligned} \quad (5.102)$$

Recall that at low SNR the CSIR capacity is $\text{SNR} \log_2 e$ bits/s/Hz. Hence, transmitter CSI increases the capacity by G_{\max} times, or a $10 \log_{10} G_{\max}$ dB gain. Moreover, since the AWGN capacity is the same as the CSIR capacity at low SNR, this leads to the interesting conclusion that with full CSI, *the capacity of the fading channel can be much larger than when there is no fading*. This is in contrast to the CSIR case where the fading channel capacity is always less than the capacity of the AWGN channel with the same average SNR. The gain is coming from the fact that in a fading channel, channel fluctuations create peaks and deep nulls, but when the energy per degree of freedom is small, the sender *opportunistically* transmits only when the

channel is near its peak. In a non-fading AWGN channel, the channel stays constant at the average level and there are no peaks to take advantage of.

For models like Rayleigh fading, the channel gain is actually unbounded. Hence, theoretically, the gain of the fading channel waterfilling capacity over the AWGN channel capacity is also unbounded. (See Figure 5.23.) However, to get very large relative gains, one has to operate at very low SNR. In this regime, it may be difficult for the receiver to track and feed back the channel state to the transmitter to implement the waterfilling strategy.

Overall, the performance gain from full CSI is not that large compared to CSIR, unless the SNR is very low. On the other hand, full CSI potentially simplifies the code design problem, as no coding across channel states is necessary. In contrast, one has to interleave and code across many channel states with CSIR.

Waterfilling versus channel inversion

The capacity of the fading channel with full CSI (by using the waterfilling power allocation) should be interpreted as a long-term average rate of flow of information, averaged over the fluctuations of the channel. While the waterfilling strategy increases the long-term throughput of the system by transmitting when the channel is good, an important issue is the *delay* entailed. In this regard, it is interesting to contrast the waterfilling power allocation strategy with the channel inversion strategy. Compared to waterfilling, channel inversion is much less power-efficient, as a huge amount of power is consumed to invert the channel when it is bad. On the other hand, the rate of flow of information is now the same in all fading states, and so the associated delay is *independent* of the time-scale of channel variations. Thus, one can view the channel inversion strategy as a *delay-limited* power allocation strategy. Given an average power constraint, the maximum achievable rate by this strategy can be thought of as a *delay-limited* capacity. For applications with very tight delay constraints, this delay-limited capacity may be a more appropriate measure of performance than the waterfilling capacity.

Without diversity, the delay-limited capacity is typically very small. With increased diversity, the probability of encountering a bad channel is reduced and the average power consumption required to support a target delay-limited rate is reduced. Put another way, a larger delay-limited capacity is achieved for a given average power constraint (Exercise 5.24).

Example 5.3 Rate adaptation in IS-856

IS-856 downlink

IS-856, also called CDMA 2000 $1 \times$ EV-DO (Enhanced Version Data Optimized) is a cellular data standard operating on the 1.25-MHz bandwidth.

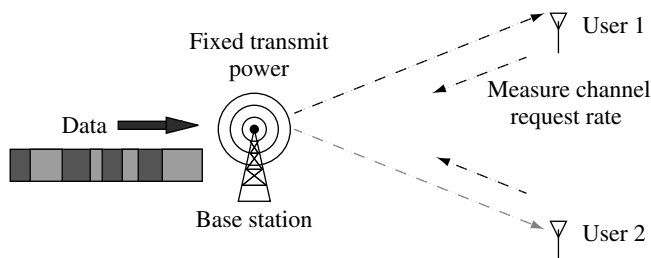


Figure 5.25 Downlink of IS-856 (CDMA 2000 1× EV-DO). Users measure their channels based on the downlink pilot and feed back requested rates to the base-station. The base-station schedules users in a time-division manner.

The uplink is CDMA-based, not too different from IS-95, but the downlink is quite different (Figure 5.25):

- Multiple access is TDMA, with one user transmission at a time. The finest granularity for scheduling the user transmissions is a *slot* of duration 1.67 ms.
- Each user is *rate*-controlled rather than *power*-controlled. The transmit power at the base-station is fixed at all times and the rate of transmission to a user is adapted based on the current channel condition.

In contrast, the uplink of IS-95 (cf. Section 4.3.2) is CDMA-based, with the total power dynamically allocated among the users to meet their individual SIR requirements. The multiple access and scheduling aspects of IS-856 are discussed in Chapter 6; here the focus is only on rate adaptation.

Rate versus power control

The contrast between power control in IS-95 and rate control in IS-856 is roughly analogous to that between the channel inversion and the waterfilling strategies discussed above. In the former, power is allocated dynamically to a user to maintain a constant target rate at all times; this is suitable for voice, which has a stringent delay requirement and requires a consistent throughput. In the latter, rate is adapted to transmit more information when the channel is strong; this is suitable for data, which have a laxer delay requirement and can take better advantage of a variable transmission rate. The main difference between IS-856 and the waterfilling strategy is that there is no dynamic power adaptation in IS-856, only rate adaptation.

Rate control in IS-856

Like IS-95, IS-856 is an FDD system. Hence, rate control has to be performed based on channel state feedback from the mobile to the base-station. The mobile measures its own channel based on a common strong pilot broadcast by the base-station. Using the measured values, the mobile predicts the SINR for the next time slot and uses that to predict the rate the base-station can send information to it. This *requested rate* is fed back to the base-station on the uplink. The transmitter then sends a packet at

the requested rate to the mobile starting at the next time slot (if the mobile is scheduled). The table below describes the possible requested rates, the SINR thresholds for those rates, the modulation used and the number of time slots the transmission takes.

Requested rate (kbits/s)	SINR threshold (dB)	Modulation	Number of slots
38.4	-11.5	QPSK	16
76.8	-9.2	QPSK	8
153.6	-6.5	QPSK	4
307.2	-3.5	QPSK	2 or 4
614.4	-0.5	QPSK	1 or 2
921.6	2.2	8-PSK	2
1228.8	3.9	QPSK or 16-QAM	1 or 2
1843.2	8.0	8-PSK	1
2457.6	10.3	16-QAM	1

To simplify the implementation of the encoder, the codes at the different rates are all derived from a basic 1/5-rate turbo code. The low-rate codes are obtained by repeating the turbo-coded symbols over a number of time slots; as demonstrated in Exercise 5.25, such repetition loses little spectral efficiency in the low SNR regime. The higher-rate codes are obtained by using higher-order constellations in the modulation.

Rate control is made possible by the presence of the strong pilot to measure the channel and the rate request feedback from the mobile to the base-station. The pilot is shared between all users in the cell and is also used for many other functions such as coherent reception and synchronization. The rate request feedback is solely for the purpose of rate control. Although each request is only 4 bits long (to specify the various rate levels), this is sent by every active user at every slot and moreover considerable power and coding is needed to make sure the information gets fed back accurately and with little delay. Typically, sending this feedback consumes about 10% of the uplink capacity.

Impact of prediction uncertainty

Proper rate adaptation relies on the accurate tracking and prediction of the channel at the transmitter. This is possible only if the coherence time of the channel is much longer than the lag between the time the channel is measured at the mobile and the time when the packet is actually transmitted at the base-station. This lag is at least two slots (2×1.67 ms) due to the delay in getting the requested rate fed back to the base-station, but can be considerably more at the low rates since the packet is transmitted over multiple slots and the predicted channel has to be valid during this time.

At a walking speed of 3 km/h and a carrier frequency $f_c = 1.9$ GHz, the coherence time is of the order of 25 ms, so the channel can be quite accurately predicted. At a driving speed of 30 km/h, the coherence time is only 2.5 ms and accurate tracking of the channel is already very difficult. (Exercise 5.26 explicitly connects the prediction error to the physical parameters of the channel.) At an even faster speed of 120 km/h, the coherence time is less than 1 ms and tracking of the channel is impossible; there is now no transmitter CSI. On the other hand, the multiple slot low rate packets essentially go through a fast fading channel with significant time diversity over the duration of the packet. Recall that the fast fading capacity is given by (5.89):

$$C = \mathbb{E}[\log(1 + |h|^2 \text{SNR})] \approx \mathbb{E}[|h|^2] \text{SNR} \log_2 e \text{ bits/s/Hz} \quad (5.103)$$

in the low SNR regime, where h follows the stationary distribution of the fading. Thus, to determine an appropriate transmission rate across this fast fading channel, it suffices for the mobile to predict the *average* SINR over the transmission time of the packet, and this average is quite easy to predict. Thus, the difficult regime is actually in between the very slow and very fast fading scenarios, where there is significant uncertainty in the channel prediction and yet not very much time diversity over the packet transmission time. This channel uncertainty has to be taken into account by being more conservative in predicting the SINR and in requesting a rate. This is similar to the outage scenario considered in Section 5.4.1, except that the randomness of the channel is conditional on the predicted value. The requested rate should be set to meet a target outage probability (Exercise 5.27).

The various situations are summarized in Figure 5.26. Note the different roles of coding in the three scenarios. In the first scenario, when the predicted SINR is accurate, the main role of coding is to combat the additive Gaussian noise; in the other two scenarios, coding combats the residual randomness in the channel by exploiting the available time diversity.

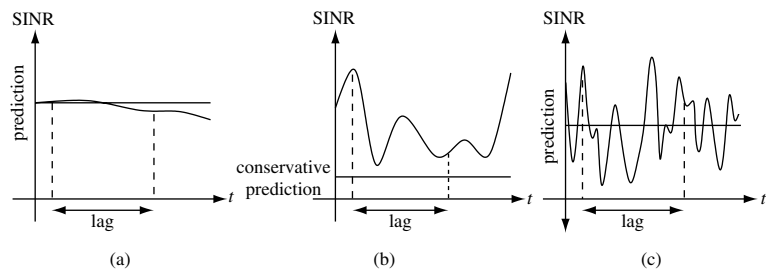


Figure 5.26 (a) Coherence time is long compared to the prediction time lag; predicted SINR is accurate. Near perfect CSI at transmitter. (b) Coherence time is comparable to the prediction time lag, predicted SINR has to be conservative to meet an outage criterion. (c) Coherence time is short compared to the prediction time lag; prediction of average SINR suffices. No CSI at the transmitter.

To reduce the loss in performance due to the conservativeness of the channel prediction, IS-856 employs an *incremental* ARQ (or hybrid-ARQ) mechanism for the repetition-coded multiple slot packets. Instead of waiting until the end of the transmission of all slots before decoding, the mobile will attempt to decode the information incrementally as it receives the repeated copies over the time slots. When it succeeds in decoding, it will send an acknowledgement back to the base-station so that it can stop the transmission of the remaining slots. This way, a rate higher than the requested rate can be achieved if the actual SINR is higher than the predicted SINR.

5.4.7 Frequency-selective fading channels

So far, we have considered flat fading channels (cf. (5.53)). In Section 5.3.3, the capacity of the *time-invariant* frequency-selective channel (5.32) was also analyzed. It is simple to extend the understanding to *underspread* time-varying frequency-selective fading channels: these are channels with the coherence time much larger than the delay spread. We model the channel as a time-invariant L -tap channel as in (5.32) over each coherence time interval and view it as N_c parallel sub-channels (in frequency). For underspread channels, N_c can be chosen large so that the cyclic prefix loss is negligible. This model is a generalization of the flat fading channel in (5.53): here there are N_c (frequency) sub-channels over each coherence time interval and multiple (time) sub-channels over the different coherence time intervals. Overall it is still a parallel channel. We can extend the capacity results from Sections 5.4.5 and 5.4.6 to the frequency-selective fading channel. In particular, the fast fading capacity with full CSI (cf. Section 5.4.6) can be generalized here to a combination of waterfilling over time and frequency: the coherence time intervals provide sub-channels in time and each coherence time interval provides sub-channels in frequency. This is carried out in Exercise 5.30.

5.4.8 Summary: a shift in point of view

Let us summarize our investigation on the performance limits of fading channels. In the slow fading scenario without transmitter channel knowledge, the amount of information that is allowed through the channel is random, and no positive rate of communication can be reliably supported (in the sense of arbitrarily small error probability). The outage probability is the main performance measure, and it behaves like $1/\text{SNR}$ at high SNR. This is due to a lack of diversity and, equivalently, the outage capacity is very small. With L branches of diversity, either over space, time or frequency, the outage

probability is improved and decays like $1/\text{SNR}^L$. The fast fading scenario can be viewed as the limit of infinite time diversity and has a capacity of $\mathbb{E}[\log(1 + |h|^2\text{SNR})]$ bits/s/Hz. This however incurs a coding delay much longer than the coherence time of the channel. Finally, when the transmitter and the receiver can both track the channel, a further performance gain can be obtained by dynamically allocating power and opportunistically transmitting when the channel is good.

The slow fading scenario emphasizes the *detrimental* effect of fading: a slow fading channel is very unreliable. This unreliability is *mitigated* by providing more diversity in the channel. This is the traditional way of viewing the fading phenomenon and was the central theme of Chapter 3. In a narrowband channel with a single antenna, the only source of diversity is through time. The capacity of the fast fading channel (5.89) can be viewed as the performance limit of any such time diversity scheme. Still, the capacity is less than the AWGN channel capacity as long as there is no channel knowledge at the transmitter. With channel knowledge at the transmitter, the picture changes. Particularly at low SNR, the capacity of the fading channel with full CSI can be larger than that of the AWGN channel. Fading can be *exploited* by transmitting near the peak of the channel fluctuations. Channel fading is now turned from a foe to a friend.

This new theme on fading will be developed further in the multiuser context in Chapter 6, where we will see that opportunistic communication will have a significant impact at *all* SNRs, and not only at low SNR.

Chapter 5 The main plot

Channel capacity

The maximum rate at which information can be communicated across a noisy channel with arbitrary reliability.

Linear time-invariant Gaussian channels

Capacity of the AWGN channel with SNR per degree of freedom is

$$C_{\text{awgn}} = \log(1 + \text{SNR}) \text{ bits/s/Hz.} \quad (5.104)$$

Capacity of the continuous-time AWGN channel with bandwidth W , average received power \bar{P} and white noise power spectral density N_0 is

$$C_{\text{awgn}} = W \log \left(1 + \frac{\bar{P}}{N_0 W} \right) \text{ bits/s.} \quad (5.105)$$

Bandwidth-limited regime: $\text{SNR} = \bar{P}/(N_0 W)$ is high and capacity is logarithmic in the SNR.

Power-limited regime: SNR is low and capacity is linear in the SNR.

Capacities of the SIMO and the MISO channels with time-invariant channel gains h_1, \dots, h_L are the same:

$$C = \log(1 + \text{SNR} \|\mathbf{h}\|^2) \text{ bits/s/Hz.} \quad (5.106)$$

Capacity of frequency-selective channel with response $H(f)$ and power constraint P per degree of freedom:

$$C = \int_0^W \log \left(1 + \frac{P^*(f) |H(f)|^2}{N_0} \right) df \text{ bits/s} \quad (5.107)$$

where $P^*(f)$ is waterfilling:

$$P^*(f) = \left(\frac{1}{\lambda} - \frac{N_0}{|H(f)|^2} \right)^+, \quad (5.108)$$

and λ satisfies:

$$\int_0^W \left(\frac{1}{\lambda} - \frac{N_0}{|H(f)|^2} \right)^+ df = P. \quad (5.109)$$

Slow fading channels with receiver CSI only

Setting: coherence time is much longer than constraint on coding delay.

Performance measures:

Outage probability $p_{\text{out}}(R)$ at a target rate R .

Outage capacity C_ϵ at a target outage probability ϵ .

Basic flat fading channel:

$$y[m] = hx[m] + w[m]. \quad (5.110)$$

Outage probability is

$$p_{\text{out}}(R) = \mathbb{P} \{ \log(1 + |h|^2 \text{SNR}) < R \}, \quad (5.111)$$

where SNR is the average signal-to-noise ratio at each receive antenna.

Outage probability with receive diversity is

$$p_{\text{out}}(R) := \mathbb{P} \left\{ \log \left(1 + \|\mathbf{h}\|^2 \text{SNR} \right) < R \right\}. \quad (5.112)$$

This provides power and diversity gains.

Outage probability with L -fold transmit diversity is

$$p_{\text{out}}(R) := \mathbb{P} \left\{ \log \left(1 + \|\mathbf{h}\|^2 \frac{\text{SNR}}{L} \right) < R \right\}. \quad (5.113)$$

This provides diversity gain only.

Outage probability with L -fold time diversity is

$$p_{\text{out}}(R) = \mathbb{P} \left\{ \frac{1}{L} \sum_{\ell=1}^L \log \left(1 + |h_{\ell}|^2 \text{SNR} \right) < R \right\}. \quad (5.114)$$

This provides diversity gain only.

Fast fading channels

Setting: coherence time is much shorter than coding delay.

Performance measure: capacity.

Basic model:

$$y[m] = h[m]x[m] + w[m]. \quad (5.115)$$

$\{h[m]\}$ is an ergodic fading process.

Receiver CSI only:

$$C = \mathbb{E} \left[\log \left(1 + |h|^2 \text{SNR} \right) \right]. \quad (5.116)$$

Full CSI:

$$C = \mathbb{E} \left[\log \left(1 + \frac{P^*(h)|h|^2}{N_0} \right) \right] \text{ bits/s/Hz} \quad (5.117)$$

where $P^*(h)$ waterfills over the fading states:

$$P^*(h) = \left(\frac{1}{\lambda} - \frac{N_0}{|h|^2} \right)^+, \quad (5.118)$$

and λ satisfies:

$$\mathbb{E} \left[\left(\frac{1}{\lambda} - \frac{N_0}{|h|^2} \right)^+ \right] = P. \quad (5.119)$$

Power gain over the receiver CSI only case. Significant at low SNR.

5.5 Bibliographical notes

Information theory and the formulation of the notions of reliable communication and channel capacity were introduced in a path-breaking paper by Shannon [109]. The underlying philosophy of using simple models to understand the essence of an engineering problem has pervaded the development of the communication field ever since. In that paper, as a consequence of his general theory, Shannon also derived the capacity of the AWGN channel. He returned to a more in-depth geometric treatment of this channel in a subsequent paper [110]. Sphere-packing arguments were used extensively in the text by Wozencraft and Jacobs [148].

The linear cellular model was introduced by Shamai and Wyner [108]. One of the early studies of wireless channels using information theoretic techniques is due to Ozarow, *et al.* [88], where they introduced the concept of outage capacity. Telatar [119] extended the formulation to multiple antennas. The capacity of fading channels with full CSI was analyzed by Goldsmith and Varaiya [51]. They observed the optimality of the waterfilling power allocation with full CSI and the corollary that full CSI over CSI at the receiver alone is beneficial only at low SNRs. A comprehensive survey of information theoretic results on fading channels was carried out by Biglieri, Proakis and Shamai [9].

The design issues in IS-856 have been elaborately discussed in Bender *et al.* [6] and by Wu and Esteves [149].

5.6 Exercises

Exercise 5.1 What is the maximum reliable rate of communication over the (complex) AWGN channel when only the I channel is used? How does that compare to the capacity of the complex channel at low and high SNR, with the same average power constraint? Relate your conclusion to the analogous comparison between uncoded schemes in Section 3.1.2 and Exercise 3.4, focusing particularly on the high SNR regime.

Exercise 5.2 Consider a linear cellular model with equi-spaced base-stations at distance $2d$ apart. With a reuse ratio of ρ , base-stations at distances of integer multiples of $2d/\rho$ reuse the same frequency band. Assuming that the interference emanates from the center of the cell, calculate the fraction f_ρ defined as the ratio of the interference to the received power from a user at the edge of the cell. You can assume that all uplink transmissions are at the same transmit power P and that the dominant interference comes from the nearest cells reusing the same frequency.

Exercise 5.3 Consider a regular hexagonal cellular model (cf. Figure 4.2) with a frequency reuse ratio of ρ .

1. Identify “appropriate” reuse patterns for different values of ρ , with the design goal of minimizing inter-cell interference. You can use the assumptions made in Exercise 5.2 on how the interference originates.
2. For the reuse patterns identified, show that $f_\rho = 6(\sqrt{\rho}/2)^\alpha$ is a good approximation to the fraction of the received power of a user at the edge of the cell that the interference represents. *Hint:* You can explicitly construct reuse patterns for $\rho = 1, 1/3, 1/4, 1/7, 1/9$ with exactly these fractions.