6

Multiuser capacity and opportunistic communication

In Chapter 4, we studied several specific multiple access *techniques* (TDMA/FDMA, CDMA, OFDM) designed to share the channel among several users. A natural question is: what are the "optimal" multiple access schemes? To address this question, one must now step back and take a fundamental look at the multiuser *channels* themselves. Information theory can be generalized from the point-to-point scenario, considered in Chapter 5, to the multiuser ones, providing limits to multiuser communications and suggesting optimal multiple access strategies. New techniques and concepts such as *successive cancellation*, *superposition coding* and *multiuser diversity* emerge.

The first part of the chapter focuses on the uplink (many-to-one) and downlink (one-to-many) AWGN channel without fading. For the uplink, an optimal multiple access strategy is for all users to spread their signal across the entire bandwidth, much like in the CDMA system in Chapter 4. However, rather than decoding every user treating the interference from other users as noise, a *successive interference cancellation* (SIC) receiver is needed to achieve capacity. That is, after one user is decoded, its signal is stripped away from the aggregate received signal before the next user is decoded. A similar strategy is optimal for the downlink, with signals for the users superimposed on top of each other and SIC done at the mobiles: each user decodes the information intended for all of the weaker users and strips them off before decoding its own. It is shown that in situations where users have very disparate channels to the base-station, CDMA together with successive cancellation can offer significant gains over the conventional multiple access techniques discussed in Chapter 4.

In the second part of the chapter, we shift our focus to multiuser *fading* channels. One of the main insights learnt in Chapter 5 is that, for fast fading channels, the ability to track the channel at the transmitter can increase point-to-point capacity by *opportunistic communication*: transmitting at high rates when the channel is good, and at low rates or not at all when the channel is poor. We extend this insight to the multiuser setting, both for the uplink

and for the downlink. The performance gain of opportunistic communication comes from exploiting the fluctuations of the fading channel. Compared to the point-to-point setting, the multiuser settings offer more opportunities to exploit. In addition to the choice of *when* to transmit, there is now an additional choice of *which user(s)* to transmit from (in the uplink) or to transmit to (in the downlink) and the amount of power to allocate between the users. This additional choice provides a further performance gain not found in the point-to-point scenario. It allows the system to benefit from a *multiuser diversity* effect: at any time in a large network, with high probability there is a user whose channel is near its peak. By allowing such a user to transmit at that time, the overall multiuser capacity can be achieved.

In the last part of the chapter, we will study the system issues arising from the implementation of opportunistic communication in a cellular system. We use as a case study IS-856, the third-generation standard for wireless data already introduced in Chapter 5. We show how multiple antennas can be used to further boost the performance gain that can be extracted from opportunistic communication, a technique known as *opportunistic beamforming*. We distill the insights into a new design principle for wireless systems based on opportunistic communication and multiuser diversity.

6.1 Uplink AWGN channel

6.1.1 Capacity via successive interference cancellation

The baseband discrete-time model for the uplink AWGN channel with two users (Figure 6.1) is

$$y[m] = x_1[m] + x_2[m] + w[m], \tag{6.1}$$

where $w[m] \sim \mathcal{CN}(0, N_0)$ is i.i.d. complex Gaussian noise. User k has an average power constraint of P_k joules/symbol (with $k = 1, 2$).

In the point-to-point case, the *capacity* of a channel provides the performance limit: reliable communication can be attained at any rate $R < C$; reliable communication is impossible at rates $R > C$. In the multiuser case, we should extend this concept to a *capacity region* $\mathcal{C}$: this is the set of all pairs (R_1, R_2) such that *simultaneously* user 1 and 2 can reliably communicate at rate R_1 and R_2, respectively. Since the two users share the same bandwidth, there is naturally a tradeoff between the reliable communication rates of the users: if one wants to communicate at a higher rate, the other user may need to lower its rate. For example, in orthogonal multiple access schemes, such as OFDM, this tradeoff can be achieved by varying the number of sub-carriers allocated to each user. The capacity region $\mathcal{C}$ characterizes the *optimal* tradeoff achievable by *any* multiple access scheme. From this
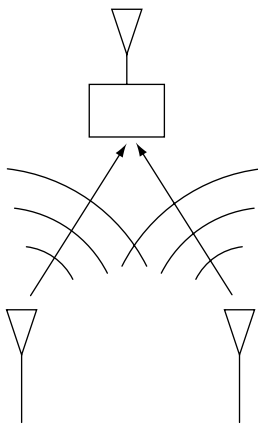

Figure 6.1 Two-user uplink.

capacity region, one can derive other scalar performance measures of interest. For example:

- The symmetric capacity:

$$C_{\text{sym}} := \max_{(R,R)\in\mathcal{C}} R \qquad (6.2)$$

  is the maximum common rate at which both the users can simultaneously reliably communicate.
- The sum capacity:

$$C_{\text{sum}} := \max_{(R_1,R_2)\in\mathcal{C}} R_1 + R_2 \qquad (6.3)$$

  is the maximum total throughput that can be achieved.

Just like the capacity of the AWGN channel, there is a very simple characterization of the capacity region $\mathcal{C}$ of the uplink AWGN channel: this is the set of all rates (R_1, R_2) satisfying the three constraints (Appendix B.9 provides a formal justification):

$$R_1 < \log\left(1 + \frac{P_1}{N_0}\right), \qquad (6.4)$$

$$R_2 < \log\left(1 + \frac{P_2}{N_0}\right), \qquad (6.5)$$

$$R_1 + R_2 < \log\left(1 + \frac{P_1 + P_2}{N_0}\right). \qquad (6.6)$$

The capacity region is the pentagon shown in Figure 6.2. All the three constraints are natural. The first two say that the rate of the individual user cannot exceed the capacity of the point-to-point link with the other user absent from
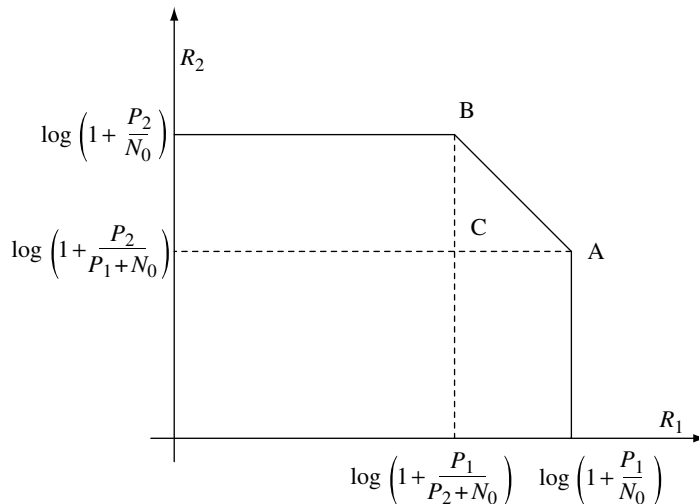

Figure 6.2 Capacity region of the two-user uplink AWGN channel.

the system (these are called single-user bounds). The third says that the total throughput cannot exceed the capacity of a point-to-point AWGN channel with the sum of the received powers of the two users. This is indeed a valid constraint since the signals the two users send are independent and hence the power of the aggregate received signal is the sum of the powers of the individual received signals.[1] Note that without the third constraint, the capacity region would have been a rectangle, and each user could simultaneously transmit at the point-to-point capacity as if the other user did not exist. This is clearly too good to be true and indeed the third constraint says this is not possible: there must be a tradeoff between the performance of the two users.

Nevertheless, something surprising does happen: user 1 can achieve its single-user bound while at the same time user 2 can get a non-zero rate; in fact as high as its rate at point A, i.e.,

$$R_2^* = \log\left(1 + \frac{P_1 + P_2}{N_0}\right) - \log\left(1 + \frac{P_1}{N_0}\right) = \log\left(1 + \frac{P_2}{P_1 + N_0}\right). \quad (6.7)$$

How can this be achieved? Each user encodes its data using a capacity-achieving AWGN channel code. The receiver decodes the information of both the users in two stages. In the first stage, it decodes the data of user 2, treating the signal from user 1 as Gaussian interference. The maximum rate user 2 can achieve is precisely given by (6.7). Once the receiver decodes the data of user 2, it can reconstruct user 2's signal and subtract it from the aggregate received signal. The receiver can then decode the data of user 1. Since there is now only the background Gaussian noise left in the system, the maximum rate user 1 can transmit at is its single-user bound $\log(1 + P_1/N_0)$. This receiver is called a *successive interference cancellation* (SIC) receiver or simply a successive cancellation decoder. If one reverses the order of cancellation, then one can achieve point B, the other corner point. All the other rate points on the segment AB can be obtained by time-sharing between the multiple access strategies in point A and point B. (We see in Exercise 6.7 another technique called *rate-splitting* that also achieves these intermediate points.)

The segment AB contains all the "optimal" operating points of the channel, in the sense that any other point in the capacity region is component-wise dominated by some point on AB. Thus one can always increase *both* users' rates by moving to a point on AB, and there is no reason not to.[2] No such domination exists *among* the points on AB, and the preferred operating point depends on the system objective. If the goal of the system is to maximize the sum rate, then any point on AB is equally fine. On the other hand, some operating points are not *fair*, especially if the received power of one user is

---

[1] This is the same argument we used for deriving the capacity of the MISO channel in Section 5.3.2.

[2] In economics terms, the points on AB are called *Pareto optimal*.

much larger than the other. In this case, consider operating at the corner point in which the strong user is decoded *first*: now the weak user gets the best possible rate.[3] In the case when the weak user is the one further away from the base-station, it is shown in Exercise 6.10 that this decoding order has the property of minimizing the total *transmit* power to meet given target rates for the two users. Not only does this lead to savings in the battery power of the users, it also translates to an increase in the system capacity of an interference-limited cellular system (Exercise 6.11).

6.1.2 Comparison with conventional CDMA

There is a certain similarity between the multiple access technique that achieves points A and B, and the CDMA technique discussed in Chapter 4. The only difference is that in the CDMA system described there, *every* user is decoded treating the other users as interference. This is sometimes called a *conventional* or a *single-user* CDMA receiver. In contrast, the SIC receiver is a *multiuser* receiver: one of the users, say user 1, is decoded treating user 2 as interference, but user 2 is decoded with the benefit of the signal of user 1 already removed. Thus, we can immediately conclude that the performance of the conventional CDMA receiver is suboptimal; in Figure 6.2, it achieves the point C which is strictly in the interior of the capacity region.

The benefit of SIC over the conventional CDMA receiver is particularly significant when the received power of one user is much larger than that of the other: by decoding and subtracting the signal of the strong user first, the weaker user can get a much higher data rate than when it has to contend with the interference of the strong user (Figure 6.3). In the context of a cellular system, this means that rather than having to keep the received powers of all users equal by transmit power control, users closer to the base-station can be allowed to take advantage of the stronger channel and transmit at a higher rate while not degrading the performance of the users in the edge of the cell. With a conventional receiver, this is not possible due to the *near–far problem*. With the SIC, we are turning the near–far problem into a near–far advantage. This advantage is less apparent in providing voice service where the required data rate of a user is constant over time, but it can be important for providing data services where users can take advantage of the higher data rates when they are closer to the base-station.

6.1.3 Comparison with orthogonal multiple access

How about orthogonal multiple access techniques? Can they be information theoretically optimal? Consider an orthogonal scheme that allocates a fraction

---

[3] This operating point is said to be *max–min fair*.

Figure 6.3 In the case when the received powers of the users are very disparate, successive cancellation (point A) can provide a significant advantage to the weaker user compared to conventional CDMA decoding (point C). The conventional CDMA solution is to control the received power of the strong user to equal that of the weak user (point D), but then the rates of both users are much lower. Here, $P_1/N_0 = 0$ dB, $P_2/N_0 = 20$ dB.
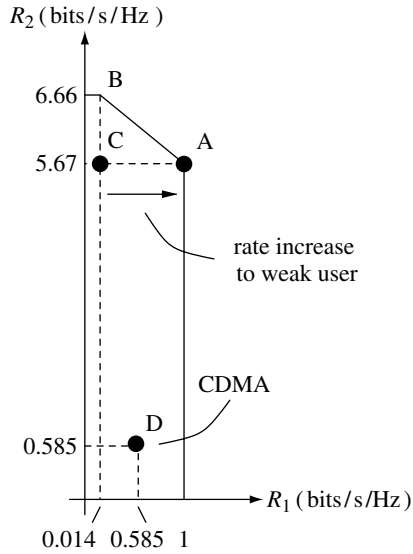
α of the degrees of freedom to user 1 and the rest, $1 - \alpha$, to user 2 (note that it is irrelevant for the capacity analysis whether the partitioning is across frequency or across time, since the power constraint is on the average across the degrees of freedom). Since the received power of user 1 is P_1, the amount of received energy is P_1/α joules per degree of freedom. The maximum rate user 1 can achieve over the total bandwidth W is

$$\alpha W \log\left(1 + \frac{P_1}{\alpha N_0}\right) \text{ bits/s.} \tag{6.8}$$

Similarly, the maximum rate user 2 can achieve is

$$(1 - \alpha)W \log\left(1 + \frac{P_2}{(1 - \alpha)N_0}\right) \text{ bits/s.} \tag{6.9}$$

Varying α from 0 to 1 yields all the rate pairs achieved by orthogonal schemes. See Figure 6.4.

Comparing these rates with the capacity region, one can see that the orthogonal schemes are in general suboptimal, except for one point: when $\alpha = P_1/(P_1 + P_2)$, i.e., the amount of degrees of freedom allocated to each user is proportional to its received power (Exercise 6.2 explores the reason why). However, when there is a large disparity between the received powers of the two users (as in the example of Figure 6.4), this operating point is highly unfair since most of the degrees of freedom are given to the strong user and the weak user has hardly any rate. On the other hand, by decoding the strong user first and then the weak user, the weak user can achieve the highest possible rate and this is therefore the most fair possible operating point (point A in Figure 6.4). In contrast, orthogonal multiple access techniques

Figure 6.4 Performance of
orthogonal multiple access
compared to capacity. The
SNRs of the two users are:
$P_1/N_0 = 0\,\text{dB}$ and
$P_2/N_0 = 20\,\text{dB}$. Orthogonal
multiple access achieves the
sum capacity at exactly one
point, but at that point the
weak user 1 has hardly any
rate and it is therefore a highly
unfair operating point. Point A
gives the highest possible rate
to user 1 and is most fair.
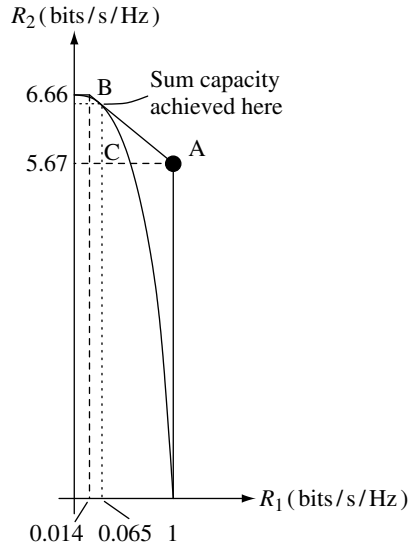
can approach this performance for the weak user only by nearly sacrificing all the rate of the strong user. Here again, as in the comparison with CDMA, SIC's advantage is in exploiting the proximity of a user to the base-station to give it high rate while protecting the far-away user.

6.1.4 General *K*-user uplink capacity

We have so far focused on the two-user case for simplicity, but the results extend readily to an arbitrary number of users. The K-user capacity region is described by $2^K - 1$ constraints, one for each possible non-empty subset $\mathcal{S}$ of users:

$$\sum_{k \in \mathcal{S}} R_k < \log \left(1 + \frac{\sum_{k \in \mathcal{S}} P_k}{N_0} \right) \qquad \text{for all } \mathcal{S} \subset \{1, \ldots, K\}. \qquad (6.10)$$

The right hand side corresponds to the maximum sum rate that can be achieved by a single transmitter with the total power of the users in $\mathcal{S}$ and with no other users in the system. The sum capacity is

$$C_{\text{sum}} = \log \left(1 + \frac{\sum_{k=1}^{K} P_k}{N_0} \right) \text{bits/s/Hz}. \qquad (6.11)$$

It can be shown that there are exactly $K!$ corner points, each one corresponding to a successive cancellation order among the users (Exercise 6.9).

The equal received power case ($P_1 = \ldots = P_K = P$) is particularly simple. The sum capacity is

$$C_{\text{sum}} = \log \left(1 + \frac{KP}{N_0} \right). \qquad (6.12)$$

The symmetric capacity is

$$C_{\text{sym}} = \frac{1}{K} \cdot \log\left(1 + \frac{KP}{N_0}\right). \qquad (6.13)$$

This is the maximum rate for each user that can be obtained if every user operates at the same rate. Moreover, this rate can be obtained via orthogonal multiplexing: each user is allocated a fraction $1/K$ of the total degrees of freedom.[4] In particular, we can immediately conclude that under equal received powers, the OFDM scheme considered in Chapter 4 has a better performance than the CDMA scheme (which uses conventional receivers.)

Observe that the sum capacity (6.12) is unbounded as the number of users grows. In contrast, if the conventional CDMA receiver (decoding every user treating all other users as noise) is used, each user will face an interference from $K - 1$ users of total power $(K - 1)P$, and thus the sum rate is only

$$K \cdot \log\left(1 + \frac{P}{(K-1)P + N_0}\right) \text{bits/s/Hz}, \qquad (6.14)$$

which approaches

$$K \cdot \frac{P}{(K-1)P + N_0} \log_2 e \approx \log_2 e = 1.442 \,\text{bits/s/Hz}, \qquad (6.15)$$

as $K \to \infty$. Thus, the total spectral efficiency is bounded in this case: the growing interference is eventually the limiting factor. Such a rate is said to be *interference-limited*.

The above comparison pertains effectively to a single-cell scenario, since the only external effect modeled is white Gaussian noise. In a cellular network, the out-of-cell interference must be considered, and as long as the out-of-cell signals cannot be decoded, the system would still be interference-limited, no matter what the receiver is.

6.2 Downlink AWGN channel

The downlink communication features a single transmitter (the base-station) sending separate information to multiple users (Figure 6.5). The baseband downlink AWGN channel with two users is

$$y_k[m] = h_k x[m] + w_k[m], \qquad k = 1, 2, \qquad (6.16)$$

where $w_k[m] \sim \mathcal{CN}(0, N_0)$ is i.i.d. complex Gaussian noise and $y_k[m]$ is the received signal at user k at time m, for both the users $k = 1, 2$. Here h_k is

---

[4] This fact is specific to the AWGN channel and does not hold in general. See Section 6.3.
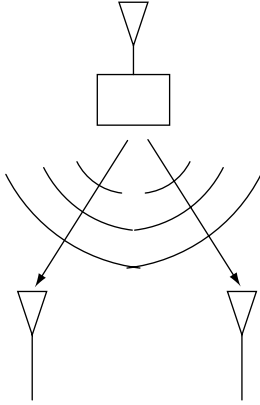
Figure 6.5 Two-user downlink.

the fixed (complex) channel gain corresponding to user k. We assume that h_k is known to both the transmitter and the user k (for $k = 1, 2$). The transmit signal $\{x[m]\}$ has an average power constraint of P joules/symbol. Observe the difference from the uplink of this overall constraint: there the power restrictions are separate for the signals of each user. The users separately decode their data using the signals they receive.

As in the uplink, we can ask for the capacity region $\mathcal{C}$, the region of the rates (R_1, R_2), at which the two users can simultaneously reliably communicate. We have the single-user bounds, as in (6.4) and (6.5),

$$R_k < \log\left(1 + \frac{P|h_k|^2}{N_0}\right), \qquad k = 1, 2. \tag{6.17}$$

This upper bound on R_k can be attained by using all the power and degrees of freedom to communicate to user k (with the other user getting zero rate). Thus, we have the two extreme points (with rate of one user being zero) in Figure 6.6. Further, we can share the degrees of freedom (time and bandwidth) between the users in an orthogonal manner to obtain any rate pair on the line joining these two extreme points. Can we achieve a rate pair outside this triangle by a more sophisticated communication strategy?

6.2.1 Symmetric case: two capacity-achieving schemes

To get more insight, let us first consider the symmetric case where $|h_1| = |h_2|$. In this symmetric situation, the SNR of both the users is the same. This means that if user 1 can successfully decode its data, then user 2 should also be
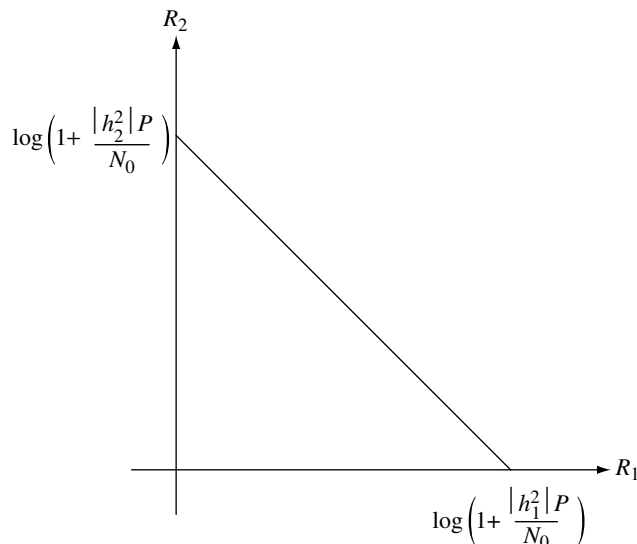

Figure 6.6 The capacity region of the downlink with two users having symmetric AWGN channels, i.e., $|h_1| = |h_2|$.

able to decode successfully the data of user 1 (and vice versa). Thus the sum information rate must also be bounded by the single-user capacity:

$$R_1 + R_2 < \log\left(1 + \frac{P|h_1|^2}{N_0}\right). \tag{6.18}$$

Comparing this with the single-user bounds in (6.17) and recalling the symmetry assumption $|h_1| = |h_2|$, we have shown the triangle in Figure 6.6 to be the capacity region of the symmetric downlink AWGN channel.

Let us continue our thought process within the realm of the symmetry assumption. The rate pairs in the capacity region can be achieved by strategies used on point-to-point AWGN channels and sharing the degrees of freedom (time and bandwidth) between the two users. However, the symmetry between the two channels (cf. (6.16)) suggests a natural, and alternative, approach. The main idea is that if user 1 can successfully decode its data from y_1, then user 2, which has the same SNR, should also be able to decode the data of user 1 from y_2. Then user 2 can *subtract* the codeword of user 1 from its received signal y_2 to better decode its own data, i.e., it can perform *successive interference cancellation*. Consider the following strategy that *superposes* the signals of the two users, much like in a spread-spectrum CDMA system. The transmit signal is the sum of two signals,

$$x[m] = x_1[m] + x_2[m], \tag{6.19}$$

where $\{x_k[m]\}$ is the signal intended for user k. The transmitter encodes the information for each user using an i.i.d. Gaussian code spread on the entire bandwidth (and powers P_1, P_2, respectively, with $P_1 + P_2 = P$). User 1 treats the signal for user 2 as noise and can hence be communicated to reliably at a rate of

$$\begin{aligned}R_1 &= \log\left(1 + \frac{P_1|h_1|^2}{P_2|h_1|^2 + N_0}\right) \\ &= \log\left(1 + \frac{(P_1 + P_2)|h_1|^2}{N_0}\right) - \log\left(1 + \frac{P_2|h_1|^2}{N_0}\right). \end{aligned} \tag{6.20}$$

User 2 performs successive interference cancellation: it first decodes the data of user 1 by treating x_2 as noise, subtracts the exactly determined (with high probability) user 1 signal from y_2 and extracts its data. Thus user 2 can support reliably a rate

$$R_2 = \log\left(1 + \frac{P_2|h_2|^2}{N_0}\right). \tag{6.21}$$

This superposition strategy is schematically represented in Figures 6.7 and 6.8. Using the power constraint $P_1 + P_2 = P$ we see directly from (6.20) and (6.21) that the rate pairs in the capacity region (Figure 6.6) can be achieved by this strategy as well. We have hence seen two coding schemes for the

Figure 6.7 Superposition encoding example. The QPSK constellation of user 2 is superimposed on that of user 1.
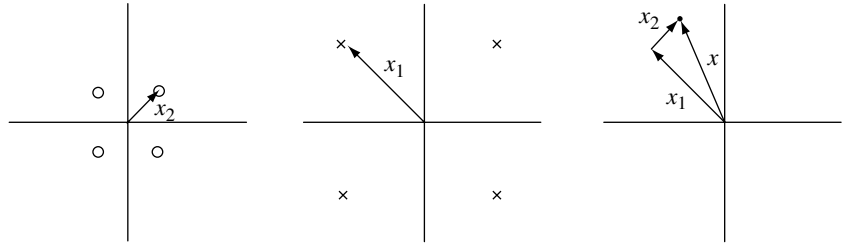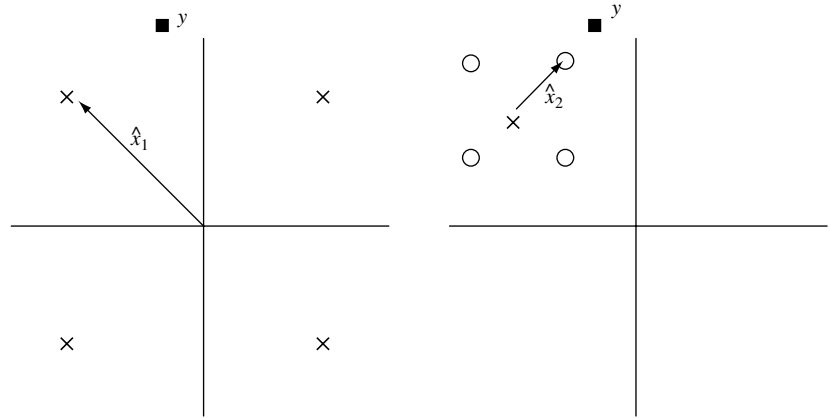

Figure 6.8 Superposition decoding example. The transmitted constellation point of user 1 is decoded first, followed by decoding of the constellation point of user 2.

symmetric downlink AWGN channel that are both optimal: single-user codes followed by orthogonalization of the degrees of freedom among the users, and the superposition coding scheme.

6.2.2 General case: superposition coding achieves capacity

Let us now return to the general downlink AWGN channel without the symmetry assumption and take $|h_1| < |h_2|$. Now user 2 has a better channel than user 1 and hence can decode any data that user 1 can successfully decode. Thus, we can use the superposition coding scheme: First the transmit signal is the (linear) superposition of the signals of the two users. Then, user 1 treats the signal of user 2 as noise and decodes its data from y_1. Finally, user 2, which has the better channel, performs SIC: it decodes the data of user 1 (and hence the transmit signal corresponding to user 1's data) and then proceeds to subtract the transmit signal of user 1 from y_2 and decode its data. As before, with each possible power split of $P = P_1 + P_2$, the following rate pair can be achieved:

$$
\begin{aligned}
R_1 &= \log\left(1 + \frac{P_1|h_1|^2}{P_2|h_1|^2 + N_0}\right) \text{ bits/s/Hz,} \\
R_2 &= \log\left(1 + \frac{P_2|h_2|^2}{N_0}\right) \text{ bits/s/Hz.}
\end{aligned}
\tag{6.22}
$$

On the other hand, orthogonal schemes achieve, for each power split $P = P_1 + P_2$ and degree-of-freedom split $\alpha \in [0, 1]$, as in the uplink (cf. (6.8) and (6.9)),

$$R_1 = \alpha \log \left(1 + \frac{P_1|h_1|^2}{\alpha N_0}\right) \text{ bits/s/Hz},$$

$$R_2 = (1-\alpha) \log \left(1 + \frac{P_2|h_2|^2}{(1-\alpha)N_0}\right) \text{ bits/s/Hz}. \qquad (6.23)$$

Here, α represents the fraction of the bandwidth devoted to user 1. Figure 6.9 plots the boundaries of the rate regions achievable with superposition coding and optimal orthogonal schemes for the asymmetric downlink AWGN channel (with $\mathsf{SNR}_1 = 0\,\text{dB}$ and $\mathsf{SNR}_2 = 20\,\text{dB}$). We observe that the performance of the superposition coding scheme is better than that of the orthogonal scheme.

One can show that the superposition decoding scheme is strictly better than the orthogonalization schemes (except for the two corner points where only one user is being communicated to). That is, for any rate pair achieved by orthogonalization schemes there is a power split for which the successive decoding scheme achieves rate pairs that are strictly larger (see Exercise 6.25). This gap in performance is more pronounced when the asymmetry between the two users deepens. In particular, superposition coding can provide a very reasonable rate to the strong user, while achieving close to the single-user bound for the weak user. In Figure 6.9, for example, while maintaining the rate of the weaker user R_1 at 0.9 bits/s/Hz, superposition coding can provide a rate of around $R_2 = 3$ bits/s/Hz to the strong user while an orthogonal scheme can provide a rate of only around 1 bits/s/Hz. Intuitively, the strong user, being at high SNR, is degree-of-freedom limited and superposition coding allows it to use the full degrees of freedom of the channel while being allocated only a small amount of transmit power, thus causing small amount

Figure 6.9 The boundary of rate pairs (in bits/s/Hz) achievable by superposition coding (solid line) and orthogonal schemes (dashed line) for the two-user asymmetric downlink AWGN channel with the user SNRs equal to 0 and 20 dB (i.e., $P|h_1|^2/N_0 = 1$ and $P|h_2|^2/N_0 = 100$). In the orthogonal schemes, both the power split $P = P_1 + P_2$ and split in degrees of freedom α are jointly optimized to compute the boundary.
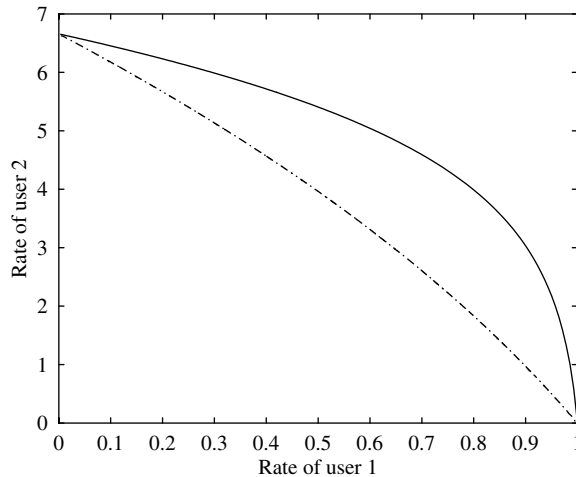
of interference to the weak user. In contrast, an orthogonal scheme has to allocate a significant fraction of the degrees of freedom to the weak user to achieve near single-user performance, and this causes a large degradation in the performance of the strong user.

So far we have considered a specific signaling scheme: linear superposition of the signals of the two users to form the transmit signal (cf. (6.19)). With this specific encoding method, the SIC decoding procedure is optimal. However, one can show that *this scheme in fact achieves the capacity* and the boundary of the capacity region of the downlink AWGN channel is given by (6.22) (Exercise 6.26).

While we have restricted ourselves to two users in the presentation, these results have natural extensions to the general K-user downlink channel. In the symmetric case $|h_k| = |h|$ for all k, the capacity region is given by the single constraint

$$\sum_{k=1}^{K} R_k < \log\left(1 + \frac{P|h|^2}{N_0}\right). \tag{6.24}$$

In general with the ordering $|h_1| \leq |h_2| \leq \cdots \leq |h_K|$, the boundary of the capacity region of the downlink AWGN channel is given by the parameterized rate tuple

$$R_k = \log\left(1 + \frac{P_k|h_k|^2}{N_0 + \left(\sum_{j=k+1}^{K} P_j\right)|h_k|^2}\right), \qquad k = 1\ldots K, \tag{6.25}$$

where $P = \sum_{k=1}^{K} P_k$ is the power split among the users. Each rate tuple on the boundary, as in (6.25), is achieved by superposition coding.

Since we have a full characterization of the tradeoff between the rates at which users can be reliably communicated to, we can easily derive specific scalar performance measures. In particular, we focused on sum capacity in the *uplink* analysis; to achieve the sum capacity we required all the users to transmit simultaneously (using the SIC receiver to decode the data). In contrast, we see from (6.25) that the sum capacity of the downlink is achieved by transmitting to a *single* user, the user with the highest SNR.

Summary 6.1 Uplink and downlink AWGN capacity

Uplink:

$$y[m] = \sum_{k=1}^{K} x_k[m] + w[m] \tag{6.26}$$

with user k having power constraint P_k.

Achievable rates satisfy:

$$\sum_{k \in \mathcal{S}} R_k \leq \log \left(1 + \frac{\sum_{k \in \mathcal{S}} P_k}{N_0} \right) \qquad \text{for all } \mathcal{S} \subset \{1, \ldots, K\} \qquad (6.27)$$

The $K!$ corner points are achieved by SIC, one corner point for each cancellation order. They all achieve the same optimal sum rate.

A natural ordering would be to decode starting from the strongest user first and move towards the weakest user.

Downlink:

$$y_k[m] = h_k x[m] + w_k[m], \qquad k = 1, \ldots K \qquad (6.28)$$

with $|h_1| \leq |h_2| \leq \ldots \leq |h_K|$.

The boundary of the capacity region is given by the rate tuples:

$$R_k = \log \left(1 + \frac{P_k |h_k|^2}{N_0 + (\sum_{j=k+1}^{K} P_j)|h_k|^2} \right), \qquad k = 1 \ldots K, \qquad (6.29)$$

for all possible splits $P = \sum_k P_k$ of the total power at the base-station.

The optimal points are achieved by superposition coding at the transmitter and SIC at each of the receivers.

The cancellation order at every receiver is *always* to decode the weaker users before decoding its own data.

Discussion 6.1 SIC: implementation issues

We have seen that successive interference cancellation plays an important role in achieving the capacities of both the uplink and the downlink channels. In contrast to the receivers for the multiple access systems in Chapter 4, SIC is a multiuser receiver. Here we discuss several potential practical issues in using SIC in a wireless system.

- **Complexity scaling with the number of users** In the uplink, the base-station has to decode the signals of every user in the cell, whether it uses the conventional single-user receiver or the SIC. In the downlink, on the other hand, the use of SIC at the mobile means that it now has to decode information intended for some of the other users, something it would not be doing in a conventional system. Then the complexity at each mobile scales with the number of users in the cell; this is not very acceptable. However, we have seen that superposition coding in conjunction with

SIC has the largest performance gain when the users have very disparate channels from the base-station. Due to the spatial geometry, typically there are only a few users close to the base-station while most of the users are near the edge of the cell. This suggests a practical way of limiting complexity: break the users in the cell into groups, with each group containing a small number of users with disparate channels. Within each group, superposition coding/SIC is performed, and across the groups, transmissions are kept orthogonal. This should capture a significant part of the performance gain.

- **Error propagation** Capacity analysis assumes error-free decoding but of course, with actual codes, errors are made. Once an error occurs for a user, all the users later in the SIC decoding order will very likely be decoded incorrectly. Exercise 6.12 shows that if $p_e^{(i)}$ is the probability of decoding the ith user incorrectly, assuming that all the previous users are decoded correctly, then the actual error probability for the kth user under SIC is at most

$$\sum_{i=1}^{k} p_e^{(i)}. \tag{6.30}$$

  So, if all the users are coded with the same target error probability assuming no propagation, the effect of error propagation degrades the error probability by a factor of at most the number of users K. If K is reasonably small, this effect can easily be compensated by using a slightly stronger code (by, say, increasing the block length by a small amount).

- **Imperfect channel estimates** To remove the effect of a user from the aggregate received signal, its contribution must be reconstructed from the decoded information. In a wireless multipath channel, this contribution depends also on the impulse response of the channel. Imperfect estimate of the channel will lead to *residual* cancellation errors. One concern is that, if the received powers of the users are very disparate (as in the example in Figure 6.3 where they differ by 20 dB), then the residual error from cancelling the stronger user can still swamp the weaker user's signal. On the other hand, it is also easier to get an accurate channel estimate when the user is strong. It turns out that these two effects compensate each other and the effect of residual errors does not grow with the power disparity (Exercise 6.13).

- **Analog-to-digital quantization error** When the received powers of the users are very disparate, the analog-to-digital (A/D) converter needs to have a very large dynamic range, and at the same time, enough resolution to quantize accurately the contribution from the weak signal. For example, if the power disparity is 20 dB, even 1-bit accuracy for the weak signal would require an 8-bit A/D converter. This may well pose an implementation constraint on how much gain SIC can offer.

6.3 Uplink fading channel

Let us now include fading. Consider the complex baseband representation of the uplink flat fading channel with K users:

$$y[m] = \sum_{k=1}^{K} h_k[m]x_k[m] + w[m], \tag{6.31}$$

where $\{h_k[m]\}_m$ is the fading process of user k. We assume that the fading processes of different users are independent of each other and $\mathbb{E}[|h_k[m]|^2] = 1$. Here, we focus on the symmetric case when each user is subject to the same average power constraint, P, and the fading processes are identically distributed. In this situation, the sum and the symmetric capacities are the key performance measures. We will see later in Section 6.7 how the insights obtained from this idealistic symmetric case can be applied to more realistic asymmetric situations. To understand the effect of the channel fluctuations, we make the simplifying assumption that the base-station (receiver) can perfectly track the fading processes of all the users.

6.3.1 Slow fading channel

Let us start with the slow fading situation where the time-scale of communication is short relative to the coherence time interval for all the users, i.e., $h_k[m] = h_k$ for all m. Suppose the users are transmitting at the same rate R bits/s/Hz. Conditioned on each realization of the channels $h_1, \ldots, h_K$, we have the standard uplink AWGN channel with received SNR of user k equal to $|h_k|^2 P/N_0$. If the symmetric capacity of this uplink AWGN channel is less than R, then the base-station can never recover *all* of the users' information accurately; this results in outage. From the expression for the capacity region of the general K-user uplink AWGN channel (cf. (6.10)), the probability of the outage event can be written as

$$p_{\text{out}}^{\text{ul}} := \mathbb{P}\left\{ \log\left(1 + \mathsf{SNR}\sum_{k \in \mathcal{S}} |h_k|^2\right) < |\mathcal{S}|R, \quad \text{for some } \mathcal{S} \subset \{1, \ldots, K\}\right\}. \tag{6.32}$$

Here $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$ and $\mathsf{SNR} := P/N_0$. The corresponding ϵ-outage symmetric capacity, C_ϵ^{sym}, is then the largest rate R such that the outage probability in (6.32) is smaller than or equal to ϵ.

In Section 5.4.1, we have analyzed the behavior of the outage capacity, $C_\epsilon(\mathsf{SNR})$, of the point-to-point slow fading channel. Since this corresponds to the performance of just a single user, it is equal to C_ϵ^{sym} with $K = 1$. With more than one user, C_ϵ^{sym} is only smaller: now each user has to deal not only

with a random channel realization but also inter-user interference. Orthogonal multiple access is designed to completely eliminate inter-user interference at the cost of lesser (by a factor of $1/K$) degrees of freedom to each user (but the SNR is boosted by a factor of K). Since the users experience independent fading, an individual outage probability of ϵ for each user translates into

$$1 - (1 - \epsilon)^K \approx K\epsilon$$

outage probability when we require *each* user's information to be successfully decoded. We conclude that the largest symmetric ϵ-outage rate with orthogonal multiple access is equal to

$$\frac{C_{\epsilon/K}(K\mathsf{SNR})}{K}. \tag{6.33}$$

How much improved are the outage performances of more sophisticated multiple access schemes, as compared to orthogonal multiple access?

At low SNRs, the outage performance for any K is just as poor as the point-to-point case (with the outage probability, p_{out}, in (5.54)): indeed, at low SNRs we can approximate (6.32) as

$$\begin{aligned}
p_{\text{out}}^{\text{ul}} &\approx \mathbb{P}\left\{\frac{|h_k|^2 P}{N_0} < R\log_e 2, \quad \text{for some } k \in \{1, \ldots, K\}\right\} \\
&\approx Kp_{\text{out}}. \tag{6.34}
\end{aligned}$$

So we can write

$$\begin{aligned}
C_{\epsilon}^{\text{sym}} &\approx C_{\epsilon/K}(\mathsf{SNR}) \\
&\approx F^{-1}\left(1 - \frac{\epsilon}{K}\right) C_{\text{awgn}}. \tag{6.35}
\end{aligned}$$

Here we used the approximation for C_{ϵ} at low SNR in (5.61). Since C_{awgn} is linear in SNR at low SNR,

$$C_{\epsilon}^{\text{sym}} \approx \frac{C_{\epsilon/K}(K\mathsf{SNR})}{K}, \tag{6.36}$$

the same performance as orthogonal multiple access (cf. (6.33)).

The analysis at high SNR is more involved, so to get a feel for the role of inter-user interference on the outage performance of optimal multiple access schemes, we plot $C_{\epsilon}^{\text{sym}}$ for $K = 2$ users as compared to C_{ϵ}, for Rayleigh fading, in Figure 6.10. As SNR increases, the ratio of $C_{\epsilon}^{\text{sym}}$ to C_{ϵ} increases; thus the effect of the inter-user interference is becoming smaller. However, as SNR becomes very large, the ratio starts to decrease; the inter-user interference begins to dominate. In fact, at very large SNRs the ratio drops back to $1/K$ (Exercise 6.14). We will obtain a deeper understanding of this behavior when we study outage in the uplink with multiple antennas in Section 10.1.4.

Figure 6.10 Plot of the symmetric ϵ-outage capacity of the two-user Rayleigh slow fading uplink as compared to C_ϵ, the corresponding performance of a point-to-point Rayleigh slow fading channel.
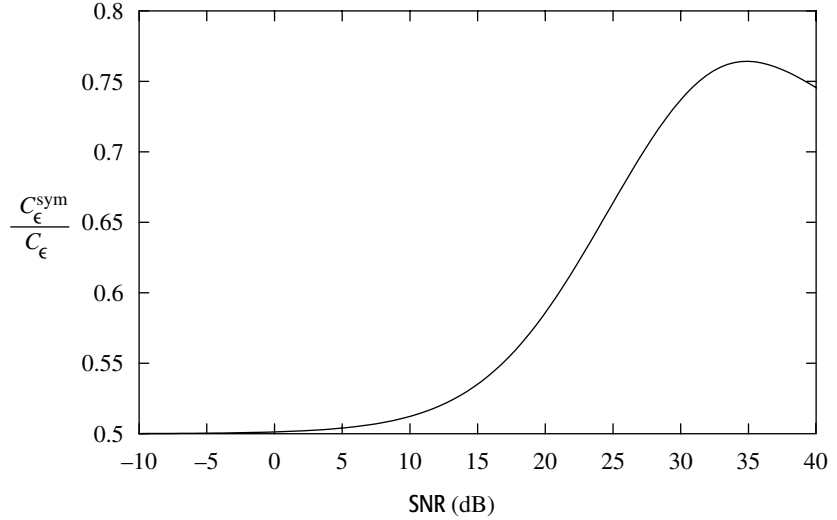
6.3.2 Fast fading channel

Let us now turn to the fast fading scenario, where each $\{h_k[m]\}_m$ is modelled as a time-varying ergodic process. With the ability to code over multiple coherence time intervals, we can have a meaningful definition of the capacity region of the uplink fading channel. With only receiver CSI, the transmitters cannot track the channel and there is no dynamic power allocation. Analogous to the discussion in the point-to-point case (cf. Section 5.4.5 and, in particular, (5.89)), the sum capacity of the uplink fast fading channel can be expressed as:

$$C_{\text{sum}} = \mathbb{E}\left[\log\left(1 + \frac{\sum_{k=1}^{K} |h_k|^2 P}{N_0}\right)\right]. \tag{6.37}$$

Here h_k is the random variable denoting the fading of user k at a particular time and the time averages are taken to converge to the same limit for all realizations of the fading process (i.e., the fading processes are ergodic). A formal derivation of the capacity region of the fast fading uplink (with potentially multiple antenna elements) is carried out in Appendix B.9.3.

How does this compare to the sum capacity of the uplink channel without fading (cf. (6.12))? Jensen's inequality implies that

$$\mathbb{E}\left[\log\left(1 + \frac{\sum_{k=1}^{K} |h_k|^2 P}{N_0}\right)\right] \leq \log\left(1 + \frac{\mathbb{E}[\sum_{k=1}^{K} |h_k|^2]P}{N_0}\right)$$

$$= \log\left(1 + \frac{KP}{N_0}\right).$$

Hence, without channel state information at the transmitter, fading always hurts, just as in the point-to-point case. However, when the number of users becomes large, $1/K \cdot \sum_{k=1}^{K} |h_k|^2 \to 1$ with probability 1, and the penalty due to fading vanishes.

To understand why the effect of fading goes away as the number of users grows, let us focus on a specific decoding strategy to achieve the sum capacity. With each user spreading their information on the entire bandwidth simultaneously, the successive interference cancellation (SIC) receiver, which is optimal for the uplink AWGN channel, is also optimal for the uplink fading channel. Consider the kth stage of the cancellation procedure, where user k is being decoded and users $k+1, \ldots, K$ are not canceled. The effective channel that user k sees is

$$y[m] = h_k[m]x_k[m] + \sum_{i=k+1}^{K} h_i[m]x_i[m] + w[m]. \qquad (6.38)$$

The rate that user k gets is

$$R_k = \mathbb{E}\left[\log\left(1 + \frac{|h_k|^2 P}{\sum_{i=k+1}^{K} |h_i|^2 P + N_0}\right)\right]. \qquad (6.39)$$

Since there are many users sharing the spectrum, the SINR for user k is low. Thus, the capacity penalty due to the fading of user k is small (cf. (5.92)). Moreover, there is also *averaging* among the interferers. Thus, the effect of the fading of the interferers also vanishes. More precisely,

$$R_k \approx \mathbb{E}\left[\frac{|h_k|^2 P}{\sum_{i=k+1}^{K} |h_i|^2 P + N_0}\right]\log_2 e$$

$$\approx \mathbb{E}\left[\frac{|h_k|^2 P}{(K-k)P + N_0}\right]\log_2 e$$

$$= \frac{P}{(K-k)P + N_0}\log_2 e,$$

which is the rate that user k would have got in the (unfaded) AWGN channel. The first approximation comes from the linearity of $\log(1 + \mathsf{SNR})$ for small SNR, and the second approximation comes from the law of large numbers.

In the AWGN case, the sum capacity can be achieved by an orthogonal multiple access scheme which gives a fraction, $1/K$, of the total degrees of freedom to each user. How about the fading case? The sum rate achieved by this orthogonal scheme is

$$\sum_{k=1}^{K} \frac{1}{K}\mathbb{E}\left[\log\left(1 + \frac{K|h_k|^2 P}{N_0}\right)\right] = \mathbb{E}\left[\log\left(1 + \frac{K|h_k|^2 P}{N_0}\right)\right], \qquad (6.40)$$

which is strictly less than the sum capacity of the uplink fading channel (6.37) for $K \geq 2$. In particular, the penalty due to fading persists even when there is a large number of users.

6.3.3 Full channel side information

We now come to a case of central interest in this chapter, the fast fading channel with tracking of the channels of all the users at the receiver and all the transmitters.[5] As opposed to the case with only receiver CSI, we can now dynamically allocate powers to the users as a function of the channel states. Analogous to the point-to-point case, we can without loss of generality focus on the simple block fading model

$$y[m] = \sum_{k=1}^{K} h_k[m] x_k[m] + w[m], \qquad (6.41)$$

where $h_k[m] = h_{k,\ell}$ remains constant over the ℓth 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 uplink channel with L sub-channels which fade independently. Each sub-channel is an uplink AWGN channel. For a given realization of the channel gains $h_{k,\ell}$, $k = 1, \ldots, K, \ell = 1, \ldots, L$, the sum capacity (in bits/symbol) of this parallel channel is, as for the point-to-point case (cf. (5.95)),

$$\max_{P_{k,\ell}: k=1, \ldots, K, \ell=1, \ldots, L} \frac{1}{L} \sum_{\ell=1}^{L} \log \left(1 + \frac{\sum_{k=1}^{K} P_{k,\ell} |h_{k,\ell}|^2}{N_0} \right), \qquad (6.42)$$

subject to the powers being non-negative and the average power constraint on each user:

$$\frac{1}{L} \sum_{\ell=1}^{L} P_{k,\ell} = P, \qquad k = 1, \ldots, K. \qquad (6.43)$$

The solution to this optimization problem as $L \to \infty$ yields the appropriate power allocation policy to be followed by the users.

As discussed in the point-to-point communication context with full CSI (cf. Section 5.4.6), we can use a variable rate coding scheme: in the ℓth sub-channel, the transmit powers dictated by the solution to the optimization problem above (6.42) are used by the users and a code designed for this fading state is used. For this code, each codeword sees a *time-invariant* uplink

---

[5] As we will see, the transmitters will not need to explicitly keep track of the channel variations of *all* the users. Only an appropriate function of the channels of all the users needs to be tracked, which the receiver can compute and feed back to the users.

AWGN channel. Thus, we can use the encoding and decoding procedures for the code designed for the uplink AWGN channel. In particular, to achieve the maximum sum rate, we can use orthogonal multiple access: this means that the codes designed for the *point-to-point* AWGN channel can be used. Contrast this with the case when only the receiver has CSI, where we have shown that orthogonal multiple access is *strictly* suboptimal for fading channels. Note that this argument on the optimality of orthogonal multiple access holds regardless of whether the users have symmetric fading statistics.

In the case of the *symmetric* uplink considered here, the optimal power allocation takes on a particularly simple structure. To derive it, let us consider the optimization problem (6.42), but with the individual power constraints in (6.43) relaxed and replaced by a total power constraint:

$$\frac{1}{L} \sum_{\ell=1}^{L} \sum_{k=1}^{K} P_{k,\ell} = KP. \tag{6.44}$$

The sum rate in the ℓth sub-channel is

$$\log\left(1 + \frac{\sum_{k=1}^{K} P_{k,\ell} |h_{k,\ell}|^2}{N_0}\right), \tag{6.45}$$

and for a given total power $\sum_{k=1}^{K} P_{k,\ell}$ allocated to the ℓth sub-channel, this quantity is maximized by giving all that power to the user with the strongest channel gain. Thus, the solution of the optimization problem (6.42) subject to the constraint (6.44) is that at each time, allow only the user with the *best* channel to transmit. Since there is just one user transmitting at any time, we have reduced to a point-to-point problem and can directly infer from our discussion in Section 5.4.6 that the best user allocates its power according to the *waterfilling* policy. More precisely, the optimal power allocation policy is

$$P_{k,\ell} = \begin{cases} \left(\dfrac{1}{\lambda} - \dfrac{N_0}{\max_i |h_{i,\ell}|^2}\right)^{+} & \text{if } |h_{k,\ell}| = \max_i |h_{i,\ell}|, \\ 0 & \text{else,} \end{cases} \tag{6.46}$$

where λ is chosen to meet the sum power constraint (6.44). Taking the number of coherence periods $L \to \infty$ and appealing to the ergodicity of the fading process, we get the optimal capacity-achieving power allocation strategy, which allocates powers to the users as a function of the joint channel state $\mathbf{h} := (h_1, \ldots, h_K)$:

$$P_k^*(\mathbf{h}) = \begin{cases} \left(\dfrac{1}{\lambda} - \dfrac{N_0}{\max_i |h_i|^2}\right)^{+} & \text{if } |h_k|^2 = \max_i |h_i|^2, \\ 0 & \text{else,} \end{cases} \tag{6.47}$$

with λ chosen to satisfy the power constraint

$$\sum_{k=1}^{K} \mathbb{E}[P_k^*(\mathbf{h})] = KP. \qquad (6.48)$$

(Rigorously speaking, this formula is valid only when there is exactly one user with the strongest channel. See Exercise 6.16 for the generalization to the case when multiple users can have the same fading state.) The resulting sum capacity is

$$\boxed{C_{\text{sum}} = \mathbb{E}\left[\log\left(1 + \frac{P_{k^*}(\mathbf{h})|h_{k^*}|^2}{N_0}\right)\right],} \qquad (6.49)$$

where $k^*(\mathbf{h})$ is the index of the user with the strongest channel at joint channel state $\mathbf{h}$.

We have derived this result assuming a *total* power constraint on all the users, but by symmetry, the power consumption of all the users is the same under the optimal solution (recall that we are assuming independent and identical fading processes across the users here). Therefore the individual power constraints in (6.43) are automatically satisfied and we have solved the original problem as well.

This result is the multiuser generalization of the idea of opportunistic communication developed in Chapter 5: resource is allocated at the times and to the user whose channel is good.

When one attempts to generalize the optimal power allocation solution from the point-to-point setting to the multiuser setting, it may be tempting to think of "users" as a new dimension, in addition to the time dimension, over which dynamic power allocation can be performed. This may lead us to guess that the optimal solution is waterfilling over the joint time/user space. This, as we have already seen, is not the correct solution. The flaw in this reasoning is that having multiple users *does not* provide additional degrees of freedom in the system: the users are just sharing the time/frequency degrees of freedom already existing in the channel. Thus, the optimal power allocation problem should really be thought of as how to partition the total resource (power) across the time/frequency degrees of freedom and how to share the resource across the users in each of those degrees of freedom. The above solution says that from the point of view of maximizing the sum capacity, the optimal sharing is just to allocate all the power to the user with the strongest channel on that degree of freedom.

We have focused on the sum capacity in the symmetric case where users have identical channel statistics and power constraints. It turns out that in the asymmetric case, the optimal strategy to achieve sum capacity is still to have one user transmitting at a time, but the criterion of choosing which user is different. This problem is analyzed in Exercise 6.15. However, in the asymmetric case, maximizing the sum rate may not be the appropriate objective,

since the user with the statistically better channel may get a much higher rate at the expense of the other users. In this case, one may be interested in operating at points in the multiuser capacity region of the uplink fading channel other than the point maximizing the sum rate. This problem is analyzed in Exercise 6.18. It turns out that, as in the time-invariant uplink, orthogonal multiple access is not optimal. Instead, users transmit simultaneously and are jointly decoded (using SIC, for example), even though the rates and powers are still dynamically allocated as a function of the channel states.

Summary 6.2 Uplink fading channel

Slow Rayleigh fading At low SNR, the symmetric outage capacity is equal to the outage capacity of the point-to-point channel, but scaled down by the number of users. At high SNR, the symmetric outage capacity for moderate number of users is approximately equal to the outage capacity of the point-to-point channel. Orthogonal multiple access is close to optimal at low SNR.

Fast fading, receiver CSI With a large number of users, each user gets the same performance as in an uplink AWGN channel with the same average SNR. Orthogonal multiple access is strictly suboptimal.

Fast fading, full CSI Orthogonal multiple access can still achieve the sum capacity. In a symmetric uplink, the policy of allowing only the best user to transmit at each time achieves the sum capacity.

6.4 Downlink fading channel

We now turn to the downlink fading channel with K users:

$$y_k[m] = h_k[m]x[m] + w_k[m], \qquad k = 1, \ldots, K, \qquad (6.50)$$

where $\{h_k[m]\}_m$ is the channel fading process of user k. We retain the average power constraint of P on the transmit signal and $w_k[m] \sim \mathcal{CN}(0, N_0)$ to be i.i.d. in time m (for each user $k = 1, \ldots, K$).

As in the uplink, we consider the symmetric case: $\{h_k[m]\}_m$ are identically distributed processes for $k = 1 \ldots K$. Further, let us also make the same assumption we did in the uplink analysis: the processes $\{h_k[m]\}_m$ are ergodic (i.e., the time average of every realization equals the statistical average).

6.4.1 Channel side information at receiver only

Let us first consider the case when the receivers can track the channel but the transmitter does not have access to the channel realizations (but has access

to a statistical characterization of the channel processes of the users). To get a feel for good strategies to communicate on this fading channel and to understand the capacity region, we can argue as in the downlink AWGN channel. We have the single-user bounds, in terms of the point-to-point fading channel capacity in (5.89):

$$R_k < \mathbb{E}\left[\log\left(1 + \frac{|h|^2 P}{N_0}\right)\right], \qquad k = 1, \ldots, K, \tag{6.51}$$

where h is a random variable distributed as the stationary distribution of the ergodic channel processes. In the symmetric downlink AWGN channel, we argued that the users have the same channel quality and hence could decode each other's data. Here, the fading statistics are symmetric and by the assumption of ergodicity, we can extend the argument of the AWGN case to say that, if user k can decode its data reliably, then all the other users can also successfully decode user k's data. Analogous to (6.18) in the AWGN downlink analysis, we obtain

$$\sum_{k=1}^{K} R_k < \mathbb{E}\left[\log\left(1 + \frac{|h|^2 P}{N_0}\right)\right]. \tag{6.52}$$

An alternative way to see that the right hand side in (6.52) is the best sum rate one can achieve is outlined in Exercise 6.27. The bound (6.52) is clearly achievable by transmitting to one user only or by time-sharing between any number of users. Thus in the symmetric fading channel, we obtain the same conclusion as in the symmetric AWGN downlink: the rate pairs in the capacity region can be achieved by both orthogonalization schemes and superposition coding.

How about the downlink fading channel with *asymmetric* fading statistics of the users? While we can use the orthogonalization scheme in this asymmetric model as well, the applicability of superposition decoding is not so clear. Superposition coding was successfully applied in the downlink AWGN channel because there is an *ordering* of the channel strength of the users from weak to strong. In the asymmetric fading case, users in general have different fading distributions and there is no longer a *complete* ordering of the users. In this case, we say that the downlink channel is *non-degraded* and little is known about good strategies for communication. Another interesting situation when the downlink channel is non-degraded arises when the transmitter has an array of multiple antennas; this is studied in Chapter 10.

6.4.2 Full channel side information

We saw in the uplink that the communication scenario becomes more interesting when the transmitters can track the channel as well. In this case, the transmitters can vary their powers as a function of the channel. Let us now

turn to the analogous situation in the downlink where the single transmitter tracks all the channels of the users it is communicating to (the users continue to track their individual channels). As in the uplink, we can allocate powers to the users as a function of the channel fade level. To see the effect, let us continue focusing on sum capacity. We have seen that without fading, the sum capacity is achieved by transmitting only to the best user. Now as the channels vary, we can pick the best user at each time and further allocate it an appropriate power, subject to a constraint on the average power. Under this strategy, the downlink channel reduces to a point-to-point channel with the channel gain distributed as

$$\max_{k=1\ldots K} |h_k|^2.$$

The optimal power allocation is the, by now familiar, waterfilling solution:

$$P^*(\mathbf{h}) = \left(\frac{1}{\lambda} - \frac{N_0}{\max_{k=1\ldots K} |h_k|^2} \right)^+, \qquad (6.53)$$

where $\mathbf{h} = (h_1, \ldots, h_K)^t$ is the joint fading state and $\lambda > 0$ is chosen such that the average power constraint is met. The optimal strategy is exactly the same as in the sum capacity of the uplink. The sum capacity of the downlink is:

$$\mathbb{E}\left[\log \left(1 + \frac{P^*(\mathbf{h})(\max_{k=1\ldots K} |h_k^2|)}{N_0} \right) \right]. \qquad (6.54)$$

6.5 Frequency-selective fading channels

The extension of the flat fading analysis in the uplink and the downlink to underspread frequency-selective fading channels is conceptually straightforward. As we saw in Section 5.4.7 in the point-to-point setting, we can think of the underspread channel as a set of parallel sub-carriers over each coherence time interval and varying independently from one coherence time interval to the other. We can see this constructively by imposing a cyclic prefix to all the transmit signals; the cyclic prefix should be of length that is larger than the largest multipath delay spread that we are likely to encounter among the different users. Since this overhead is fixed, the loss is amortized when communicating over a long block length.

We can apply exactly the same OFDM transformation to the multiuser channels. Thus on the nth sub-carrier, we can write the uplink channel as

$$\tilde{y}_n[i] = \sum_{k=1}^{K} \tilde{h}_n^{(k)}[i]\, \tilde{d}_n^{(k)}[i] + \tilde{w}_n[i], \qquad (6.55)$$

where $\widetilde{\mathbf{d}}^{(k)}[i]$, $\widetilde{\mathbf{h}}^{(k)}[i]$ and $\widetilde{\mathbf{y}}[i]$, respectively, represent the DFTs of the transmitted sequence of user k, of the channel and of the received sequence at OFDM symbol time i.

The flat fading uplink channel can be viewed as a set of parallel multiuser sub-channels, one for each coherence time interval. With full CSI, the optimal strategy to maximize the sum rate in the symmetric case is to allow only the user with the best channel to transmit at each coherence time interval. The frequency-selective fading uplink channel can also be viewed as a set of parallel multiuser sub-channels, one for each sub-carrier and each coherence time interval. Thus, the optimal strategy is to allow the best user to transmit on each of these sub-channels. The power allocated to the best user is waterfilling over time and frequency. As opposed to the flat fading case, multiple users can now transmit at the same time, but over different sub-carriers. Exactly the same comments apply to the downlink.

6.6 Multiuser diversity

6.6.1 Multiuser diversity gain

Let us consider the sum capacity of the uplink and downlink flat fading channels (see (6.49) and (6.54), respectively). Each can be interpreted as the waterfilling capacity of a point-to-point link with a power constraint equal to the total transmit power (in the uplink this is equal to KP and in the downlink it is equal to P), and a fading process whose magnitude varies as $\{\max_k |h_k[m]|\}$. Compared to a system with a single transmitting user, the multiuser gain comes from two effects:

1. the increase in total transmit power in the case of the uplink;
2. the effective channel gain at time m that is improved from $|h_1[m]|^2$ to $\max_{1 \leq k \leq K} |h_k[m]|^2$.

The first effect already appeared in the uplink AWGN channel and also in the fading channel with channel side information only at the receiver. The second effect is entirely due to the ability to dynamically schedule resources among the users as a function of the channel state.

The sum capacity of the uplink Rayleigh fading channel with full CSI is plotted in Figure 6.11 for different numbers of users. The performance curves are plotted as a function of total $\mathsf{SNR} := KP/N_0$ so as to focus on the second effect. The sum capacity of the channel with only CSI at the receiver is also plotted for different numbers of users. The capacity of the point-to-point AWGN channel with received power KP (which is also the sum capacity of a K-user uplink AWGN channel) is shown as a baseline. Figure 6.12 focuses on the low SNR regime.

Figure 6.11 Sum capacity of
the uplink Rayleigh fading
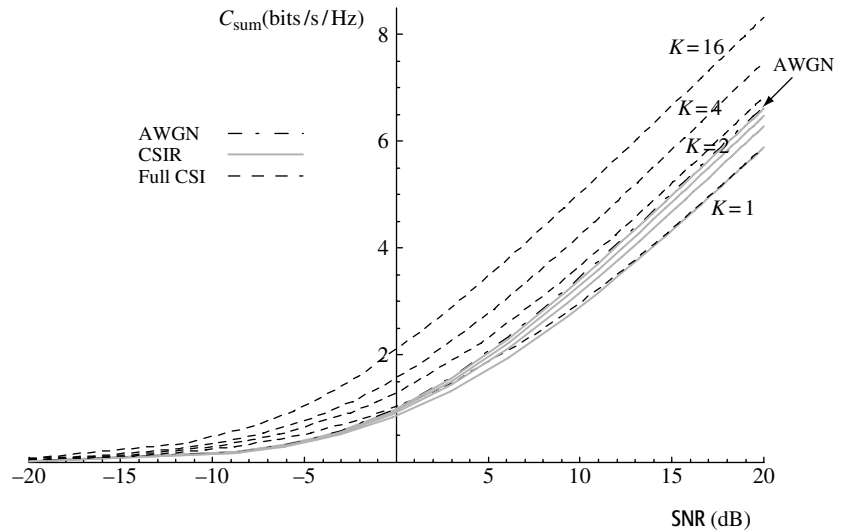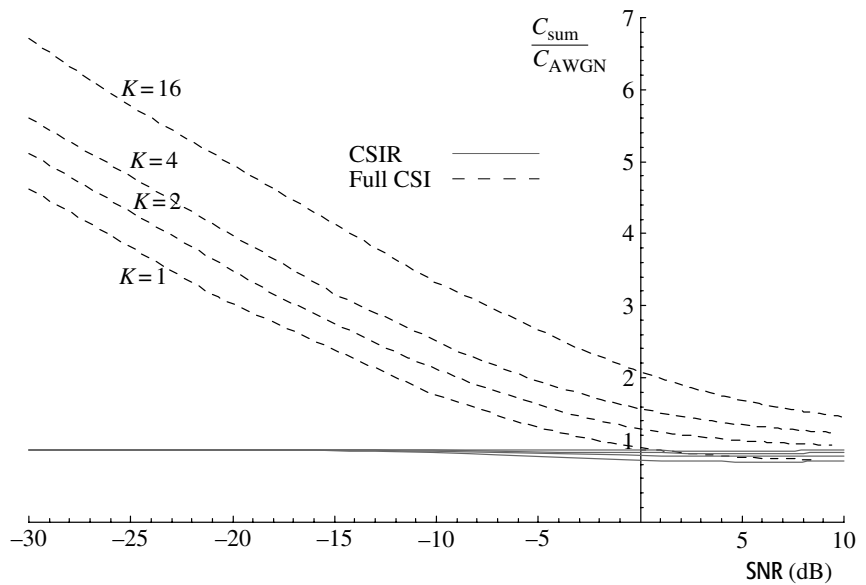channel plotted as a function
of SNR $= KP/N_0$.

Figure 6.12 Sum capacity of
the uplink Rayleigh fading
channel plotted as a function
of SNR $= KP/N_0$ in the low
SNR regime. Everything is
plotted as a fraction of the
AWGN channel capacity.

Several observations can be made from the plots:

- The sum capacity without transmitter CSI increases with the number of the users, but not significantly. This is due to the multiuser averaging effect explained in the last section. This sum capacity is always bounded by the capacity of the AWGN channel.
- The sum capacity with full CSI increases significantly with the number of users. In fact, with even two users, this sum capacity already exceeds that

of the AWGN channel. At 0 dB, the capacity with $K = 16$ users is about a factor of 2.5 of the capacity with $K = 1$. The corresponding power gain is about 7 dB. Compared to the AWGN channel, the capacity gain for $K = 16$ is about a factor of 2.2 and an SNR gain of 5.5 dB.

- For $K = 1$, the capacity benefit of transmitter CSI only becomes apparent at quite low SNR levels; at high SNR there is no gain. For $K > 1$ the benefit is apparent throughout the entire SNR range, although the relative gain is still more significant at low SNR. This is because the gain is still primarily a *power gain*.

The increase in the full CSI sum capacity comes from a *multiuser diversity* effect: when there are many users that fade *independently*, at any one time there is a high probability that one of the users will have a strong channel. By allowing only that user to transmit, the shared channel resource is used in the most efficient manner and the total system throughput is maximized. The larger the number of users, the stronger tends to be the strongest channel, and the more the multiuser diversity gain.

The amount of multiuser diversity gain depends crucially on the *tail* of the fading distribution $|h_k|^2$: the heavier the tail, the more likely there is a user with a very strong channel, and the larger the multiuser diversity gain. This is shown in Figure 6.13, where the sum capacity is plotted as a function of the number of users for both Rayleigh and Rician fading with κ-factor equal to 5, with the total SNR, equal to KP/N_0, fixed at 0 dB. Recall from
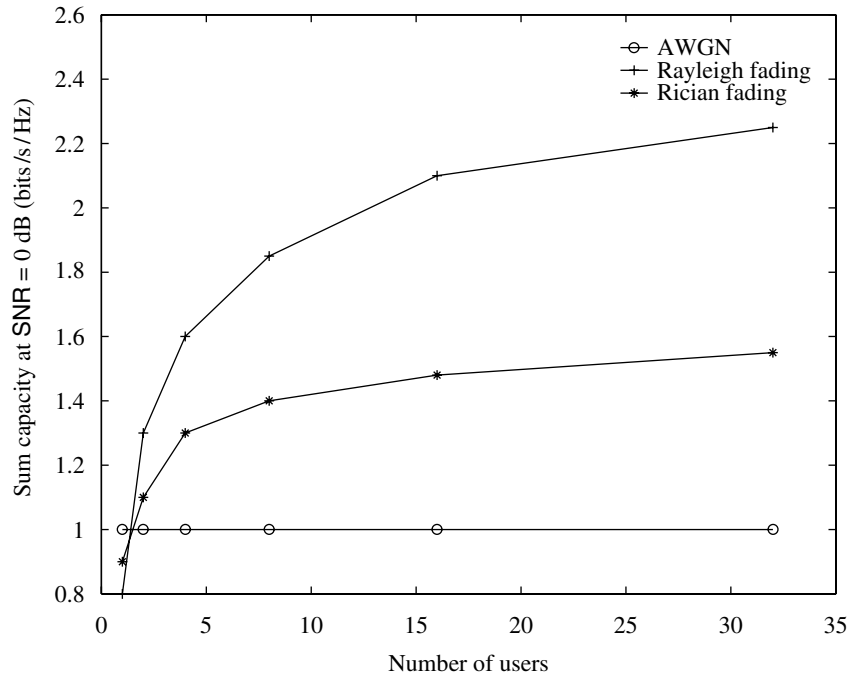

Figure 6.13 Multiuser diversity gain for Rayleigh and Rician fading channels ($\kappa = 5$); $KP/N_0 = 0$ dB.

Section 2.4 that, Rician fading models the situation when there is a strong specular line-of-sight path plus many small reflected paths. The parameter κ is defined as the ratio of the energy in the specular line-of-sight path to the energy in the diffused components. Because of the line-of-sight component, the Rician fading distribution is less "random" and has a lighter tail than the Rayleigh distribution with the same average channel gain. As a consequence, it can be seen that the multiuser diversity gain is significantly smaller in the Rician case compared to the Rayleigh case (Exercise 6.21).

6.6.2 Multiuser versus classical diversity

We have called the above explained phenomenon multiuser diversity. Like the diversity techniques discussed in Chapter 3, multiuser diversity also arises from the existence of independently faded signal paths, in this case from the multiple users in the network. However, there are several important differences. First, the main objective of the diversity techniques in Chapter 3 is to improve the *reliability* of communication in slow fading channels; in contrast, the role of multiuser diversity is to increase the total throughput over fast fading channels. Under the sum-capacity-achieving strategy, a user has no guarantee of a high rate in any particular slow fading state; only by averaging over the variations of the channel is a high long-term average throughput attained. Second, while the diversity techniques are designed to *counteract* the adverse effect of fading, multiuser diversity improves system performance by *exploiting* channel fading: channel fluctuations due to fading ensure that with high probability there is a user with a channel strength much larger than the mean level; by allocating all the system resources to that user, the benefit of this strong channel is fully capitalized. Third, while the diversity techniques in Chapter 3 pertain to a point-to-point link, the benefit of multiuser diversity is *system-wide*, across the users in the network. This aspect of multiuser diversity has ramifications on the implementation of multiuser diversity in a cellular system. We will discuss this next.

6.7 Multiuser diversity: system aspects

The cellular system requirements to extract the multiuser diversity benefits are:

- the base-station has access to channel quality measurements: in the downlink, we need each receiver to track its own channel SNR, through say a common downlink pilot, and feed back the instantaneous channel quality to the base-station (assuming an FDD system); and in the uplink, we need transmissions from the users so that their channel qualities can be tracked;

- the ability of the base-station to schedule transmissions among the users as well as to adapt the data rate as a function of the instantaneous channel quality.

These features are already present in the designs of many third-generation systems. Nevertheless, in practice there are several considerations to take into account before realizing such gains. In this section, we study three main hurdles towards a system implementation of the multiuser diversity idea and some prominent ways of addressing these issues.

1. **Fairness and delay** To implement the idea of multiuser diversity in a real system, one is immediately confronted with two issues: fairness and delay. In the ideal situation when users' fading *statistics* are the same, the strategy of communicating with the user having the best channel maximizes not only the total throughput of the system but also that of individual users. In reality, the statistics are not symmetric; there are users who are closer to the base-station with a better average SNR; there are users who are stationary and some that are moving; there are users who are in a rich scattering environment and some with no scatterers around them. Moreover, the strategy is only concerned with maximizing long-term average throughputs; in practice there are latency requirements, in which case the average throughput over the delay time-scale is the performance metric of interest. The challenge is to address these issues while at the same time exploiting the multiuser diversity gain inherent in a system with users having independent, fluctuating channel conditions. As a case study, we will look at one particular scheduler that harnesses multiuser diversity while addressing the real-world fairness and delay issues.

2. **Channel measurement and feedback** One of the key system requirements to harness multiuser diversity is to have scheduling decisions by the base-station be made as a function of the channel states of the users. In the uplink, the base-station has access to the user transmissions (over trickle channels which are used to convey control information) and has an estimate of the user channels. In the downlink, the users have access to their channel states but need to feedback these values to the base-station. Both the error in channel state measurement and the delay in feeding it back constitute a significant bottleneck in extracting the multiuser diversity gains.

3. **Slow and limited fluctuations** We have observed that the multiuser diversity gains depend on the distribution of channel fluctuations. In particular, larger and faster variations in a channel are preferred over slow ones. However, there may be a line-of-sight path and little scattering in the environment, and hence the dynamic range of channel fluctuations may be small. Further, the channel may fade very slowly compared to the delay constraints of the application so that transmissions cannot wait until the channel reaches its peak. Effectively, the dynamic range of channel fluctuations is small within the time-scale of interest. Both are important

sources of hindrance to implementing multiuser diversity in a real system. We will see a simple and practical scheme using an antenna array at the base-station that creates fast and large channel fluctuations even when the channel is originally slow fading with a small range of fluctuation.

6.7.1 Fair scheduling and multiuser diversity

As a case study, we describe a simple scheduling algorithm, called the *proportional fair* scheduler, designed to meet the challenges of delay and fairness constraints while harnessing multiuser diversity. This is the baseline scheduler for the downlink of IS-856, the third-generation data standard, introduced in Chapter 5. Recall that the downlink of IS-856 is TDMA-based, with users scheduled on time slots of length 1.67 ms based on the requested rates from the users (Figure 5.25). We have already discussed the rate adaptation mechanism in Chapter 5; here we will study the scheduling aspect.

Proportional fair scheduling: hitting the peaks

The *scheduler* decides which user to transmit information to at each time slot, based on the requested rates the base-station has previously received from the mobiles. The simplest scheduler transmits data to each user in a *round-robin* fashion, regardless of the channel conditions of the users. The scheduling algorithm used in IS-856 schedules in a *channel-dependent* manner to exploit multiuser diversity. It works as follows. It keeps track of the average throughput $T_k[m]$ of each user in an exponentially weighted window of length t_c. In time slot m, the base-station receives the "requested rates" $R_k[m]$, $k = 1, \ldots, K$, from all the users and the scheduling algorithm simply transmits to the user k^* with the largest

$$\frac{R_k[m]}{T_k[m]}$$

among all active users in the system. The average throughputs $T_k[m]$ are updated using an exponentially weighted low-pass filter:

$$T_k[m+1] = \begin{cases} (1 - 1/t_c)T_k[m] + (1/t_c)R_k[m] & k = k^*, \\ (1 - 1/t_c)T_k[m] & k \neq k^*. \end{cases} \tag{6.56}$$

One can get an intuitive feel of how this algorithm works by inspecting Figures 6.14 and 6.15. We plot the sample paths of the requested data rates of two users as a function of time slots (each time slot is 1.67 ms in IS-856). In Figure 6.14, the two users have identical fading *statistics*. If the scheduling time-scale t_c is much larger than the coherence time of the channels, then by symmetry the throughput of each user $T_k[m]$ converges to the same quantity. The scheduling algorithm reduces to always picking the user with the highest

Figure 6.14 For symmetric channel statistics of users, the scheduling algorithm reduces to serving each user with the largest requested rate.
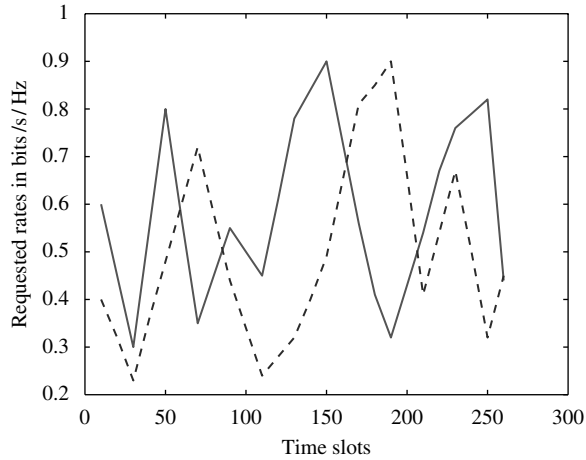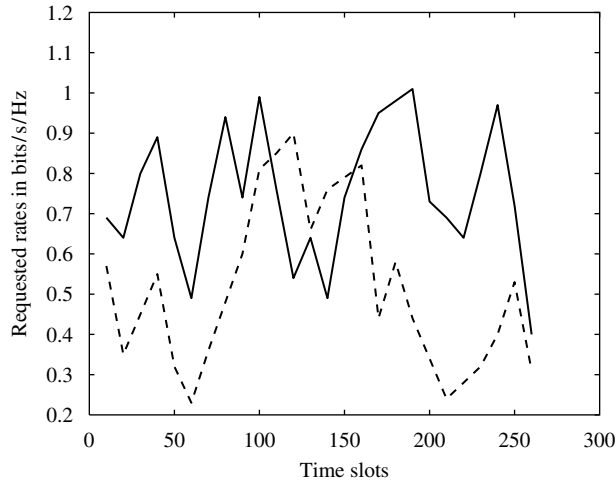
Figure 6.15 In general, with asymmetric user channel statistics, the scheduling algorithm serves each user when it is near its peak within the latency time-scale t_c.

requested rate. Thus, each user is scheduled when its channel is good and at the same time the scheduling algorithm is perfectly fair in the long-term.

In Figure 6.15, due perhaps to different distances from the base-station, one user's channel is much stronger than that of the other user on average, even though both channels fluctuate due to multipath fading. Always picking the user with the highest requested rate means giving all the system resources to the statistically stronger user, and would be highly unfair. In contrast, under the scheduling algorithm described above, users compete for resources not directly based on their requested rates but based on the rates normalized by their respective average throughputs. The user with the statistically stronger channel will have a higher average throughput.

Thus, the algorithm schedules a user when its instantaneous channel quality is high *relative* to its own average channel condition over the time-scale t_c.

In short, data are transmitted to a user when its channel is *near its own peaks*. Multiuser diversity benefit can still be extracted because channels of different users fluctuate independently so that if there is a sufficient number of users in the system, most likely there will be a user near its peak at any one time.

The parameter t_c is tied to the latency time-scale of the application. Peaks are defined with respect to this time-scale. If the latency time-scale is large, then the throughput is averaged over a longer time-scale and the scheduler can afford to wait longer before scheduling a user when its channel hits a really high peak.

The main theoretical property of this algorithm is the following: With a very large t_c (approaching ∞), the algorithm maximizes

$$\sum_{k=1}^{K} \log T_k, \qquad (6.57)$$

among all schedulers (see Exercise 6.28). Here, T_k is the long-term average throughput of user k.

Multiuser diversity and superposition coding

Proportional fair scheduling is an approach to deal with fairness among asymmetric users within the orthogonal multiple access constraint (TDMA in the case of IS-856). But we understand from Section 6.2.2 that for the AWGN channel, superposition coding in conjunction with SIC can yield significantly better performance than orthogonal multiple access in such asymmetric environments. One would expect similar gains in fading channels, and it is therefore natural to combine the benefits of superposition coding with multiuser diversity scheduling.

One approach is to divide the users in a cell into, say, two classes depending on whether they are near the base-station or near the cell edge, so that users in each class have statistically comparable channel strengths. Users whose current channel is instantaneously strongest in their own class are scheduled for simultaneous transmission via superposition coding (Figure 6.16). The user near the base-station can decode its own signal after stripping off the signal destined for the far-away user. By transmitting to the strongest user in each class, multiuser diversity benefits are captured. On the other hand, the nearby user has a very strong channel and the full degrees of freedom available (as opposed to only a fraction under orthogonal multiple access), and thus only needs to be allocated a small fraction of the power to enjoy very good rates. Allocating a small fraction of power to the nearby user has a salutary effect: the presence of this user will minimally affect the performance of the cell edge user. Hence, fairness can be maintained by a suitable allocation of power. The efficiency of this approach over proportional fair TDMA scheduling is quantified in Exercise 6.20. Exercise 6.19 shows that this strategy is in fact optimal in achieving *any* point on the boundary of

Figure 6.16 Superposition coding in conjunction with multiuser diversity scheduling. The strongest user from each cluster is scheduled and they are simultaneously transmitted to, via superposition coding.
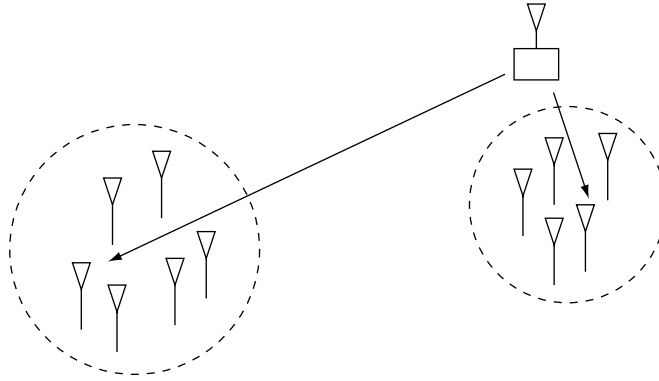
the downlink fading channel capacity region (as opposed to the strategy of transmitting to the user with the best channel overall, which is only optimal for the sum rate and which is an unfair operating point in this asymmetric scenario).

Multiuser diversity gain in practice

We can use the proportional fair algorithm to get some more insights into the issues involved in realizing multiuser diversity benefits in practice. Consider the plot in Figure 6.17, showing the total simulated throughput of the 1.25 MHz IS-856 downlink under the proportional fair scheduling algorithm in three environments:

- **Fixed** Users are fixed, but there are movements of objects around them (2 Hz Rician, $\kappa := E_{\text{direct}}/E_{\text{specular}} = 5$). Here E_{direct} is the energy in the direct
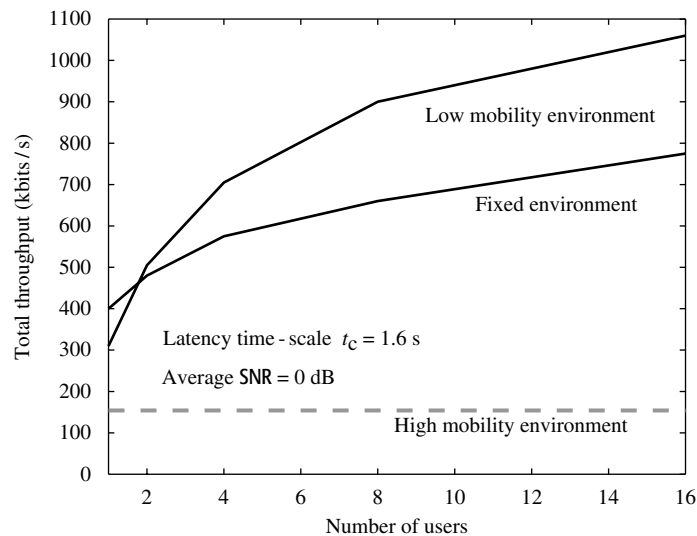

Figure 6.17 Multiuser diversity gain in fixed and mobile environments.

path that is not varying, while E_{specular} refers to the energy in the specular or time-varying component that is assumed to be Rayleigh distributed. The Doppler spectrum of this component follows Clarke's model with a Doppler spread of 2 Hz.

- **Low mobility** Users move at walking speeds (3 km/hr, Rayleigh).
- **High mobility** Users move at 30 km/hr, Rayleigh.

The average channel gain $\mathbb{E}[|h|^2]$ is kept the same in all the three scenarios for fairness of comparison. The total throughput increases with the number of users in both the fixed and low mobility environments, but the increase is more dramatic in the low mobility case. While the channel varies in both cases, the dynamic range and the rate of the variations is larger in the mobile environment than in the fixed one (Figure 6.18). This means that over the latency time-scale ($t_{\text{c}} = 1.67$ s in these examples) the peaks of the channel fluctuations are likely to be higher in the mobile environment, and the peaks are what determines the performance of the scheduling algorithm. Thus, the inherent multiuser diversity is more limited in the fixed environment.

Should one then expect an even higher throughput gain in the high mobility environment? In fact quite the opposite is true. The total throughput hardly increases with the number of users! It turns out that at this speed the receiver has trouble tracking and predicting the channel variations, so that the predicted channel is a low-pass smoothed version of the actual fading process. Thus, even though the actual channel fluctuates, opportunistic communication is impossible without knowing when the channel is actually good.

In the next section, we will discuss how the tracking of the channel can be improved in high mobility environments. In Section 6.7.3, we will discuss a scheme that *boosts* the inherent multiuser diversity in fixed environments.

6.7.2 Channel prediction and feedback

The prediction error is due to two effects: the error in measuring the channel from the pilot and the delay in feeding back the information to the base-station.
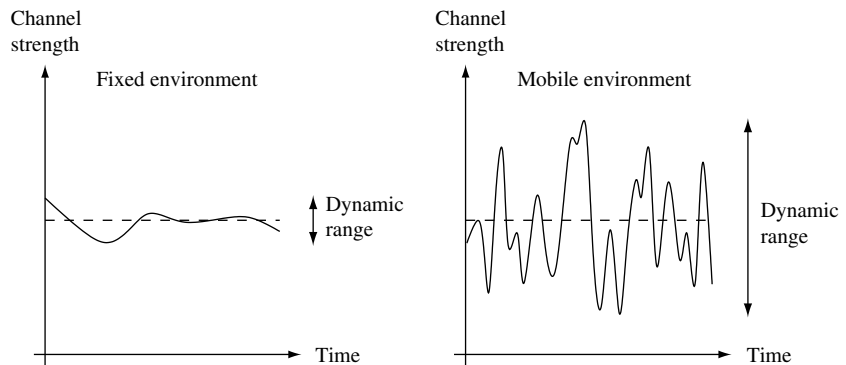
Figure 6.18 The channel varies much faster and has larger dynamic range in the mobile environment.

In the downlink, the pilot is shared between many users and is strong; so, the measurement error is quite small and the prediction error is mainly due to the feedback delay. In IS-856, this delay is about two time slots, i.e., 3.33 ms. At a vehicular speed of 30 km/h and carrier frequency of 1.9 GHz, the coherence time is approximately 2.5 ms; the channel coherence time is comparable to the delay and this makes prediction difficult.

One remedy to reduce the feedback delay is to shrink the size of the scheduling time slot. However, this increases the requested rate feedback frequency in the uplink and thus increases the system overhead. There are ways to reduce this feedback though. In the current system, *every* user feeds back the requested rates, but in fact only users whose channels are near their peaks have any chance of getting scheduled. Thus, an alternative is for each user to feed back the requested rate only when its current requested rate to average throughput ratio, $R_k[m]/T_k[m]$, exceeds a threshold γ. This threshold, γ, can be chosen to trade off the average aggregate amount of feedback the users send with the probability that none of the users sends any feedback in a given time slot (thus wasting the slot) (Exercise 6.22).

In IS-856, multiuser diversity scheduling is implemented in the downlink, but the same concept can be applied to the uplink. However, the issues of prediction error and feedback are different. In the uplink, the base-station would be measuring the channels of the users, and so a separate pilot would be needed for each user. The downlink has a single pilot and this amortization among the users is used to have a strong pilot. However, in the uplink, the fraction of power devoted to the pilot is typically small. Thus, it is expected that the *measurement* error will play a larger role in the uplink. Moreover, the pilot will have to be sent continuously even if the user is not currently scheduled, thus causing some interference to other users. On the other hand, the base-station only needs to broadcast which user is scheduled at that time slot, so the amount of feedback is much smaller than in the downlink (unless the selective feedback scheme is implemented).

The above discussion pertains to an FDD system. You are asked to discuss the analogous issues for a TDD system in Exercise 6.23.

6.7.3 Opportunistic beamforming using dumb antennas

The amount of multiuser diversity depends on the rate and dynamic range of channel fluctuations. In environments where the channel fluctuations are small, a natural idea comes to mind: why not amplify the multiuser diversity gain by *inducing* faster and larger fluctuations? Focusing on the downlink, we describe a technique that does this using multiple transmit antennas at the base-station as illustrated in Figure 6.19.

Consider a system with n_t transmit antennas at the base-station. Let $h_{lk}[m]$ be the complex channel gain from antenna l to user k in time m. In time m, the same symbol $x[m]$ is transmitted from all of the antennas except that it is

Figure 6.19 Same signal is transmitted over the two antennas with time-varying phase and powers.
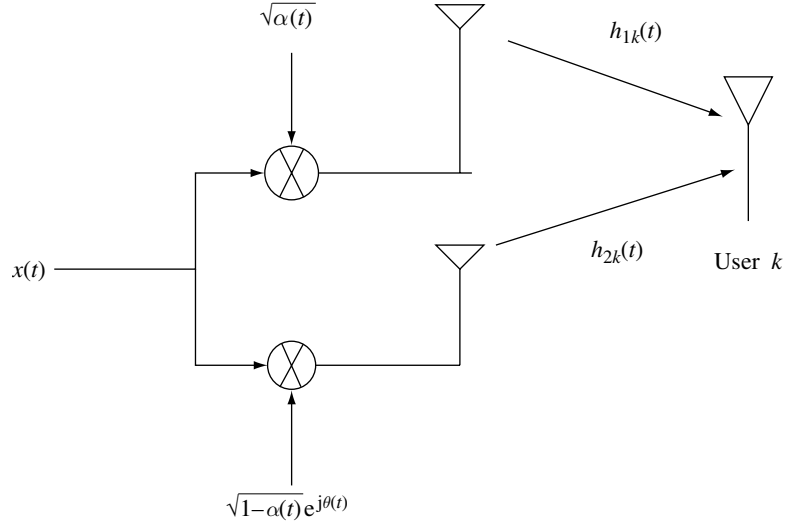
multiplied by a complex number $\sqrt{\alpha_l[m]}\,e^{j\theta_l[m]}$ at antenna l, for $l = 1, \ldots, n_t$, such that $\sum_{l=1}^{n_t} \alpha_l[m] = 1$, preserving the total transmit power. The received signal at user k (see the basic downlink fading channel model in (6.50) for comparison) is given by

$$y_k[m] = \left(\sum_{l=1}^{n_t} \sqrt{\alpha_l[m]}\, e^{j\theta_l[m]} h_{lk}[m] \right) x[m] + w_k[m]. \tag{6.58}$$

In vector form, the scheme transmits $\mathbf{q}[m]x[m]$ at time m, where

$$\mathbf{q}[m] := \begin{bmatrix} \sqrt{\alpha_1[m]}\, e^{j\theta_1[m]} \\ \vdots \\ \sqrt{\alpha_{n_t}[m]}\, e^{j\theta_{n_t}[m]} \end{bmatrix} \tag{6.59}$$

is a unit vector and

$$y_k[m] = (\mathbf{h}_k[m]^* \mathbf{q}[m])x[m] + w_k[m] \tag{6.60}$$

where $\mathbf{h}_k[m]^* := (h_{1k}[m], \ldots, h_{n_t,k}[m])$ is the channel vector from the transmit antenna array to user k.

The overall channel gain seen by user k is now

$$\mathbf{h}_k[m]^* \mathbf{q}[m] = \sum_{l=1}^{n_t} \sqrt{\alpha_l[m]}\, e^{j\theta_l[m]} h_{lk}[m]. \tag{6.61}$$

The $\alpha_l[m]$ denote the fractions of power allocated to each of the transmit antennas, and the $\theta_l[m]$ denote the phase shifts applied at each antenna to the

Figure 6.20 Pictorial representation of the slow fading channels of two users before (left) and after (right) applying opportunistic beamforming.
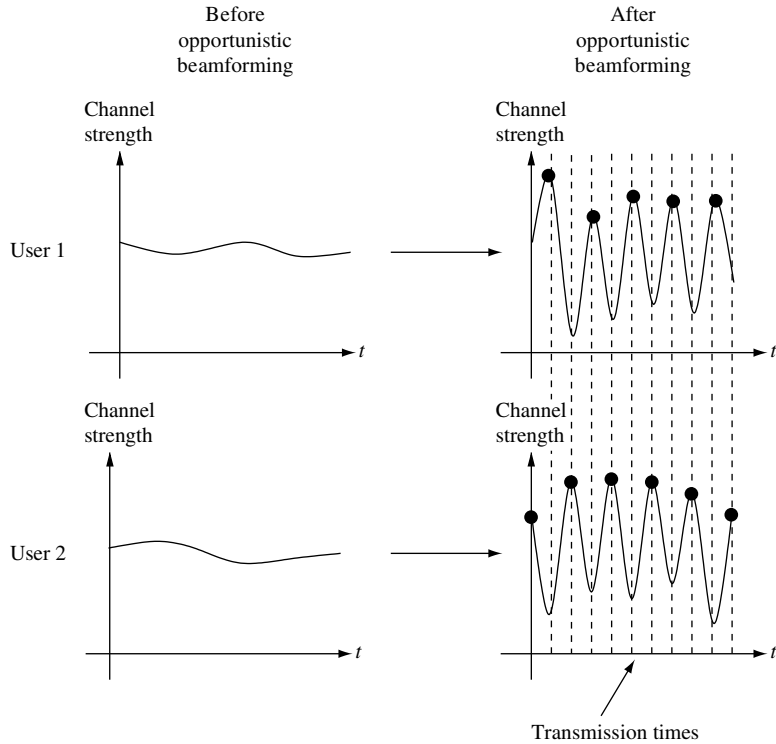
signal. By varying these quantities over time ($\alpha_l[m]$ from 0 to 1 and $\theta_l[m]$ from 0 to 2π) , the antennas transmit signals in a time-varying direction, and fluctuations in the overall channel can be induced even if the physical channel gains $\{h_{lk}[m]\}$ have very little fluctuation (Figure 6.20).

As in the single transmit antenna system, each user k feeds back the overall received SNR of its own channel, $|\mathbf{h}_k[m]^*\mathbf{q}[m]|^2/N_0$, to the base-station (or equivalently the data rate that the channel can currently support) and the base-station schedules transmissions to users accordingly. There is no need to measure the individual channel gains $h_{lk}[m]$ (phase or magnitude); in fact, the existence of multiple transmit antennas is completely transparent to the users. Thus, only a single pilot signal is needed for channel measurement (as opposed to a pilot to measure each antenna gain). The pilot symbols are repeated at each transmit antenna, exactly like the data symbols.

The rate of variation of $\{\alpha_l[m]\}$ and $\{\theta_l[m]\}$ in time (or, equivalently, of the transmit direction $\mathbf{q}[m]$) is a design parameter of the system. We would like it to be as fast as possible to provide full channel fluctuations within the latency time-scale of interest. On the other hand, there is a practical limitation to how fast this can be. The variation should be slow enough and should happen at a time-scale that allows the channel to be reliably estimated by the users and the SNR fed back. Further, the variation should be slow enough

to ensure that the channel seen by a user does not change abruptly and thus maintains stability of the channel tracking loop.

Slow fading: opportunistic beamforming

To get some insight into the performance of this scheme, consider the case of slow fading where the channel gain vector of each user k remains constant, i.e., $\mathbf{h}_k[m] = \mathbf{h}_k$, for all m. (In practice, this means for all m over the latency time-scale of interest.) The received SNR for this user would have remained constant if only one antenna were used. If all users in the system experience such slow fading, no multiuser diversity gain can be exploited. Under the proposed scheme, on the other hand, the overall channel gain $\mathbf{h}_k[m]^*\mathbf{q}[m]$ for each user k varies in time and provides opportunity for exploiting multiuser diversity.

Let us focus on a particular user k. Now if $\mathbf{q}[m]$ varies across all directions, the amplitude squared of the channel $|\mathbf{h}_k^*\mathbf{q}[m]|^2$ seen by user k varies from 0 to $\|\mathbf{h}_k\|^2$. The peak value occurs when the transmission is aligned along the direction of the channel of user k, i.e., $\mathbf{q}[m] = \mathbf{h}_k/\|\mathbf{h}_k\|$ (recall Example 5.2 in Section 5.3). The power and phase values are then in the *beamforming configuration*:

$$\alpha_l = \frac{|h_{lk}|^2}{\|\mathbf{h}_k\|^2}, \qquad l = 1, \ldots, n_t,$$

$$\theta_l = -\arg(h_{lk}), \qquad l = 1, \ldots, n_t.$$

To be able to beamform to a particular user, the base-station needs to know individual channel amplitude and phase responses from all the antennas, which requires much more information to feedback than just the overall SNR. However, if there are many users in the system, the proportional fair algorithm will schedule transmission to a user only when its overall channel SNR is near its peak. Thus, it is plausible that in a slow fading environment, the technique can approach the performance of coherent beamforming but with only overall SNR feedback (Figure 6.21). In this context, the technique can be interpreted as *opportunistic beamforming*: by varying the phases and powers allocated to the transmit antennas, a beam is randomly swept and at any time transmission is scheduled to the user currently closest to the beam. With many users, there is likely to be a user very close to the beam at any time. This intuition has been formally justified (see Exercise 6.29).

Fast fading: increasing channel fluctuations

We see that opportunistic beamforming can significantly improve performance in slow fading environments by adding fast time-scale fluctuations on the overall channel quality. The rate of channel fluctuation is artificially sped up. Can opportunistic beamforming help if the underlying channel variations are already fast (fast compared to the latency time-scale)?

Figure 6.21 Plot of spectral
efficiency under opportunistic
beamforming as a function of
the total number of users in
the system. The scenario is for
slow Rayleigh faded channels
for the users and the channels
are fixed in time. The spectral
efficiency plotted is the
performance averaged over
the Rayleigh distribution. As
the number of users grows,
the performance approaches
the performance of true
beamforming.
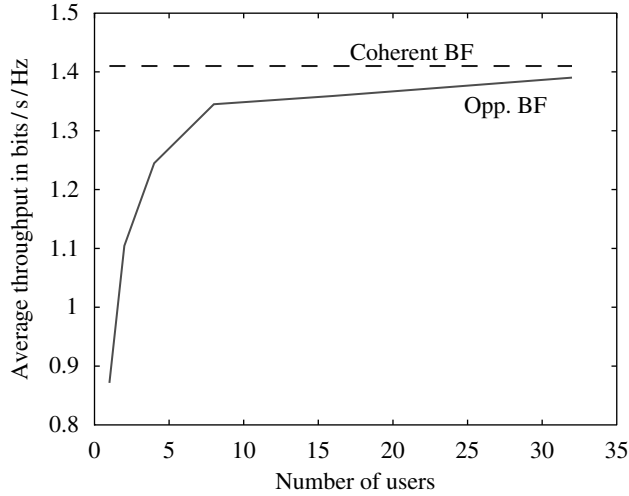
The long-term throughput under fast fading depends only on the stationary distribution of the channel gains. The impact of opportunistic beamforming in the fast fading scenario then depends on how the stationary distributions of the overall channel gains can be modified by power and phase randomization. Intuitively, better multiuser diversity gain can be exploited if the dynamic range of the distribution of h_k can be increased, so that the maximum SNRs can be larger. We consider two examples of common fading models.

- **Independent Rayleigh fading** In this model, appropriate for an environment where there is full scattering and the transmit antennas are spaced sufficiently, the channel gains $h_{1k}[m], \ldots, h_{n_t k}[m]$ are i.i.d. $\mathcal{CN}$ random variables. In this case, the channel vector $\mathbf{h}_k[m]$ is isotropically distributed, and $\mathbf{h}_k[m]^* \mathbf{q}[m]$ is circularly symmetric Gaussian for any choice of $\mathbf{q}[m]$; moreover the overall gains are independent across the users. Hence, the stationary statistics of the channel are *identical* to the original situation with one transmit antenna. Thus, in an independent fast Rayleigh fading environment, the opportunistic beamforming technique does not provide any performance gain.

- **Independent Rician fading** In contrast to the Rayleigh fading case, opportunistic beamforming has a significant impact in a Rician environment, particularly when the κ-factor is large. In this case, the scheme can significantly increase the dynamic range of the fluctuations. This is because the fluctuations in the underlying Rician fading process come from the diffused component, while with randomization of phase and powers, the fluctuations are from the coherent addition and cancellation of the direct path components in the signals from the different transmit antennas, in addition to the fluctuation of the diffused components. If the direct path

Figure 6.22 Total throughput
as a function of the number of
users under Rician fast fading,
with and without opportunistic
beamforming. The power
allocations $\alpha_l[m]$ are
uniformly distributed in $[0, 1]$
and the phases $\theta_l[m]$ uniform
in $[0, 2\pi]$.
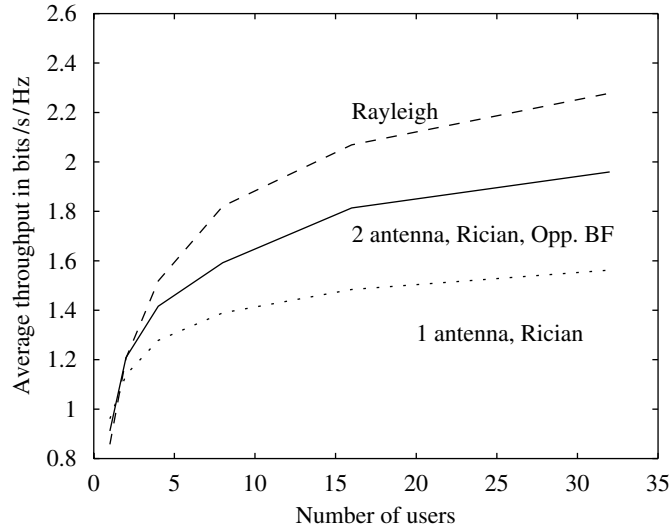
is much stronger than the diffused part (large κ values), then much larger
fluctuations can be created with this technique.

This intuition is substantiated in Figure 6.22, which plots the total
throughput with the proportional fair algorithm (large t_c, of the order of 100
time slots) for Rician fading with $\kappa = 10$. We see that there is a considerable
improvement in performance going from the single transmit antenna case
to dual transmit antennas with opportunistic beamforming. For comparison,
we also plot the analogous curves for pure Rayleigh fading; as expected,
there is no improvement in performance in this case. Figure 6.23 compares
the stationary distributions of the overall channel gain $|\mathbf{h}_k[m]^*\mathbf{q}[m]|$ in the
single-antenna and dual-antenna cases; one can see the increase in dynamic
range due to opportunistic beamforming.

Antennas: dumb, smart and smarter

In this section so far, our discussion has focused on the use of multiple
transmit antennas to induce larger and faster channel fluctuations for multiuser
diversity benefits. It is insightful to compare this with the two other point-
to-point transmit antenna techniques we have already discussed earlier in the
book:

- **Space-time codes** like the Alamouti scheme (Section 3.3.2). They are
  primarily used to increase the diversity in slow fading point-to-point links.
- **Transmit beamforming** (Section 5.3.2). In addition to providing diversity,
  a power gain is also obtained through the coherent addition of signals at
  the users.

Figure 6.23 Comparison of the distribution of the overall channel gain with and without opportunistic beamforming using two transmit antennas, Rician fading. The Rayleigh distribution is also shown.
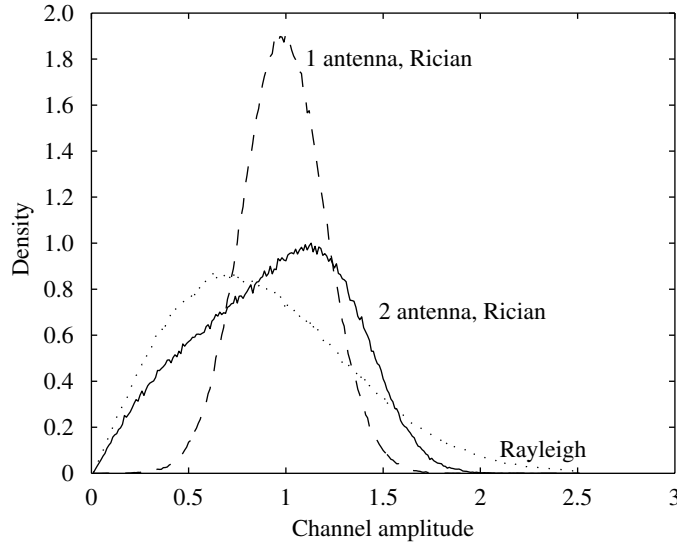
The three techniques have different system requirements. Coherent space-time codes like the Alamouti scheme require the users to track all the *individual* channel gains (amplitude and phase) from the transmit antennas. This requires separate pilot symbols on each of the transmit antennas. Transmit beamforming has an even stronger requirement that the channel should be known at the transmitter. In an FDD system, this means feedback of the individual channel gains (amplitude and phase). In contrast to these two techniques, the opportunistic beamforming scheme requires no knowledge of the individual channel gains, neither at the users nor at the transmitter. In fact, the users are *completely ignorant* of the fact that there are multiple transmit antennas and the receiver is identical to that in the single transmit antenna case. Thus, they can be termed *dumb antennas*. Opportunistic beamforming does rely on multiuser diversity scheduling, which requires the feedback of the overall SNR of each user. However, this only needs a *single* pilot to measure the overall channel.

What is the performance of these techniques when used in the downlink? In a slow fading environment, we have already remarked that opportunistic beamforming approaches the performance of transmit beamforming when there are many users in the system. On the other hand, space-time codes do not perform as well as transmit beamforming since they do not capture the array power gain. This means, for example, using the Alamouti scheme on dual transmit antennas in the downlink is 3 dB worse than using opportunistic beamforming combined with multiuser diversity scheduling when there are many users in the system. Thus, dumb antennas together with smart scheduling can surpass the performance of smart space-time codes and approach that of the even smarter transmit beamforming.

Table 6.1 A comparison between three methods of using transmit antennas.

	Dumb antennas (Opp. beamform)	Smart antennas (Space-time codes)	Smarter antennas (Transmit beamform)
Channel knowledge	Overall SNR	Entire CSI at Rx	Entire CSI at Rx, Tx
Slow fading performance gain	Diversity and power gains	Diversity gain only	Diversity and power gains
Fast fading performance gain	No impact	Multiuser diversity ↓	Multiuser diversity ↓ power ↑

How about in a fast Rayleigh fading environment? In this case, we have observed that dumb antennas have no effect on the overall channel as the full multiuser diversity gain has already been realized. Space-time codes, on the other hand, *increase* the diversity of the point-to-point links and consequently *decrease* the channel fluctuations and hence the multiuser diversity gain. (Exercise 6.31 makes this more precise.) Thus, the use of space-time codes as a point-to-point technology in a multiuser downlink with rate control and scheduling can actually be *harmful*, in the sense that even the naturally present multiuser diversity is removed. The performance impact of using transmit beamforming is not so clear: on the one hand it reduces the channel fluctuation and hence the multiuser diversity gain, but on the other hand it provides an array power gain. However, in an FDD system the fast fading channel may make it very difficult to feed back so much information to enable coherent beamforming.

The comparison between the three schemes is summarized in Table 6.1. All three techniques use the multiple antennas to transmit to only one user at a time. With full channel knowledge at the transmitter, an even smarter scheme can transmit to multiple users simultaneously, exploiting the multiple degrees of freedom existing inherently in the multiple antenna channel. We will discuss this in Chapter 10.

6.7.4 Multiuser diversity in multicell systems

So far we have considered a single-cell scenario, where the noise is assumed to be white Gaussian. For wideband cellular systems with full frequency reuse (such as the CDMA and OFDM based systems in Chapter 4), it is important to consider the effect of inter-cell interference on the performance of the system, particularly in interference-limited scenarios. In a cellular system, this effect is captured by measuring the channel quality of a user by the SINR, signal-to-interference-plus-noise ratio. In a fading environment, the energies in both the received signal and the received interference fluctuate over time. Since the multiuser diversity scheduling algorithm allocates resources based

on the channel SINR (which depends on both the channel amplitude and the amplitude of the interference), it automatically exploits both the fluctuations in the energy of the received signal and those of the interference: the algorithm tries to schedule resource to a user whose instantaneous channel is good *and* the interference is weak. Thus, multiuser diversity naturally takes advantage of the time-varying interference to increase the spatial reuse of the network.

From this point of view, amplitude and phase randomization at the base-station transmit antennas plays an additional role: it increases not only the amount of fluctuations of the received signal to the intended users *within* the cells, it also increases the fluctuations of the interference that the base-station causes in *adjacent* cells. Hence, opportunistic beamforming has a dual benefit in an interference-limited cellular system. In fact, opportunistic beam-forming performs *opportunistic nulling* simultaneously: while randomization of amplitude and phase in the transmitted signals from the antennas allows near coherent beamforming to some user within the cell, it will create near nulls at some other user in an adjacent cell. This in effect allows *interference avoidance* for that user if it is currently being scheduled.

Let us focus on the downlink and slow flat fading scenario to get some insight into the performance gain from opportunistic beamforming and nulling. Under amplitude and phase randomization at all base-stations, the received signal of a typical user that is interfered by J adjacent base-stations is given by

$$y[m] = (\mathbf{h}^*\mathbf{q}[m])x[m] + \sum_{j=1}^{J}(\mathbf{g}_j^*\mathbf{q}_j[m])u_j[m] + z[m]. \qquad (6.62)$$

Here, $x[m], \mathbf{h}, \mathbf{q}[m]$ are respectively the signal, channel vector and random transmit direction from the base-station of interest; $u_j[m], \mathbf{g}_j, \mathbf{q}_j[m]$ are respectively the interfering signal, channel vector and random transmit direction from the jth base-station. All base-stations have the same transmit power, P, and n_t transmit antennas and are performing amplitude and phase randomization independently.

By averaging over the signal $x[m]$ and the interference $u_j[m]$, the (time-varying) SINR of the user k can be computed to be

$$\mathsf{SINR}_k[m] = \frac{P|\mathbf{h}^*\mathbf{q}[m]|^2}{P\sum_{j=1}^{J}|\mathbf{g}_j^*\mathbf{q}_j[m]|^2 + N_0}. \qquad (6.63)$$

As the random transmit directions $\mathbf{q}[m], \mathbf{q}_j[m]$ vary, the overall SINR changes over time. This is due to the variations of the overall gain from the base-station of interest as well as those from the interfering base-stations. The SINR is high when $\mathbf{q}[m]$ is closely aligned to the channel vector $\mathbf{h}$, and/or for many j, $\mathbf{q}_j[m]$ is nearly orthogonal to $\mathbf{g}_j$, i.e., the user is near a null of the interference pattern from the jth base-station. In a system with many other users, the proportional fair scheduler will serve this user while its SINR

is at its peak $P\|\mathbf{h}\|^2/N_0$, i.e., when the received signal is the strongest and the interference is completely nulled out. Thus, the opportunistic nulling and beamforming technique has the potential of shifting a user from a low SINR, interference-limited regime to a high SINR, noise-limited regime. An analysis of the tail of the distribution of SINR is conducted in Exercise 6.30.

6.7.5 A system view

A new design principle for wireless systems can now be seen through the lens of multiuser diversity. In the three systems in Chapter 4, many of the design techniques centered on making the *individual* point-to-point links as close to AWGN channels as possible, with a reliable channel quality that is constant over time. This is accomplished by *channel averaging*, and includes the use of diversity techniques such as multipath combining, time-interleaving and antenna diversity that attempt to keep the channel fading constant in time, as well as interference management techniques such as interference averaging by means of spreading.

However, if one shifts from the view of the wireless system as a set of point-to-point links to the view of a system with multiple users sharing the same resources (spectrum and time), then quite a different design objective suggests itself. Indeed, the results in this chapter suggest that one should instead try to *exploit* the channel fluctuations. This is done through an appropriate scheduling algorithm that "rides the peaks", i.e., each user is scheduled when it has a very strong channel, while taking into account real world traffic constraints such as delay and fairness. The technique of dumb antennas goes one step further by *creating* variations when there are none. This is accomplished by varying the strengths of *both* the signal and the interference that a user receives through opportunistic beamforming and nulling.

The viability of the opportunistic communication scheme depends on traffic that has some tolerance to scheduling delays. On the other hand, there are some forms of traffic that are not so flexible. The functioning of the wireless systems is supported by the overhead control channels, which are "circuit-switched" and hence have very tight latency requirements, unlike data, which have the flexibility to allow dynamic scheduling. From the perspective of these signals, it is preferable that the channel remain unfaded; a requirement that is contradictory to our scheduler-oriented observation that we would prefer the channel to have fast and large variations.

This issue suggests the following design perspective: separate very-low latency signals (such as control signals) from flexible latency data. One way to achieve this separation is to split the bandwidth into two parts. One part is made as flat as possible (by using the principles we saw in Chapter 4 such as spreading over this part of the bandwidth) and is used to transmit flows with very low latency requirements. The performance metric here is to make the channel as reliable as possible (equivalently keeping the probability

of outage low) for some fixed data rate. The second part uses opportunistic beamforming to induce large and fast channel fluctuations and a scheduler to harness the multiuser diversity gains. The performance metric on this part is to maximize the multiuser diversity gain.

The gains of the opportunistic beamforming and nulling depend on the probability that the received signal is near beamformed *and* all the interference is near null. In the interference-limited regime and when $P/N_0 \gg 1$, the performance depends mainly on the probability of the latter event (see Exercise 6.30). In the downlink, this probability is large since there are only one or two base-stations contributing most of the interference. The uplink poses a contrasting picture: there is interference from many mobiles allowing interference averaging. Now the probability that the *total* interference is near null is much smaller. Interference averaging, which is one of the principle design features of the wideband full reuse systems (such as the ones we saw in Chapter 4 based on CDMA and OFDM), is actually unfavorable for the opportunistic scheme described here, since it reduces the likelihood of the nulling of the interference and hence the likelihood of the peaks of the SINR.

In a typical cell, there will be a distribution of users, some closer to the base-station and some closer to the cell boundaries. Users close to the base-station are at high SINR and are noise-limited; the contribution of the inter-cell interference is relatively small. These users benefit mainly from opportunistic beamforming. Users close to the cell boundaries, on the other hand, are at low SINR and are interference-limited; the average interference power can be much larger than the background noise. These users benefit both from opportunistic beamforming and from opportunistic nulling of inter-cell interference. Thus, the cell edge users benefit more in this system than users in the interior. This is rather desirable from a system fairness point-of-view, as the cell edge users tend to have poorer service. This feature is particularly important for a system without soft handoff (which is difficult to implement in a packet data scheduling system). To maximize the opportunistic nulling benefits, the transmit power at the base-station should be set as large as possible, subject to regulatory and hardware constraints. (See Exercise 6.30(5) where this is explored in more detail.)

We have seen the multiuser diversity as primarily a form of power gain. The opportunistic beamforming technique of using an array of multiple transmit antennas has approximately an n_t-fold improvement in received SNR to a user in a slow fading environment, as compared to the single-antenna case. With an array of n_r *receive* antennas at each mobile (and say a single transmit antenna at the base-station), the received SNR of any user gets an n_r-fold improvement as compared to a single receive antenna; this gain is realized by *receiver beamforming*. This operation is easy to accomplish since the mobile has full channel information at each of the antenna elements. Hence the gains of opportunistic beamforming are about the same order as that of installing a receive antenna array at *each* of the mobiles.

Thus, for a system designer, the opportunistic beamforming technique provides a compelling case for implementation, particularly in view of the constraints of space and cost of installing multiple antennas on *each* mobile device. Further, this technique needs neither any extra processing on the part of any user, nor any updates to an existing air-link interface standard. In other words, the mobile receiver can be completely ignorant of the use or non-use of this technique. This means that it does not have to be "designed in" (by appropriate inclusions in the air interface standard and the receiver design) and can be added/removed at any time. This is one of the important benefits of this technique from an overall system design point of view.

In the cellular wireless systems studied in Chapter 4, the cell is sectorized to allow better focusing of the power transmitted from the antennas and also to reduce the interference seen by mobile users from transmissions of the same base-station but intended for users in different sectors. This technique is particularly gainful in scenarios when the base-station is located at a fairly large height and thus there is limited scattering around the base-station. In contrast, in systems with far denser deployment of base-stations (a strategy that can be expected to be a good one for wireless systems aiming to provide mobile, broadband data services), it is unreasonable to stipulate that the base-stations be located high above the ground so that the local scattering (around the base-station) is minimal. In an urban environment, there is substantial local scattering around a base-station and the gains of sectorization are minimal; users in a sector also see interference from the same base-station (due to the local scattering) intended for another sector. The opportunistic beamforming scheme can be thought of as sweeping a random beam and scheduling transmissions to users when they are beamformed. Thus, the gains

Table 6.2 Contrast between conventional multiple access and opportunistic communication.

	Conventional multiple access	Opportunistic communication
Guiding principle	Averaging out fast channel fluctuations	Exploiting channel fluctuations
Knowledge at Tx	Track slow fluctuations No need to track fast ones	Track as many fluctuations as possible
Control	Power control the slow fluctuations	Rate control to all fluctuations
Delay requirement	Can support tight delay	Needs some laxity
Role of Tx antennas	Point-to-point diversity	Increase fluctuations
Power gain in downlink	Multiple Rx antennas	Opportunistic beamform via multiple Tx antennas
Interference management	Averaged	Opportunistically avoided

of sectorization are automatically realized. We conclude that the opportunistic beamforming technique is particularly suited to harness sectorization gains even in low-height base-stations with plenty of local scattering. In a cellular system, the opportunistic beamforming scheme also obtains the gains of nulling, a gain traditionally obtained by coordinated transmissions from neighboring base-stations in a full frequency reuse system or by appropriately designing the frequency reuse pattern.

The discussion is summarized in Table 6.2.

Chapter 6 The main plot

This chapter looked at the capacities of uplink and downlink channels. Two important sets of concepts emerged:

- successive interference cancellation (SIC) and superposition coding;
- multiuser opportunistic communication and multiuser diversity.

SIC and superposition coding
Uplink

Capacity is achieved by allowing users to simultaneously transmit on the full bandwidth and the use of SIC to decode the users.

SIC has a significant performance gain over conventional multiple access techniques in near–far situations. It takes advantage of the strong channel of the nearby user to give it high rate while providing the weak user with the best possible performance.

Downlink

Capacity is achieved by superimposing users' signals and the use of SIC at the receivers. The strong user decodes the weak user's signal first and then decodes its own.

Superposition coding/SIC has a significant gain over orthogonal techniques. Only a small amount of power has to be allocated to the strong user to give it a high rate, while delivering near-optimal performance to the weak user.

Opportunistic communication
Symmetric uplink fading channel:

$$y[m] = \sum_{k=1}^{K} h_k[m]x_k[m] + w[m]. \tag{6.64}$$

Sum capacity with CSI at receiver only:

$$C_{\text{sum}} = \mathbb{E}\left[\log\left(1 + \frac{\sum_{k=1}^{K}|h_k|^2 P}{N_0}\right)\right]. \tag{6.65}$$

Very close to AWGN capacity for large number of users. Orthogonal multiple access is strictly suboptimal.

Sum capacity with full CSI:

$$C_{\text{sum}} = \mathbb{E}\left[\log\left(1 + \frac{P_{k^*}(\mathbf{h})|h_{k^*}|^2}{N_0}\right)\right], \tag{6.66}$$

where k^* is the user with the strongest channel at joint channel state $\mathbf{h}$. This is achieved by transmitting only to the user with the best channel and a waterfilling power allocation $P_{k^*}(\mathbf{h})$ over the fading state.

Symmetric downlink fading channel:

$$y_k[m] = h_k[m]x[m] + w_k[m], \qquad k = 1, \ldots, K. \tag{6.67}$$

Sum capacity with CSI at receiver only:

$$C_{\text{sum}} = \mathbb{E}\left[\log\left(1 + \frac{|h_k|^2 P}{N_0}\right)\right]. \tag{6.68}$$

Can be achieved by orthogonal multiple access.
Sum capacity with full CSI: same as uplink.

Multiuser diversity
Multiuser diversity gain: under full CSI, capacity increases with the number of users: in a large system with high probability there is always a user with a very strong channel.
System issues in implementing multiuser diversity:
- **Fairness** Fair access to the channel when some users are statistically stronger than others.
- **Delay** Cannot wait too long for a good channel.
- **Channel tracking** Channel has to be measured and fed back fast enough.
- **Small and slow channel fluctuations** Multiuser diversity gain is limited when channel varies too slowly and/or has a small dynamic range.

The solutions discussed were:
- Proportional fair scheduler transmits to a user when its channel is near its peak within the delay constraint. Every user has access to the channel for roughly the same amount of time.
- Channel feedback delay can be reduced by having shorter time slots and feeding back more often. Aggregate feedback can be reduced by each user selectively feeding channel state back only when its channel is near its peak.

> • Channel fluctuations can be sped up and their dynamic range increased
>   by the use of multiple transmit antennas to perform opportunistic beam-
>   forming. The scheme sweeps a random beam and schedules transmis-
>   sions to users when they are beamformed.
>
> In a cellular system, multiuser diversity scheduling performs interference
> avoidance as well: a user is scheduled transmission when its channel is
> strong *and* the out-of-cell interference is weak.
>
> Multiple transmit antennas can perform opportunistic beamforming as well
> as nulling.

6.8 Bibliographical notes

Classical treatment of the general multiple access channel was initiated by Ahlswede
[2] and Liao [73] who characterized the capacity region. The capacity region of the
Gaussian multiple access channel is derived as a special case. A good survey of
the literature on MACs was done by Gallager [45]. Hui [59] first observed that the
sum capacity of the uplink channel with single-user decoding is bounded by 1.442
bits/s/Hz.

The general broadcast channel was introduced by Cover [25] and a complete
characterization of its capacity is one of the famous open problems in information
theory. Degraded broadcast channels, where the users can be "ordered" based on their
channel quality, are fully understood with superposition coding being the optimal
strategy; a textbook reference is Chapter 14.6 in Cover and Thomas [26]. The best
inner and outer bounds are by Marton [81] and a good survey of the literature
appears in [24].

The capacity region of the uplink fading channel with receiver CSI was derived
by Gallager [44], where he also showed that orthogonal multiple access schemes
are strictly suboptimal in fading channels. Knopp and Humblet [65] studied the sum
capacity of the uplink fading channel with full CSI. They noted that transmitting to
only one user is the optimal strategy. An analogous result was obtained earlier by
Cheng and Verdú [20] in the context of the time-invariant uplink frequency-selective
channels. Both these channels are instances of the parallel Gaussian multiple access
channel, so the two results are mathematically equivalent. The latter authors also
derived the capacity region in the two-user case. The solution for arbitrary number of
users was obtained by Tse and Hanly [122], exploiting a basic polymatroid property
of the region.

The study of downlink fading channels with full CSI was carried out by Tse [124]
and Li and Goldsmith [74]. The key aspect of the study was to observe that the fading
downlink is really a parallel degraded broadcast channel, the capacity of which has
been fully understood (El Gamal [33]). There is an intriguing similarity between the
downlink resource allocation solution and the uplink one. This connection is studied
further in Chapter 10.

Multiuser diversity is a key distinguishing feature of the uplink and the downlink
fading channel study as compared to our understanding of the point-to-point fading

channel. The term multiuser diversity was coined by Knopp and Humblet [66]. The multiuser diversity concept was integrated into the downlink design of IS-856 (CDMA 2000 EV-DO) via the proportional fair scheduler by Tse [19]. In realistic scenarios, performance gains of 50% to 100% have been reported (Wu and Esteves [149]).

If the channels are slowly varying, then the multiuser diversity gains are limited. The opportunistic beamforming idea mitigates this defect by creating variations while maintaining the same average channel quality; this was proposed by Viswanath *et al.* [137], who also studied its impact on system design.

Several works have studied the design of schedulers that harness the multiuser diversity gain. A theoretical analysis of the proportional fair scheduler has appeared in several places including a work by Borst and Whiting [12].

6.9 Exercises

Exercise 6.1 The sum constraint in (6.6) applies because the two users send independent information and cannot cooperate in the encoding. If they could cooperate, what is the maximum sum rate they could achieve, still assuming individual power constraints P_1 and P_2 on the two users? In the case $P_1 = P_2$, quantify the cooperation gain at low and at high SNR. In which regime is the gain more significant?

Exercise 6.2 Consider the basic uplink AWGN channel in (6.1) with power constraints P_k on user k (for $k = 1, 2$). In Section 6.1.3, we stated that orthogonal multiple access is optimal when the degrees of freedom are split in direct proportion to the powers of the users. Verify this. Show also that any other split of degrees of freedom is strictly suboptimal, i.e., the corresponding rate pair lies strictly inside the capacity region given by the pentagon in Figure 6.2. *Hint*: Think of the sum rate as the performance of a point-to-point channel and apply the insight from Exercise 5.6.

Exercise 6.3 Calculate the symmetric capacity, (6.2), for the two-user uplink channel. Identify scenarios where there are definitely superior operating points.

Exercise 6.4 Consider the uplink of a single IS-95 cell where all the users are controlled to have the same received power P at the base-station.
1. In the IS-95 system, decoding is done by a conventional CDMA receiver which treats the interference of the other users as Gaussian noise. What is the maximum number of voice users that can be accommodated, assuming capacity-achieving point-to-point codes? You can assume a total bandwidth of 1.25 MHz and a data rate per user of 9.6 kbits/s. You can also assume that the background noise is negligible compared to the intra-cell interference.
2. Now suppose one of the users is a data user and it happens to be close to the base-station. By not controlling its power, its received power can be 20 dB above the rest. Propose a receiver that can give this user a higher rate while still delivering 9.6 kbits/s to the other (voice) users. What rate can it get?

Exercise 6.5 Consider the uplink of an IS-95 system.
1. A single cell is modeled as a disk of radius 1 km. If a mobile at the edge of the cell transmits at its maximum power limit, its received SNR at the base-station is 15 dB when no one else is transmitting. Estimate (via numerical simulations)

the average sum capacity of the uplink with 16 users that are independently and uniformly located in the disk. Compare this to the corresponding average total throughput in a system with conventional CDMA decoding and each user perfectly power controlled at the base-station. What is the potential percentage gain in spectral efficiency by using the more sophisticated receiver? You can assume that all mobiles have the same transmit power constraint and the path loss (power) attenuation is proportional to r^{-4}.

2. Part (1) ignores out-of-cell interference. With out-of-cell interference taken into consideration, the received SINR of the cell edge user is only $-10\,\text{dB}$. Redo part (1). Is the potential gain from using a more sophisticated receiver still as impressive?

Exercise 6.6 Consider the downlink of the IS-856 system.

1. Suppose there are two users on the cell edge. Users are scheduled on a TDMA basis, with equal time for each user. The received SINR of each user is $0\,\text{dB}$ when it is transmitted to. Find the rate that each user gets. The total bandwidth is $1.25\,\text{MHz}$ and you can assume an AWGN channel and the use of capacity-achieving codes.

2. Now suppose there is an extra user which is near the base-station with a $20\,\text{dB}$ SINR advantage over the other two users. Consider two ways to accommodate this user:
   - Give a fraction of the time slots to this user and divide the rest equally among the two cell edge users.
   - Give a fraction of the power to this user and superimpose its signal on top of the signals of both users. The two cell edge users are still scheduled on a TDMA basis with equal time, and the strong user uses a SIC decoder to extract its signal after decoding the other users' signals at each time slot.

   Since the two cell edge users have weak reception, it is important to maintain the best possible quality of service to them. So suppose the constraint is that we want each of them to have 95% of the rates they were getting before this strong user joined. Compare the performance that the strong user gets in the two schemes above.

Exercise 6.7 The capacity region of the two-user AWGN uplink channel is shown in Figure 6.2. The two corner points A and B can be achieved using successive cancellation. Points inside the line segment AB can be achieved by time sharing. In this exercise we will see another way to achieve every point (R_1, R_2) on the line segment AB using successive cancellation. By definition we must have

$$R_k < \log\left(1 + \frac{P_k}{N_0}\right) \quad k = 1, 2, \tag{6.69}$$

$$R_1 + R_2 = \log\left(1 + \frac{P_1 + P_2}{N_0}\right). \tag{6.70}$$

Define $\delta > 0$ by

$$R_2 = \log\left(1 + \frac{P_2}{\delta + N_0}\right). \tag{6.71}$$

Now consider the situation when user 1 splits itself into two users, say users 1a and 1b, with power constraints $P_1 - \delta$ and δ respectively. We decode the users with successive cancellation in the order user 1a, 2, 1b, i.e., user 1a is decoded first, user 2 is decoded next (with user 1a cancelled) and finally user 1b is decoded (seeing no interference from users 1a and 2).

1. Calculate the rates of reliable communication (r_{1a}, r_2, r_{1b}) for the users 1a, 2 and 1b using the successive cancellation just outlined.
2. Show that $r_2 = R_2$ and $r_{1a} + r_{1b} = R_1$. This means that the point (R_1, R_2) on the line segment AB can be achieved by successive cancellation of three users formed by one of the users "splitting" itself into two virtual users.

Exercise 6.8 In Exercise 6.7, we studied rate splitting multiple access for two users. A reading exercise is to study [101], where this result was introduced and generalized to the K-user uplink: $K - 1$ users can split themselves into two users each (with appropriate power splits) so that any rate vector on the boundary of the capacity region that meets the sum power constraint can be achieved via successive cancellation (with appropriate ordering of the $2K - 1$ users).

Exercise 6.9 Consider the K-user AWGN uplink channel with user power constraints $P_1, \ldots, P_K$. The capacity region is the set of rate vectors that lie in the intersection of the constraints (cf. (6.10)):

$$\sum_{k \in \mathcal{S}} R_k < \log\left(1 + \frac{\sum_{k \in \mathcal{S}} P_k}{N_0}\right), \tag{6.72}$$

for every subset $\mathcal{S}$ of the K users.

1. Fix an ordering of the users $\pi_1, \ldots, \pi_K$ (here π represents a permutation of set $\{1, \ldots, K\}$). Show that the rate vector $\left(R_1^{(\pi)}, \ldots, R_K^{(\pi)}\right)$:

$$R_{\pi_k}^{(\pi)} := \log\left(1 + \frac{P_{\pi_k}}{\sum_{i=k+1}^{K} P_{\pi_i} + N_0}\right) \quad k = 1, \ldots, K, \tag{6.73}$$

is in the capacity region. This rate vector can be interpreted using the successive cancellation viewpoint: the users are successively decoded in the order $\pi_1, \ldots, \pi_K$ with cancellation after each decoding step. So, user π_k has no interference from the previously decoded users $\pi_1, \ldots, \pi_{k-1}$, but experiences interference from the users following it (namely $\pi_{k+1}, \ldots, \pi_K$). In Figure 6.2, the point A corresponds to the permutation $\pi_1 = 2, \pi_2 = 1$ and the point B corresponds to the identity permutation $\pi_1 = 1, \pi_2 = 2$.
2. Consider maximizing the linear objective function $\sum_{k=1}^{K} a_k R_k$ with non-negative $a_1, \ldots, a_K$ over the rate vectors in the capacity region. (a_k can be interpreted as the *revenue* per unit rate for user k.) Show that the maximum occurs at the rate vector of the form in (6.73) with the permutation π defined by the property:

$$a_{\pi_1} \leq a_{\pi_2} \leq \cdots \leq a_{\pi_K}. \tag{6.74}$$

This means that optimizing linear objective functions on the capacity region can be done in a greedy way: we order the users based on their priority (a_k for user k). This ordering is denoted by the permutation π in (6.74). Next, the receiver decodes via successive cancellation using this order: the user with the least priority is decoded first (seeing full interference from all the other users) and the user with the highest priority decoded last (seeing no interference from the other users). *Hint*: Show that if the ordering is not according to (6.74), then one can always improve the objective function by changing the decoding order.

3. Since the capacity region is the intersection of hyperplanes, it is a convex polyhedron. An equivalent representation of a convex polyhedron is through enumerating its *vertices*: points which cannot be expressed as a strict convex combination of any subset of other points in the polyhedron. Show that $\left(R_1^{(\pi)}, \ldots, R_K^{(\pi)} \right)$ is a *vertex* of the capacity region. *Hint*: Consider the following fact: a linear object function is maximized on a convex polyhedron at one of the vertices. Further, every vertex must be optimal for some linear objective function.

4. Show that vertices of the form (6.73) (one for each permutation, so there are $K!$ of them) are the only *interesting* vertices of the capacity region. (This means that any other vertex of the capacity region is component-wise dominated by one of these $K!$ vertices.)

Exercise 6.10 Consider the K-user uplink AWGN channel. In the text, we focus on the capacity region $\mathcal{C}(\mathbf{P})$: the set of achievable rates for given power constraint vector $\mathbf{P} := (P_1, \ldots, P_K)^t$. A "dual" characterization is the *power region* $\mathcal{P}(\mathbf{R})$: set of all feasible received power vectors that can support a given target rate vector $\mathbf{R} := (R_1, \ldots, R_K)^t$.

1. Write down the constraints describing $\mathcal{P}(\mathbf{R})$. Sketch the region for $K = 2$.
2. What are the vertices of $\mathcal{P}(\mathbf{R})$?
3. Find a decoding strategy and a power allocation that minimizes $\sum_{k=1}^{K} b_k P_k$ while meeting the given target rates. Here, the constants b_k are positive and should be interpreted as "power prices". *Hint*: Exercise 6.9 may be useful.
4. Suppose users are at different distances from the base-station so that the transmit power of user k is attenuated by a factor of γ_i. Find a decoding strategy and a power allocation that minimizes the total transmit power of the users while meeting the target rates $\mathbf{R}$.
5. In IS-95, the code used by each user is not necessarily capacity-achieving but communication is considered reliable as long as a $\mathcal{E}_b/I_0$ requirement of 7 dB is met. Suppose these codes are used in conjunction with SIC. Find the optimal decoding order to minimize the total transmit power in the uplink.

Exercise 6.11 (Impact of using SIC on interference-limited capacity) Consider the two-cell system in Exercise 4.11. The interference-limited spectral efficiency in the many-user regime was calculated for both CDMA and OFDM. Now suppose SIC is used instead of the conventional receiver in the CDMA system. In the context of SIC, the interference I_0 in the target $\mathcal{E}_b/I_0$ requirement refers to the interference from the uncancelled users. Below you can always assume that interference cancellation is perfect.

1. Focus on a single cell first and assume a background noise power of N_0. Is the system interference-limited under the SIC receiver? Was it interference-limited under the conventional CDMA receiver?
2. Suppose there are K users with user k at a distance r_k from the base-station. Give an expression for the total transmit power saving (in dB) in using SIC with the optimal decoding order as compared to the conventional CDMA receiver (with an $\mathcal{E}_b/I_0$ requirement of β).
3. Give an expression for the power saving in the asymptotic regime with a large number of users and large bandwidth. The users are randomly located in the single cell as specified in Exercise 4.11. What is this value when $\beta = 7$ dB and the power decay is r^{-2} (i.e., $\alpha = 2$)?

4. Now consider the two-cell system. Explain why in this case the system is interference-limited even when using SIC.

5. Nevertheless, SIC increases the interference-limited capacity because of the reduction in transmit power, which translates into a reduction of out-of-cell interference. Give an expression for the asymptotic interference-limited spectral efficiency under SIC in terms of β and α. You can ignore the background noise and assume that users closer to the base-station are always decoded before the users further away.

6. For $\beta = 7\,\text{dB}$ and $\alpha = 2$, compare the performance with the conventional CDMA system and the OFDM system.

7. Is the cancellation order in part 5 optimal? If not, find the optimal order and give an expression for the resulting asymptotic spectral efficiency. *Hint*: You might find Exercise 6.10 useful.

Exercise 6.12 Verify the bound (6.30) on the actual error probability of the kth user in the SIC, accounting for error propagation.

Exercise 6.13 Consider the two-user uplink fading channel,

$$y[m] = h_1[m]x_1[m] + h_2[m]x_2[m] + w[m]. \tag{6.75}$$

Here the user channels $\{h_1[m]\}, \{h_2[m]\}$ are statistically independent. Suppose that $h_1[m]$ and $h_2[m]$ are $\mathcal{CN}(0, 1)$ and user k has power P_k, $k = 1, 2$, with $P_1 \gg P_2$. The background noise $w[m]$ is i.i.d. $\mathcal{CN}(0, N_0)$. An SIC receiver decodes user 1 first, removes its contribution from $\{y[m]\}$ and then decodes user 2. We would like to assess the effect of channel estimation error of h_1 on the performance of user 2.

1. Assuming that the channel coherence time is T_c seconds and user 1 spends 20% of its power on sending a training signal, what is the mean square estimation error of h_1? You can assume the same setup as in Section 3.5.2. You can ignore the effect of user 2 in this estimation stage, since $P_1 \gg P_2$.

2. The SIC receiver decodes the transmitted signal from user 1 and subtracts its contribution from $\{y[m]\}$. Assuming that the information is decoded correctly, the residual error is due to the channel estimation error of h_1. Quantify the degradation in SINR of user 2 due to this channel estimation error. Plot this degradation as a function of P_1/N_0 for $T_c = 10\,\text{ms}$. Does the degradation worsen if the power P_1 of user 1 increases? Explain.

3. In part (2), user 2 still faced some interference due to the presence of user 1 despite decoding the information meant for user 1 accurately. This is due to the error in the channel estimate of user 1. In the calculation in part (2), we used the expression for the error of user 1's channel estimate as derived from the training symbol. However, conditioned on the event that the first user's information has been correctly decoded, the channel estimate of user 1 can be improved. Model this situation appropriately and arrive at an approximation of the error in user 1's channel estimate. Now redo part (2). Does your answer change qualitatively?

Exercise 6.14 Consider the probability of the outage event ($p_{\text{out}}^{\text{ul}}$, cf. (6.32)) in a symmetric slow Rayleigh fading uplink with the K users operating at the symmetric rate R bits/s/Hz.

1. Suppose p_{out}^{ul} is fixed to be ϵ. Argue that at very high SNR (with SNR defined to be P/N_0), the dominating event is the one on the sum rate:

$$KR > \log\left(1 + \frac{\sum_{k=1}^{K} P|h_k|^2}{N_0}\right).$$

2. Show that the ϵ-outage symmetric capacity, C_ϵ^{sym}, can be approximated at very high SNR as

$$C_\epsilon^{sym} \approx \frac{1}{K} \log_2\left(1 + \frac{P\epsilon^{\frac{1}{K}}}{N_0}\right).$$

3. Argue that at very high SNR, the ratio of C_ϵ^{sym} to C_ϵ (the ϵ-outage capacity with just a single user in the uplink) is approximately $1/K$.

Exercise 6.15 In Section 6.3.3, we have discussed the optimal multiple access strategy for achieving the sum capacity of the uplink fading channel when users have identical channel statistics and power constraints.

1. Solve the problem for the general case when the channel statistics and the power constraints of the users are arbitrary. *Hint*: Construct a Lagrangian for the convex optimization problem (6.42) with a separate Lagrange multiplier for each of the individual power constraints (6.43).

2. Do you think the sum capacity is a reasonable performance measure in the asymmetric case?

Exercise 6.16 In Section 6.3.3, we have derived the optimal power allocation with full CSI in the symmetric uplink with the assumption that there is always a unique user with the strongest channel at any one time. This assumption holds with probability 1 when the fading distributions are continuous. Moreover, under this assumption, the solution is *unique*. This is in contrast to the uplink AWGN channel where there is a continuum of solutions that achieves the optimal sum rate, of which only one is orthogonal. We will see in this exercise that transmitting to only one user at a time is not necessarily the unique optimal solution even for fading channels, if the fading distribution is discrete (to model measurement realities, such as the feedback of a finite number of rate levels).

Consider the full CSI two-user uplink with identical, independent, stationary and ergodic flat fading processes for the two users. The stationary distribution of the flat fading for both of the users takes one of just two values: channel amplitude is either at 0 or at 1 (with equal probability). Both of the users are individually average power constrained (by $\bar{P}$). Calculate explicitly *all* the optimal joint power allocation and decoding policies to maximize the sum rate. Is the optimal solution unique? *Hint*: Clearly there is no benefit by allocating power to a user whose channel is fully faded (the zero amplitude state).

Exercise 6.17 In this exercise we further study the nature of the optimal power and rate control strategy that achieves the sum capacity of the symmetric uplink fading channel.

1. Show that the optimal power/rate allocation policy for achieving the sum capacity of the symmetric uplink fading channel can be obtained by solving for each fading state the optimization problem:

$$\max_{\mathbf{r},\mathbf{p}} \sum_{k=1}^{K} r_k - \lambda \sum_{k=1}^{K} p_k, \tag{6.76}$$

subject to the constraint that

$$\mathbf{r} \in \mathcal{C}(\mathbf{p}, \mathbf{h}), \tag{6.77}$$

where $\mathcal{C}(\mathbf{p}, \mathbf{h})$ is the uplink AWGN channel capacity region with received power $p_k|h_k|^2$. Here λ is chosen to meet the average power constraint of P for each user.

2. What happens when the channels are not symmetric but we are still interested in the sum rate?

Exercise 6.18 [122] In the text, we focused on computing the power/rate allocation policy that maximizes the *sum* rate. More generally, we can look for the policy that maximizes a weighted sum of rates $\sum_k \mu_k R_k$. Since the uplink fading channel capacity region is convex, solving this for all non-negative μ_i will enable us to characterize the entire capacity region (as opposed to just the sum capacity point).

In analogy with Exercise 6.17, it can be shown that the optimal power/rate allocation policy can be computed by solving for each fading state $\mathbf{h}$ the optimization problem:

$$\max_{\mathbf{r},\mathbf{p}} \sum_{k=1}^{K} \mu_k r_k - \sum_{k=1}^{K} \lambda_k p_k, \tag{6.78}$$

subject to the constraint that

$$\mathbf{r} \in \mathcal{C}(\mathbf{p}, \mathbf{h}), \tag{6.79}$$

where the λ_k are chosen to meet the average power constraints P_k of the users (averaged over the fading distribution). If we define $q_k := p_k|h_k|^2$ as the received power, then we can rewrite the optimization problem as

$$\max_{\mathbf{r},\mathbf{q}} \sum_{k=1}^{K} \mu_k r_k - \sum_{k=1}^{K} \frac{\lambda_k}{|h_k|^2} p_k \tag{6.80}$$

subject to the constraint that

$$\mathbf{r} \in \mathcal{C}(\mathbf{q}), \tag{6.81}$$

where $\mathcal{C}(\mathbf{q})$ is the uplink AWGN channel capacity region. You are asked to solve this optimization problem in several steps below.

1. Verify that the capacity of a point-to-point AWGN channel can be written in the integral form:

$$C_{\text{awgn}} = \log\left(1 + \frac{P}{N_0}\right) = \int_0^P \frac{1}{N_0 + z}\, \mathrm{d}z. \tag{6.82}$$

Give an interpretation in terms of splitting the single user into many infinitesimally small virtual users, each with power dz (cf. Exercise 6.7). What is the interpretation of the quantity $1/(N_0 + z) \, dz$?

2. Consider first $K = 1$ in the uplink fading channel above, i.e., the point-to-point scenario. Define the *utility function*:

$$u_1(z) = \frac{\mu_1}{N_0 + z} - \frac{\lambda_1}{|h_1|^2}, \qquad (6.83)$$

where N_0 is the background noise power. Express the optimal solution in terms of the graph of $u_1(z)$ against z. Interpret the solution as a greedy solution and also give an interpretation of $u_1(z)$. *Hint*: Make good use of the rate-splitting interpretation in part 1.

3. Now, for $K > 1$, define the utility function of user k to be

$$u_k(z) = \frac{\mu_k}{N_0 + z} - \frac{\lambda_k}{|h_k|^2}. \qquad (6.84)$$

Guess what the optimal solution should be in terms of the graphs of $u_k(z)$ against z for $k = 1, \ldots, K$.

4. Show that each pair of the utility functions intersects at most once for non-negative z.

5. Using the previous parts, verify your conjecture in part (3).

6. Can the optimal solution be achieved by successive cancellation?

7. Verify that your solution reduces to the known solution for the sum capacity problem (i.e., when $\mu_1 = \cdots = \mu_K$).

8. What does your solution look like when there are two groups of users such that within each group, users have the same μ_k and λ_k (but not necessarily the same h_k).

9. Using your solution to the optimization problem (6.78), compute numerically the boundary of the capacity region of the two-user Rayleigh uplink fading channel with average received SNR of $0 \, dB$ for each of the two users.

Exercise 6.19 [124] Consider the downlink fading channel.

1. Formulate and solve the downlink version of Exercise 6.18.

2. The total transmit power varies as a function of time in the optimal solution. But now suppose we fix the total transmit power to be P at all times (as in the IS-856 system). Re-derive the optimal solution.

Exercise 6.20 Within a cell in the IS-856 system there are eight users on the edge and one user near the base-station. Every user experiences independent Rayleigh fading, but the average SNR of the user near the base-station is γ times that of the users on the edge. Suppose the average SNR of a cell edge user is $0 \, dB$ when all the power of the base-station is allocated to it. A fixed transmit power of P is used at all times.

1. Simulate the proportional fair scheduling algorithm for t_c large and compute the performance of each user for a range of γ from 1 to 100. You can assume the use of capacity-achieving codes.

2. Fix γ. Show how you would compute the *optimal* achievable rate among *all strategies* for the user near the base-station, given a (equal) rate for all the users on the edge. *Hint*: Use the results in Exercise 6.19.

3. Plot the potential gain in rate for the strong user over what it gets under the proportional fair algorithm, for the same rate for the weak users.

Exercise 6.21 In Section 6.6, we have seen that the multiuser diversity gain comes about because the effective channel gain becomes the maximum of the channel gains of the K users:

$$|h|^2 := \max_{k=1\ldots K} |h_k|^2.$$

1. Let $h_1, \ldots, h_K$ be i.i.d. $\mathcal{CN}(0,1)$ random variables. Show that

$$\mathbb{E}[|h|^2] = \sum_{k=1}^{K} \frac{1}{k}. \tag{6.85}$$

Hint: You might find it easier to prove the following stronger result (using induction):

$$|h|^2 \text{ has the same distribution as } \sum_{k=1}^{K} \frac{|h_k|^2}{k}. \tag{6.86}$$

2. Using the previous part, or directly, show that

$$\frac{\mathbb{E}[|h|^2]}{\log_e K} \to 1 \qquad \text{as } K \to \infty; \tag{6.87}$$

thus the mean of the effective channel grows logarithmically with the number of users.

3. Now suppose $h_1, \ldots, h_K$ are i.i.d. $\mathcal{CN}(\sqrt{\kappa}/\sqrt{1+\kappa}, 1/(1+\kappa))$ (i.e., Rician random variables with the ratio of specular path power to diffuse path power equal to κ). Show that

$$\frac{\mathbb{E}[|h|^2]}{\log_e K} \to \frac{1}{1+\kappa} \qquad \text{as } K \to \infty; \tag{6.88}$$

i.e., the mean of the effective channel is now reduced by a factor $1+\kappa$ compared to the Rayleigh fading case. Can you see this result intuitively as well? *Hint*: You might find the following limit theorem (p. 261 of [28]) useful for this exercise. Let $h_1, \ldots, h_K$ be i.i.d. real random variables with a common cdf $F(\cdot)$ and pdf $f(\cdot)$ satisfying $F(h)$ is less than 1 and is twice differentiable for all h, and is such that

$$\lim_{h \to \infty} \frac{d}{dh} \left[\frac{1 - F(h)}{f(h)} \right] = 0. \tag{6.89}$$

Then

$$\max_{1 \le k \le K} K f(l_K)(h_k - l_K)$$

converges in distribution to a limiting random variable with cdf

$$\exp(-e^{-x}).$$

In the above, l_K is given by $F(l_K) = 1 - 1/K$. This result states that the maximum of K such i.i.d. random variables grows like l_K.

Exercise 6.22 (Selective feedback) The downlink of IS-856 has K users each experiencing i.i.d. Rayleigh fading with average SNR of 0 dB. Each user selectively feeds back the requested rate only if its channel is greater than a threshold γ. Suppose γ is chosen such that the probability that no one sends a requested rate is ϵ. Find the expected number of users that sends in a requested rate. Plot this number for $K = 2, 4, 8, 16, 32, 64$ and for $\epsilon = 0.1$ and $\epsilon = 0.01$. Is selective feedback effective?

Exercise 6.23 The discussions in Section 6.7.2 about channel measurement, prediction and feedback are based on an FDD system. Discuss the analogous issues for a TDD system, both in the uplink and in the downlink.

Exercise 6.24 Consider the two-user downlink AWGN channel (cf. (6.16)):

$$y_k[m] = h_k x[m] + z_k[m], \quad k = 1, 2. \tag{6.90}$$

Here $z_k[m]$ are i.i.d. $\mathcal{CN}(0, N_0)$ Gaussian processes marginally ($k = 1, 2$). Let us take $|h_1| > |h_2|$ for this problem.

1. Argue that the capacity region of this downlink channel does not depend on the correlation between the additive Gaussian noise processes $z_1[m]$ and $z_2[m]$. *Hint*: Since the two users cannot cooperate, it should be intuitive that the error probability for user k depends only on the marginal distribution of $z_k[m]$ (for both $k = 1, 2$).
2. Now consider the following specific correlation between the two additive noises of the users. The pair $(z_1[m], z_2[m])$ is i.i.d. with time m with the distribution $\mathcal{CN}(0, \mathbf{K}_z)$. To preserve the marginals, the diagonal entries of the covariance matrix $\mathbf{K}_z$ have to be both equal to N_0. The only parameter that is free to be chosen is the off-diagonal element (denoted by ρN_0 with $|\rho| \leq 1$):

$$\mathbf{K}_z = \begin{bmatrix} N_0 & \rho N_0 \\ \rho N_0 & N_0 \end{bmatrix}.$$

Let us now allow the two users to cooperate, in essence creating a point-to-point AWGN channel with a single transmit but two receive antennas. Calculate the capacity ($C(\rho)$) of this channel as a function of ρ and show that if the rate pair (R_1, R_2) is within the capacity region of the downlink AWGN channel, then

$$R_1 + R_2 \leq C(\rho). \tag{6.91}$$

3. We can now choose the correlation ρ to minimize the upper bound in (6.91). Find the minimizing ρ (denoted by $\rho_{\min}$) and show that the corresponding (minimal) $C(\rho_{\min})$ is equal to $\log(1 + |h_1|^2 P / N_0)$.
4. The result of the calculation in the previous part is rather surprising: the rate $\log(1 + |h_1|^2 P / N_0)$ can be achieved by simply user 1 alone. This means that with a specific correlation ($\rho_{\min}$), cooperation among the users is not gainful. Show this formally by proving that for every time m with the correlation given by $\rho_{\min}$, the sequence of random variables $x[m], y_1[m], y_2[m]$ form a Markov chain (i.e., conditioned on $y_1[m]$, the random variables $x[m]$ and $y_2[m]$ are independent). This technique is useful in characterizing the capacity region of more involved downlinks, such as when there are multiple antennas at the base station.

Exercise 6.25 Consider the rate vectors in the downlink AWGN channel (cf. (6.16)) with superposition coding and orthogonal signaling as given in (6.22) and (6.23),

respectively. Show that superposition coding is strictly better than the orthogonal schemes, i.e., for every non-zero rate pair achieved by an orthogonal scheme, there is a superposition coding scheme which allows *each* user to strictly increase its rate.

Exercise 6.26 A reading exercise is to study [8], where the sufficiency of superposition encoding and decoding for the downlink AWGN channel is shown.

Exercise 6.27 Consider the two-user symmetric downlink fading channel with receiver CSI alone (cf. (6.50)). We have seen that the capacity region of the downlink channel does not depend on the correlation between the additive noise processes $z_1[m]$ and $z_2[m]$ (cf. Exercise 6.24(1)). Consider the following specific correlation: $(z_1[m], z_2[m])$ are $\mathcal{CN}(0, \mathbf{K}[m])$ and independent in time m. To preserve the marginal variance, the diagonal entries of the covariance matrix $\mathbf{K}[m]$ must be N_0 each. Let us denote the off-diagonal term by $\rho[m]N_0$ (with $|\rho[m]| \leq 1$). Suppose now we let the two users cooperate.

1. Show that by a careful choice of $\rho[m]$ (as a function of $h_1[m]$ and $h_2[m]$), cooperation is not gainful: that is, for any reliable rates R_1, R_2 in the downlink fading channel,

$$R_1 + R_2 \leq \mathbb{E}\left[\log\left(1 + \frac{|h|^2 P}{N_0}\right)\right], \tag{6.92}$$

    the same as can be achieved by a single user alone (cf. (6.51)). Here distribution of h is the symmetric stationary distribution of the fading processes $\{h_k[m]\}$ (for $k = 1, 2$). *Hint*: You will find Exercise 6.24(3) useful.
2. Conclude that the capacity region of the symmetric downlink fading channel is that given by (6.92).

Exercise 6.28 Show that the proportional fair algorithm with an infinite time-scale window maximizes (among all scheduling algorithms) the sum of the logarithms of the throughputs of the users. This justifies (6.57). This result has been derived in the literature at several places, including [12].

Exercise 6.29 Consider the opportunistic beamforming scheme in conjunction with a proportional fair scheduler operating in a slow fading environment. A reading exercise is to study Theorem 1 of [137], which shows that the rate available to each user is approximately equal to the instantaneous rate when it is being transmit beamformed, scaled down by the number of users.

Exercise 6.30 In a cellular system, the multiuser diversity gain in the downlink is expressed through the maximum SINR (cf. (6.63))

$$\text{SINR}_{\text{max}} := \max_{k=1\ldots K} \text{SINR}_k = \frac{P|h_k|^2}{N_0 + P\sum_{j=1}^{J}|g_{kj}|^2}, \tag{6.93}$$

where we have denoted P by the average received power at a user. Let us denote the ratio P/N_0 by SNR. Let us suppose that $h_1, \ldots, h_K$ are i.i.d. $\mathcal{CN}(0, 1)$ random variables, and $\{g_{kj}, k = 1 \ldots K, j = 1 \ldots J\}$ are i.i.d. $\mathcal{CN}(0, 0.2)$ random variables independent of h. (A factor of 0.2 is used to model the average scenario of the mobile user being closer to the base-station it is communicating with as opposed to all the other base-stations it is hearing interference from, cf. Section 4.2.3.)

1. Show using the limit theorem in Exercise 6.21 that

$$\frac{\mathbb{E}[\text{SINR}_{\text{max}}]}{x_K} \to 1, \qquad \text{as } K \to \infty, \tag{6.94}$$

where x_K satisfies the non-linear equation:

$$\left(1 + \frac{x_K}{5}\right)^J = K \exp\left(-\frac{x_K}{\text{SNR}}\right). \tag{6.95}$$

2. Plot x_K for $K = 1, \ldots, 16$ for different values of SNR (ranging from 0 dB to 20 dB). Can you intuitively justify the observation from the plot that x_K increases with increasing SNR values? *Hint*: The probability that $|h_k|^2$ is less than or equal to a small positive number ϵ is approximately equal to ϵ itself, while the probability that $|h_k|^2$ is larger than a large number $1/\epsilon$ is $\exp(-1/\epsilon)$. Thus the likely way SINR becomes large is by the denominator being small as opposed to the numerator becoming large.

3. Show using part (1), or directly, that at small values of SNR the mean of the effective SINR grows like $\log K$. You can also see this directly from (6.93): at small values of SNR, the effective SINR is simply the maximum of K Rayleigh distributed random variables and from Exercise 6.21(2) we know that the mean value grows like $\log K$.

4. At very high values of SNR, we can approximate $\exp(-x_K/\text{SNR})$ in (6.95) by 1. With this approximation, show, using part (1), that the scaling x_K is approximately like $K^{1/J}$. This is a faster growth rate than the one at low SNR.

5. In a cellular system, typically the value of P is chosen such that the background noise N_0 and the interference term are of the same order. This makes sense for a system where there is no scheduling of users: since the system is interference plus noise limited, there is no point in making one of them (interference or background noise) much smaller than the other. In our notation here, this means that SNR is approximately 0 dB. From the calculations of this exercise what design setting of P can you infer for a system using the multiuser diversity harnessing scheduler? Thus, conventional transmit power settings will have to be revisited in this new system point of view.

Exercise 6.31 (Interaction between space-time codes and multiuser diversity scheduling) A design is proposed for the downlink IS-856 using dual transmit antennas at the base-station. It employs the Alamouti scheme when transmitting to a single user and among the users schedules the user with the best *effective* instantaneous SNR under the Alamouti scheme. We would like to compare the performance gain, if any, of using this scheme as opposed to using just a single transmit antenna and scheduling to the user with the best instantaneous SNR. Assume independent Rayleigh fading across the transmit antennas.

1. Plot the distribution of the instantaneous effective SNR under the Alamouti scheme, and compare that to the distribution of the SNR for a single antenna.

2. Suppose there is only a single user (i.e., $K = 1$). From your plot in part (1), do you think the dual transmit antennas provide any gain? Justify your answer. *Hint*: Use Jensen's inequality.

3. How about when $K > 1$? Plot the achievable throughput under both schemes at average SNR $= 0$ dB and for different values of K.

4. Is the proposed way of using dual transmit antennas smart?