10

MIMO IV: multiuser communication

In Chapters 8 and 9, we have studied the role of multiple transmit and receive antennas in the context of point-to-point channels. In this chapter, we shift the focus to multiuser channels and study the role of multiple antennas in both the uplink (many-to-one) and the downlink (one-to-many). In addition to allowing spatial multiplexing and providing diversity to each user, multiple antennas allow the base-station to simultaneously transmit or receive data from multiple users. Again, this is a consequence of the increase in degrees of freedom from having multiple antennas.

We have considered several MIMO transceiver architectures for the point-to-point channel in Chapter 8. In some of these, such as linear receivers with or without successive cancellation, the complexity is mainly at the receiver. Independent data streams are sent at the different transmit antennas, and no cooperation across transmit antennas is needed. Equating the transmit antennas with users, these receiver structures can be directly used in the uplink where the users have a single transmit antenna each but the base-station has multiple receive antennas; this is a common configuration in cellular wireless systems.

It is less apparent how to come up with good strategies for the *downlink*, where the *receive* antennas are at the different users; thus the receiver structure has to be separate, one for each user. However, as will see, there is an interesting duality between the uplink and the downlink, and by exploiting this duality, one can map each receive architecture for the uplink to a corresponding transmit architecture for the downlink. In particular, there is an interesting *precoding* strategy, which is the "transmit dual" to the receiver-based successive cancellation strategy. We will spend some time discussing this.

The chapter is structured as follows. In Section 10.1, we first focus on the uplink with a single transmit antenna for each user and multiple receive antennas at the base-station. We then, in Section 10.2, extend our study to the MIMO uplink where there are multiple transmit antennas for each user. In Sections 10.3 and 10.4, we turn our attention to the use of multiple antennas in the downlink. We study precoding strategies that achieve the capacity of

the downlink. We conclude in Section 10.5 with a discussion of the system implications of using MIMO in cellular networks; this will link up the new insights obtained here with those in Chapters 4 and 6.

10.1 Uplink with multiple receive antennas

We begin with the narrowband time-invariant uplink with each user having a single transmit antenna and the base-station equipped with an array of antennas (Figure 10.1). The channels from the users to the base-station are time-invariant. The baseband model is

$$\mathbf{y}[m] = \sum_{k=1}^{K} \mathbf{h}_k x_k[m] + \mathbf{w}[m], \qquad (10.1)$$

with $\mathbf{y}[m]$ being the received vector (of dimension n_r, the number of receive antennas) at time m, and $\mathbf{h}_k$ the spatial signature of user k impinged on the receive antenna array at the base-station. User k's scalar transmit symbol at time m is denoted by $x_k[m]$ and $\mathbf{w}[m]$ is i.i.d. $\mathcal{CN}(0, N_0\mathbf{I}_{n_\mathrm{r}})$ noise.

10.1.1 Space-division multiple access

In the literature, the use of multiple receive antennas in the uplink is often called *space-division multiple access* (SDMA): we can discriminate amongst the users by exploiting the fact that different users impinge different spatial signatures on the receive antenna array.

An easy observation we can make is that this uplink is very similar to the MIMO point-to-point channel in Chapter 5 except that the signals sent out on the transmit antennas cannot be coordinated. We studied precisely such a signaling scheme using separate data streams on each of the transmit antennas in Section 8.3. We can form an analogy between users and transmit antennas (so n_t, the number of transmit antennas in the MIMO point-to-point channel in Section 8.3, is equal to the number of users K). Further, the equivalent MIMO point-to-point channel $\mathbf{H}$ is $[\mathbf{h}_1, \ldots, \mathbf{h}_K]$, constructed from the SIMO channels of the users.

Thus, the transceiver architecture in Figure 8.1 in conjunction with the receiver structures in Section 8.3 can be used as an SDMA strategy. For example, each of the user's signal can be demodulated using a linear decorrelator or an MMSE receiver. The MMSE receiver is the optimal compromise between maximizing the signal strength from the user of interest and suppressing the interference from the other users. To get better performance, one can also augment the linear receiver structure with successive cancellation to yield the MMSE–SIC receiver (Figure 10.2). With successive cancellation, there is also a further choice of cancellation ordering. By choosing a
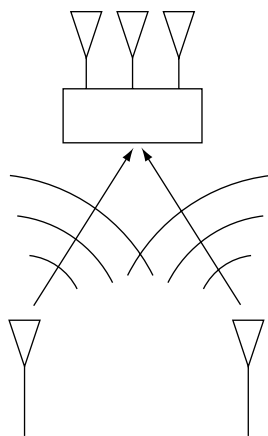

Figure 10.1 The uplink with single transmit antenna at each user and multiple receive antennas at the base-station.
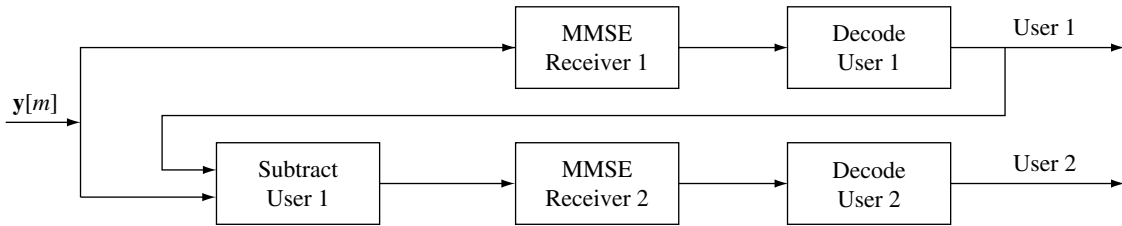
Figure 10.2 The MMSE–SIC receiver: user 1's data is first decoded and then the corresponding transmit signal is subtracted off before the next stage. This receiver structure, by changing the ordering of cancellation, achieves the two corner points in the capacity region.

different order, users are prioritized differently in the sharing of the common resource of the uplink channel, in the sense that users canceled later are treated better.

Provided that the overall channel matrix $\mathbf{H}$ is well-conditioned, all of these SDMA schemes can fully exploit the total number of degrees of freedom $\min\{K, n_r\}$ of the uplink channel (although, as we have seen, different schemes have different power gains). This translates to being able to simultaneously support multiple users, each with a data rate that is not limited by interference. Since the users are geographically separated, their transmit signals arrive in different directions at the receive array even when there is limited scattering in the environment, and the assumption of a well-conditioned $\mathbf{H}$ is usually valid. (Recall Example 7.4 in Section 7.2.4.) Contrast this to the point-to-point case when the transmit antennas are co-located, and a rich scattering environment is needed to provide a well-conditioned channel matrix $\mathbf{H}$.

Given the power levels of the users, the achieved SINR of each user can be computed for the different SDMA schemes using the formulas derived in Section 8.3 (Exercise 10.1). Within the class of linear receiver architecture, we can also formulate a power control problem: given target SINR requirements for the users, how does one optimally choose the powers and linear filters to meet the requirements? This is similar to the uplink CDMA power control problem described in Section 4.3.1, except that there is a further flexibility in the choice of the receive filters as well as the transmit powers. The first observation is that for *any* choice of transmit powers, one always wants to use the MMSE filter for each user, since that choice maximizes the SINR for every user. Second, the power control problem shares the basic *monotonicity* property of the CDMA problem: when a user lowers its transmit power, it creates less interference and benefits all other users in the system. As a consequence, there is a component-wise optimal solution for the powers, where every user is using the minimum possible power to support the SINR requirements. (See Exercise 10.2.) A simple distributed power control algorithm will converge to the optimal solution: at each step, each user first updates its MMSE filter as a function of the current power levels of the other users, and then updates its own transmit power so that its SINR requirement is just met. (See Exercise 10.3.)

10.1.2 SDMA capacity region

In Section 8.3.4, we have seen that the MMSE–SIC receiver achieves the best total rate among all the receiver structures. The performance limit of the uplink channel is characterized by the notion of a *capacity region*, introduced in Chapter 6. How does the performance achieved by MMSE–SIC compare to this limit?

With a *single* receive antenna at the base-station, the capacity region of the two-user uplink channel was presented in Chapter 6; it is the pentagon in Figure 6.2:

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

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

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

where P_1 and P_2 are the average power constraints on users 1 and 2 respectively. The individual rate constraints correspond to the maximum rate that each user can get if it has the entire channel to itself; the sum rate constraint is the total rate of a point-to-point channel with the two users acting as two transmit antennas of a single user, but sending independent signals.

The SDMA capacity region, for the *multiple* receive antenna case, is the natural extension (Appendix B.9 provides a formal justification):

$$R_1 < \log\left(1 + \frac{\|\mathbf{h}_1\|^2 P_1}{N_0}\right), \tag{10.2}$$

$$R_2 < \log\left(1 + \frac{\|\mathbf{h}_2\|^2 P_2}{N_0}\right), \tag{10.3}$$

$$R_1 + R_2 < \log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right), \tag{10.4}$$

where $\mathbf{K}_x = \mathrm{diag}(P_1, P_2)$. The capacity region is plotted in Figure 10.3.

The capacities of the point-to-point SIMO channels from each user to the base-station serve as the maximum rate each user can reliably communicate at if it has the entire channel to itself. These yield the constraints (10.2) and (10.3). The point-to-point capacity for user $k(k = 1, 2)$ is achieved by receive beamforming (projecting the received vector $\mathbf{y}$ in the direction of $\mathbf{h}_k$), converting the effective channel into a SISO one, and then decoding the data of the user.

Inequality (10.4) is a constraint on the sum of the rates that the users can communicate at. The right hand side is the total rate achieved in a point-to-point channel with the two users acting as two transmit antennas of one user with independent inputs at the antennas (cf. (8.2)).

Figure 10.3 Capacity region of the two-user SDMA uplink.
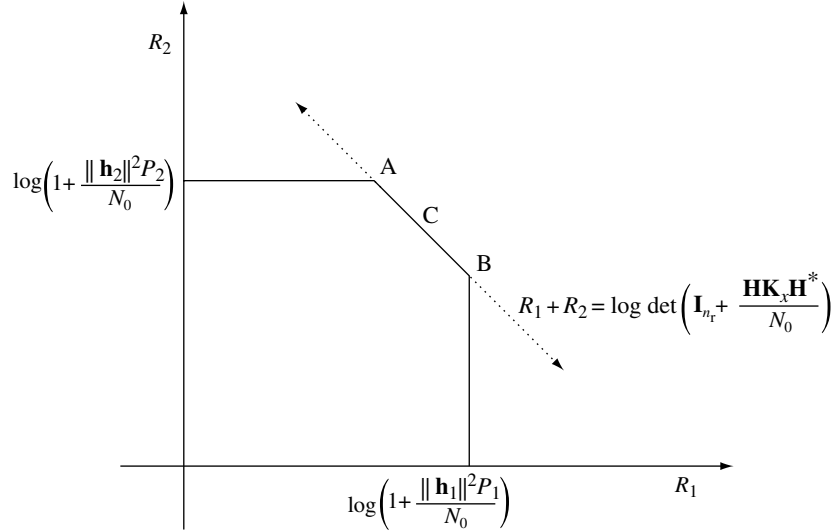
Since MMSE–SIC receivers (in Figure 10.2) are optimal with respect to achieving the total rate of the point-to-point channel with the two users acting as two transmit antennas of one user, it follows that the rates for the two users that this architecture can achieve in the uplink meets inequality (10.4) with equality. Moreover, if we cancel user 1 first, user 2 only has to contend with the background Gaussian noise and its performance meets the single-user bound (10.2). Hence, we achieve the corner point A in Figure 10.3. By reversing the cancellation order, we achieve the corner point B. Thus, MMSE–SIC receivers are information theoretically optimal for SDMA in the sense of achieving rate pairs corresponding to the two corner points A and B. Explicitly, the rate point A is given by the rate tuple (R_1, R_2):

$$R_2 = \log\left(1 + \frac{P_2\|\mathbf{h}_2\|^2}{N_0}\right),$$
$$R_1 = \log(1 + P_1\mathbf{h}_1^*(N_0\mathbf{I}_{n_r} + P_2\mathbf{h}_2\mathbf{h}_2^*)^{-1}\mathbf{h}_1), \tag{10.5}$$

where $P_1\mathbf{h}_1^*(N_0\mathbf{I}_{n_r} + P_2\mathbf{h}_2^*\mathbf{h}_2^*)^{-1}\mathbf{h}_1$ is the output SIR of the MMSE receiver for user 1 treating user 2's signal as colored Gaussian interference (cf. (8.62)).

For the single receive antenna (scalar) uplink channel, we have already seen in Section 6.1 that the corner points are also achievable by the SIC receiver, where at each stage a user is decoded treating all the uncanceled users as Gaussian noise. In the vector case with multiple receive antennas, the uncanceled users are also treated as Gaussian noise, but now this is a colored vector Gaussian noise. The MMSE filter is the optimal demodulator for a user in the face of such colored noise (cf. Section 8.3.3). Thus, we see that successive cancellation with MMSE filtering at each stage is the natural generalization of the SIC receiver we developed for the single antenna channel. Indeed, as explained in

Section 8.3.4, the SIC receiver is really just a special case of the MMSE–SIC receiver when there is only one receive antenna, and they are optimal for the same reason: they "implement" the chain rule of mutual information.

A comparison between the capacity regions of the uplink with and without multiple receive antennas (Figure 6.2 and Figure 10.3, respectively) highlights the importance of having multiple receive antennas in allowing SDMA. Let us focus on the high SNR scenario when N_0 is very small as compared with P_1 and P_2. With a single receive antenna at the base-station, we see from Figure 6.2 that there is a total of only one spatial degree of freedom, shared between the users. In contrast, with multiple receive antennas we see from Figure 10.3 that while the individual rates of the users have no more than one spatial degree of freedom, the sum rate has *two* spatial degrees of freedom. This means that both users can simultaneously enjoy one spatial degree of freedom, a scenario made possible by SDMA and not possible with a single receive antenna. The intuition behind this is clear when we look back at our discussion of the decorrelator (cf. Section 8.3.1). The received signal space has more dimensions than that spanned by the transmit signals of the users. Thus in decoding user 1's signal we can project the received signal in a direction orthogonal to the transmit signal of user 2, completely eliminating the inter-user interference (the analogy between streams and users carries forth here as well). This allows two effective parallel channels at high SNR. Improving the simple decorrelator by using the MMSE–SIC receiver allows us to *exactly* achieve the information theoretic limit.

In the light of this observation, we can take a closer look at the two corner points in the boundary of the capacity region (points A and B in Figure 10.3). If we are operating at point A we see that both users 1 and 2 have one spatial degree of freedom each. The point C, which corresponds to the symmetric capacity of the uplink (cf. (6.2)), also allows both users to have unit spatial degree of freedom. (In general, the symmetric capacity point C need not lie on the line segment joining points A and B; however it will be the center of this line segment when the channels are symmetric, i.e., $\|\mathbf{h}_1\| = \|\mathbf{h}_2\|$.) While the point C cannot be achieved directly using the receiver structure in Figure 10.2, we can achieve that rate pair by time-sharing between the operating points A and B (these two latter points can be achieved by the MMSE–SIC receiver).

Our discussion has been restricted to the two-user uplink. The extension to K users is completely natural. The capacity region is now a K-dimensional polyhedron: the set of rates $(R_1, \ldots, R_K)$ such that

$$\sum_{k \in S} R_k < \log \det \left(\mathbf{I}_{n_r} + \frac{1}{N_0} \sum_{k \in S} P_k \mathbf{h}_k \mathbf{h}_k^* \right), \quad \text{for each } S \subset \{1, \ldots, K\}. \quad (10.6)$$

There are $K!$ corner points on the boundary of the capacity region and each corner point is specified by an ordering of the K users and the corresponding rates are achieved by an MMSE–SIC receiver with that ordering of cancelling users.

10.1.3 System implications

What are the practical ways of exploiting multiple receive antennas in the uplink, and how does their performance compare to capacity? Let us first consider the narrowband system from Chapter 4 where the allocation of resources among the users is orthogonal. In Section 6.1 we studied orthogonal multiple access for the uplink with a single receive antenna at the base-station. Analogous to (6.8) and (6.9), the rates achieved by two users, when the base-station has multiple receive antennas and a fraction α of the degrees of freedom is allocated to user 1, are

$$\left(\alpha \log \left(1 + \frac{P_1 \|\mathbf{h}_1\|^2}{\alpha N_0} \right), (1-\alpha) \log \left(1 + \frac{P_2 \|\mathbf{h}_2\|^2}{(1-\alpha) N_0} \right) \right). \qquad (10.7)$$

It is instructive to compare this pair of rates with the one obtained with orthogonal multiple access in the single receive antenna setting (cf. (6.8) and (6.9)). The difference is that the received SNR of user k is boosted by a factor $\|\mathbf{h}_k\|^2$; this is the receive beamforming power gain. There is however no gain in the degrees of freedom: the total is still one. The power gain allows the users to reduce their transmit power for the same received SNR level. However, due to orthogonal resource allocation and sparse reuse of the bandwidth, narrowband systems already operate at high SNR and in this situation a power gain is not much of a system benefit. A degree-of-freedom gain would have made a larger impact.

At high SNR, we have already seen that the two-user SDMA sum capacity has two spatial degrees of freedom as opposed to the single one with only one receive antenna at the base-station. Thus, orthogonal multiple access makes very poor use of the available spatial degrees of freedom when there are multiple receive antennas. Indeed, this can be seen clearly from a comparison of the orthogonal multiple access rates with the capacity region. With a single receive antenna, we have found that we can get to exactly one point on the boundary of the uplink capacity region (see Figure 6.4); the gap is not too large unless there is a significant power disparity. With multiple receive antennas, Figure 10.4 shows that the orthogonal multiple access rates are strictly suboptimal at all points[1] and the gap is also larger.

Intuitively, to exploit the available degrees of freedom *both* users must access the channel simultaneously and their signals should be separable at the base-station (in the sense that $\mathbf{h}_1$ and $\mathbf{h}_2$, the receive spatial signatures of the users at the base-station, are linearly independent). To get this benefit, more complex signal processing is required at the receiver to extract the signal of each user from the aggregate. The complexity of SDMA grows with the number of users K when there are more users in the system. On the

---

[1] Except for the degenerate case when $\mathbf{h}_1$ and $\mathbf{h}_2$ are multiples of each other; see Exercise 10.4.

Figure 10.4 The two-user
uplink with multiple receive
antennas at the base-station:
performance of orthogonal
multiple access is strictly
inferior to the capacity.
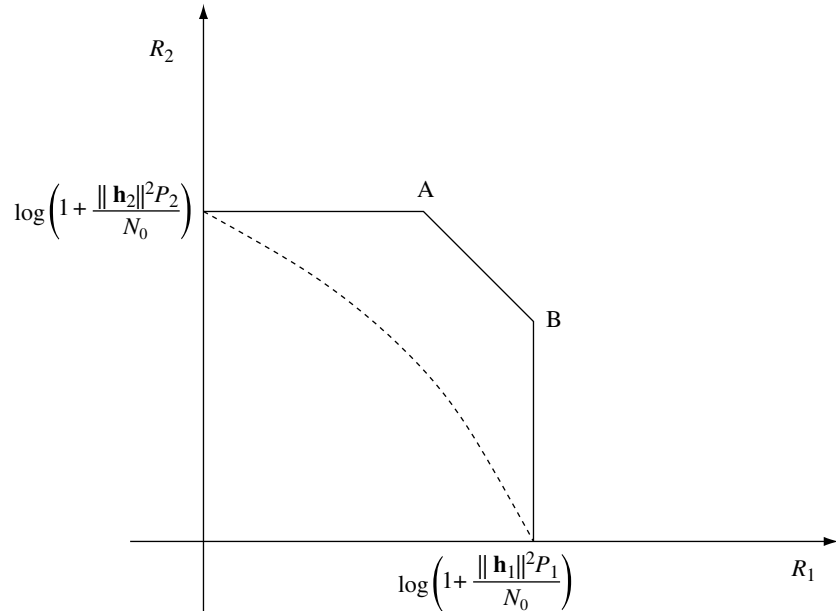
other hand, the available degrees of freedom are limited by the number of
receive antennas, n_r, and so there is no further degree-of-freedom gain beyond
having n_r users performing SDMA simultaneously. This suggests a nearly
optimal multiple access strategy where the users are divided into groups of n_r
users with SDMA *within* each group and orthogonal multiple access *between*
the groups. Exercise 10.5 studies the performance of this scheme in greater
detail.

On the other hand, at low SNR, the channel is power-limited rather than
degrees-of-freedom-limited and SDMA provides little performance gain over
orthogonal multiple access. This can be observed by an analysis as in the char-
acterization of the capacity of MIMO channels at low SNR, cf. Section 8.2.2,
and is elaborated in Exercise 10.6.

In general, multiple receive antennas can be used to provide beamforming
gain for the users. While this power gain is not of much benefit to the
narrowband systems, both the wideband CDMA and wideband OFDM uplink
operate at low SNR and the power gain is more beneficial.

Summary 10.1 SDMA and orthogonal multiple access

The MMSE–SIC receiver is optimal for achieving SDMA capacity.

SDMA with n_r receive antennas and K users provides $\min(n_r, K)$ spatial
degrees of freedom.

> Orthogonal multiple access with n_r receive antennas provides only one spatial degree of freedom but n_r-fold power gain.
>
> Orthogonal multiple access provides comparable performance to SDMA at low SNR but is far inferior at high SNR.

10.1.4 Slow fading

We introduce fading first in the scenario when the delay constraint is small relative to the coherence time of all the users: the slow fading scenario. The uplink fading channel can be written as an extension of (10.1), as

$$\mathbf{y}[m] = \sum_{k=1}^{K} \mathbf{h}_k[m] x_k[m] + \mathbf{w}[m]. \qquad (10.8)$$

In the slow fading model, for every user k, $\mathbf{h}_k[m] = \mathbf{h}_k$ for all time m. As in the uplink with a single antenna (cf. Section 6.3.1), we will analyze only the symmetric uplink: the users have the same transmit power constraint, P, and further, the channels of the users are statistically independent and identical. In this situation, symmetric capacity is a natural performance measure and we suppose the users are transmitting at the same rate R bits/s/Hz.

Conditioned on a realization of the received spatial signatures $\mathbf{h}_1, \ldots, \mathbf{h}_K$, we have the time-invariant uplink studied in Section 10.1.2. When the symmetric capacity of this channel is less than R, an outage results. The probability of the outage event is, from (10.6),

$$p_{\text{out}}^{\text{ul}-\text{mimo}} := \quad \mathbb{P}\left\{ \log \det \left(\mathbf{I}_{n_r} + \text{SNR} \sum_{k \in \mathcal{S}} \mathbf{h}_k \mathbf{h}_k^* \right) < |\mathcal{S}| R, \right.$$

$$\left. \text{for some } \mathcal{S} \subset \{1, \ldots, K\} \right\}. \qquad (10.9)$$

Here we have written $\text{SNR} := P/N_0$. The corresponding largest rate R such that $p_{\text{out}}^{\text{ul}-\text{mimo}}$ is less than or equal to ϵ is the ϵ-outage symmetric capacity C_ϵ^{sym}. With a single user in the system, C_ϵ^{sym} is simply the ϵ-outage capacity, $C_\epsilon(\text{SNR})$, of the point-to-point channel with receive diversity studied in Section 5.4.2. More generally, with $K > 1$, C_ϵ^{sym} is upper bounded by this quantity: with more users, inter-user interference is another source of error.

Orthogonal multiple access completely eliminates inter-user interference and the corresponding largest symmetric outage rate is, as in (6.33),

$$\frac{C_{\epsilon/K}(K\text{SNR})}{K}. \qquad (10.10)$$

We can see, just as in the situation when the base-station has a single receive antenna (cf. Section 6.3.1), that orthogonal multiple access at low SNR is

close to optimal. At low SNR, we can approximate $p_{\text{out}}^{\text{ul}-\text{mimo}}$ (with $n_{\text{r}} = 1$, a similar approximation is in (6.34)):

$$p_{\text{out}}^{\text{ul}-\text{mimo}} \approx K p_{\text{out}}^{\text{rx}}, \tag{10.11}$$

where $p_{\text{out}}^{\text{rx}}$ is the outage probability of the point-to-point channel with receive diversity (cf. (5.62)). Thus $C_{\epsilon}^{\text{sym}}$ is approximately $C_{\epsilon/K}(\text{SNR})$. On the other hand, the rate in (10.10) is also approximately equal to $C_{\epsilon/K}(\text{SNR})$ at low SNR.

At high SNR, we have seen that orthogonal multiple access is suboptimal, both in the context of outage performance with a single receive antenna and the capacity region of SDMA. A better baseline performance can be obtained by considering the outage performance of the bank of decorrelators: this receiver structure performed well in terms of the capacity of the point-to-point MIMO channel, cf. Figure 8.9. With the decorrelator bank, the inter-user interference is completely nulled out (assuming $n_{\text{r}} \geq K$). Further, with i.i.d. Rayleigh fading, each user sees an effective point-to-point channel with $n_{\text{r}} - K + 1$ receive diversity branches (cf. Section 8.3.1). Thus, the largest symmetric outage rate is exactly the ϵ-outage capacity of the point-to-point channel with $n_{\text{r}} - K + 1$ receive diversity branches, leading to the following interpretation:

> Using the bank of decorrelators, increasing the number of receive antennas, n_{r}, by 1 allows us to *either* admit one extra user with the same outage performance for each user, *or* increase the effective number of diversity branches seen by each user by 1.

How does the outage performance improve if we replace the bank of decor-relators with the joint ML receiver? The direct analysis of $C_{\epsilon}^{\text{sym}}$ at high SNR is quite involved, so we resort to the use of the coarser diversity–multiplexing tradeoff introduced in Chapter 9 to answer this question. For the bank of decorrelators, the diversity gain seen by each user is $(n_{\text{r}} - K + 1)(1 - r)$ where r is the multiplexing gain of each user (cf. Exercise 9.5). This provides a lower bound to the diversity–multiplexing performance of the joint ML receiver. On the other hand, the outage performance of the uplink cannot be better than the situation when there is no inter-user interference, i.e., each user sees a point-to-point channel with receiver diversity of n_{r} branches. This is the *single-user* upper bound. The corresponding single-user tradeoff curve is $n_{\text{r}}(1 - r)$. These upper and lower bounds to the outage performance are plotted in Figure 10.5.

The tradeoff curve with the joint ML receiver in the uplink can be evaluated: with more receive antennas than the number of users (i.e., $n_{\text{r}} \geq K$), the tradeoff curve is the *same* as the upper bound derived with each user seeing no inter-user interference. In other words, the tradeoff curve is $n_{\text{r}}(1 - r)$ and single-user performance is achieved even though there are other users in

Figure 10.5 The diversity–multiplexing tradeoff curves for the uplink with a bank of decorrelators (equal to $(n_r - K + 1)(1 - r)$, a lower bound to the outage performance with the joint ML receiver) and that when there is no inter-user interference (equal to $n_r(1 - r)$, the single-user upper bound to the outage performance of the uplink). The latter is actually achievable.
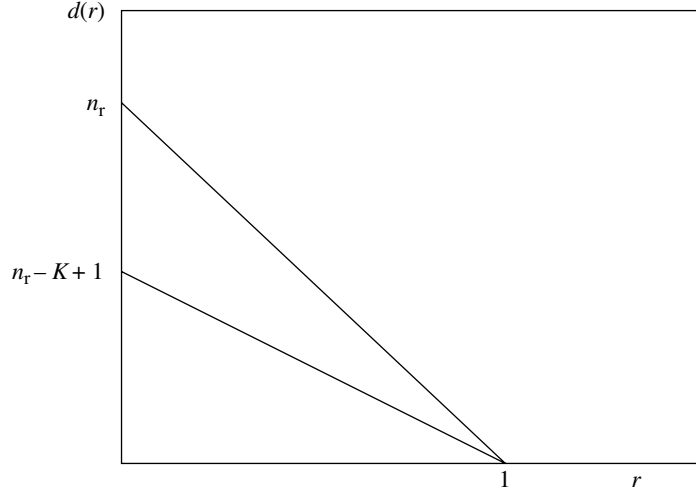
the system. This allows the following interpretation of the performance of the joint ML receiver, in contrast to the decorrelator bank:

Using the joint ML receiver, increasing the number of receive antennas, n_r, by 1 allows us to *both* admit one extra user *and* simultaneously increase the effective number of diversity branches seen by each user by 1.

With $n_r < K$, the optimal uplink tradeoff curve is more involved. We can observe that the total spatial degrees of freedom in the uplink is now limited by n_r and thus the largest multiplexing rate *per user* can be no more than n_r/K. On the other hand, with no inter-user interference, each user can have a multiplexing gain up to 1; thus, this upper bound can never be attained for large enough multiplexing rates. It turns out that for slightly smaller multiplexing rates $r \leq n_r/(K+1)$ per user, the diversity gain obtained is still equal to the single-user bound of $n_r(1 - r)$. For r larger than this threshold (but still smaller than n_r/K), the diversity gain is that of a $K \times n_r$ MIMO channel at a total multiplexing rate of Kr; this is as if the K users pooled their total rate together. The overall optimal uplink tradeoff curve is plotted in Figure 10.6: it has two line segments joining the points

$$(0, n_r), \qquad \left(\frac{n_r}{K+1}, \frac{n_r(K - n_r + 1)}{K+1} \right), \qquad \text{and } \left(\frac{n_r}{K}, 0 \right).$$

Exercise 10.7 provides the justification to the calculation of this tradeoff curve.

In Section 6.3.1, we plotted the ratio of C_ϵ^{sym} for a single receive antenna uplink to $C_\epsilon(\text{SNR})$, the outage capacity of a point-to-point channel with no inter-user interference. For a fixed outage probability ϵ, increasing the SNR

Figure 10.6 The diversity–
multiplexing tradeoff curve for
the uplink with the joint ML
receiver for $n_r < K$. The
multiplexing rate r is measured
per user. Up to a multiplexing
gain of $n_r/(K+1)$, single-user
tradeoff performance of
$n_r(1-r)$ is achieved. The
maximum number of degrees
of freedom per user is n_r/K,
limited by the number of
receive antennas.
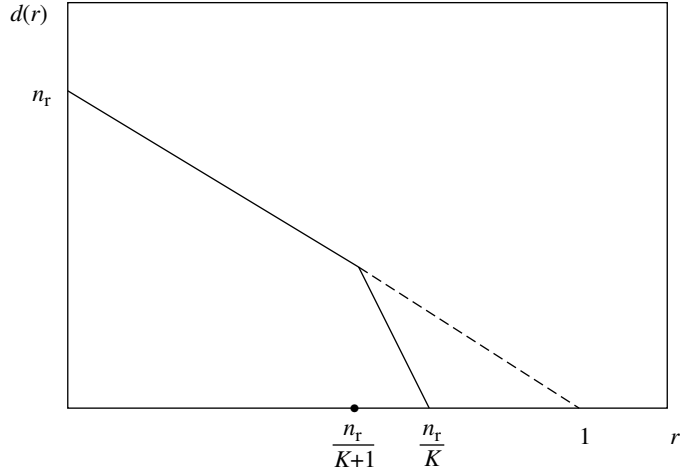
corresponds to decreasing the required diversity gain. Substituting $n_r = 1$ and
$K = 2$, in Figure 10.6, we see that as long as the required diversity gain
is larger than 2/3, the corresponding multiplexing gain is as if there is no
inter-user interference. This explains the behavior in Figure 6.10, where the
ratio of C_ϵ^{sym} to $C_\epsilon(\text{SNR})$ increases initially with SNR. With a further increase
in SNR, the corresponding desired diversity gain drops below 2/3 and now
there is a penalty in the achievable multiplexing rate due to the inter-user
interference. This penalty corresponds to the drop of the ratio in Figure 6.10
as SNR increases further.

10.1.5 Fast fading

Here we focus on the case when communication is over several coherence
intervals of the user channels; this way most channel fade levels are experi-
enced. This is the *fast fading* assumption studied for the single antenna uplink
in Section 6.3 and the point-to-point MIMO channel in Section 8.2. As usual,
to simplify the analysis we assume that the base-station can perfectly track
the channels of all the users.

Receiver CSI

Let us first consider the case when the users have only a statistical model of
the channel (taken to be stationary and ergodic, as in the earlier chapters). In
our notation, this is the case of receiver CSI. For notational simplicity, let us
consider only two users in the uplink (i.e., $K = 2$). Each user's rate cannot be
larger than when it is the only user transmitting (an extension of (5.91) with
multiple receive antennas):

$$R_k \leq \mathbb{E}\left[\log\left(1 + \frac{\|\mathbf{h}_k\|^2 P_k}{N_0}\right)\right], \qquad k = 1, 2. \tag{10.12}$$

Here the figure shows the capacity region pentagon with axes R_2 (vertical) and R_1 (horizontal). The vertical axis intercept is labeled $\mathrm{E}\left[\log\left(1+\frac{\|\mathbf{h}_2\|^2 P_2}{N_0}\right)\right]$ and the horizontal axis intercept is labeled $\mathrm{E}\left[\log\left(1+\frac{\|\mathbf{h}_1\|^2 P_1}{N_0}\right)\right]$. Points A and B mark the corner points, and the sloped boundary is labeled $R_1+R_2=\mathrm{E}\left[\log\det\left(\mathbf{I}_{n_r}+\frac{\mathbf{H}\mathbf{K}_x\mathbf{H}^*}{N_0}\right)\right]$.
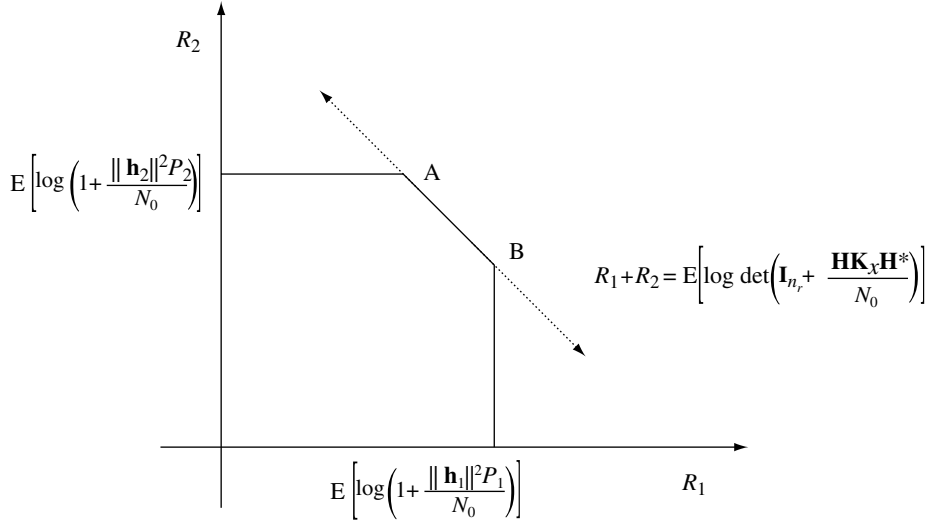
Figure 10.7 Capacity region of the two-user SIMO uplink with receiver CSI.

We also have the sum constraint (an extension of (6.37) with multiple receive antennas, cf.(8.10)):

$$R_1 + R_2 \leq \mathbb{E}\left[\log\det\left(\mathbf{I}_{n_\mathrm{r}} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right)\right]. \qquad (10.13)$$

Here we have written $\mathbf{H} = [\mathbf{h}_1\mathbf{h}_2]$ and $\mathbf{K}_x = \mathrm{diag}\{P_1, P_2\}$. The capacity region is a pentagon (see Figure 10.7). The two corner points are achieved by the receiver architecture of linear MMSE filters followed by successive cancellation of the decoded user. Appendix B.9.3 provides a formal justification.

Let us focus on the sum capacity in (10.13). This is exactly the capacity of a point-to-point MIMO channel with receiver CSI where the covariance matrix is chosen to be diagonal. The performance gain in the sum capacity over the single receive antenna case (cf. (6.37)) is of the same nature as that of a point-to-point MIMO channel over a point-to-point channel with only a single receive antenna. With a sufficiently random and well-conditioned channel matrix $\mathbf{H}$, the performance gain is significant (cf. our discussion in Section 8.2.2). Since there is a strong likelihood of the users being geographically far apart, the channel matrix is likely to be well-conditioned (recall our discussion in Example 7.4 in Section 7.2.4). In particular, the important observation we can make is that each of the users has one spatial degree of freedom, while with a single receive antenna, the sum capacity itself has one spatial degree of freedom.

Full CSI

We now move to the other scenario, full CSI both at the base-station and at each of the users.[2] We have studied the full CSI case in the uplink for single transmit and receive antennas in Section 6.3 and here we will see the role played by an array of receive antennas.

Now the users can vary their transmit power as a function of the channel realizations; still subject to an average power constraint. If we denote the transmit power of user k at time m by $P_k(\mathbf{h}_1[m], \mathbf{h}_2[m])$, i.e., it is a function of the channel states $\mathbf{h}_1[m], \mathbf{h}_2[m]$ at time m, then the rate pairs (R_1, R_2) at which the users can jointly reliably communicate to the base-station satisfy (analogous to (10.12) and (10.13)):

$$R_k \leq \mathbb{E}\left[\log\left(1 + \frac{\|\mathbf{h}_k\|^2 P_k(\mathbf{h}_1, \mathbf{h}_2)}{N_0}\right)\right], \qquad k = 1, 2, \quad (10.14)$$

$$R_1 + R_2 \leq \mathbb{E}\left[\log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right)\right]. \qquad (10.15)$$

Here we have written $\mathbf{K}_x = \text{diag}\{P_1(\mathbf{h}_1, \mathbf{h}_2), P_2(\mathbf{h}_1, \mathbf{h}_2)\}$. By varying the power allocations, the users can communicate at rate pairs in the *union* of the pentagons of the form defined in (10.14) and (10.15). By time sharing between two different power allocation policies, the users can also achieve every rate pair in the *convex hull*[3] of the union of these pentagons; this is the capacity region of the uplink with full CSI. The power allocations are still subject to the average constraint, denoted by P (taken to be the same for each user for notational convenience):

$$\mathbb{E}[P_k(\mathbf{h}_1, \mathbf{h}_2)] \leq P, \qquad k = 1, 2. \qquad (10.16)$$

In the point-to-point channel, we have seen that the power variations are waterfilling over the channel states (cf. Section 5.4.6). To get some insight into how the power variations are done in the uplink with multiple receive antennas, let us focus on the sum capacity

$$C_{\text{sum}} = \max_{P_k(\mathbf{h}_1, \mathbf{h}_2),\ k=1,2} \mathbb{E}\left[\log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right)\right], \qquad (10.17)$$

where the power allocations are subject to the average constraint in (10.16). In the uplink with a single receive antenna at the base-station (cf. Section 6.3.3), we have seen that the power allocation that maximizes sum capacity allows only the best user to transmit (a power that is waterfilling over the best user's

---

[2] In an FDD system, the base-station need not feedback all the channel states of all the users to every user. Instead, only the amount of power to be transmitted needs be relayed to the users.

[3] The convex hull of a set is the collection of all points that can be represented as convex combinations of elements of the set.

channel state, cf. (6.47)). Here each user is received as a vector ($\mathbf{h}_k$ for user k) at the base-station and there is no natural ordering of the users to bring this argument forth here. Still, the optimal allocation of powers can be found using the Lagrangian techniques, but the solution is somewhat complicated and is studied in Exercise 10.9.

10.1.6 Multiuser diversity revisited

One of the key insights from the study of the performance of the uplink with full CSI in Chapter 6 was the discovery of multiuser diversity. How do multiple receive antennas affect multiuser diversity? With a single receive antenna and i.i.d. user channel statistics, we have seen (see Section 6.6) that the sum capacity in the uplink can be interpreted as the capacity of the following point-to-point channel with full CSI:

- The power constraint is the sum of the power constraints of the users (equal to KP with equal power constraints for the users $P_i = P$).
- The channel quality is $|h_{k^*}|^2 := \max_{k=1\ldots K} |h_k|^2$, that corresponding to the strongest user k^*.

The corresponding sum capacity is (see (6.49))

$$C_{\text{sum}} = \mathbb{E}\left[\log\left(1 + \frac{P^*(h_{k^*})|h_{k^*}|^2}{N_0}\right)\right], \tag{10.18}$$

where P^* is the waterfilling power allocation (see (5.100) and (6.47)). With multiple receive antennas, the optimal power allocation does not allow a simple characterization. To get some insight, let us first consider (the suboptimal strategy of) transmitting from only one user at a time.

One user at a time policy

In this case, the multiple antennas at the base-station translate into receive beamforming gain for the users. Now we can order the users based on the beamforming power gain due to the multiple receive antennas at the base-station. Thus, as an analogy to the strongest user in the single antenna situation, here we can choose that user which has the largest receive beamforming gain: the user with the largest $\|\mathbf{h}_k\|^2$. Assuming i.i.d. user channel statistics, the sum rate with this policy is

$$\mathbb{E}\left[\log\left(1 + \frac{P^*_{k^*}(\|\mathbf{h}_{k^*}\|)\|\mathbf{h}_{k^*}\|^2}{N_0}\right)\right]. \tag{10.19}$$

Comparing (10.19) with (10.18), we see that the only difference is that the scalar channel gain $|h_k|^2$ is replaced by the receive beamforming gain $\|\mathbf{h}_k\|^2$.

The multiuser diversity gain depends on the probability that the maximum of the users' channel qualities becomes large (the *tail* probability). For

example, we have seen (cf. Section 6.7) that the multiuser diversity gain with
Rayleigh fading is larger than that in Rician fading (with the same average
channel quality). With i.i.d. channels to the receive antenna array (with unit
average channel quality), we have by the law of large numbers

$$\frac{\|\mathbf{h}_k\|^2}{n_{\mathrm{r}}} \to 1, \qquad n_{\mathrm{r}} \to \infty. \tag{10.20}$$

So, the receive beamforming gain can be approximated as $\|\mathbf{h}_k\|^2 \approx n_{\mathrm{r}}$ for
large enough n_{r}. This means that the tail of the receive beamforming gain
decays rapidly for large n_{r}.

   As an illustration, the density of $\|\mathbf{h}_k\|^2$ for i.i.d. Rayleigh fading (i.e., it is
a $\chi^2_{2n_{\mathrm{r}}}$ random variable) scaled by n_{r} is plotted in Figure 10.8. We see that the
larger the n_{r} value is, the more concentrated the density of the scaled random
variable $\chi^2_{2n_{\mathrm{r}}}$ is around its mean. This illustration is similar in nature to that
in Figure 6.23 in Section 6.7 where we have seen the plot of the densities of
the channel quality with Rayleigh and Rician fading. Thus, while the array of
receive antennas provides a beamforming gain, the multiuser diversity gain is
restricted. This effect is illustrated in Figure 10.9 where we see that the sum
capacity does not increase much with the number of users, when compared
to the corresponding AWGN channel.

Optimal power allocation policy

We have discussed the impact of multiple receive antennas on multiuser diver-
sity under the suboptimal strategy of allowing only one user (the best user)
to transmit at any time. Let us now consider how the sum capacity benefits
from multiuser diversity; i.e., we have to study the power allocation policy
that is optimal for the sum of user rates. In our previous discussions, we have
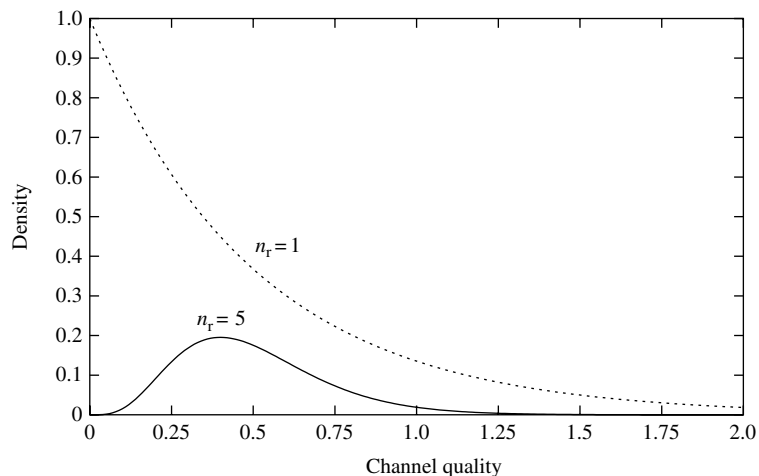found a simple form for this power allocation policy: for a point-to-point single

Figure 10.8 Plot of the density of a $\chi^2_{2n_{\mathrm{r}}}$ random variable divided by n_{r} for $n_{\mathrm{r}} = 1, 5$. The larger the n_{r}, the more concentrated the normalized random variable is around its mean of one.

Figure 10.9 Sum capacities of the uplink Rayleigh fading channel with n_r the number of receive antennas, for $n_r = 1, 5$. Here SNR = 1 (0 dB) and the Rayleigh fading channel is $\mathbf{h} \sim \mathcal{CN}(0, \mathbf{I}_{n_r})$. Also plotted for comparison is the corresponding performance for the uplink AWGN channel with $n_r = 5$ and SNR = 5 (7 dB).
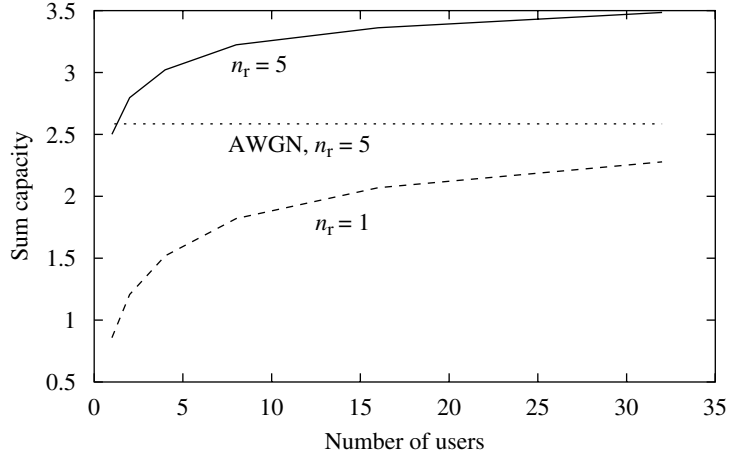
antenna channel, the allocation is waterfilling. For the single antenna uplink, the policy is to allow only the best user to transmit and, further, the power allocated to the best user is waterfilling over its channel quality. In the uplink with multiple receive antennas, there is no such simple expression in general. However, with both n_r and K large and comparable, the following simple policy is very close to the optimal one. (See Exercise 10.10.) *Every* user transmits and the power allocated is waterfilling over its own channel state, i.e.,

$$P_k(\mathbf{H}) = \left(\frac{1}{\lambda} - \frac{I_0}{\|\mathbf{h}_k\|^2} \right)^+, \qquad k = 1, \ldots, K. \tag{10.21}$$

As usual the water level, λ, is chosen such that the average power constraint is met.

It is instructive to compare the waterfilling allocation in (10.21) with the one in the uplink with a single receive antenna (see (6.47)). The important difference is that when there is only one user transmitting, waterfilling is done over the channel quality with respect to the background noise (of power density N_0). However, here all the users are simultaneously transmitting, using a similar waterfilling power allocation policy. Hence the waterfilling in (10.21) is done over the channel quality (the receive beamforming gain) with respect to the background *interference* plus noise: this is denoted by the term I_0 in (10.21). In particular, at high SNR the waterfilling policy in (10.21) simplifies to the constant power allocation at all times (under the condition that there are more receive antennas than the number of users).

Now the impact on multiuser diversity is clear: it is reduced to the basic opportunistic communication gain by waterfilling in a point-to-point channel. This gain depends solely on how the individual channel qualities of the users fluctuate with time and thus the multiuser nature of the gain is lost. As we have seen earlier (cf. Section 6.6), the gain of opportunistic communication in a point-to-point context is much more limited than that in the multiuser context.

Summary 10.2 Opportunistic communication and multiple receive antennas

Orthogonal multiple access: scheduled user gets a power gain but reduced multiuser diversity gain.

SDMA: multiple users simultaneously transmit.

- Optimal power allocation approximated by *waterfilling* with respect to an intra-cell interference level.
- Multiuser nature of the opportunistic gain is lost.
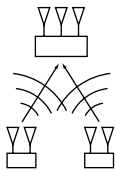
10.2 MIMO uplink

Figure 10.10 The MIMO uplink with multiple transmit antennas at each user and multiple receive antennas at the base-station.

Now we move to consider the role of multiple transmit antennas (at the mobiles) along with the multiple receive antennas at the base-station (Figure 10.10). Let us denote the number of transmit antennas at user k by $n_{tk}, k = 1, \ldots, K$. We begin with the time-invariant channel; the corresponding model is an extension of (10.1):

$$\mathbf{y}[m] = \sum_{k=1}^{K} \mathbf{H}_k \mathbf{x}_k[m] + \mathbf{w}[m], \qquad (10.22)$$

where $\mathbf{H}_k$ is a fixed n_r by n_{tk} matrix.

10.2.1 SDMA with multiple transmit antennas

There is a natural extension of our SDMA discussion in Section 10.1.2 to multiple transmit antennas. As before, we start with $K = 2$ users.

- **Transmitter architecture** Each user splits its data and encodes them into independent streams of information with user k employing $n_k := \min(n_{tk}, n_r)$ streams (just as in the point-to-point MIMO channel). Powers $P_{k1}, P_{k2}, \ldots, P_{kn_k}$ are allocated to the n_k data streams, passed through a rotation $\mathbf{U}_k$ and sent over the transmit antenna array at user k. This is analogous to the transmitter structure we have seen in the point-to-point MIMO channel in Chapter 5. In the time-invariant *point-to-point* MIMO channel, the rotation matrix $\mathbf{U}$ was chosen to correspond to the right rotation in the singular value decomposition of the channel and the powers allocated to the data streams correspond to the waterfilling allocations over the squared singular values of the channel matrix (cf. Figure 7.2). The transmitter architecture is illustrated in Figure 10.11.
- **Receiver architecture** The base-station uses the MMSE–SIC receiver to decode the data streams of the users. This is an extension of the receiver
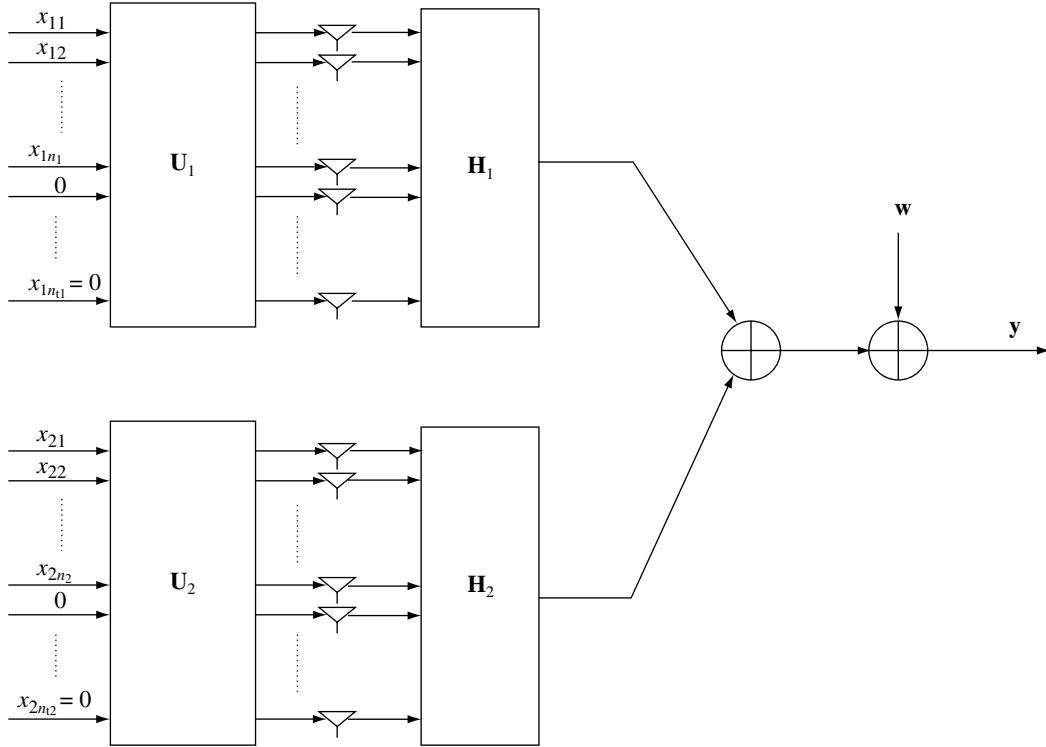
architecture in Chapter 8 (cf. Figure 8.16). This architecture is illustrated in Figure 10.12.

The rates R_1, R_2 achieved by this transceiver architecture must satisfy the constraints, analogous to (10.2), (10.3) and (10.4):

$$R_k \leq \log\det\left(\mathbf{I}_{n_\mathrm{r}} + \frac{1}{N_0}\mathbf{H}_k\mathbf{K}_{xk}\mathbf{H}_k^*\right), \qquad k = 1, 2, \quad (10.23)$$

$$R_1 + R_2 \leq \log\det\left(\mathbf{I}_{n_\mathrm{r}} + \frac{1}{N_0}\sum_{k=1}^{2}\mathbf{H}_k\mathbf{K}_{xk}\mathbf{H}_k^*\right). \qquad (10.24)$$

Here we have written $\mathbf{K}_{xk} := \mathbf{U}_k\Lambda_k\mathbf{U}_k^*$ and Λ_k to be a diagonal matrix with the n_{tk} diagonal entries equal to the power allocated to the data streams $P_{k1}, \ldots, P_{kn_k}$ (if $n_k < n_{tk}$ then the remaining diagonal entries are equal to zero, see Figure 10.11). The rate region defined by the constraints in (10.23) and (10.24) is a pentagon; this is similar to the one in Figure 10.3 and illustrated in Figure 10.13. The receiver architecture in Figure 10.2, where the data streams of user 1 are decoded first, canceled, and then the data streams of user 2 are decoded, achieves the corner point A in Figure 10.13.
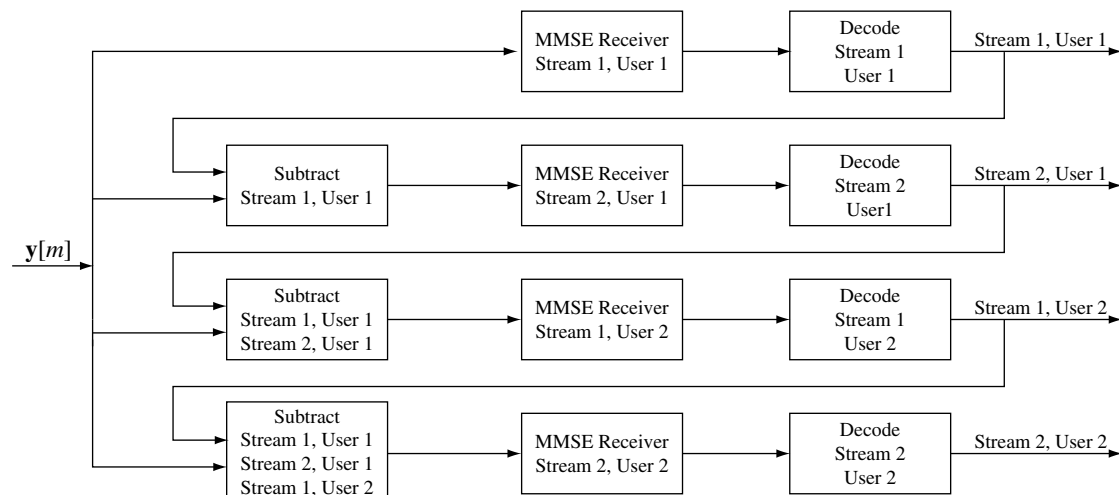
Figure 10.12 Receiver architecture for the two-user MIMO uplink. In this figure, each user has two transmit antennas and splits their data into two data streams each. The base-station decodes the data streams of the users using the linear MMSE filter, successively canceling them as they are decoded.

With a single transmit antenna at each user, the transmitter architecture simplifies considerably: there is only one data stream and the entire power is allocated to it. With multiple transmit antennas, we have a choice of power splits among the data streams and also the choice of the rotation $\mathbf{U}$ before sending the data streams out of the transmit antennas. In general, different choices of power splits and rotations lead to different pentagons (see Figure 10.14), and the capacity region is the convex hull of the union of all these pentagons; thus the capacity region in general is not a pentagon. This is because, unlike the single transmit antenna case, there are no covariance matrices $\mathbf{K}_{x1}, \mathbf{K}_{x2}$ that simultaneously maximize the right hand side of all the three constraints in (10.23) and (10.24). Depending on how one wants to trade off the performance of the two users, one would use different input strategies. This is formulated as a convex programming problem in Exercise 10.12.

Throughout this section, our discussion has been restricted to the two-user uplink. The extension to K users is completely natural. The capacity region is now K dimensional and for fixed transmission filters $\mathbf{K}_{xk}$ modulating the streams of user k (here $k = 1, \ldots, K$) there are $K!$ corner points on the boundary region of the achievable rate region; each corner point is specified by an ordering of the K users and the corresponding rate tuple is achieved by the linear MMSE filter bank followed by successive cancellation of users (and streams within a user's data). The transceiver structure is a K user extension of the pictorial depiction for two users in Figures 10.11 and 10.12.

10.2.2 System implications

Simple engineering insights can be drawn from the capacity results. Consider an uplink channel with K mobiles, each with a single transmit antenna. There
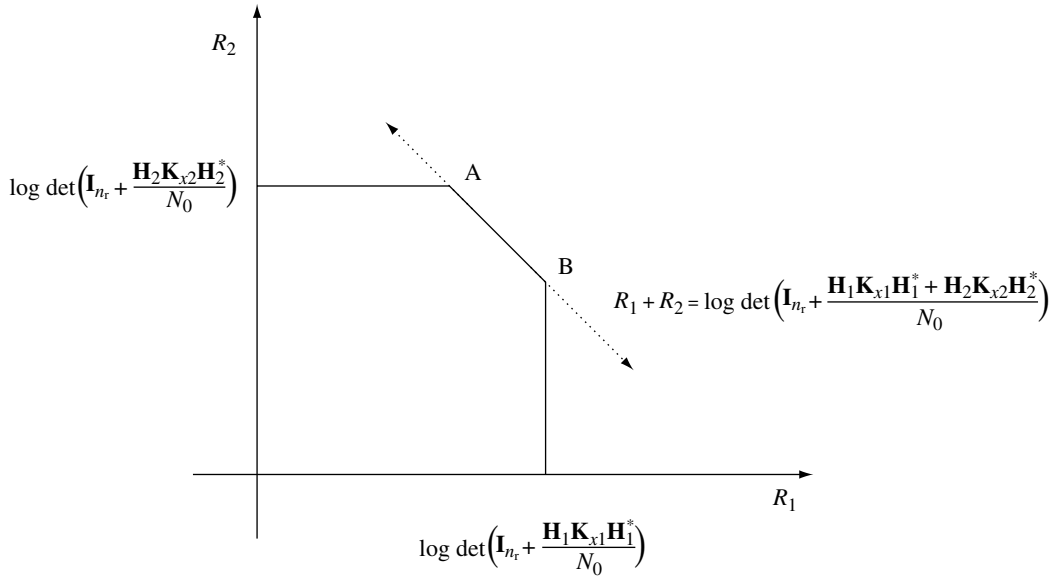
Figure 10.13 The rate region of the two-user MIMO uplink with transmitter strategies (power allocations to the data streams and the choice of rotation before sending over the transmit antenna array) given by the covariance matrices $\mathbf{K}_{x1}$ and $\mathbf{K}_{x2}$.

are n_r receive antennas at the base-station. Suppose the system designer wants to add one more transmit antenna at each mobile. How does this translate to increasing the number of spatial degrees of freedom?

If we look at each user in isolation and think of the uplink channel as a set of isolated SIMO point-to-point links from each user to the base-station, then adding one extra antenna at the mobile increases by one the available spatial degrees of freedom in each such link. However, this is misleading. Due to the sum rate constraint, the *total* number of spatial degrees of freedom is limited by the minimum of K and n_r. Hence, if K is larger than n_r, then the number of spatial degrees of freedom is already limited by the number of receive antennas at the base-station, and increasing the number of transmit antennas at the mobiles will not increase the total number of spatial degrees of freedom further. This example points out the importance of looking at
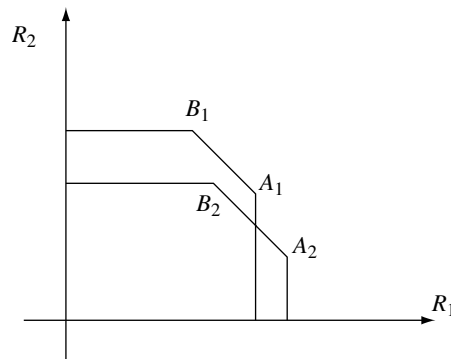

Figure 10.14 The achievable rate region for the two-user MIMO MAC with two specific choices of transmit filter covariances: $\mathbf{K}_{xk}$ for user k, for $k = 1, 2$.

the uplink channel as a whole rather than as a set of isolated point-to-point links.

On the other hand, multiple transmit antennas at each of the users significantly benefit the performance of orthogonal multiple access (which, however, is suboptimal to start with when $n_r > 1$). With a single transmit antenna, the total number of spatial degrees of freedom with orthogonal multiple access is just one. Increasing the number of transmit antennas at the users boosts the number of spatial degrees of freedom; user k has $\min(n_{tk}, n_r)$ spatial degrees of freedom when it is transmitting.

10.2.3 Fast fading

Our channel model is an extension of (10.22):

$$\mathbf{y}[m] = \sum_{k=1}^{K} \mathbf{H}_k[m]\mathbf{x}_k[m] + \mathbf{w}[m]. \qquad (10.25)$$

The channel variations $\{\mathbf{H}_k[m]\}_m$ are independent across users k and stationary and ergodic in time m.

Receiver CSI

In the receiver CSI model, the users only have access to the statistical characterization of the channels while the base-station tracks all the users' channel realizations. The users can still follow the SDMA transmitter architecture in Figure 10.11: splitting the data into independent data streams, splitting the total power across the streams and then sending the rotated version of the data streams over the transmit antenna array. However, the power allocations and the choice of rotation can only depend on the channel statistics and not on the explicit realization of the channels at any time m.

In our discussion of the point-to-point MIMO channel with receiver CSI in Section 8.2.1, we have seen some additional structure to the transmit signal. With linear antenna arrays and sufficiently rich scattering so that the channel elements can be modelled as zero mean uncorrelated entries, the capacity achieving transmit signal sends independent data streams over the different *angular windows*; i.e., the covariance matrix is of the form (cf. (8.11)):

$$\mathbf{K}_x = \mathbf{U}_t \mathbf{\Lambda} \mathbf{U}_t^*, \qquad (10.26)$$

where $\mathbf{\Lambda}$ is a diagonal matrix with non-negative entries (representing the power transmitted in each of the transmit angular windows). The rotation matrix $\mathbf{U}_t$ represents the transformation of the signal sent over the angular windows to the actual signal sent out of the linear antenna array (cf. (7.68)).

A similar result holds in the uplink MIMO channel as well. When each of the users' MIMO channels (viewed in the angular domain) have zero mean, uncorrelated entries then it suffices to consider covariance matrices of the form in (10.26); i.e., user k has the transmit covariance matrix:

$$\mathbf{K}_{xk} = \mathbf{U}_{tk} \Lambda_k \mathbf{U}_{tk}^*, \tag{10.27}$$

where the diagonal entries of Λ_k represent the powers allocated to the data streams, one in each of the angular windows (so their sum is equal to P_k, the power constraint for user k). (See Exercise 10.13.) With this choice of transmit strategy, the pair of rates (R_1, R_2) at which users can jointly reliably communicate is constrained, as in (10.12) and (10.13), by

$$R_k \leq \mathbb{E}\left[\log \det\left(\mathbf{I}_{n_r} + \frac{1}{N_0} \mathbf{H}_k \mathbf{K}_{xk} \mathbf{H}_k^*\right)\right], \qquad k = 1, 2, \tag{10.28}$$

$$R_1 + R_2 \leq \mathbb{E}\left[\log \det\left(\mathbf{I}_{n_r} + \frac{1}{N_0} \sum_{k=1}^{2} \mathbf{H}_k \mathbf{K}_{xk} \mathbf{H}_k^*\right)\right]. \tag{10.29}$$

This constraint forms a pentagon and the corner points are achieved by the architecture of the linear MMSE filter combined with successive cancellation of data streams (cf. Figure 10.12).

The capacity region is the convex hull of the union of these pentagons, one for each power allocation to the data streams of the users (i.e., the diagonal entries of Λ_1, Λ_2). In the point-to-point MIMO channel, with some additional symmetry (such as in the i.i.d. Rayleigh fading model), we have seen that the capacity achieving power allocation is equal powers to the data streams (cf. (8.12)). An analogous result holds in the MIMO uplink as well. With i.i.d. Rayleigh fading for all the users, the equal power allocation to the data streams, i.e.,

$$\mathbf{K}_{xk} = \frac{P_k}{n_{tk}} \mathbf{I}_{n_{tk}}, \tag{10.30}$$

achieves the entire capacity region; thus in this case the capacity region is simply a pentagon. (See Exercise 10.14.)

The analysis of the capacity region with full CSI is very similar to our previous analysis (cf. Section 10.1.5). Due to the increase in number of parameters to feedback (so that the users can change their transmit strategies as a function of the time-varying channels), this scenario is also somewhat less relevant in engineering practice, at least for FDD systems.

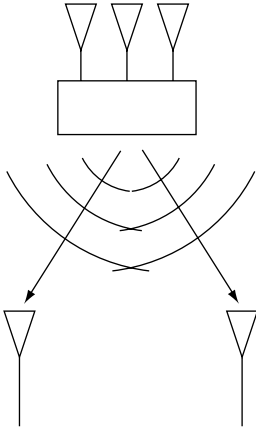## 10.3 Downlink with multiple transmit antennas

Figure 10.15 The downlink with multiple transmit antennas at the base-station and single receive antenna at each user.

We now turn to the downlink channel, from the base-station to the multiple users. This time the base-station has an array of transmit antennas but each user has a single receive antenna (Figure 10.15). It is often a practically interesting situation since it is easier to put multiple antennas at the base-station than at the mobile users. As in the uplink case we first consider the time-invariant scenario where the channel is fixed. The baseband model of the narrowband downlink with the base-station having n_t antennas and K users with each user having a single receive antenna is

$$y_k[m] = \mathbf{h}_k^* \mathbf{x}[m] + w_k[m], \qquad k = 1, \ldots, K, \tag{10.31}$$

where $\mathbf{y}_k[m]$ is the received vector for user k at time m, $\mathbf{h}_k^*$ is an n_t dimensional row vector representing the channel from the base-station to user k. Geometrically, user k observes the projection of the transmit signal in the spatial direction $\mathbf{h}_k$ in additive Gaussian noise. The noise $w_k[m] \sim \mathcal{CN}(0, N_0)$ and is i.i.d. in time m. An important assumption we are implicitly making here is that the channel's $\mathbf{h}_k$ are known to the base-station as well as to the users.

10.3.1 Degrees of freedom in the downlink

If the users could cooperate, then the resulting MIMO point-to- point channel would have $\min(n_t, K)$ spatial degrees of freedom, assuming that the rank of the matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$ is full. Can we attain this full spatial degrees of freedom even when users cannot cooperate?

Let us look at a special case. Suppose $\mathbf{h}_1, \ldots, \mathbf{h}_K$ are orthogonal (which is only possible if $K \leq n_t$). In this case, we can transmit independent streams of data to each user, such that the stream for the kth user $\{\tilde{x}_k[m]\}$ is along the transmit spatial signature $\mathbf{h}_k$, i.e.,

$$\mathbf{x}[m] = \sum_{k=1}^{K} \tilde{x}_k[m]\mathbf{h}_k. \tag{10.32}$$

The overall channel decomposes into a set of parallel channels; user k receives

$$y_k[m] = \|\mathbf{h}_k\|^2 \tilde{x}_k[m] + w_k[m]. \tag{10.33}$$

Hence, one can transmit K parallel non-interfering streams of data to the users, and attain the full number of spatial degrees of freedom in the channel.

What happens in general, when the channels of the users are not orthogonal? Observe that to obtain non-interfering channels for the users in the example above, the key property of the transmit signature $\mathbf{h}_k$ is that $\mathbf{h}_k$ is orthogonal

to the spatial direction's $\mathbf{h}_i$ of all the other users. For general channels (but still assuming linear independence among $\mathbf{h}_1, \ldots, \mathbf{h}_K$; thus $K \leq n_t$), we can preserve the same property by replacing the signature $\mathbf{h}_k$ by a vector $\mathbf{u}_k$ that lies in the subspace V_k orthogonal to all the other $\mathbf{h}_i$; the resulting channel for user k is

$$y_k[m] = (\mathbf{h}_k^*\mathbf{u}_k)\tilde{x}_k[m] + w_k[m]. \qquad (10.34)$$

Thus, in the general case too, we can get K spatial degrees of freedom. We can further choose $\mathbf{u}_k \in V_k$ to maximize the SNR of the channel above; geometrically, this is given by the projection of $\mathbf{h}_k$ onto the subspace V_k. This transmit filter is precisely the decorrelating receive filter used in the uplink and also in the point-to-point setting. (See Section 8.3.1 for the geometric derivation of the decorrelator.)

The above discussion is for the case when $K \leq n_t$. When $K \geq n_t$, one can apply the same scheme but transmitting only to n_t users at a time, achieving n_t spatial degrees of freedom. Thus, in all cases, we can achieve a total spatial degrees of freedom of $\min(n_t, K)$, the same as that of the point-to-point link when all the receivers can cooperate.

An important point to observe is that this performance is achieved assuming knowledge of the channels $\mathbf{h}_k$ at the base-station. We required the same channel side information at the base-station when we studied SDMA and showed that it achieves the same spatial degrees of freedom as when the users cooperate. In a TDD system, the base-station can exploit channel reciprocity and measure the uplink channel to infer the downlink channel. In an FDD system, the uplink and downlink channels are in general quite different, and feedback would be required: quite an onerous task especially when the users are highly mobile and the number of transmit antennas is large. Thus the requirement of channel state information at the base-station is quite asymmetric in the uplink and the downlink: it is more onerous in the downlink.

10.3.2 Uplink–downlink duality and transmit beamforming

In the *uplink*, we understand that the decorrelating receiver is the optimal linear filter at high SNR when the interference from other streams dominates over the additive noise. For general SNR, one should use the linear MMSE receiver to balance optimally between interference and noise suppression. This was also called *receive beamforming*. In the previous section, we found a downlink transmission strategy that is the analog of the decorrelating receive strategy. It is natural to look for a downlink transmission strategy analogous to the linear MMSE receiver. In other words, what is "optimal" transmit beamforming?

For a given set of powers, the *uplink* performance of the kth user is a function of only the receive filter $\mathbf{u}_k$. Thus, it is simple to formulate what

we mean by an "optimal" linear receiver: the one that maximizes the output SINR. The solution is the MMSE receiver. In the downlink, however, the SINR of each user is a function of *all* of the transmit signatures $\mathbf{u}_1, \ldots, \mathbf{u}_K$ of the users. Thus, the problem is seemingly more complex. However, there is in fact a downlink transmission strategy that is a natural "dual" to the MMSE receive strategy and is optimal in a certain sense. This is in fact a consequence of a more general duality between the uplink and the downlink, which we now explain.

Uplink–downlink duality

Suppose transmit signatures $\mathbf{u}_1, \ldots, \mathbf{u}_K$ are used for the K users. The transmitted signal at the antenna array is

$$\mathbf{x}[m] = \sum_{k=1}^{K} \tilde{x}_k[m]\mathbf{u}_k, \tag{10.35}$$

where $\{\tilde{x}_k[m]\}$ is the data stream of user k. Substituting into (10.31) and focusing on user k, we get

$$\mathbf{y}_k[m] = (\mathbf{h}_k^*\mathbf{u}_k)\tilde{x}_k[m] + \sum_{j \neq k}(\mathbf{h}_k^*\mathbf{u}_j)\tilde{x}_j[m] + w_k[m]. \tag{10.36}$$

The SINR for user k is given by

$$\text{SINR}_k := \frac{P_k \mid \mathbf{u}_k^*\mathbf{h}_k \mid^2}{N_0 + \sum_{j \neq k} P_j \mid \mathbf{u}_j^*\mathbf{h}_k \mid^2}. \tag{10.37}$$

where P_k is the power allocated to user k.

Denote $\mathbf{a} := (a_1, \ldots, a_K)^t$ where

$$a_k := \frac{\text{SINR}_k}{(1 + \text{SINR}_k) \mid \mathbf{h}_k^*\mathbf{u}_k \mid^2},$$

and we can rewrite (10.37) in matrix notation as

$$(\mathbf{I}_K - \text{diag}\{a_1, \ldots, a_K\}\mathbf{A})\mathbf{p} = N_0\mathbf{a}. \tag{10.38}$$

Here we denoted $\mathbf{p}$ to be the vector of transmitted powers $(P_1, \ldots, P_K)$. We also denoted the $K \times K$ matrix $\mathbf{A}$ to have component (k, j) equal to $\mid \mathbf{u}_j^*\mathbf{h}_k \mid^2$.

We now consider an uplink channel that is naturally "dual" to the given downlink channel. Rewrite the downlink channel (10.31) in matrix form:

$$\mathbf{y}_{\text{dl}}[m] = \mathbf{H}^*\mathbf{x}_{\text{dl}}[m] + \mathbf{w}_{\text{dl}}[m], \tag{10.39}$$

where $\mathbf{y}_{\text{dl}}[m] := (y_1[m], \ldots, y_K[m])^t$ is the vector of the received signals at the K users and $\mathbf{H} := [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K]$ is an n_t by K matrix. We added the
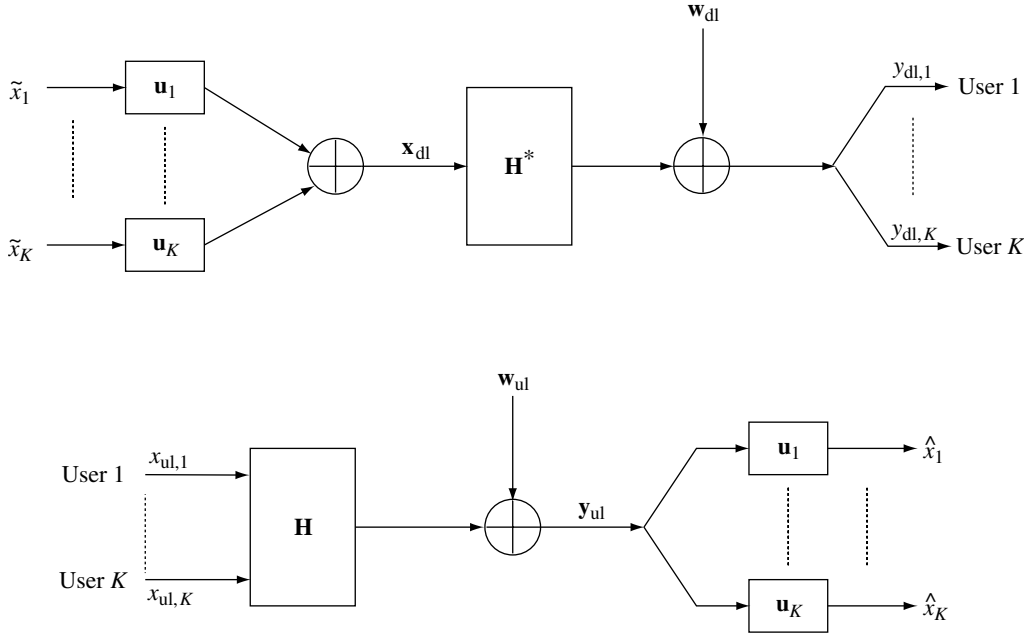
Figure 10.16 The original downlink with linear transmit strategy and its uplink dual with linear reception strategy.

subscript "dl" to emphasize that this is the downlink. The dual uplink channel has K users (each with a single transmit antenna) and n_t receive antennas:

$$\mathbf{y}_{ul}[m] = \mathbf{H}\mathbf{x}_{ul}[m] + \mathbf{w}_{ul}[m], \tag{10.40}$$

where $\mathbf{x}_{ul}[m]$ is the vector of transmitted signals from the K users, $\mathbf{y}_{ul}[m]$ is the vector of received signals at the n_t receive antennas, and $\mathbf{w}_{ul}[m] \sim \mathcal{CN}(0, N_0)$. To demodulate the kth user in this uplink channel, we use the receive filter $\mathbf{u}_k$, which is the transmit filter for user k in the downlink. The two dual systems are shown in Figure 10.16.

In this uplink, the SINR for user k is given by

$$\mathsf{SINR}_k^{ul} := \frac{Q_k \mid \mathbf{u}_k^* \mathbf{h}_k \mid^2}{N_0 + \sum_{j \neq k} Q_j \mid \mathbf{u}_k^* \mathbf{h}_j \mid^2}, \tag{10.41}$$

where Q_k is the transmit power of user k. Denoting $\mathbf{b} := (b_1, \ldots, b_K)^t$ where

$$b_k := \frac{\mathsf{SINR}_k^{ul}}{(1 + \mathsf{SINR}_k^{ul}) \mid \mathbf{u}_k^* \mathbf{h}_k \mid^2},$$

we can rewrite (10.41) in matrix notation as

$$(\mathbf{I}_K - \mathrm{diag}\{b_1, \ldots, b_K\} \mathbf{A}^t)\mathbf{q} = N_0 \mathbf{b}. \tag{10.42}$$

Here, $\mathbf{q}$ is the vector of transmit powers of the users and $\mathbf{A}$ is the same as in (10.38).

What is the relationship between the performance of the downlink transmission strategy and its dual uplink reception strategy? We claim that to achieve the same SINR for the users in both the links, the *total transmit power* is the same in the two systems. To see this, we first solve (10.38) and (10.42) for the transmit powers and we get

$$\mathbf{p} = N_0(\mathbf{I}_K - \mathrm{diag}\{a_1, \ldots, a_K\}\mathbf{A})^{-1}\mathbf{a} = N_0(D_a - \mathbf{A})^{-1}\mathbf{1}, \quad (10.43)$$

$$\mathbf{q} = N_0(\mathbf{I}_K - \mathrm{diag}\{b_1, \ldots, b_K\}\mathbf{A}')^{-1}\mathbf{b} = N_0(D_b - \mathbf{A}')^{-1}\mathbf{1}, \quad (10.44)$$

where $D_a := \mathrm{diag}(1/a_1, \ldots, 1/a_K)$, $D_b := \mathrm{diag}(1/b_1, \ldots, 1/b_K)$ and $\mathbf{1}$ is the vector of all 1's. To achieve the same SINR in the downlink and its dual uplink, $\mathbf{a} = \mathbf{b}$, and we conclude

$$\sum_{k=1}^{K} P_k = N_0\mathbf{1}^t(D_a - \mathbf{A})^{-1}\mathbf{1} = N_0\mathbf{1}^t\left[(D_a - \mathbf{A})^{-1}\right]^t\mathbf{1}$$

$$= N_0\mathbf{1}^t(D_a - \mathbf{A}')^{-1}\mathbf{1} = \sum_{k=1}^{K} Q_k. \quad (10.45)$$

It should be emphasized that the *individual* powers P_k and Q_k to achieve the same SINR are not the same in the downlink and the uplink dual; only the *total* power is the same.

Transmit beamforming and optimal power allocation

As observed earlier, the SINR of each user in the downlink depends in general on *all* the transmit signatures of the users. Hence, it is not meaningful to pose the problem of choosing the transmit signatures to maximize each of the SINR separately. A more sensible formulation is to minimize the total transmit power needed to meet a *given* set of SINR requirements. The optimal transmit signatures balance between focusing energy in the direction of the user of interest and minimizing the interference to other users. This transmit strategy can be thought of as performing *transmit beamforming*. Implicit in this problem formulation is also a problem of allocating powers to each of the users.

Armed with the uplink–downlink duality established above, the transmit beamforming problem can be solved by looking at the uplink dual. Since for any choice of transmit signatures, the same SINR can be met in the uplink dual using the transmit signatures as receive filters and the same total transmit power, the downlink problem is solved if we can find receive filters that minimize the total transmit power in the uplink dual. But this problem was already solved in Section 10.1.1. The receive filters are always chosen to be the MMSE filters given the transmit powers of the users; the transmit powers are iteratively updated so that the SINR requirement of each user is just met. (In fact, this algorithm not only minimizes the total

transmit power, it minimizes the transmit powers of every user simultaneously.) The MMSE filters at the optimal solution for the uplink dual can now be used as the optimal transmit signatures in the downlink, and the corresponding optimal power allocation **p** for the downlink can be obtained via (10.43).

It should be noted that the MMSE filters are the ones associated with the minimum powers used in the *uplink dual*, not the ones associated with the optimal transmit powers **p** in the *downlink*. At high SNR, each MMSE filter approaches a decorrelator, and since the decorrelator, unlike the MMSE filter, does not depend on the powers of the other interfering users, the same filter is used in the uplink and in the downlink. This is what we have already observed in Section 10.3.1.

Beyond linear strategies

In our discussion of receiver architectures for point-to-point communication in Section 8.3 and the uplink in Section 10.1.1, we boosted the performance of linear receivers by adding successive cancellation. Is there something analogous in the downlink as well?

In the case of the downlink with *single* transmit antenna at the base-station, we have already seen such a strategy in Section 6.2: superposition coding and decoding. If multiple users' signals are superimposed, the user with the strongest channel can decode the signals of the weaker users, strip them off and then decode its own. This is a natural analog to successive cancellation in the uplink. In the multiple transmit antenna case, however, there is no natural ordering of the users. In particular, if a linear superposition of signals is transmitted at the base-station:

$$\mathbf{x}[m] = \sum_{k=1}^{K} \tilde{x}_k[m]\mathbf{u}_k,$$

then each user's signal will be projected differently onto different users, and there is no guarantee that there is a single user who would have sufficient SINR to decode everyone else's data.

In both the uplink and the point-to-point MIMO channel, successive cancellation was possible because there was a single entity (the base-station) that had access to the entire vector of received signals. In the downlink we do not have that luxury since the users cannot cooperate. This was overcome in the special case of single transmit antenna because, from a decodability point of view, it is *as though* a given user has access to the received signals of all the users with weaker channels. In the general multiple transmit antenna case, this property does not hold and a "cancellation" scheme has to be necessarily *at the base-station*, which does indeed have access to the data of all the users. But how does one cancel a signal of a user even before it has been transmitted? We turn to this topic next.

10.3.3 Precoding for interference known at transmitter

Let us consider the precoding problem in a simple point-to-point context:

$$y[m] = x[m] + s[m] + w[m], \qquad (10.46)$$

where $x[m], y[m], w[m]$ are the real transmitted symbol, received symbol and $\mathcal{N}(0, \sigma^2)$ noise at time m respectively. The noise is i.i.d. in time. The interference sequence $\{s[m]\}$ is known in its entirety at the transmitter but not at the receiver. The transmitted signal $\{x[m]\}$ is subject to a power constraint. For simplicity, we have assumed all the signals to be real-valued for now. When applied to the downlink problem, $\{s[m]\}$ is the signal intended for another user, hence known at the transmitter (the base-station) but not necessary at the receiver of the user of interest. This problem also appears in many other scenarios. For example, in *data hiding* applications, $\{s[m]\}$ is the "host" signal in which one wants to hide digital information; typically the encoder has access to the host signal but not the decoder. The power constraint on $\{x[m]\}$ in this case reflects a constraint on how much the host signal can be distorted, and the problem here is to embed as much information as possible given this constraint.[4]

How can the transmitter precode the information onto the sequence $\{x[m]\}$ taking advantage of its knowledge of the interference? How much power penalty must be paid when compared to the case when the interference is also known at the receiver, or equivalently, when the interference does not exist? To get some intuition about the problem, let us first look at symbol-by-symbol precoding schemes.

Symbol-by-symbol precoding: Tomlinson–Harashima

For concreteness, suppose we would like to modulate information using uncoded $2M$-PAM: the constellation points are $\{a(1 + 2i)/2, i = -M, \ldots, M-1\}$, with a separation of a. We consider only symbol-by-symbol precoding in this subsection, and so to simplify notations below, we drop the index m. Suppose we want to send a symbol u in this constellation. The simplest way to compensate for the interference s is to transmit $x = u - s$ instead of u, so that the received signal is $y = u + w$.[5] However, the price to pay is an increase in the required energy by s^2. This power penalty grows unbounded with s^2. This is depicted in Figure 10.17.

The problem with the naive pre-cancellation scheme is that the PAM symbol may be arbitrarily far away from the interference. Consider the following

---

[4] A good application of data hiding is embedding digital information in analog television broadcast.

[5] This strategy will not work for the downlink channel at all because s contains the message of the other user and cancellation of s at the transmitter means that the other user will get nothing.
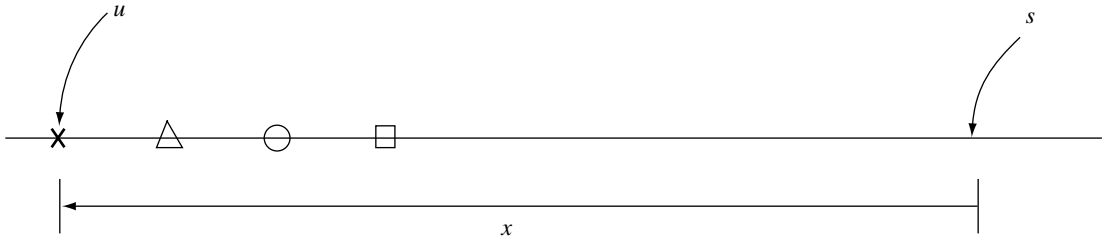
Figure 10.17 The transmitted signal is the difference between the PAM symbol and the interference. The larger the interference, the more the power that is consumed.

precoding scheme which performs better. The idea is to replicate the PAM constellation along the entire length of the real line to get an infinite extended constellation (Figures 10.18 and 10.19). Each of the $2M$ information symbols now corresponds to the equivalence class of points at the same relative position in the replicated constellations. Given the information symbol u, the precoding scheme chooses that representation p in its equivalence class which is closest to the interference s. We then transmit the difference $x = p - s$. Unlike the naive scheme, this difference can be much smaller and does not grow unbounded with s. A visual representation of the precoding scheme is provided in Figure 10.20.

One way to interpret the precoding operation is to think of the equivalence class of any one PAM symbol u as a (uniformly spaced) *quantizer* $q_u(\cdot)$ of the real line. In this context, we can think of the transmitted signal x to be the *quantization error*: the difference between the interference s and the quantized value $p = q_u(s)$, with u being the information symbol to be transmitted.

The received signal is

$$y = (q_u(s) - s) + s + w = q_u(s) + w.$$

The receiver finds the point in the infinite replicated constellation that is closest to s and then decodes to the equivalence class containing that point.

Let us look at the probability of error and the power consumption of this scheme, and how they compare to the corresponding performance when there is no interference. The probability of error is approximately[6]

$$2Q\left(\frac{a}{2\sigma}\right), \tag{10.47}$$

When there is no interference and a $2M$-PAM is used, the error probability of the interior points is the same as (10.47) but for the two exterior points, the error probability is $Q(a/2\sigma)$, smaller by a factor of $1/2$. The probability of error is larger for the exterior points in the precoding case because there is an

---

[6] The reason why this is not exact is because there is a chance that the noise will be so large that the closest point to y just happens to be in the same equivalence class of the information symbol, thus leading to a correct decision. However, the probability of this event is negligible.

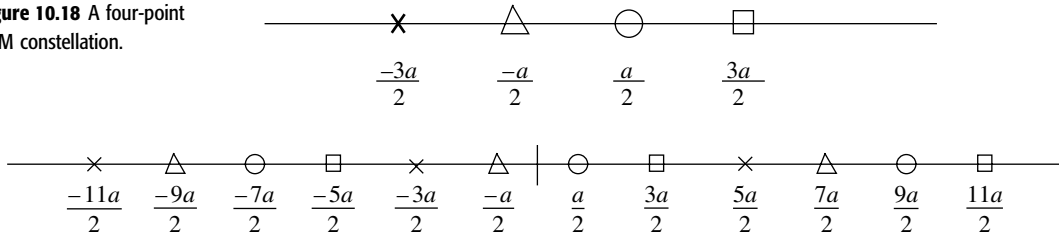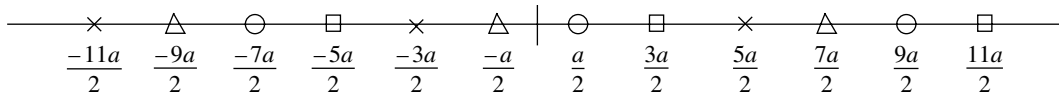**Figure 10.18** A four-point PAM constellation.

Figure 10.19 The four-point PAM constellation is replicated along the entire real line. Points marked by the same sign correspond to the same information symbol (one of the four points in the original constellation).

additional possibility of confusion *across* replicas. However, the difference is negligible when error probabilities are small.[7]

What about the power consumption of the precoding scheme? The distance between adjacent points in each equivalence class is $2Ma$; thus, unlike in the naive interference pre-cancellation scheme, the quantization error does not grow unbounded with s:

$$|x| \leq Ma.$$

If we assume that s is totally random so that this quantization error is uniform between zero and this value, then the average transmit power is

$$\mathbb{E}[x^2] = \frac{a^2 M^2}{3}. \tag{10.48}$$

In comparison, the average transmit power of the original $2M$-PAM constellation is $a^2 M^2/3 - a^2/12$. Hence, the precoding scheme requires a factor of

$$\frac{4M^2}{4M^2 - 1}$$

Figure 10.20 Depiction of the precoding operation for $M = 2$ and PAM information symbol $u = -3a/2$. The crosses form the equivalence class for this symbol. The difference between s and the closest cross p is transmitted.
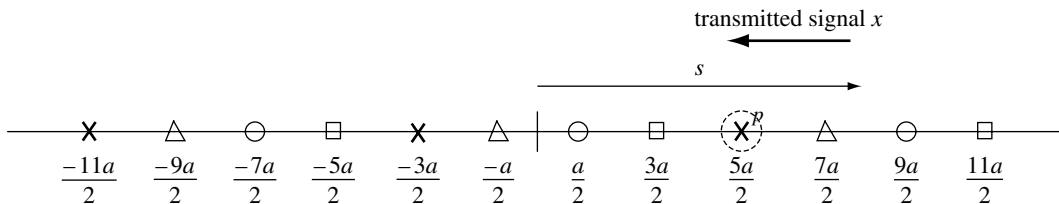
more transmit power. Thus, there is still a gap from AWGN detection performance. However, this power penalty is negligible when the constellation size M is large.

Our description is motivated from a similar precoding scheme for the point-to-point frequency-selective (ISI) channel, devised independently by

---

[7] This factor of 2 can easily be compensated for by making the symbol separation slightly larger.

Tomlinson [121] and Harashima and Miyakawa [57]. In this context, the interference is inter-symbol interference:

$$s[m] = \sum_{\ell \geq 0} h_\ell x[m - \ell],$$

where h is the impulse response of the channel. Since the previous transmitted symbols are known to the transmitter, the interference is known if the transmitter has knowledge of the channel. In Discussion 8.1 we have alluded to connections between MIMO and frequency-selective channels and precoding is yet another import from one knowledge base to the other. Indeed, Tomlinson–Harashima precoding was devised as an alternative to receiver-based decision-feedback equalization for the frequency-selective channel, the analog to the SIC receiver in MIMO and uplink channels. The precoding approach has the advantage of avoiding the error propagation problem of decision-feedback equalizers, since in the latter the cancellation is based on detected symbols, while the precoding is based on known symbols at the transmitter.

Dirty-paper precoding: achieving AWGN capacity

The precoding scheme in the last section is only for a single-dimensional constellation (such as PAM), while spectrally efficient communication requires coding over multiple dimensions. Moreover, in the low SNR regime, uncoded transmission yields very poor error probability performance and coding is necessary. There has been much work in devising block precoding schemes and it is still a very active research area. A detailed discussion of specific schemes is beyond the scope of this book. Here, we will build on the insights from symbol-by-symbol precoding to give a plausibility argument that *appropriate precoding can in fact completely obliviate the impact of the interference and achieve the capacity of the AWGN channel*. Thus, the power penalty we observed in the symbol-by-symbol precoding scheme can actually be avoided with high-dimensional coding. In the literature, the precoding technique presented here is also called *Costa precoding* or dirty-paper precoding.[8]

A first attempt

Consider communication over a block of length N symbols:

$$\mathbf{y} = \mathbf{x} + \mathbf{s} + \mathbf{w}. \tag{10.49}$$

In the symbol-by-symbol precoding scheme earlier, we started with a basic PAM constellation and replicated it to cover uniformly the entire (one-dimensional) range the interference s spans. For block coding, we would like

---

[8] This latter name comes from the title of Costa's paper: "Writing on dirty-paper" [23]. The writer of the message knows where the dirt is and can adapt his writing to help the reader decipher the message without knowing where the dirt is.
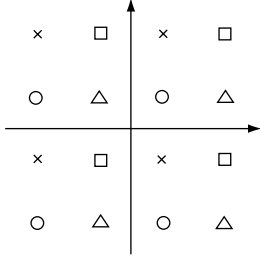
Figure 10.21 A replicated constellation in high dimension. The information specifies an equivalence class of points corresponding to replicas of a codeword (here with the same marking).

to mimic this strategy by starting with a basic AWGN constellation and replicating it to cover the N-dimensional space uniformly. Using a sphere-packing argument, we give an estimate of the maximum rate of reliable communication using this type of scheme.

Consider a domain of volume V in $\Re^N$. The exact size of the domain is not important, as long as we ensure that the domain is large enough for the received signal $\mathbf{y}$ to lie inside. This is the domain on which we replicate the basic codebook. We generate a codebook with M codewords, and replicate each of the codewords K times and place the extended constellation $\mathcal{C}_e$ of MK points on the domain sphere (Figure 10.21). Each codeword then corresponds to an equivalence class of points in $\Re^N$. Equivalently, the given information bits $\mathbf{u}$ define a quantizer $q_u(\cdot)$. The natural generalization of the symbol-by-symbol precoding procedure simply quantizes the known interference $\mathbf{s}$ using this quantizer to a point $\mathbf{p} = q_u(\mathbf{s})$ in $\mathcal{C}_e$ and transmits the quantization error

$$\mathbf{x}_1 = \mathbf{p} - \mathbf{s}. \tag{10.50}$$

Based on the received signal $\mathbf{y}$, the decoder finds the point in the extended constellation that is closest to $\mathbf{y}$ and decodes to the information bits corresponding to its equivalence class.

Performance

To estimate the maximum rate of reliable communication for a given average power constraint P using this scheme, we make two observations:

- **Sphere-packing** To avoid confusing $\mathbf{x}_1$ with any of the other $K(M-1)$ points in the extended constellation $\mathcal{C}_e$ that belong to other equivalence classes, the noise spheres of radius $\sqrt{N\sigma^2}$ around each of these points should be disjoint. This means that

$$KM < \frac{V}{\text{Vol}[B_N(\sqrt{N\sigma^2})]}, \tag{10.51}$$

the ratio of the volume of the domain sphere to that of the noise sphere.

- **Sphere-covering** To maintain the average transmit power constraint of P, the quantization error should be no more than $\sqrt{NP}$ for any interference vector $\mathbf{s}$. Thus, the spheres of radius $\sqrt{NP}$ around the K replicas of a codeword should be able to cover the whole domain such that any point is within a distance of $\sqrt{NP}$ from a replica. To ensure that,

$$K > \frac{V}{\text{Vol}[B_N(\sqrt{NP})]}. \tag{10.52}$$

This in effect imposes a constraint on the *minimal density* of the replication.

Putting the two constraints (10.51) and (10.52) together, we get

$$M < \frac{\text{Vol}[B_N(\sqrt{NP})]}{\text{Vol}[B_N(\sqrt{N\sigma^2})]} = \frac{\left(\sqrt{NP}\right)^N}{\left(\sqrt{N\sigma^2}\right)^N}, \tag{10.53}$$

which implies that the maximum rate of reliable communication is, at most,

$$R := \frac{\log M}{N} = \frac{1}{2}\log\frac{P}{\sigma^2}. \tag{10.54}$$

This yields an upper bound on the rate of reliable communication. Moreover, it can be shown that if the *MK* constellation points are independently and uniformly distributed on the domain, then with high probability, communication is reliable if condition (10.51) holds and the average power constraint is satisfied if condition (10.52) holds. Thus, the rate (10.54) is also achievable. The proof of this is along the lines of the argument in Appendix B.5.2, where the achievability of the AWGN capacity is shown.

Observe that the rate (10.54) is close to the AWGN capacity $1/2\log(1 + P/\sigma^2)$ at high SNR. However, the scheme is strictly suboptimal at finite SNR. In fact, it achieves zero rate if the SNR is below 0 dB. How can the performance of this scheme be improved?

### Performance enhancement via MMSE estimation

The performance of the above scheme is limited by the two constraints (10.51) and (10.52). To meet the average power constraint, the density of replication cannot be reduced beyond (10.52). On the other hand, constraint (10.51) is a direct consequence of the nearest neighbor decoding rule, and this rule is in fact suboptimal for the problem at hand. To see why, consider the case when the interference vector $\mathbf{s}$ is 0 and the noise variance σ^2 is significantly larger than P. In this case, the transmitted vector $\mathbf{x}_1$ is roughly at a distance $\sqrt{NP}$ from the origin while the received vector $\mathbf{y}$ is at a distance $\sqrt{N(P+\sigma^2)}$, much further away. Blindly decoding to the point in $\mathcal{C}_e$ nearest to $\mathbf{y}$ makes no use of the prior information that the transmitted vector $\mathbf{x}_1$ is of (relatively short) length $\sqrt{NP}$ (Figure 10.22). Without using this prior information, the transmitted vector is thought of by the receiver as anywhere in a large *uncertainty sphere* of radius $\sqrt{N\sigma^2}$ around $\mathbf{y}$ and the extended constellation points have to be spaced that far apart to avoid confusion. By making use of the prior information, the size of the uncertainty sphere can be reduced. In particular, we can consider a linear estimate $\alpha\mathbf{y}$ of $\mathbf{x}_1$. By the law of large numbers, the squared error in the estimate is

$$\|\alpha\mathbf{y} - \mathbf{x}_1\|^2 = \|\alpha\mathbf{w} + (\alpha - 1)\mathbf{x}_1\|^2 \approx N\left[\alpha^2\sigma^2 + (1-\alpha)^2 P\right] \tag{10.55}$$

and by choosing

$$\alpha = \frac{P}{P+\sigma^2}, \tag{10.56}$$

Figure 10.22 MMSE decoding yields a much smaller uncertainty sphere than does nearest neighbor decoding.
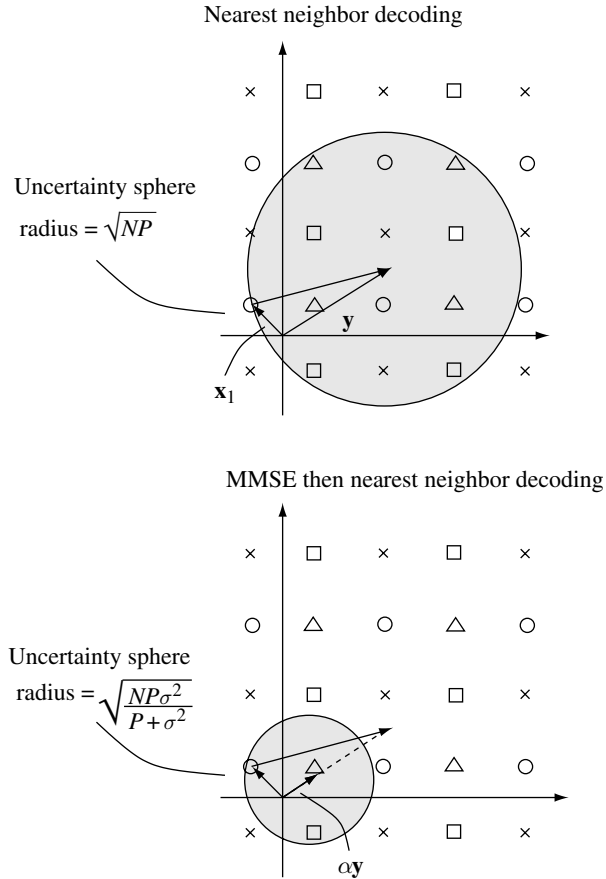
this error is minimized, equalling

$$\frac{NP\sigma^2}{P+\sigma^2}. \tag{10.57}$$

In fact $\alpha\mathbf{y}$ is nothing but the linear MMSE estimate $\hat{\mathbf{x}}_{\text{mmse}}$ of $\mathbf{x}_1$ from $\mathbf{y}$ and $NP\sigma^2/(P+\sigma^2)$ is the MMSE estimation error. If we now use a decoder that decodes to the constellation point nearest to $\alpha\mathbf{y}$ (as opposed to $\mathbf{y}$), then an error occurs only if there is another constellation point closer than this distance to $\alpha\mathbf{y}$. Thus, the uncertainty sphere is now of radius

$$\sqrt{\frac{NP\sigma^2}{P+\sigma^2}}. \tag{10.58}$$

We can now redo the analysis in the above subsection, but with the radius $\sqrt{N\sigma^2}$ of the noise sphere replaced by this radius of the MMSE uncertainty sphere. The maximum achievable rate is now

$$\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right), \tag{10.59}$$
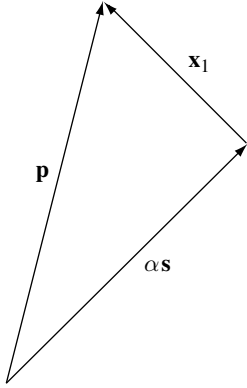
thus achieving the AWGN capacity.

Figure 10.23 The precoding process with the α factor.

In the above, we have simplified the problem by assuming $\mathbf{s} = 0$, to focus on how the decoder has to be modified. For a general interference vector $\mathbf{s}$,

$$\alpha \mathbf{y} = \alpha(\mathbf{x}_1 + \mathbf{s} + \mathbf{w}) = \alpha(\mathbf{x}_1 + \mathbf{w}) + \alpha \mathbf{s} = \hat{\mathbf{x}}_{\text{mmse}} + \alpha \mathbf{s}, \tag{10.60}$$

i.e., the linear MMSE estimate of $\mathbf{x}_1$ but shifted by $\alpha \mathbf{s}$. Since the receiver does not know $\mathbf{s}$, this shift has to be pre-compensated for at the transmitter. In the earlier scheme, we were using the nearest neighbor rule and we compensated for the effect of $\mathbf{s}$ by pre-subtracting $\mathbf{s}$ from the constellation point $\mathbf{p}$ representing the information, i.e., we sent the error in quantizing $\mathbf{s}$. But now we are using the MMSE rule and hence we should compensate by pre-subtracting $\alpha \mathbf{s}$ instead. Specifically, given the data $\mathbf{u}$, we find within the equivalence class representing $\mathbf{u}$ the point $\mathbf{p}$ that is closest to $\alpha \mathbf{s}$, and transmit $\mathbf{x}_1 = \mathbf{p} - \alpha \mathbf{s}$ (Figure 10.23). Then,

$$\mathbf{p} = \mathbf{x}_1 + \alpha \mathbf{s}$$
$$\alpha \mathbf{y} = \hat{\mathbf{x}}_{\text{mmse}} + \alpha \mathbf{s} = \hat{\mathbf{p}}$$

and

$$\mathbf{p} - \alpha \mathbf{y} = \mathbf{x}_1 - \hat{\mathbf{x}}_{\text{mmse}}. \tag{10.61}$$

The receiver finds the constellation point nearest to $\alpha \mathbf{y}$ and decodes the information (Figure 10.24). An error occurs only if there is another constellation point closer to $\alpha \mathbf{y}$ than $\mathbf{p}$, i.e., if it lies in the MMSE uncertainty sphere. This is exactly the same situation as in the case of zero interference.
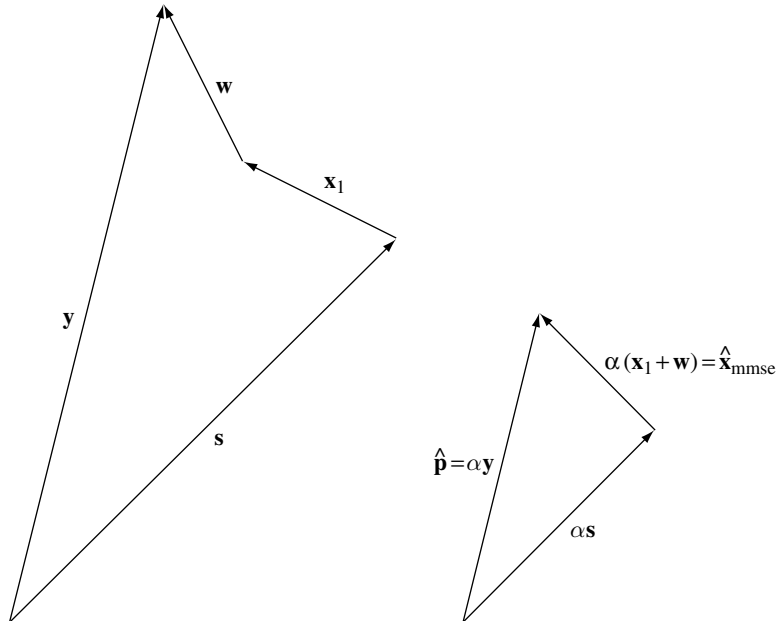

Figure 10.24 The decoding process with the α factor.

Transmitter knowledge of interference is enough

Something quite remarkable has been accomplished: even though the interference is known only at the transmitter and not at the receiver, the performance that can be achieved is as though there were no interference at all. The comparison between the cases with and without interference is depicted in Figure 10.25.

For the plain AWGN channel without interference, the codewords lie in a sphere of radius $\sqrt{NP}$ (x-sphere). When a codeword $\mathbf{x}_1$ is transmitted, the received vector $\mathbf{y}$ lies in the y-sphere, outside the x-sphere. The MMSE rule scales down $\mathbf{y}$ to $\alpha\mathbf{y}$, and the uncertainty sphere of radius $\sqrt{NP\sigma^2/(P+\sigma^2)}$ around $\alpha\mathbf{y}$ lies inside the x-sphere. The maximum reliable rate of communication is given by the number of uncertainty spheres that can be packed into the x-sphere:

$$\frac{1}{N} \log \frac{\mathrm{Vol}[B_N(\sqrt{NP})]}{\mathrm{Vol}[B_N(\sqrt{NP\sigma^2/(P+\sigma^2)})]} = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right), \qquad (10.62)$$

the capacity of the AWGN channel. In fact, this is how achievability of the AWGN capacity is shown in Appendix B.5.2.
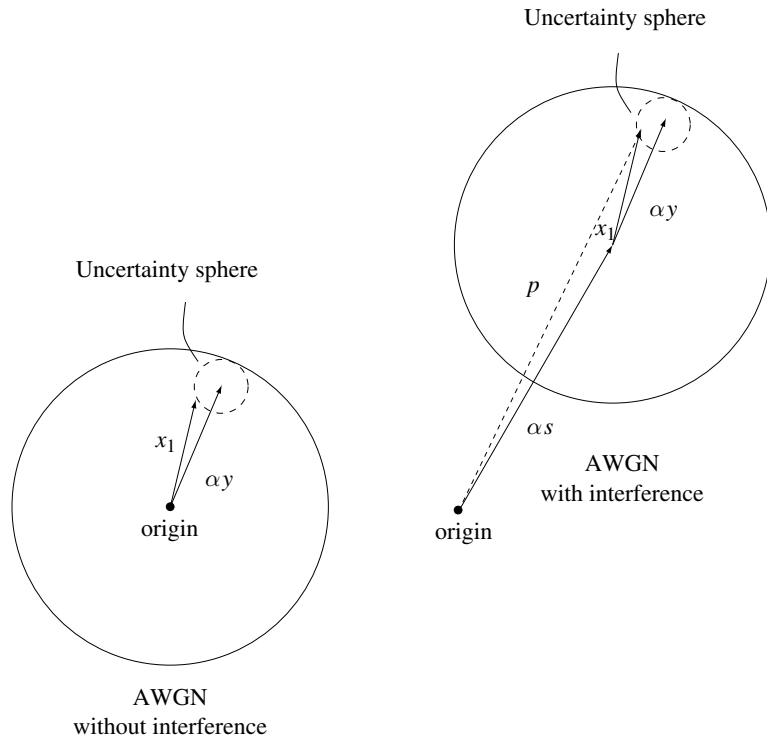

Figure 10.25 Pictorial representation of the cases with and without interference.

With interference, the codewords have to be replicated to cover the entire domain where the interference vector can lie. For any interference vector $\mathbf{s}$, consider a sphere of radius $\sqrt{NP}$ around $\alpha\mathbf{s}$; this can be thought of as the AWGN x-sphere whose center is shifted to $\alpha\mathbf{s}$. A constellation point $\mathbf{p}$ representing the given information bits lies inside this sphere. The vector $\mathbf{p} - \alpha\mathbf{s}$ is transmitted. By using the MMSE rule, the uncertainty sphere around $\alpha\mathbf{y}$ again lies inside this shifted x-sphere. Thus, we have the same situation as in the case without interference: the same information rate can be supported.

In the case without interference and where the codewords lie in a sphere of radius $\sqrt{NP}$, both the nearest neighbor rule and the MMSE rule achieve capacity. This is because although $\mathbf{y}$ lies outside the x-sphere, there are no codewords outside the x-sphere and the nearest neighbor rule will automatically find the codeword in the x-sphere closest to $\mathbf{y}$. However, in the precoding problem when there *are* constellation points lying outside the shifted x-sphere, the nearest neighbor rule will lead to confusion with these other points and is therefore strictly suboptimal.

Dirty-paper code design

We have given a plausibility argument of how the AWGN capacity can be achieved without knowledge of the interference at the receiver. It can be shown that randomly chosen codewords can achieve this performance. Construction of practical codes is the subject of current research. One such class of codes is called *nested lattice codes* (Figure 10.26). The design requirements of this nested lattice code are:

- Each sub-lattice should be a good vector quantizer for the scaled interference $\alpha\mathbf{s}$, to minimize the transmit power.
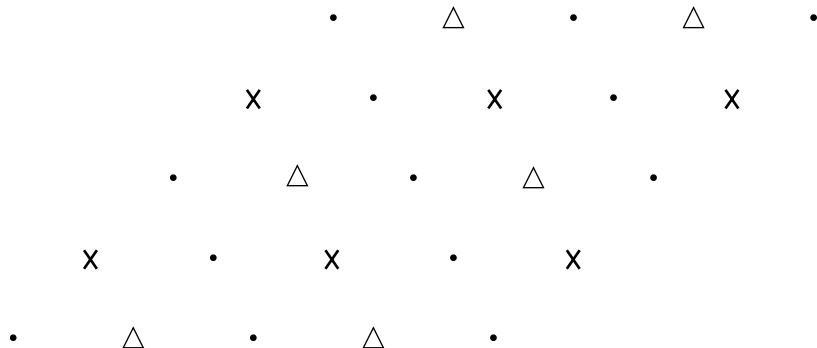- The entire extended constellation should behave as a good AWGN channel code.

Figure 10.26 A nested lattice code. All the points in each sub-lattice represent the same information bits.

The discussion of such codes is beyond the scope of this book. The design problem, however, simplifies in the low SNR regime. We discuss this below.

Low SNR: opportunistic orthogonal coding

In the infinite bandwidth channel, the SNR per degree of freedom is zero and we can use this as a concrete channel to study the nature of precoding at low SNR. Consider the infinite bandwidth real AWGN channel with additive interference $s(t)$ modelled as real white Gaussian (with power spectral density $N_s/2$) and known non-causally to the transmitter. The interference is independent of both the background real white Gaussian noise and the real transmit signal, which is power constrained, but not bandwidth constrained. Since the interference is known non-causally only to the transmitter, the minimum $\mathcal{E}_b/N_0$ for reliable communication on this channel can be no smaller than that in the plain AWGN channel without interference; thus a lower bound on the minimum $\mathcal{E}_b/N_0$ is -1.59 dB.

We have already seen for the AWGN channel (cf. Section 5.2.2 and Exercises 5.8 and 5.9) that orthogonal codes achieve the capacity in the infinite bandwidth regime. Equivalently, orthogonal codes achieve the minimum $\mathcal{E}_b/N_0$ of -1.59 dB over the AWGN channel. Hence, we start with an orthogonal set of codewords representing M messages. Each of the codewords is replicated K times so that the overall constellation with MK vectors forms an orthogonal set. Each of the M messages corresponds to a set of K orthogonal signals. To convey a specific message, the encoder transmits that signal, among the set of K orthogonal signals corresponding to the message selected, that is closest to the interference $s(t)$, i.e., the one that has the largest correlation with the $s(t)$. This signal is the constellation point to which $s(t)$ is quantized. Note that, in the general scheme, the signal $q_u(\alpha \mathbf{s}) - \alpha \mathbf{s}$ is transmitted, but since $\alpha \to 0$ in the low SNR regime, we are transmitting $q_u(\alpha \mathbf{s})$ itself.

An equivalent way of seeing this scheme is as *opportunistic pulse position modulation*: classical PPM involves a pulse that conveys information based on the position when it is not zero. Here, every K of the pulse positions corresponds to one message and the encoder opportunistically chooses the position of the pulse among the K possible pulse positions (once the desired message to be conveyed is picked) where the interference is the *largest*.

The decoder first picks the most likely position of the transmit pulse (among the MK possible choices) using the standard largest amplitude detector. Next, it picks the message corresponding to the set in which the most likely pulse occurs. Choosing K large allows the encoder to harness the opportunistic gains afforded by the knowledge of the additive interference. On the other hand, decoding gets harder as K increases since the number of possible pulse positions, MK, grows with K. An appropriate choice of K as a function of the number of messages, M, and the noise and interference powers, N_0 and N_s respectively, trades off the opportunistic gains on the one hand with

the increased difficulty in decoding on the other. This tradeoff is evaluated in Exercise 10.16 where we see that the correct choice of K allows the opportunistic orthogonal codes to achieve the infinite bandwidth capacity of the AWGN channel *without* interference. Equivalently, the minimum $\mathcal{E}_b/N_0$ is the *same* as that in the plain AWGN channel and is achieved by opportunistic orthogonal coding.

10.3.4 Precoding for the downlink

We now apply the precoding technique to the downlink channel. We first start with the single transmit antenna case and then discuss the multiple antenna case.

Single transmit antenna

Consider the two-user downlink channel with a single transmit antenna:

$$y_k[m] = h_k x[m] + w_k[m], \qquad k = 1, 2, \tag{10.63}$$

where $w_k[m] \sim \mathcal{CN}(0, N_0)$. Without loss of generality, let us assume that user 1 has the stronger channel: $|h_1|^2 \geq |h_2|^2$. Write $x[m] = x_1[m] + x_2[m]$, where $\{x_k[m]\}$ is the signal intended for user k, $k = 1, 2$. Let P_k be the power allocated to user k. We use a standard i.i.d. Gaussian codebook to encode information for user 2 in $\{x_2[m]\}$. Treating $\{x_2[m]\}$ as interference that is known at the transmitter, we can apply Costa precoding for user 1 to achieve a rate of

$$R_1 = \log\left(1 + \frac{|h_1|^2 P_1}{N_0}\right), \tag{10.64}$$

the capacity of an AWGN channel for user 1 with $\{x_2[m]\}$ completely absent. What about user 2? It can be shown that $\{x_1[m]\}$ can be made to appear like independent Gaussian noise to user 2. (See Exercise 10.17.) Hence, user 2 gets a reliable data rate of

$$R_2 = \log\left(1 + \frac{|h_2|^2 P_2}{|h_2|^2 P_1 + N_0}\right). \tag{10.65}$$

Since we have assumed that user 1 has the stronger channel, these same rates can in fact be achieved by superposition coding and decoding (cf. Section 6.2): we superimpose independent i.i.d. Gaussian codebook for user 1 and 2, with user 2 decoding the signal $\{x_2[m]\}$ treating $\{x_1[m]\}$ as Gaussian noise, and user 1 decoding the information for user 2, canceling it off, and then decoding the information intended for it. Thus, precoding is another approach to achieve rates on the boundary of the capacity region in the single antenna downlink channel.

Superposition coding is a *receiver-centric* scheme: the base-station simply adds the codewords of the users while the stronger user has to do the decoding job of both the users. In contrast, precoding puts a substantial computational burden on the base-station with receivers being regular nearest neighbor decoders (though the user whose signal is being precoded needs to decode the extended constellation, which has more points than the rate would entail). In this sense we can think of precoding as a *transmitter-centric* scheme.

However, there is something curious about this calculation. The precoding strategy described above encodes information for user 1 treating user 2's signal as known interference. But certainly we can reverse the role of user 1 and user 2, and encode information for user 2, treating user 1's signal as interference. This strategy achieves rates

$$R_1' = \log\left(1 + \frac{|h_1|^2 P_1}{|h_1|^2 P_2 + N_0}\right), \qquad R_2' = \log\left(1 + \frac{|h_2|^2 P_2}{N_0}\right). \qquad (10.66)$$

But these rates *cannot* be achieved by superposition coding/decoding under the power allocations P_1, P_2: the weak user cannot remove the signal intended for the strong user. Is this rate tuple then outside the capacity region? It turns out that there is no contradiction and this rate pair is strictly contained inside the capacity region (Exercise 10.19).

In this discussion, we have restricted ourselves to just two users, but the extension to K users is obvious. See Exercise 10.19.

Multiple transmit antennas

We now return to the scenario of real interest, multiple transmit antennas (10.31):

$$y_k[m] = \mathbf{h}_k^* \mathbf{x}[m] + w_k[m], \qquad k = 1, 2, \ldots, K. \qquad (10.67)$$

The precoding technique can be applied to upgrade the performance of the linear beamforming technique described in Section 10.3.2. Recall from (10.35), the transmitted signal is

$$\mathbf{x}[m] = \sum_{k=1}^{K} \tilde{x}_k[m] \mathbf{u}_k, \qquad (10.68)$$

where $\{\tilde{x}_k[m]\}$ is the signal for user k and $\mathbf{u}_k$ is its transmit beamforming vector. The received signal of user k is given by

$$\mathbf{y}_k[m] = (\mathbf{h}_k^* \mathbf{u}_k)\tilde{x}_k[m] + \sum_{j \neq k}(\mathbf{h}_k^* \mathbf{u}_j)\tilde{x}_j[m] + w_k[m], \qquad (10.69)$$

$$= (\mathbf{h}_k^* \mathbf{u}_k)\tilde{x}_k[m] + \sum_{j < k}(\mathbf{h}_k^* \mathbf{u}_j)\tilde{x}_j[m]$$

$$+ \sum_{j > k}(\mathbf{h}_k^* \mathbf{u}_j)\tilde{x}_j[m] + w_k[m]. \qquad (10.70)$$

Applying Costa precoding for user k, treating the interference $\sum_{j<k}(\mathbf{h}_k^*\mathbf{u}_j)\tilde{x}_j[m]$ from users $1, \ldots, k-1$ as known and $\sum_{j>k}(\mathbf{h}_k^*\mathbf{u}_j)\tilde{x}_j[m]$ from users $k+1, \ldots K$ as Gaussian noise, the rate that user k gets is

$$R_k = \log(1 + \mathsf{SINR}_k), \tag{10.71}$$

where SINR_k is the effective signal-to-interference-plus-noise ratio after precoding:

$$\mathsf{SINR}_k = \frac{P_k \mid \mathbf{u}_k^*\mathbf{h}_k \mid^2}{N_0 + \sum_{j>k} P_j \mid \mathbf{u}_j^*\mathbf{h}_k \mid^2}. \tag{10.72}$$

Here P_j is the power allocated to user j. Observe that unlike the single transmit antenna case, this performance may not be achievable by superposition coding/decoding.

For linear beamforming strategies, an interesting uplink–downlink duality is identified in Section 10.3.2. We can use the downlink transmit signatures (denoted by $\mathbf{u}_1, \ldots, \mathbf{u}_K$) to be the same as the receive filters in the dual uplink channel (10.40) and the same SINR for the users can be achieved in both the uplink and the downlink with appropriate user power allocations such that the sum of these power allocations is the same for both the uplink and the downlink. We now extend this observation to a duality between transmit beamforming with precoding in the downlink and receive beamforming with SIC in the uplink.

Specifically, suppose we use Costa precoding in the downlink and SIC in the uplink, and the transmit signatures of the users in the downlink are the same as the receive filters of the users in the uplink. Then it turns out that the same set SINR of the users can be achieved by appropriate user power allocations in the uplink and the downlink and, further, the sum of these power allocations is the same. This duality holds provided that the order of SIC in the uplink is the *reverse* of the Costa precoding order in the downlink. For example, in the Costa precoding above we employed the order $1, \ldots, K$; i.e., we precoded the user k signal so as to cancel the interference from the signals of users $1, \ldots, k-1$. For this duality to hold, we need to *reverse* this order in the SIC in the uplink; i.e., the users are successively canceled in the order $K, \ldots, 1$ (with user k seeing no interference from the canceled user signals $K, K-1, \ldots, k+1$).

The derivation of this duality follows the same lines as for linear strategies and is done in Exercise 10.20. Note that in this SIC ordering, user 1 sees the least uncanceled interference and user K sees the most. This is exactly the opposite to that under the Costa precoding strategy. Thus, we see that in this duality the ordering of the users is *reversed*. Identifying this duality facilitates the computation of good transmit filters in the downlink. For example, we know that in the uplink the optimal filters for a given set of powers are MMSE filters; the same filters can be used in the downlink transmission.

In Section 10.1.2, we saw that receive beamforming in conjunction with SIC achieves the capacity region of the uplink channel with multiple receive antennas. It has been shown that transmit beamforming in conjunction with Costa precoding achieves the capacity of the downlink channel with multiple transmit antennas.

10.3.5 Fast fading

The time-varying downlink channel is an extension of (10.31):

$$y_k[m] = \mathbf{h}_k^*[m]\mathbf{x}[m] + w_k[m], \qquad k = 1, \ldots, K. \tag{10.73}$$

Full CSI

With full CSI, both the base-station and the users track the channel fluctuations and, in this case, the extension of the linear beamforming strategies combined with Costa precoding to the fading channel is natural. Now we can vary the power and transmit signature allocations of the users, and the Costa precoding order as a function of the channel variations. Linear beamforming combined with Costa precoding achieves the capacity of the fast fading downlink channel with full CSI, just as in the time-invariant downlink channel.

It is interesting to compare this sum capacity achieving strategy with that when the base-station has just one transmit antenna (see Section 6.4.2). In this basic downlink channel, we identified the structure of the sum capacity achieving strategy: transmit only to the best user (using a power that is waterfilling over the best user's channel quality, see (6.54)). The linear beamforming strategy proposed here involves in general transmitting to all the users simultaneously and is quite different from the one user at a time policy. This difference is analogous to what we have seen in the uplink with single and multiple receive antennas at the base-station.

Due to the duality, we have a connection between the strategies for the downlink channel and its dual uplink channel. Thus, the impact of multiple transmit antennas at the base-station on multiuser diversity follows the discussion in the uplink context (see Section 10.1.6): focusing on the one user at a time policy, the multiple transmit antennas provide a beamforming power gain; this gain is the same as in the point-to-point context and the multiuser nature of the gain is lost. With the sum capacity achieving strategy, the multiple transmit antennas provide multiple spatial degrees of freedom allowing the users to be transmitted to simultaneously, but the opportunistic gains are of the same form as in the point-to-point case; the multiuser nature of the gain is diminished.

Receiver CSI

So far we have made the full CSI assumption. In practice, it is often very hard for the base-station to have access to the user channel fluctuations and

the receiver CSI model is more natural. The major difference here is that now the transmit signatures of the users cannot be allocated as a function of the channel variations. Furthermore, the base-station is not aware of the interference caused by the other users' signals for any specific user k (since the channel to the kth user is unknown) and Costa precoding is ruled out.

Exercise 10.21 discusses how to use the multiple antennas at the base-station without access to the channel fluctuations. One of the important conclusions is that time sharing among the users achieves the capacity region in the symmetric downlink channel with receiver CSI alone. This implies that the total spatial degrees of freedom in the downlink are restricted to one, the same as the degrees of freedom of the channel from the base-station to any individual user. On the other hand, with full CSI at the base-station we have seen (Section 10.3.1) that the spatial degrees of freedom are equal to $\min(n_t, K)$. Thus lack of CSI at the base-station causes a drastic reduction in the degrees of freedom of the channel.

Partial CSI at the base-station: opportunistic beamforming with multiple beams

In many practical systems, there is some form of partial CSI fed back to the base-station from the users. For example, in the IS-856 standard discussed in Chapter 6 each user feeds back the overall SINR of the link to the base-station it is communicating with. Thus, while the base-station does not have exact knowledge of the channel (phase and amplitude) from the transmit antenna array to the users, it does have partial information: the overall quality of the channel (such as $\|\mathbf{h}_k[m]\|^2$ for user k at time m).

In Section 6.7.3 we studied opportunistic beamforming that induces time fluctuations in the channel to increase the multiuser diversity. The multiple transmit antennas were used to induce time fluctuations and the partial CSI was used to schedule the users at appropriate time slots. However, the gain from multiuser diversity is a power gain (boost in the SINR of the user being scheduled) and with just a single user scheduled at any time slot, only one of the spatial degrees of freedom is being used. This basic scheme can be modified, however, allowing multiple users to be scheduled and thus increasing the utilized spatial degrees of freedom.

The conceptual idea is to have *multiple beams*, each orthogonal to one another, at the same time (Figure 10.27). Separate pilot symbols are introduced on each of the beams and the users feedback the SINR of *each* beam. Transmissions are scheduled to as many users as there are beams at each time slot. If there are enough users in the system, the user who is beamformed with respect to a specific beam (and orthogonal to the other beams) is scheduled on the specific beam. Let us consider $K \geq n_t$ (if $K < n_t$ then we use only K of the transmit antennas), and at each time m, let $\mathbf{Q}[m] = [\mathbf{q}_1[m], \ldots, \mathbf{q}_{n_t}[m]]$ be an $n_t \times n_t$ unitary matrix, with the columns $\mathbf{q}_1[m], \ldots, \mathbf{q}_{n_t}[m]$ orthonormal. The vector $\mathbf{q}_i[m]$ represents the ith beam at time m.
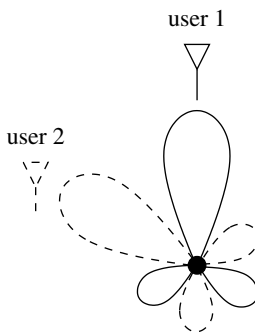

user 1

user 2

Figure 10.27 Opportunistic beamforming with two orthogonal beams. The user "closest" to a beam is scheduled on that beam, resulting in two parallel data streams to two users.

The vector signal sent out from the antenna array at time m is

$$\sum_{i=1}^{n_t} \tilde{x}_i[m]\mathbf{q}_i[m]. \tag{10.74}$$

Here $\tilde{x}_1, \ldots, \tilde{x}_{n_t}$ are the n_t independent data streams (in the case of coherent downlink reception, these signals include pilot symbols as well). The unitary matrix $\mathbf{Q}[m]$ is varied such that the individual components do not change abruptly in time. Focusing on the kth user, the signal it receives at time m is (substituting (10.74) in (10.73))

$$y_k[m] = \sum_{i=1}^{n_t} \tilde{x}_i[m]\mathbf{h}_k^*[m]\mathbf{q}_i[m] + w_k[m]. \tag{10.75}$$

For simplicity, let us consider the scenario when the channel coefficients are not varying over the time-scale of communication (slow fading), i.e., $\mathbf{h}_k[m] = \mathbf{h}_k$. When the ith beam takes on the value

$$\mathbf{q}_i[m] = \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}, \tag{10.76}$$

then user k is in beamforming configuration with respect to the ith beam; moreover, it is simultaneously orthogonal to the other beams. The received signal at user k is

$$y_k[m] = \|\mathbf{h}_k\|\tilde{x}_i[m] + w_k[m]. \tag{10.77}$$

If there are enough users in the system, for every beam i some user will be nearly in beamforming configuration with respect to it (and simultaneously nearly orthogonal to the other beams). Thus n_t data streams are transmitted simultaneously in orthogonal spatial directions and the full spatial degrees of freedom are utilized. The limited feedback from the users allows opportunistic scheduling of the user transmissions in the appropriate beams at the appropriate time slots. To achieve close to the beamforming performance and corresponding nulling to all the other beams requires a user population that is larger than in the scenario of Section 6.7.3. In general, depending on the number of the users in the system, the number of spatially orthogonal beams can be designed.

There are extra system requirements to support multiple beams (as compared to just the single time-varying beam introduced in Section 6.7.3). First, multiple pilot symbols have to be inserted (one for each beam) to enable coherent downlink reception; thus the fraction of pilot symbol power increases. Second, the receivers now track n_t separate beams and feedback SINR of each on each of the beams. On a practical note, the receivers could feedback only the *best* SINR and the identification of the beam that yields this SINR; this

restriction probably will not degrade the performance by much. Thus, with almost the same amount of feedback as the single beam scheme, the modified opportunistic beamforming scheme utilizes all the spatial degrees of freedom.
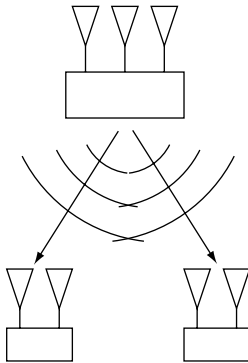
10.4 MIMO downlink

Figure 10.28 The downlink with multiple transmit antennas at the base-station and multiple receive antennas at each user.

We have seen so far how downlink is affected by the availability of multiple transmit antennas at the base-station. In this section, we study the downlink with multiple receive antennas (at the users) (see Figure 10.28). To focus on the role of multiple receive antennas, we begin with a single transmit antenna at the base-station.

The downlink channel with a single transmit and multiple receive antennas at each user can be written as

$$\mathbf{y}_k[m] = \mathbf{h}_k x[m] + \mathbf{w}_k[m], \qquad k = 1, 2, \tag{10.78}$$

where $\mathbf{w}_k[m] \sim \mathcal{CN}(0, N_0 I_{n_r})$ and i.i.d. in time m. The receive spatial signature at user k is denoted by $\mathbf{h}_k$. Let us focus on the time-invariant model first and fix this vector. If there is only one user, then we know from Section 7.2.1 that the user should do receive beamforming: project the received signal in the direction of the vector channel. Let us try this technique here, with both users matched filtering their received signals w.r.t. their channels. This is illustrated in Figure 10.29 and can be shown to be the optimal strategy for both the users (Exercise 10.22). With the matched filter front-end at each user, we have an effective AWGN downlink with a single antenna:

$$\tilde{y}_k[m] := \frac{\mathbf{h}_k^* \mathbf{y}_k[m]}{\|\mathbf{h}_k\|} = \|\mathbf{h}_k\| x[m] + w_k[m], \qquad k = 1, 2. \tag{10.79}$$

Here $w_k[m]$ is $\mathcal{CN}(0, N_0)$ and i.i.d. in time m and the downlink channel in (10.79) is very similar to the basic single antenna downlink channel model of (6.16) in Section 6.2. The only difference is that user k's channel quality $|h_k|^2$ is replaced by $\|\mathbf{h}_k\|^2$.

Thus, to study the downlink with multiple receive antennas, we can now carry over all our discussions from Section 6.2 for the single antenna scenario. In particular, we can order the two users based on their received SNR (suppose $\|\mathbf{h}_1\| \leq \|\mathbf{h}_2\|$) and do superposition coding: the transmit signal is the linear superposition of the signals to the two users. User 1 treats the signal of user 2 as noise and decodes its data from $\tilde{y}_1$. User 2, which has the better SNR, decodes the data of user 1, subtracts the transmit signal of user 1 from $\tilde{y}_2$ and then decodes its data. With a total power constraint of P and splitting this among the two users $P = P_1 + P_2$ we can write the

Figure 10.29 Each user with a
front-end matched filter
converting the SIMO downlink
into a SISO downlink.

Base station
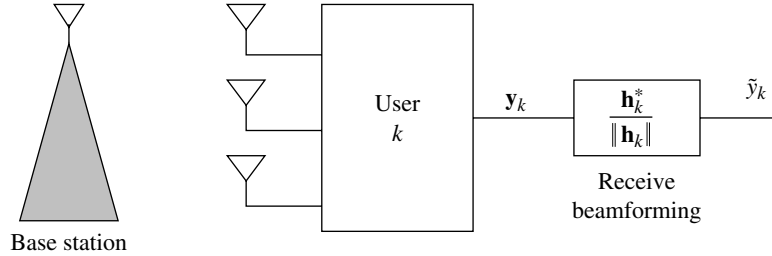
rate tuple that is achieved with the receiver architecture in Figure 10.29 and
superposition coding (cf. (6.22)),

$$R_1 = \log\left(1 + \frac{P_1\|\mathbf{h}_1\|^2}{P_2\|\mathbf{h}_1\|^2 + N_0}\right), \quad R_2 = \log\left(1 + \frac{P_2\|\mathbf{h}_2\|^2}{N_0}\right). \qquad (10.80)$$

Thus we have combined the techniques of Sections 7.2.1 and 6.2, namely
receive beamforming and superposition coding into a communication strategy
for the single transmit and multiple receive antenna downlink.

The matched filter operation by the users in Figure 10.29 only requires
tracking of their channels by the users, i.e., CSI is required at the receivers.
Thus, even with fast fading, the architecture in Figure 10.29 allows us to trans-
form the downlink with multiple receive antennas to the basic single antenna
downlink channel as long as the users have their channel state information.
In particular, analyzing receiver CSI and full CSI for the downlink in (10.78)
simplifies to the basic single antenna downlink discussion (in Section 6.4).

In particular, we can ask what impact multiple receive antennas have on
multiuser diversity, an important outcome of our discussion in Section 6.4. The
only difference here is the distribution of the channel quality: $\|\mathbf{h}_k\|^2$ replacing
$|h_k|^2$. This was also the same difference in the uplink when we studied the role
of multiple receive antennas in multiuser diversity gain (in Section 10.1.6).
We can carry over our main observation: the multiple receive antennas provide
a beamforming gain but the tail of $\|\mathbf{h}_k\|^2$ decays more rapidly (Figure 10.8)
and the multiuser diversity gain is restricted (Figure 10.9). To summarize,
the traditional receive beamforming power gain is balanced by the loss of the
benefit of the multiuser diversity gain (which is also a power gain) due to the
"hardening" of the effective fading distribution: $\|\mathbf{h}_k\|^2 \approx n_{\mathrm{r}}$ (cf. (10.20)).

With multiple transmit antennas at the base-station and multiple receive
antennas at each of the users, we can extend our set of linear strategies from
the discussion in Section 10.3.2: now the base-station splits the information
for user k into independent data streams, modulates them on different spatial
signatures and then transmits them. With full CSI, we can vary these spatial
signatures and powers allocated to the users (and the further allocation among
the data streams within a user) as a function of the channel fluctuations. We
can also embellish the linear strategies with Costa precoding, successively

precanceling the data streams. The performance of this scheme (linear beam-forming strategies with and without Costa precoding) can be related to the corresponding performance of a dual MIMO uplink channel (much as in the discussion of Section 10.3.2 with multiple antennas at the base-station alone). This scheme achieves the capacity of the MIMO downlink channel.

10.5 Multiple antennas in cellular networks: a system view

We have discussed the system design implications of multiple antennas in both the uplink and the downlink. These discussions have been in the context of multiple access *within a single cell* and are spread throughout the chapter (Sections 10.1.3, 10.1.6, 10.2.2, 10.3.5 and 10.4). In this section we take stock of these implications and consider the role of multiple antennas in cellular networks with multiple cells. Particular emphasis is on two points:

- the use of multiple antennas in suppressing inter-cell interference;
- how the use of multiple antennas within cells impacts the optimal amount of frequency reuse in the network.

Summary 10.3 System implications of multiple antennas on multiple access

Three ways of using multiple receive antennas in the uplink:
- **Orthogonal multiple access** Each user gets a power gain, but no change in degrees of freedom.
- **Opportunistic communication, one user at a time** Power gain but the multiuser diversity gain is reduced.
- **Space division multiple access** is capacity achieving: users simultaneously transmit and are jointly decoded at the base-station.

Comparison between orthogonal multiple access and SDMA
- Low SNR: performance of orthogonal multiple access comparable to that of SDMA.
- High SNR: SDMA allows up to n_r users to simultaneously transmit with a single degree of freedom each. Performance is significantly better than that with orthogonal multiple access.
- An intermediate access scheme with moderate complexity performs comparably to SDMA at all SNR levels: blocks of approximately n_r users in SDMA mode and orthogonal access for different blocks.

MIMO uplink
- Orthogonal multiple access: each user has multiple degrees of freedom.
- SDMA: the overall degrees of freedom are still restricted by the number of receive antennas.

> Downlink with multiple receive antennas
> Each user gets receive beamforming gain but reduced multiuser diversity gain.
> Downlink with multiple transmit antennas
> - No CSI at the base-station: single spatial degree of freedom.
> - Full CSI: the uplink–downlink duality principle makes this situation analogous to the uplink with multiple receive antennas and now there are up to n_t spatial degrees of freedom.
> - Partial CSI at the base-station: the same spatial degrees of freedom as the full CSI scenario can be achieved by a modification of the opportunistic beamforming scheme: multiple spatially orthogonal beams are sent out and multiple users are simultaneously scheduled on these beams.

10.5.1 Inter-cell interference management

Consider the multiple receive antenna uplink with users operating in SDMA mode. We have seen that successive cancellation is an optimal way to handle interference among the users within the same cell. However, this technique is not suitable to handle interference from neighboring cells: the out-of-cell transmissions are meant to be decoded by their nearest base-stations and the received signal quality is usually too poor to allow decoding at base-stations further away. On the other hand, linear receivers such as the MMSE do not decode the information from the interference and can be used to suppress out-of-cell interference.

The following model captures the essence of out-of-cell interference: the received signal at the antenna array ($\mathbf{y}$) comprises the signal (x) of the user of interest (with the signals of other users in the same cell successfully canceled) and the out-of-cell interference ($\mathbf{z}$):

$$\mathbf{y} = \mathbf{h}x + \mathbf{z}. \tag{10.81}$$

Here $\mathbf{h}$ is the received spatial signature of the user of interest. One model for the random interference $\mathbf{z}$ is as $\mathcal{CN}(0, \mathbf{K}_z)$, i.e., it is *colored* Gaussian noise with covariance matrix $\mathbf{K}_z$. For example, if the interference originates from just one out-of-cell transmission (with transmit power, say, q) and the base-station has an estimate of the received spatial signature of the interfering transmission (say, $\mathbf{g}$), then the covariance matrix is

$$q\mathbf{g}\mathbf{g}^* + N_0\mathbf{I}, \tag{10.82}$$

taking into account the structure of the interference and the background additive Gaussian noise.

Once such a model has been adopted, the multiple receive antennas can be used to suppress interference: we can use the linear MMSE receiver developed in Section 8.3.3 to get the soft estimate (cf. (8.61)):

$$\hat{x} = \mathbf{v}_{\text{mmse}}^{*}\mathbf{y} = \mathbf{h}^{*}\mathbf{K}_{z}^{-1}\mathbf{y}. \tag{10.83}$$

The expression for the corresponding SINR is in (8.62). This is the best SINR possible with a linear estimate. When the interfering noise is white, the operation is simply traditional receive beamforming. On the other hand, when the interference is very large and not white then the operation reduces to a decorrelator: this corresponds to nulling out the interference. The effect of channel estimation error on interference suppression is explored in Exercise 10.23.

In the uplink, the model for the interference depends on the type of multiple access. In many instances, a natural model for the interference is that it is white. For example, if the out-of-cell interference comes from many geographically spread out users (this situation occurs when there are many users in SDMA mode), then the overall interference is averaged over the multiple users' spatial locations and white noise is a natural model. In this case, the receive antenna array does not explicitly suppress out-of-cell interference. To be able to exploit the interference suppression capability of the antennas, two things must happen:

- The number of simultaneously transmitting users in each cell should be small. For example,in a hybrid SDMA/TDMA strategy, the total number of users in each cell may be large but the number of users simultaneously in SDMA mode is small (equal to or less than the number of receive antennas).
- The out-of-cell interference has to be trackable. In the SDMA/TDMA system, even though the interference at any time comes from a small number of users, the interference depends on the geographic location of the interfering user(s), which changes with the time slot. So either each slot has to be long enough to allow enough time to estimate the color of the interference based only on the pilot signal received in that time slot, or the users are scheduled in a periodic manner and the interference can be tracked across different time slots.

An example of such a system is described in Example 10.1.

On the other hand, interference suppression in the downlink using multiple receive antennas at the mobiles is different. Here the interference comes from a few base-stations of the neighboring cells that reuse the same frequency, i.e., from fixed specific geographic locations. Now, an estimate of the covariance of the interference can be formed and the linear MMSE can be used to manage the inter-cell interference.

We now turn to the role of multiple antennas in deciding the optimal amount of frequency reuse in the cellular network. We consider the effect

on both the uplink and the downlink and the role of multiple receive and multiple transmit antennas separately.

10.5.2 Uplink with multiple receive antennas

We begin with a discussion of the impact of multiple antennas at the base-station on the two orthogonal cellular systems studied in Chapter 4 and then move to SDMA.

Orthogonal multiple access

The array of multiple antennas is used to boost the received signal strength from the user within the cell via receive beamforming. One immediate benefit is that each user can lower its transmit power by a factor equal to the beamforming gain (proportional to n_r) to maintain the same signal quality at the base-station. This reduction in transmit power also helps to reduce inter-cell interference, so the effective SINR with the power reduction is in fact more than the SINR achieved in the original setting.

In Example 5.2 we considered a linear array of base-stations and analyzed the tradeoff between reuse and data rates per user for a given cell size and transmit power setting. With an array of antennas at each base-station, the SNR of every user improves by a factor equal to the receive beamforming gain. Much of the insight derived in Example 5.2 on how much to reuse can be naturally extended to the case here with the operating SNR boosted by the receive beamforming gain.

SDMA

If we do not impose the constraint that uplink communication be orthogonal among the users in the cell, we can use the SDMA strategy where many users simultaneously transmit and are jointly decoded at the base-station. We have seen that this scheme significantly betters orthogonal multiple access at high SNR due to the increased spatial degrees of freedom. At low SNR, both orthogonal multiple access and SDMA benefit comparably, with the users getting a receive beamforming gain. Thus, for SDMA to provide significant performance improvement over orthogonal multiple access, we need the operating SNR to be large; in the context of a cellular system, this means less frequency reuse.

Whether the loss in spectral efficiency due to less frequency reuse is fully compensated for by the increase in spatial degrees of freedom depends on the specific physical situation. The frequency reuse ratio ρ represents the loss in spectral efficiency. The corresponding reduction in interference is represented by the fraction f_ρ: this is the fraction of the received power from a user at the edge of the cell that the interference constitutes. For example, in a linear cellular system f_ρ decays roughly as ρ^α, but for a hexagonal cellular system the decay is much slower: f_ρ decays roughly as $\rho^{\alpha/2}$ (cf. Example 5.2).

Suppose all the K users are at the edge of the cell (a worst case scenario) and communicating via SDMA to the base-station with receiver CSI. W is the total bandwidth allotted to the cellular system scaled down by the number of simultaneous SDMA users sharing it within a cell (as with orthogonal multiple access, cf. Example 5.2). With SDMA used in each cell, K users simultaneously transmit over the entire bandwidth $K\rho W$.

The SINR of the user at the edge of the cell is, as in (5.20),

$$\mathsf{SINR} = \frac{\mathsf{SNR}}{\rho K + f_\rho \mathsf{SNR}}, \quad \text{with} \quad \mathsf{SNR} := \frac{P}{N_0 W d^\alpha}. \tag{10.84}$$

The SNR at the edge of the cell is SNR, a function of the transmit power P, the cell size d, and the power decay rate α (cf. (5.21)). The notation for the fraction f_ρ is carried over from Example 5.2. The largest symmetric rate each user gets is, the MIMO extension of (5.22),

$$R_\rho = \rho W \mathbb{E}[\log \det(\mathbf{I}_{n_r} + \mathsf{SINR}\, \mathbf{H}\mathbf{H}^*)]\ \text{bits/s}. \tag{10.85}$$

Here the columns of $\mathbf{H}$ represent the receive spatial signatures of the users at the base-station and the log det expression is the sum of the rates at which users can simultaneously communicate reliably.

We can now address the engineering question of how much to reuse using the simple formula for the rate in (10.85). At low SNR the situation is analogous to the single receive antenna scenario studied in Example 5.2: the rate is insensitive to the reuse factor and this can be verified directly from (10.85). On the other hand, at large SNR the interference grows as well and the SINR peaks at $1/f_\rho$. The largest rate then is, as in (5.23),

$$\rho W \mathbb{E}\left[\log \det\left(\mathbf{I}_{n_r} + \frac{1}{f_\rho} \mathbf{H}\mathbf{H}^*\right)\right]\ \text{bits/s}, \tag{10.86}$$

and goes to zero for small values of ρ: thus as in Example 5.2, less reuse does not lead to a favorable situation.

How do multiple receive antennas affect the optimal reuse ratio? Setting $K = n_r$ (a rule of thumb arrived at in Exercise 10.5), we can use the approximation in (8.29) to simplify the expression for the rate in (10.86):

$$R_\rho \approx \rho W n_r c^*\left(\frac{1}{f_\rho}\right). \tag{10.87}$$

The first observation we can make is that since the rate grows linearly in n_r, the optimal reuse ratio does not depend on the number of receive antennas. The optimal reuse ratio thus depends only on how the inter-cell interference f_ρ decays with the reuse parameter ρ, as in the single antenna situation studied in Example 5.2.

Figure 10.30 The symmetric
rate for every user (in bps/Hz)
with $K = 5$ users in SDMA
model in an uplink with $n_r = 5$
receive antennas plotted as a
function of the power decay
rate α for the linear cellular
system. The rates are plotted
for reuse ratios 1, 1/2 and 1/3.
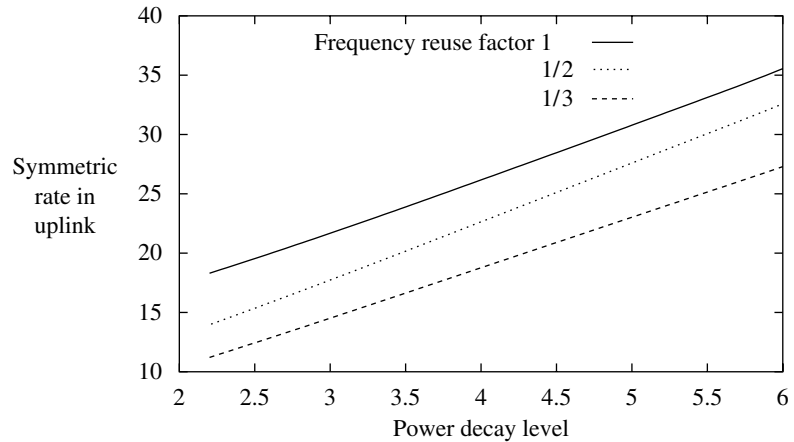
The rates at high SNR with reuse ratios 1, 1/2 and 1/4 are plotted in
Figure 10.30 for $n_r = K = 5$ in the linear cellular system. We observe the
optimality of universal reuse at all power decay rates: the gain in SINR from
less reuse is not worth the loss in spectral reuse. Comparing with the single
receive antenna example, the receive antennas provide a performance boost
(the rate increases linearly with n_r). We also observe that universal reuse is
now preferred. The hexagonal cellular system provides even less improvement
in SINR and thus universal reuse is optimal; this is unchanged from the single
receive antenna example.

10.5.3 MIMO uplink

An implementation of SDMA corresponds to altering the nature of medium
access. For example, there is no simple way of incorporating SDMA in any
of the three cellular systems introduced in Chapter 4 without altering the
fundamental way resource allocation is done among users. On the other hand,
the use of multiple antennas at the base-station to do receive beamforming
for each user of interest is a scheme based at the level of a point-to-point
communication link and can be implemented regardless of the nature of the
medium access. In some contexts where the medium access scheme cannot be
altered, a scheme based on improving the quality of individual point-to-point
links is preferred. However, an array of multiple antennas at the base-station
used to receive beamform provides only a power gain and not an increase in
degrees of freedom. If each user has multiple transmit antennas as well, then
an increase in the degrees of freedom of each individual point-to-point link
can be obtained.

In an orthogonal system, the point-to-point MIMO link provides each user
with multiple degrees of freedom and added diversity. With receiver CSI,
each user can use its transmit antenna array to harness the spatial degrees of

freedom when it is scheduled. The discussion of the role of frequency reuse earlier now carries over to this case. The nature of the tradeoff is similar: there is a loss in spectral degrees of freedom (due to less reuse) but an increase in the spatial degrees of freedom (due to the availability of multiple transmit antennas at the users).

10.5.4 Downlink with multiple receive antennas

In the downlink the interference comes from a few specific locations at fixed transmit powers: the neighboring base-stations that reuse the same frequency. Thus, the interference pattern can be empirically measured at each user and the array of receive antennas used to do linear MMSE (as discussed in Section 10.5.1) and boost the received SINR. For orthogonal systems, the impact on frequency reuse analysis is similar to that in the uplink with the SINR from the MMSE receiver replacing the earlier simpler expression (as in (5.20), for the uplink example).

If the base-station has multiple transmit antennas as well, the interference could be harder to suppress: in the presence of substantial scattering, each of the base-station transmit antennas could have a distinct receive spatial signature at the mobile, and in this case an appropriate model for the interference is white noise. On the other hand, if the scattering is only local (at the base-station and at the mobile) then all the base-station antennas have the same receive spatial signature (cf. Section 7.2.3) and interference suppression via the MMSE receiver is still possible.

10.5.5 Downlink with multiple transmit antennas

With full CSI (i.e., both at the base-station and at the users), the uplink–downlink duality principle (see Section 10.3.2) allows a comparison to the reciprocal uplink with the multiple *receive* antennas and receiver CSI. In particular, there is a one-to-one relationship between linear schemes (with and without successive cancellation) for the uplink and that for the downlink. Thus, many of our inferences in the uplink with multiple receive antennas hold in the downlink as well. However, full CSI may not be so practical in an FDD system: having CSI at the base-station in the downlink requires substantial CSI feedback via the uplink.

Example 10.1 SDMA in ArrayComm systems
ArrayComm Inc. is one of the early companies implementing SDMA technology. Their products include an SDMA overlay on Japan's PHS cellular system, a fixed wireless local loop system, and a mobile cellular system (iBurst).

An ArrayComm SDMA system exemplifies many of the design features that multiple antennas at the base-station allow. It is TDMA based and is much like the narrowband system we studied in Chapter 4. The main difference is that within each narrowband channel in each time slot, a small number of users are in SDMA mode (as opposed to just a single user in the basic narrowband system of Section 4.2). The array of antennas at the base-station is also used to suppress out-of-cell interference, thus allowing denser frequency reuse than a basic narrowband system. To enable successful SDMA operation and interference suppression in both the uplink and the downlink, the ArrayComm system has several key design features.

- The time slots for TDMA are synchronized across different cells. Further, the time slots are long enough to allow accurate estimation of the interference using the training sequence. The estimate of the color of the interference is then in the same time slot to suppress out-of-cell interference. Channel state information is not kept across slots.

- The small number of SDMA users within each narrowband channel are demodulated using appropriate linear filters: for each user, this operation suppresses both the out-of-cell interference and the in-cell interference from the other users in SDMA mode sharing the same narrowband channel.

- The uplink and the downlink operate in TDD mode with the downlink transmission immediately *following* the uplink transmission and to the *same* set of users. The uplink transmission provides the base-station CSI that is used in the immediately following downlink transmission to perform SDMA and to suppress out-of-cell interference via transmit beamforming and nulling. TDD operation avoids the expensive channel state feedback required for downlink SDMA in FDD systems.

To get a feel for the performance improvement with SDMA over the basic narrowband system, we can consider a specific implementation of the ArrayComm system. There are up to twelve antennas per sector at the base-station with up to four users in SDMA mode over each narrowband channel. This is an improvement of roughly a factor of four over the basic narrowband system, which schedules only a single user over each narrowband channel. Since there are about three antennas per user, substantial out-of-cell interference suppression is possible. This allows us to increase the frequency reuse ratio; this is a further benefit over the basic narrowband system. For example, the SDMA overlay on the PHS system increases the frequency reuse ratio of 1/8 to 1.

In the Flash OFDM example in Chapter 4, we have mentioned that one advantage of orthogonal multiple access systems over CDMA systems is that users can get access to the system without the need to slowly ramp up

the power. The interference suppression capability of adaptive antennas provides another way to allow users who are not power controlled to get access to the system quickly without swamping the existing active users. Even in a near–far situation of 40–50 dB, SDMA still works successfully; this means that potentially many users can be kept in the hold state when there are no active transmissions.

These improvements come at an increased cost to certain system design features. For example, while downlink transmissions meant for specific users enjoy a power gain via transmit beamforming, the pilot signal is intended for all users and has to be isotropic, thus requiring a proportionally larger amount of power. This reduces the traditional amortization benefit of the downlink pilot. Another aspect is the forced symmetry between the uplink and the downlink transmissions. To successfully use the uplink measurements (of the channels of the users in SDMA mode and the color of the out-of-cell interference) in the following downlink transmission, the transmission power levels in the uplink and the downlink have to be comparable (see Exercise 10.24). This puts a strong constraint on the system designer since the mobiles operate on batteries and are typically much more power constrained than the base-station, which is powered by an AC supply. Further, the pairing of the uplink or downlink transmissions is ideal when the flow of traffic is symmetric in both directions; this is usually true in the case of voice traffic. On the other hand, data traffic can be asymmetric and leads to wasted uplink (downlink) transmissions if only downlink (uplink) transmissions are desired.

Chapter 10 The main plot

Uplink with multiple receive antennas
Space division multiple access (SDMA) is capacity-achieving: all users simultaneously transmit and are jointly decoded by the base-station.
- Total spatial degrees of freedom limited by number of users and number of receive antennas.
- Rule of thumb is to have a group of n_r users in SDMA mode and different groups in orthogonal access mode.
- *Each* of the n_r user transmissions in a group obtains the full receive diversity gain equal to n_r.

Uplink with multiple transmit and receive antennas
The overall spatial degrees of freedom are still restricted by the number of receive antennas, but the diversity gain is enhanced.

Downlink with multiple transmit antennas
Uplink–downlink duality identifies a correspondence between the downlink and the *reciprocal* uplink.

Precoding is the analogous operation to successive cancelation in the uplink. A precoding scheme that perfectly cancels the intra-cell interference caused to a user was described.

Precoding operation requires full CSI; hard to justify in an FDD system. With only partial CSI at the base-station, an opportunistic beamforming scheme with multiple orthogonal beams utilizes the full spatial degrees of freedom.

Downlink with multiple receive antennas
Each user's link is enhanced by receive beamforming: both a power gain and a diversity gain equal to the number of receive antennas are obtained.

10.6 Bibliographical notes

The precoding technique for communicating on a channel where the transmitter is aware of the channel was first studied in the context of the ISI channel by Tomlinson [121] and Harashima and Miyakawa [57]. More sophisticated precoders for the ISI channel (designed for use in telephone modems) were developed by Eyuboglu and Forney [36] and Laroia *et al.* [71]. A survey on precoding and shaping for ISI channels is contained in an article by Forney and Ungerböck [39].

Information theoretic study of a state-dependent channel where the transmitter has non-causal knowledge of the state was studied, and the capacity characterized, by Gelfand and Pinsker [46]. The calculation of the capacity for the important special case of additive Gaussian noise and an additive Gaussian state was done by Costa [23], who concluded the surprising result that the capacity is the same as that of the channel where the state is known to the receiver also. Practical construction of the binning schemes (involving two steps: a vector quantization step and a channel coding step) is still an ongoing effort and the current progress is surveyed by Zamir *et al.* [154]. The performance of the opportunistic orthogonal signaling scheme, which uses orthogonal signals as both channel codes and vector quantizers, was analyzed by Liu and Viswanath [76].

The Costa precoding scheme was used in the multiple antenna downlink channel by Caire and Shamai [17]. The optimality of these schemes for the sum rate was shown in [17, 135, 138, 153]. Weingarten, *et al.* [141] proved that the Costa precoding scheme achieves the entire capacity region of the multiple antenna downlink.

The reciprocity between the uplink and the downlink was observed in different contexts: linear beamforming (Visotsky and Madhow [134], Farrokhi *et al.* [37]), capacity of the point-to-point MIMO channel (Telatar [119]), and achievable rates of

the single antenna Gaussian MAC and BC (Jindal *et al.* [63]). The presentation here is based on a unified understanding of these results (Viswanath and Tse [138]).

10.7 Exercises

Exercise 10.1 Consider the time-invariant uplink with multiple receive antennas (10.1). Suppose user k transmits data at power P_k, $k = 1, \ldots, K$. We would like to employ a bank of linear MMSE receivers at the base-station to decode the data of the users:

$$\hat{x}_k[m] = \mathbf{c}_k^* \mathbf{y}[m], \qquad (10.88)$$

is the estimate of the data symbol $x_k[m]$.

1. Find an explicit expression for the linear MMSE filter $\mathbf{c}_k$ (for user k). *Hint*: Recall the analogy between the uplink here with independent data streams being transmitted on a point-to-point MIMO channel and see (8.66) in Section 8.3.3.
2. Explicitly calculate the SINR of user k using the linear MMSE filter. *Hint*: See (8.67).

Exercise 10.2 Consider the bank of linear MMSE receivers at the base-station decoding the user signals in the uplink (as in Exercise 10.1). We would like to tune the transmit powers of the users $P_1, \ldots, P_K$ such that the SINR of each user (calculated in Exercise 10.1(2)) is at least equal to a target level β. Show that, if it is possible to find a set of power levels that meet this requirement, then there exists a component-wise minimum power setting that meets the SINR target level. This result is on similar lines to the one in Exercise 4.5 and is proved in [128].

Exercise 10.3 In this problem, a sequel to Exercise 10.2, we will see an adaptive algorithm that updates the transmit powers and linear MMSE receivers for each user in a greedy fashion. This algorithm is closely related to the one we studied in Exercise 4.8 and is adapted from [128].

Users begin (at time 1) with an arbitrary power setting $p_1^{(1)}, \ldots, p_K^{(1)}$. The bank of linear MMSE receivers $(\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_K^{(1)})$ at the base-station is tuned to these transmit powers. At time $m + 1$, each user updates its transmit power and its MMSE filter as a function of the power levels of the other users at time m so that its SINR is exactly equal to β. Show that if there exists a set of powers such that the SINR requirement can be met, then this synchronous update algorithm will converge to the component-wise minimal power setting identified in Exercise 10.2.

In this exercise, the update of the user powers (and corresponding MMSE filters) is synchronous among the users. An asynchronous algorithm, analogous to the one in Exercise 4.9, works as well.

Exercise 10.4 Consider the two-user uplink with multiple receive antennas (10.1):

$$\mathbf{y}[m] = \sum_{k=1}^{2} \mathbf{h}_k x_k[m] + \mathbf{w}[m]. \qquad (10.89)$$

Suppose user k has an average power constraint P_k, $k = 1, 2$.

1. Consider orthogonal multiple access: with α the fraction of the degrees of freedom allocated to user 1 (and $1 - \alpha$ the fraction to user 2), the reliable communication rates of the two users are given in Eq. (10.7). Calculate the fraction α that yields the largest sum rate achievable by orthogonal multiple access and the corresponding sum rate. *Hint*: Recall the result for the uplink with a single receive antenna in Section 6.1.3 that the largest sum rate with orthogonal multiple access is *equal* to the sum capacity of the uplink, cf. Figure 6.4.

2. Consider the difference between the sum capacity of the uplink with multiple receive antennas (see (10.4)) with the largest sum rate of this uplink with orthogonal multiple access.

   (a) Show that this difference is zero *exactly* when $\mathbf{h}_1 = c\mathbf{h}_2$ for some (complex) constant c.

   (b) Suppose $\mathbf{h}_1$ and $\mathbf{h}_2$ are not scalar complex multiples of each other. Show that at high SNR (N_0 goes to zero) the difference between the two sum rates becomes arbitrarily large. With $P_1 = P_2 = P$, calculate the rate of growth of this difference with SNR (P/N_0). We conclude that at high SNR (large values of P_1, P_2 as compared to N_0) orthogonal multiple access is very suboptimal in terms of the sum of the rates of the users .

Exercise 10.5 Consider the K-user uplink and focus on the *sum* and *symmetric* capacities. The base-station has an array of n_r receive antennas. With receiver CSI and fast fading, we have the following expression: the symmetric capacity is

$$C_{\text{sym}} = \frac{1}{K} \mathbb{E}[\log_2 \det(\mathbf{I}_{n_r} + \text{SNR}\,\mathbf{H}\mathbf{H}^*)] \text{ bits/s/Hz.} \qquad (10.90)$$

and the sum capacity C_{sum} is KC_{sym}. Here the columns of $\mathbf{H}$ represent the receive spatial signatures of the users and are modeled as i.i.d. $\mathcal{CN}(0, 1)$. Each user has an identical transmit power constraint P, and the common SNR is equal to P/N_0.

1. Show that the sum capacity increases monotonically with the number of users.

2. Show that the symmetric capacity, on the other hand, goes to zero as the number of users K grows large, for every fixed SNR value and n_r. *Hint*: You can use Jensen's inequality to get a bound.

3. Show that the sum capacity increases linearly in K at low SNR. Thus the symmetric capacity is independent of K at low SNR values.

4. Argue that at high SNR the sum capacity only grows logarithmically in K as K increases beyond n_r.

5. Plot C_{sum} and C_{sym} as a function of K for sample SNR values (from 0 dB to 30 dB) and sample n_r values (3 through 6). Can you conclude some general trends from your plots? In particular, focus on the following issues.

   (a) How does the value of K at which the sum capacity starts to grow slowly depend on n_r?

   (b) How does the value of K beyond which the symmetric capacity starts to decay rapidly depend on n_r?

   (c) How does the answer to the previous two questions change with the operating SNR value?

   You should be able to arrive at the following rule of thumb: $K = n_r$ is a good operating point at most SNR values in the sense that increasing K beyond it does

not increase the sum capacity by much, and in fact reduces the symmetric capacity by quite a bit.

Exercise 10.6 Consider the K-user uplink with n_r multiple antennas at the base-station as in Exercise 10.5. The expression for the symmetric capacity is in (10.90). Argue that the symmetric capacity at low SNR is comparable to the symmetric rate with orthogonal multiple access. *Hint*: Recall the discussion on the low SNR MIMO performance gain in Section 8.2.2.

Exercise 10.7 In a slow fading uplink, the multiple receive antennas can be used to improve the reliability of reception (diversity gain), improve the rate of communication at a fixed reliability level (multiplexing gain), and also spatially separate the signals of the users (multiple access gain). A reading exercise is to study [86] and [125] which derive the fundamental tradeoff between these gains.

Exercise 10.8 In this exercise, we further study the comparison between orthogonal multiple access and SDMA with multiple receive antennas at the base-station. While orthogonal multiple access is simple to implement, SDMA is the capacity achieving scheme and outperforms orthogonal multiple access in certain scenarios (cf. Exercise 10.4) but requires complex joint decoding of the users at the base-station.

Consider the following access mechanism, which is a cross between purely orthogonal multiple access (where all the users' signals are orthogonal) and purely SDMA (where all the K users share the bandwidth and time simultaneously). Divide the K users into groups of approximately n_r users each. We provide orthogonal resource allocation (time, frequency or a combination) to each of the groups but within each group the users (approximately n_r of them) operate in an SDMA mode.

We would like to compare this intermediate scheme with orthogonal multiple access and SDMA. Let us use the largest symmetric rate achievable with each scheme as the performance criterion. The uplink model (same as the one in Exercise 10.5) is the following: receiver CSI with i.i.d. Rayleigh fast fading. Each user has the same average transmit power constraint P, and SNR denotes the ratio of P to the background complex Gaussian noise power N_0.

1. Write an expression for the symmetric rate with the intermediate access scheme (the expression for the symmetric rate with SDMA is in (10.90)).
2. Show that the intermediate access scheme has performance comparable to both orthogonal multiple access and SDMA at low SNR, in the sense that the ratio of the performances goes to 1 as SNR $\rightarrow 0$.
3. Show that the intermediate access scheme has performance comparable to SDMA at high SNR, in the sense that the ratio of the performances goes to 1 as SNR $\rightarrow \infty$.
4. Fix the number of users K (to, say, 30) and the number of receive antennas n_r (to, say, 5). Plot the symmetric rate with SDMA, orthogonal multiple access and the intermediate access scheme as a function of SNR (0 dB to 30 dB). How does the intermediate access scheme compare with SDMA and orthogonal multiple access for the intermediate SNR values?

Exercise 10.9 Consider the K-user uplink with multiple receive antennas (10.1):

$$\mathbf{y}[m] = \sum_{k=1}^{K} \mathbf{h}_k x_k[m] + \mathbf{w}[m]. \qquad (10.91)$$

Consider the sum capacity with full CSI (10.17):

$$C_{\text{sum}} = \max_{P_k(\mathbf{H}), k=1, \ldots, K} \mathbb{E}\left[\log \det\left(\mathbf{I}_{n_r} + \sum_{k=1}^{K} P_k(\mathbf{H})\mathbf{h}_k\mathbf{h}_k^*\right)\right], \tag{10.92}$$

where we have assumed the noise variance $N_0 = 1$ and have written $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$. User k has an average power constraint P; due to the ergodicity in the channel fluctuations, the average power is equal to the ensemble average of the power transmitted at each fading state ($P_k(\mathbf{H})$ when the channel state is $\mathbf{H}$). So the average power constraint can be written as

$$\mathbb{E}[P_k(\mathbf{H})] \leq P. \tag{10.93}$$

We would like to understand what power allocations maximize the sum capacity in (10.92).

1. Consider the map from a set of powers to the corresponding sum rate in the uplink:

$$(P_1, \ldots, P_K) \mapsto \log \det\left(\mathbf{I}_{n_r} + \sum_{k=1}^{K} P_k\mathbf{h}_k\mathbf{h}_k^*\right). \tag{10.94}$$

   Show that this map is jointly concave in the set of powers. *Hint*: You will find useful the following generalization (to higher dimensions) of the elementary observation that the map $x \mapsto \log x$ is concave for positive real x:

$$\mathbf{A} \mapsto \log \det(\mathbf{A}) \tag{10.95}$$

   is concave in the set of positive definite matrices $\mathbf{A}$.

2. Due to the concavity property, we can characterize the optimal power allocation policy using the Lagrangian:

$$\mathcal{L}(P_1(\mathbf{H}), \ldots, P_K(\mathbf{H})) := \mathbb{E}\left[\log \det\left(\mathbf{I}_{n_r} + \sum_{k=1}^{K} P_k(\mathbf{H})\mathbf{h}_k\mathbf{h}_k^*\right)\right]$$

$$- \sum_{k=1}^{K} \lambda_k \mathbb{E}[P_k(\mathbf{H})]. \tag{10.96}$$

   The optimal power allocation policy $P_k^*(\mathbf{H})$ satisfies the Kuhn–Tucker equations:

$$\frac{\partial L}{\partial P_k(\mathbf{H})} \begin{cases} = 0 & \text{if } P_k^*(\mathbf{H}) > 0, \\ \leq 0 & \text{if } P_k^*(\mathbf{H}) = 0. \end{cases} \tag{10.97}$$

   Calculate the partial derivative explicitly to arrive at:

$$\mathbf{h}_k^*\left(\mathbf{I}_{n_r} + \sum_{j=1}^{K} P_j^*(\mathbf{H})\mathbf{h}_j\mathbf{h}_j^*\right)^{-1}\mathbf{h}_k \begin{cases} = \lambda_k & \text{if } P_k^*(\mathbf{H}) > 0, \\ \leq \lambda_k & \text{if } P_k^*(\mathbf{H}) = 0. \end{cases} \tag{10.98}$$

   Here $\lambda_1, \ldots, \lambda_K$ are constants such that the average power constraint in (10.93) is met. With i.i.d. channel fading statistics (i.e., $\mathbf{h}_1, \ldots, \mathbf{h}_K$ are i.i.d. random vectors), these constants can be taken to be equal.

3. The optimal power allocation $P_k^*(\mathbf{H})$, $k = 1, \ldots, K$ satisfying (10.98) is also the solution to the following optimization problem:

$$\max_{P_1,\ldots,P_K \geq 0} \log\det\left(\mathbf{I}_{n_r} + \sum_{k=1}^{K} P_k \mathbf{h}_k \mathbf{h}_k^*\right) - \sum_{k=1}^{K} \lambda_k P_k. \tag{10.99}$$

In general, no closed form solution to this problem is known. However, efficient algorithms yielding numerical solutions have been designed; see [15]. Solve numerically an instance of the optimization problem in (10.99) with $n_r = 2$, $K = 3$,

$$\mathbf{h}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{h}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \mathbf{h}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{10.100}$$

and $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$. You might find the software package [82] useful.

4. To get a feel for the optimization problem in (10.99) let us consider a few illustrative examples.

   (a) Consider the uplink with a single receive antenna, i.e., $n_r = 1$. Further suppose that each of the $|h_k|^2/\lambda_k$, $k = 1, \ldots, K$ are distinct. Show that an optimal solution to the problem in (10.99) is to allocate positive power to at most one user:

   $$P_k^* = \begin{cases} \left(\frac{1}{\lambda_k} - \frac{1}{|h_k|^2}\right)^+ & \text{if } \frac{|h_k|^2}{\lambda_k} = \max_{j=1\ldots K} \frac{|h_j|^2}{\lambda_j}, \\ 0 & \text{else.} \end{cases} \tag{10.101}$$

   This calculation is a reprise of that in Section 6.3.3.

   (b) Now suppose there are three users in the uplink with two receive antennas, i.e., $K = 3$ and $n_r = 2$. Suppose $\lambda_k = \lambda$, $k = 1, 2, 3$ and

   $$\mathbf{h}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{h}_2 = \begin{bmatrix} 1 \\ \exp(j2\pi/3) \end{bmatrix}, \quad \mathbf{h}_3 = \begin{bmatrix} 1 \\ \exp(j4\pi/3) \end{bmatrix}. \tag{10.102}$$

   Show that the optimal solution to (10.99) is

   $$P_k^* = \frac{2}{9}\left(\frac{3}{\lambda} - 1\right)^+, \qquad k = 1, 2, 3. \tag{10.103}$$

   Thus for $n_r > 1$ the optimal solution in general allocates positive power to more than one user. *Hint*: First show that for any set of powers P_1, P_2, P_3 with their sum constrained (to say P), it is always optimal to choose them all equal (to $P/3$).

Exercise 10.10 In this exercise, we look for an approximation to the optimal power allocation policy derived in Exercise 10.9. To simplify our calculations, we take i.i.d. fading statistics of the users so that $\lambda_1, \ldots, \lambda_K$ can all be taken equal (and denoted by λ).

1. Show that

$$\mathbf{h}_k^*\left(\mathbf{I}_{n_r} + \sum_{j=1}^{K} P_j \mathbf{h}_j \mathbf{h}_j^*\right)^{-1} \mathbf{h}_k = \frac{\mathbf{h}_k^*\left(\mathbf{I}_{n_r} + \sum_{j \neq k} P_j \mathbf{h}_j \mathbf{h}_j^*\right)^{-1} \mathbf{h}_k}{1 + \mathbf{h}_k^*\left(\mathbf{I}_{n_r} + \sum_{j \neq k} P_j \mathbf{h}_j \mathbf{h}_j^*\right)^{-1} \mathbf{h}_k P_k}. \tag{10.104}$$

Hint: You will find the matrix inversion lemma (8.124) useful.

2. Starting from (10.98), use (10.104) to show that the optimal power allocation policy can be rewritten as

$$P_k^*(\mathbf{H}) = \left(\frac{1}{\lambda} - \frac{1}{\mathbf{h}_k^*(\mathbf{I}_{n_r} + \sum_{j \neq k} P_j^*(\mathbf{H})\mathbf{h}_j\mathbf{h}_j^*)^{-1}\mathbf{h}_k} \right)^+. \qquad (10.105)$$

3. The quantity

$$\mathsf{SINR}_k := \mathbf{h}_k^* \left(\mathbf{I}_{n_r} + \sum_{j \neq k} P_j^*(\mathbf{H})\mathbf{h}_j\mathbf{h}_j^* \right)^{-1} \mathbf{h}_k P_k^*(\mathbf{H}) \qquad (10.106)$$

can be interpreted as the SINR at the output of an MMSE filter used to demodulate user k's data (cf. (8.67)). If we define

$$I_0 := \frac{P_k^*(\mathbf{H}) \|\mathbf{h}_k\|^2}{\mathsf{SINR}_k}, \qquad (10.107)$$

then I_0 can be interpreted as the *interference* plus noise seen by user k. Substituting (10.107) in (10.105) we see that the optimal power allocation policy can be written as

$$P_k(\mathbf{H}) = \left(\frac{1}{\lambda} - \frac{I_0}{\|\mathbf{h}_k\|^2} \right)^+. \qquad (10.108)$$

While this power allocation appears to be the same as that of waterfilling, we have to be careful since I_0 itself is a function of the power allocations of the other users (which themselves depend on the power allocated to user k, cf. (10.105)). However, in a large system with K and n_r large enough (but the ratio of K and n_r being fixed) I_0 converges to a constant in probability (with i.i.d. zero mean entries of $\mathbf{H}$, the constant it converges to depends only on the variance of the entries of $\mathbf{H}$, the ratio between K and n_r and the background noise density N_0). This convergence result is essentially an application of a general convergence result that is of the same nature as the singular values of a large random matrix (discussed in Section 8.2.2). This justifies (10.21) and the details of this result can be found in [136].

Exercise 10.11 Consider the two-user MIMO uplink (see Section 10.2.1) with input covariances $\mathbf{K}_{x1}, \mathbf{K}_{x2}$.

1. Consider the corner point A in Figure 10.13, which depicts the achievable rate region using this input strategy. Show (as an extension of (10.5)) that at the point A the rates of the two users are

$$R_2 = \log \det(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}_2\mathbf{K}_{x2}\mathbf{H}_2^*), \qquad (10.109)$$

$$R_1 = \log \det(\mathbf{I}_{n_r} + \mathbf{H}_1\mathbf{K}_{x1}(N_0\mathbf{I}_{n_r} + \mathbf{H}_2\mathbf{K}_{x2}\mathbf{H}_2^*)^{-1}\mathbf{H}_1^*). \qquad (10.110)$$

2. Analogously, calculate the rate pair represented by the point B.

Exercise 10.12 Consider the capacity region of the two-user MIMO uplink (the convex hull of the union of the pentagon in Figure 10.13 for all possible input strategies parameterized by $\mathbf{K}_{x1}$ and $\mathbf{K}_{x2}$). Let us fix positive weights $a_1 \leq a_2$ and consider maximizing $a_1 R_1 + a_2 R_2$ over all rate pairs (R_1, R_2) in the capacity region.

1. Fix an input strategy ($\mathbf{K}_{xk}, k = 1, 2$) and consider the value of $a_1 R_1 + a_2 R_2$ at the two corner points A and B of the corresponding pentagon (evaluated in Exercise 10.12). Show that the value of the linear functional is always no less at the vertex A than at the vertex B. You can use the expression for the rate pairs at the two corner points A and B derived in Exercise 10.11. This result is analogous to the polymatroid property derived in Exercise 6.9 for the capacity region of the single antenna uplink.

2. Now we would like to optimize $a_1 R_1 + a_2 R_2$ over all possible input strategies. Since the linear functional will always be optimized at one of the two vertices A or B in one of the pentagons, we only need to evaluate $a_1 R_1 + a_2 R_2$ at the corner point A (cf. (10.110) and (10.109)) and then maximize over the different input strategies:

$$
\max_{\mathbf{K}_{xk}, \mathrm{Tr}\mathbf{K}_{xk} \leq P_k, k=1,2} \quad a_1 \log\det(\mathbf{I}_{n_r} + \mathbf{H}_1 \mathbf{K}_{x1} (N_0 \mathbf{I}_{n_r} + \mathbf{H}_2 \mathbf{K}_{x2} \mathbf{H}_2^*)^{-1} \mathbf{H}_1^*)
$$

$$
+ a_2 \log\det(\mathbf{I}_{n_r} + \frac{1}{N_\phi} \mathbf{H}_2 \mathbf{K}_{x2} \mathbf{H}_2^*). \tag{10.111}
$$

   Show that the function being maximized above is jointly concave in the input $\mathbf{K}_{x1}, \mathbf{K}_{x2}$. *Hint*: Show that $a_1 R_1 + a_2 R_2$ evaluated at the point A can also be written as

$$
a_1 \log\det(\mathbf{I}_{n_r} + \frac{1}{N_\phi} \mathbf{H}_1 \mathbf{K}_{x1} \mathbf{H}_1^* + \frac{1}{N_0} \mathbf{H}_2 \mathbf{K}_{x2} \mathbf{H}_2^*) + (a_2 - a_1) \log\det(\mathbf{I}_{n_r} + \frac{1}{N_\phi} \mathbf{H}_2 \mathbf{K}_{x2} \mathbf{H}_2^*).
$$
$$\tag{10.112}$$

   Now use the concavity property in (10.95) to arrive at the desired result.

3. In general there is no closed-form solution to the optimization problem in (10.111). However, the concavity property of the function being maximized has been used to design efficient algorithms that arrive at numerical solutions to this problem, [15].

Exercise 10.13 Consider the two-user fast fading MIMO uplink (see (10.25)). In the angular domain representation (see (7.70))

$$
\mathbf{H}_k^a[m] = \mathbf{U}_r^* \mathbf{H}_k[m] \mathbf{U}_t, \qquad k = 1, 2, \tag{10.113}
$$

suppose that the stationary distribution of $\mathbf{H}_k^a[m]$ has entries that are zero mean and uncorrelated (and further independent across the two users). Now consider maximizing the linear functional $a_1 R_1 + a_2 R_2$ (with $a_1 \leq a_2$) over all rate pairs (R_1, R_2) in the capacity region.

1. As in Exercise 10.12, show that the maximal value of the linear functional is attained at the vertex A in Figure 10.7 for some input covariances. Thus conclude that, analogous to (10.112), the maximal value of the linear functional over the capacity region can be written as

$$
\max_{\mathbf{K}_{xk}, \mathrm{Tr}\mathbf{K}_{xk} \leq P_k, k=1,2} \quad a_1 \mathbb{E}[\log\det(\mathbf{I}_{n_r} + \frac{1}{N_\phi} \mathbf{H}_1 \mathbf{K}_{x1} \mathbf{H}_1^* + \frac{1}{N_\phi} \mathbf{H}_2 \mathbf{K}_{x2} \mathbf{H}_2^*)]
$$

$$
+ (a_2 - a_1) \mathbb{E}[\log\det(\mathbf{I}_{n_r} + \frac{1}{N_\phi} \mathbf{H}_2 \mathbf{K}_{x2} \mathbf{H}_2^*)]. \tag{10.114}
$$

2. Analogous to Exercise 8.3 show that the input covariances of the form in (10.27) achieve the maximum above in (10.114).

Exercise 10.14 Consider the two-user fast fading MIMO uplink under i.i.d. Rayleigh fading. Show that the input covariance in (10.30) achieves the maximal value of every linear functional $a_1 R_1 + a_2 R_2$ over the capacity region. Thus the capacity region in this case is simply a pentagon. *Hint*: Show that the input covariance in (10.30) *simultaneously* maximizes each of the constraints (10.28) and (10.29).

Exercise 10.15 Consider the (primal) point-to-point MIMO channel

$$\mathbf{y}[m] = \mathbf{H}\mathbf{x}[m] + \mathbf{w}[m], \tag{10.115}$$

and its reciprocal

$$\mathbf{y}_{\text{rec}}[m] = \mathbf{H}^*\mathbf{x}_{\text{rec}}[m] + \mathbf{w}_{\text{rec}}[m]. \tag{10.116}$$

The MIMO channel $\mathbf{H}$ has n_t transmit antennas and n_r receive antennas (so the reciprocal channel $\mathbf{H}^*$ is n_t times n_r). Here $\mathbf{w}[m]$ is i.i.d. $\mathcal{CN}(0, N_0\mathbf{I}_{n_r})$ and $\mathbf{w}_{\text{rec}}[m]$ is i.i.d. $\mathcal{CN}(0, N_0\mathbf{I}_{n_t})$. Consider sending $n_{\min}$ independent data streams on both these channels. The data streams are transmitted on the channels after passing through linear transmit filters (represented by unit norm vectors): $\mathbf{v}_1, \ldots, \mathbf{v}_{n_{\min}}$ for the primal channel and $\mathbf{u}_1, \ldots, \mathbf{u}_{n_{\min}}$ for the reciprocal channel. The data streams are then recovered from the received signal after passing through linear receive filters: $\mathbf{u}_1, \ldots, \mathbf{u}_{n_{\min}}$ for the primal channel and $\mathbf{v}_1, \ldots, \mathbf{v}_{n_{\min}}$ for the reciprocal channel. This process is illustrated in Figure 10.31.

1. Suppose powers $Q_1, \ldots, Q_{n_{\min}}$ are allocated to the data streams on the primal channel and powers $P_1, \ldots, P_{n_{\min}}$ are allocated to the data streams on the reciprocal channel. Show that the SINR for data stream k on the primal channel is

$$\text{SINR}_k = \frac{Q_k \mathbf{u}_k^* \mathbf{H}\mathbf{v}_k}{N_0 + \sum_{j \neq k} Q_j \mathbf{u}_k^* \mathbf{H}\mathbf{v}_j}, \tag{10.117}$$
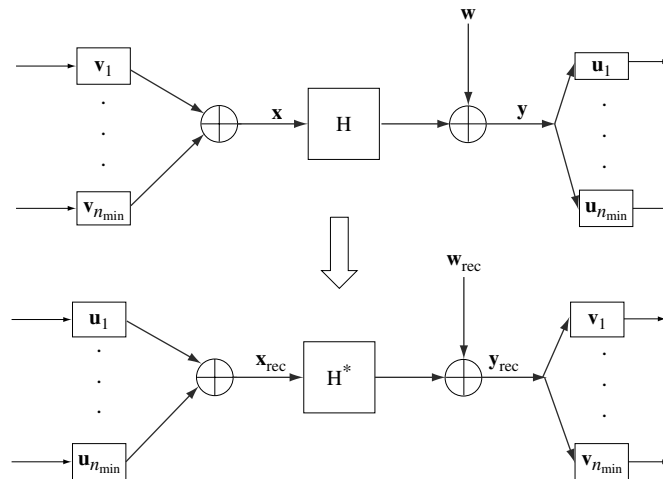

Figure 10.31 The data streams transmitted and received via linear filters on the primal (top) and reciprocal (bottom) channels.

and that on the reciprocal channel is

$$\text{SINR}_k^{\text{rec}} = \frac{P_k \mathbf{v}_k^* \mathbf{H}^* \mathbf{u}_k}{N_0 + \sum_{j \neq k} P_j \mathbf{v}_k^* \mathbf{H}^* \mathbf{u}_j}. \tag{10.118}$$

2. Suppose we fix the linear transmit and receive filters and want to allocate powers to meet a target SINR for each data stream (in both the primal and reciprocal channels). Find an expression analogous to (10.43) for the component-wise minimal set of power allocations.
3. Show that to meet the same SINR requirement for a given data stream on both the primal and reciprocal channels, the sum of the minimal set of powers is the same in both the primal and reciprocal channels. This is a generalization of (10.45).
4. We can use this general result to see earlier results in a unified way.
    (a) With the filters $\mathbf{v}_k = [0, \ldots, 0, 1, 0, \ldots, 0]^t$ (with the single 1 in the kth position), show that we capture the uplink–downlink duality result in (10.45).
    (b) Suppose $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$ is the singular value decomposition. With the filters $\mathbf{u}_k$ equal to the first $n_{\min}$ rows of $\mathbf{U}$ and the filters $\mathbf{v}_k$ equal to the first $n_{\min}$ columns of $\mathbf{V}$, show that this transceiver architecture achieves the capacity of the point-to-point MIMO primal and reciprocal channels with the same overall transmit power constraint, cf. Figure 7.2. Thus conclude that this result captures the reciprocity property discussed in Exercise 8.1.

Exercise 10.16 [76] Consider the opportunistic orthogonal signaling scheme described in Section 10.3.3. Each of the M messages corresponds to K (real) orthogonal signals. The encoder transmits the signal that has the largest correlation (among the K possible choices corresponding to the message to be conveyed) with the interference (real white Gaussian process with power spectral density $N_s/2$). The decoder decides the most likely transmit signal (among the MK possible choices) and then decides on the message corresponding to the most likely transmit signal. Fix the number of messages, M, and the number of signals for each message, K. Suppose that message 1 is to be conveyed.

1. Derive a good upper bound on the error probability of opportunistic orthogonal signaling. Here you can use the technique developed in the upper bound on the error probability of regular orthogonal signaling in Exercise 5.9. What is the appropriate choice of the threshold, γ, as a function of M, K and the power spectral densities $N_s/2, N_0/2$?
2. By an appropriate choice of K as a function of M, N_s, N_0 show that the upper bound you have derived converges to zero as M goes to infinity as long as $\mathcal{E}_b/N_0$ is larger than $-1.59\,\text{dB}$.
3. Can you explain why opportunistic orthogonal signaling achieves the capacity of the infinite bandwidth AWGN channel with no interference by interpreting the correct choice of K?
4. We have worked with the assumption that the interference $s(t)$ is white Gaussian. Suppose $s(t)$ is still white but not Gaussian. Can you think of a simple way to modify the opportunistic orthogonal signaling scheme presented in the text so that we still achieve the same minimal $\mathcal{E}_b/N_0$ of $-1.59\,\text{dB}$?

Exercise 10.17 Consider a real random variable x_1 that is restricted to the range $[0,1]$ and x_2 is another random variable that is jointly distributed with x_1. Suppose u is a

uniform random variable on $[0,1]$ and is jointly independent of x_1 and x_2. Consider the new random variable

$$\tilde{x}_1 = \begin{cases} x_1 + u & \text{if } x_1 + u \leq 1, \\ x_1 + u - 1 & \text{if } x_1 + u > 1. \end{cases} \tag{10.119}$$

The random variable $\tilde{x}_1$ can be thought of as the right cyclic addition of x_1 and u.

1. Show that $\tilde{x}_1$ is uniformly distributed on $[0,1]$.
2. Show that $\tilde{x}_1$ and (x_1, x_2) are independent.

Now suppose x_1 is the Costa-precoded signal containing the message to user 1 in a two-user single antenna downlink based on x_2, the signal of user 2 (cf. Section 10.3.4). If the realization of the random variable u is known to user 1 also, then $\tilde{x}_1$ and x_1 contain the same information (since the operation in (10.119) is invertible). Thus we could transmit $\tilde{x}_1$ in place of x_1 without any change in the performance of user 1. But the important change is that the transmit signal $\tilde{x}_1$ is now independent of x_2.

The common random variable u, shared between the base-station and user 1, is called the *dither*. Here we have focused on a single time symbol and made $\tilde{x}_1$ uniform. With a large block length, this basic argument can be extended to make the transmit vector $\tilde{\mathbf{x}}_1$ appear Gaussian and independent of $\mathbf{x}_2$; this dithering idea is used to justify (10.65).

Exercise 10.18 Consider the two-user single antenna downlink (cf. (10.63)) with $|h_1| > |h_2|$. Consider the rate tuple (R'_1, R'_2) achieved via Costa precoding in (10.66). In this exercise we show that this rate pair is strictly inside the capacity region of the downlink. Suppose we allocate powers Q_1, Q_2 to the two users and do superposition encoding and decoding (cf. Figures 6.7 and 6.8) and aim to achieve the same rates as the pair in (10.66).

1. Calculate Q_1, Q_2 such that

$$R'_1 = \log\left(1 + \frac{|h_1|^2 Q_1}{N_0}\right), \qquad R'_2 = \log\left(1 + \frac{|h_2|^2 Q_2}{N_0 + |h_2|^2 Q_1}\right), \tag{10.120}$$

   where R'_1 and R'_2 are the rate pair in (10.66).
2. Using the fact that user 1 has a stronger channel than user 2 (i.e., $|h_1| > |h_2|$) show that the total power used in the superposition strategy to achieve the same rate pair (i.e., $Q_1 + Q_2$ from the previous part) is strictly smaller than $P_1 + P_2$, the transmit power in the Costa precoding strategy.
3. Observe that an increase in transmit power strictly increases the capacity region of the downlink. Hence conclude that the rate pair in (10.66) achieved by the Costa precoding strategy is strictly within the capacity region of the downlink.

Exercise 10.19 Consider the K-user downlink channel with a single antenna (an extension of the two-user channel in (10.63)):

$$y_k[m] = h_k x[m] + w_k[m], \qquad k = 1, \ldots, K. \tag{10.121}$$

Show that the following rates are achievable using Costa precoding, extending the argument in Section 10.3.4:

$$R_k = \log\left(1 + \frac{|h_k|^2 P_k}{\sum_{j=k+1}^{K} |h_j|^2 P_j + N_0}\right), \qquad k = 1, \ldots, K. \qquad (10.122)$$

Here $P_1, \ldots, P_K$ are some non-negative numbers that sum to P, the transmit power constraint at the base-station. You should not need to assume any specific ordering of the channels qualities $|h_1|, |h_2|, \ldots, |h_K|$ in arriving at your result. On the other hand, if we have

$$|h_1| \le |h_2| \le \cdots \le |h_K|, \qquad (10.123)$$

then the superposition coding approach, discussed in Section 6.2, achieves the rates in (10.122).

Exercise 10.20 Consider the reciprocal uplink channel in (10.40) with the receive filters $\mathbf{u}_1, \ldots, \mathbf{u}_K$ as in Figure 10.16. This time we embellish the receiver with successive cancellation, canceling users in the order K through 1 (i.e., user k does not see any interference from users $K, K-1, \ldots, k+1$). With powers $Q_1, \ldots, Q_K$ allocated to the users, show that the SINR for user k can be written as

$$\mathsf{SINR}_k^{\mathsf{ul}} = \frac{Q_k \, | \, \mathbf{u}_k^* \mathbf{h}_k \, |^2}{N_0 + \sum_{j<k} Q_j \, | \, \mathbf{u}_k^* \mathbf{h}_j \, |^2}. \qquad (10.124)$$

To meet the same SINR requirement as in the downlink with Costa precoding in the reverse order (the expression for the corresponding SINR is in (10.72)) show that the sum of the minimal powers required is the same for the uplink and the downlink. This is an extension of the conservation of sum-of-powers property seen without cancellation in (10.45).

Exercise 10.21 Consider the fast fading multiple transmit antenna downlink (cf. (10.73)) where the channels from antenna i to user k are modeled as i.i.d. $\mathcal{CN}(0, 1)$ random variables (for each antenna $i = 1, \ldots, n_t$ and for each user $k = 1, \ldots, K$). Each user has a single receive antenna. Further suppose that the channel fluctuations are i.i.d. over time as well. Each user has access to the realization of its channel fluctuations, while the base-station only has knowledge of the statistics of the channel fluctuations (the receiver CSI model). There is an overall power constraint P on the transmit power.

1. With just one user in the downlink, we have a MIMO channel with receiver only CSI. Show that the capacity of this channel is equal to

$$\mathbb{E}\left[\log\left(1 + \frac{\mathsf{SNR}\|\mathbf{h}\|^2}{n_t}\right)\right], \qquad (10.125)$$

where $\mathbf{h} \sim \mathcal{CN}(0, \mathbf{I}_{n_t})$ and $\mathsf{SNR} = P/N_0$. *Hint*: Recall (8.15) and Exercise 8.4.

2. Since the statistics of the user channels are identical, argue that if user k can decode its data reliably, then all the other users can also successfully decode user k's data (as we did in Section 6.4.1 for the single antenna downlink). Conclude that the

sum of the rates at which the users are being simultaneously reliably transmitted to is bounded as

$$\sum_{k=1}^{K} R_k \leq \mathbb{E}\left[\log\left(1 + \frac{\text{SNR}\|\mathbf{h}\|^2}{n_t}\right)\right], \tag{10.126}$$

analogous to (6.52).

Exercise 10.22 Consider the downlink with multiple receive antennas (cf. (10.78)). Show that the random variables $x[m]$ and $\mathbf{y}_k[m]$ are independent conditioned on $\tilde{y}_k[m]$. Hence conclude that

$$I(x; \mathbf{y}_k) = I(x; \tilde{y}_k), \qquad k = 1, 2. \tag{10.127}$$

Thus there is no loss in information by having a matched filter front end at each of the users converting the SIMO downlink into a single antenna channel to each user.

Exercise 10.23 Consider the two-user uplink fading channel with multiple antennas at the base-station:

$$\mathbf{y}[m] = \mathbf{h}_1[m]x_1[m] + \mathbf{h}_2[m]x_2[m] + \mathbf{w}[m]. \tag{10.128}$$

Here the user channels $\{\mathbf{h}_1[m]\}, \{\mathbf{h}_2[m]\}$ are statistically independent. Suppose that $\mathbf{h}_1[m]$ and $\mathbf{h}_2[m]$ are $\mathcal{CN}(0, N_0\mathbf{I}_{n_r})$. We operate the uplink in SDMA mode with the users having the same power P. The background noise $\mathbf{w}[m]$ is i.i.d. $\mathcal{CN}(0, N_0\mathbf{I}_{n_r})$. An SIC receiver decodes user 1 first, removes its contribution from $\{\mathbf{y}[m]\}$ and then decodes user 2. We would like to assess the effect of channel estimation error of $\mathbf{h}_2$ on the performance of user 1.

1. Suppose the users send training symbols using orthogonal multiple access and they spend 20% of their power on sending the training signal, repeated every T_c seconds, which is the channel coherence time of the users. What is the mean square estimation error of $\mathbf{h}_1$ and $\mathbf{h}_2$?

2. The first step of the SIC receiver is to decode user 1's information suppressing the user 2's signal. Using the linear MMSE filter to suppress the interference, numerically evaluate the average output SINR of the filter due to the channel estimation error, as compared to that with perfect channel estimation (cf. (8.62)). Plot the degradation (ratio of the SINR with imperfect and perfect channel estimates) as a function of the SNR, P/N_0, with $T_c = 10$ ms.

3. Argue using the previous calculation that better channel estimates are required to fully harness the gains of interference suppression. This means that the pilots in the uplink with SDMA have to be stronger than in the uplink with a single receive antenna.

Exercise 10.24 In this exercise, we explore the effect of channel measurement error on the reciprocity relationship between the uplink and the downlink. To isolate the situation of interest, consider just a single user in the uplink and the downlink (this is the natural model whenever the multiple access is orthogonal) with only the base-station having an array of antennas. The uplink channel is (cf. (10.40))

$$\mathbf{y}_{ul}[m] = \mathbf{h}x_{ul}[m] + \mathbf{w}_{ul}[m], \tag{10.129}$$

with a power constraint of P_{ul} on the uplink transmit symbol x_{ul}. The downlink channel is (cf. (10.39))

$$y_{\mathrm{dl}}[m] = \mathbf{h}^* \mathbf{x}_{\mathrm{dl}}[m] + w_{\mathrm{dl}}[m], \qquad (10.130)$$

with a power constraint of P_{dl} on the downlink transmit vector $\mathbf{x}_{\mathrm{dl}}$.

1. Suppose a training symbol is sent with the full power P_{ul} over one symbol time in the uplink to estimate the channel $\mathbf{h}$ at the base-station. What is the mean square error in the best estimate $\hat{\mathbf{h}}$ of the channel $\mathbf{h}$?

2. Now suppose the channel estimate $\hat{\mathbf{h}}$ from the previous part is used to beamform in the downlink, i.e., the transmit signal is

$$\mathbf{x}_{\mathrm{dl}} = \frac{\hat{\mathbf{h}}}{\|\hat{\mathbf{h}}\|} x_{\mathrm{dl}},$$

with the power in the data symbol x_{dl} equal to P_{dl}. What is the average received SNR in the downlink? The degradation in SNR is measured by the ratio of the average received SNR with imperfect and perfect channel estimates. For a fixed uplink SNR, P_{ul}/N_0, plot the average degradation for different values of the downlink SNR, P_{dl}/N_0.

3. Argue using your calculations that using the reciprocal channel estimate in the downlink is most beneficial when the uplink power P_{ul} is larger than or of the same order as the downlink power P_{dl}. Further, there is a huge degradation in performance when P_{dl} is much larger than P_{ul}.