1

Linear and Dirty-Paper Techniques for the Multi-User MIMO Downlink

Christian B. Peel[1], Quentin H. Spencer[2], A.Lee Swindlehurst[3], Martin Haardt[4], and Bertrand M. Hochwald[5]

[1] **Swiss Federal Institute of Technology, Zürich, Switzerland.**
[2] **Distribution Control Systems, Inc. Hazelwood, MO USA.**
[3] **Brigham Young University, Provo, UT, USA.**
[4] **Ilmenau University of Technology, Ilmenau, Germany**
[5] **Lucent Technologies (Bell Laboratories) Murray Hill, NJ, USA.**

Multi-input, multi-output (MIMO) communications systems have attracted considerable attention over the past decade, mostly for single-user, point-to-point scenarios. The multiple-user MIMO case has attracted less attention, and most of the research on this problem has focused on uplink communications. Only recently has the multi-user MIMO downlink been addressed, beginning with information-theoretic capacity results [1–5], and followed by practical implementations, including those based on linear techniques [6, 7] and non-linear precoding [8–11]. In this chapter we review these techniques and discuss some important open problems.

1.1   Introduction

1.1.1   Problem Overview

The term *multi-user MIMO downlink* typically refers to situations where a multi-antenna transmitter (*e.g.*, a basestation) simultaneously communicates with several co-channel users. In the communications and information theory literature, this scenario is referred to as the MIMO broadcast channel. We will also use the term *spatial multiplexing* to describe this problem, although we note that this term is also used in connection with point-to-point MIMO links when multiple independent data streams are transmitted to a single user (*e.g.*, as in vertical Bell-Labs space-time (V-BLAST) techniques [12, 13]). The users in a multi-user MIMO network may have a single antenna, and hence no ability for spatial discrimination, or they may have multiple antennas and the ability to perform some type of interference suppression. This is to be contrasted with the MIMO uplink problem, where a multiple antenna receiver must separate the signals arriving from several different users. This scenario is often referred to as the MIMO multiple access channel (MAC) or Space-Division Multiple Access (SDMA). In this paper, we focus on the multi-user MIMO downlink or broadcast channel. Although less frequently addressed in the literature, there is still a considerable body of work on the topic that is too extensive to adequately cover in this chapter. As discussed below, we will focus on two classes of approaches to this problem: linear beamforming techniques and non-linear precoding.

Single-user MIMO systems have generated considerable excitement in the wireless communications literature due to their potential for significant gains in capacity over single-antenna links. Of particular note is that these gains are often independent of whether or not channel state information (CSI) is available at the transmitter. The situation is considerably different in the multi-user case, where interference must be taken into account and balanced against the need for high throughput. A transmission scheme that maximizes the capacity for one user in the network might result in unacceptably high interference for the other users, rendering their links useless. If high throughput is the goal, a better approach might be to maximize the *sum* capacity of the network, or the maximum sum transmission rate, where the inter-user interference is taken into consideration. Transmit CSI is the key to achieving such a goal. While in principle the receivers themselves could perform some interference cancellation via multi-user detection, for example, the desire to keep costs low and preserve battery life for the end user in cellular networks usually leads to simpler receiver architectures.

Maximizing the sum capacity of a multi-user downlink channel does not always lead to a desirable solution. For example, if one of the users has a channel with considerably higher SNR than the others, the sum capacity solution might come at the expense of the weaker users who will receive little or no throughput. An alternative in such cases is to attempt to guarantee that each user achieves some minimum acceptable Quality of Service (QoS), *e.g.*, measured in terms of signal-to-interference-plus-noise ratio (SINR) or bit-error rate (BER). The problem of meeting QoS constraints with minimum transmit power is often referred to as the downlink *power control* or *interference-balancing* problem. As with sum capacity maximization, channel knowledge at the transmitter is crucial to finding a solution.

Channel state information is most often obtained by means of uplink training data, as in a time-division duplex system, or via feedback from the users, as in the frequency-division

duplex case. Each approach has its advantages and disadvantages in terms of throughput penalty and latency. CSI can be in the form of deterministic channel estimates, or it can be described in probabilistic terms (*e.g.,* channel mean and covariance). While we will focus on the deterministic case in this chapter, statistical CSI may be directly applied in most cases. For an excellent and comprehensive treatment of the issues involved with different types of CSI, see [14].

1.1.2   Literature Survey

Algorithms for multi-user MIMO downlink processing can be classified according to a number of criteria: whether they attempt to approach the sum capacity bound, eliminate inter-user interference or achieve minimum QoS constraints, whether the users have single or multiple antennas, whether or not multiple data streams are transmitted to each user, *etc.*. We begin with the case that has received the most attention: users with single-antenna receivers. The most direct approach in this case is referred to as *channel inversion* [15, 16], which amounts to using a set of transmit beamformers that "pre-inverts" the channel and ideally removes all inter-user interference at the receivers. One can think of this approach as zero-forcing transmit beamforming. As with zero-forcing receive beamformers, problems arise when the channel is nearly rank deficient, although we will see it is not noise amplification that occurs, but rather signal attenuation. Minimum mean-squared error (MMSE) or regularized transmit beamforming can be used as an alternative to reduce sensitivity to low-rank channels; dramatically improved performance is obtained [6, 17]. Although the gain of regularized channel inversion is significant, there is still a considerable gap between its performance and the capacity bound. Algorithms from the class of so-called "dirty paper" coding techniques have recently been shown to more closely approach the sum capacity for the multi-user channel, and in some cases achieve it [3, 4, 18–20]. We will describe one such technique, referred to as *vector modulo precoding* [8, 17, 21, 22], that can be framed as an extension of the channel inversion algorithms described earlier.

The algorithms mentioned above attempt to maximize the overall throughput of the network for a fixed transmit power, under the constraint of zero (or nearly zero) interference. On the other hand, power control or interference-balancing algorithms relax the zero interference constraint and minimize the total transmitted power subject to meeting given QoS constraints. Iterative methods have been found that are guaranteed to find the optimal solution to this problem, assuming a solution exists [23, 24]. The problem can also be posed as a semidefinite optimization with convex constraints, and solved using more efficient numerical procedures [25].

To this point, the research cited has assumed that each user possesses only one receive antenna. These algorithms can be trivially extended to multiple antenna receivers by viewing each as a separate "user," provided that the total number of receive antennas for all users is no greater than the number of transmit antennas. While this allows for extremely simple receiver architectures, it ignores the ability of the receivers to perform spatial discrimination of their own, and is only practical for networks with a small number of co-channel users. The result can be either (1) a significant gap between the achievable throughput of these techniques and the capacity of the system in cases where the receivers can obtain CSI, or (2) dramatic increase in required transmit power to achieve a desired QoS, especially in situations where the channels to adjacent receive antennas are not uncorrelated.

Instead of completely diagonalizing the channel as some of the techniques above attempt to do, one could find an optimal *block*-diagonalization when the users have multiple antennas. Such an approach removes inter-user interference, but leaves the receiver responsible for separating the multiple data streams sent to it [7, 26–31]. This approach still has the drawback of requiring more transmit antennas than the total number of receive antennas among all the users. As a means of relaxing this constraint, suppose that each user employs a beamformer or beamformers of its own to receive the data stream(s) destined for it. If the transmitter knew what those beamformers were in advance, then it could consider the *effective* channel to each user to be the combination of the propagation channel for that user and the beamformers that user employs. As long as the total number of data streams to all users does not exceed the number of transmit antennas, then any of the algorithms discussed above could be used. The problem of course is that the optimal receive beamformers depend on the choice of the transmit beamformers, and vice versa. Iterative techniques have been proposed in which the transmitter postulates a set of receive beamformers, designs a corresponding set of transmit weights, updates the receive beamformers accordingly, and so on [7, 31–37].

1.1.3   Chapter Organization

In the next section, we describe the mathematical model we will assume for our discussion of the multi-user MIMO downlink, and establish a common notation. Section 1.3 describes algorithms for the case where each user has only a single receive antenna and presents some simulation results illustrating their performance. Section 1.4 does the same for cases involving multiple antennas per user. We finally summarize and review open problems in the area in Section 1.5, including references to related work that we did not address in this chapter.

1.2   Background and Notation

We will consider a standard scenario involving a basestation that simultaneously transmits data to K users, whose channels have been determined earlier either through the use of uplink training data (as in a time-division duplex system) or via a feedback channel (as in a frequency-division duplex system). The basestation is assumed to have n_T antennas, user j has n_{R_j} antennas, and the total number of receive antennas is $n_R = \sum_{j=1}^{K} n_{R_j}$. In a flat-fading propagation environment, the channel between the base and user j is described by the $n_{R_j} \times n_T$ matrix $\mathbf{H}_j$, whose rows we denote by $\mathbf{h}_{ij}^*$ as follows:

$$\mathbf{H}_j^* = [\, \mathbf{h}_{1j} \; \cdots \; \mathbf{h}_{n_{R_j}j} \,] \,.$$

The symbol $(\cdot)^*$ is used to denote the complex conjugate (Hermitian) transpose. In Section 1.3, we will focus on cases where $n_{R_j} = 1$, in which case we will simply denote the channel as $\mathbf{H}_j = \mathbf{h}_j^*$. We will follow the convention of denoting matrices by capital boldface letters, vectors in lowercase boldface, and scalars as either upper or lowercase letters without boldface.

The basestation may desire to send data at different rates to each of the K active users. This can be accomplished by an appropriate choice of the symbol constellation for each user, or by changing the number of independent data streams that are simultaneously sent to each

user. We will let m_j denote the number of data streams transmitted to user j. Suitable values for $m_1, \cdots, m_K$ will not only depend on the desired data rate for user j, but also on the available transmit power, the achievable SINR, and the number of transmit and receive antennas. We will see that, typically, $m_j \leq n_{R_j}$ without some type of additional coding or multiplexing, and that $\sum m_k \leq n_T$. We will assume that m_j has been determined beforehand, recognizing the fact that this resource allocation step is critical if optimal system performance is required. Thus, at symbol time t, the transmitter desires to send the $m_j \times 1$ vector of symbols $\mathbf{d}_j(t)$ to user j. The signal destined for user j that is actually broadcast from the transmit antennas at time t is denoted by the $n_T \times 1$ vector $\mathbf{s}_j(t)$. In many cases, the transmitted signal is a linear function of the symbols, *i.e.,* $\mathbf{s}_j(t) = \mathbf{B}_j \mathbf{d}_j(t)$, where the columns of $\mathbf{B}_j$, denoted $\mathbf{B}_j = [\mathbf{b}_{1j} \cdots \mathbf{b}_{m_j j}]$, correspond to the transmit beamformers for each symbol. In cases where $m_j = 1$, we will simply write $\mathbf{B}_j = \mathbf{b}_j$, $\mathbf{d}_j(t) = d_j(t)$, and $\mathbf{s}_j(t) = \mathbf{b}_j d_j(t)$. We will also consider algorithms that employ a nonlinear mapping of the symbols to the transmitted data: $\mathbf{s}_j(t) = \mathbf{f}_j(\mathbf{d}_j(t))$.

User j not only receives its desired signal through the channel $\mathbf{H}_j$, but also contributions from the signals destined for other users:

$$\mathbf{x}_j(t) = \sum_{k=1}^{K} \mathbf{H}_j \mathbf{s}_k(t) + \mathbf{e}_j(t) , \tag{1.1}$$

where $\mathbf{e}_j(t)$ is assumed to represent spatially white noise and interference with covariance $\mathcal{E}\{\mathbf{e}_j(t)\mathbf{e}_j^*(t)\} = \mathbf{I}$. If linear beamforming is used on the transmit side, then stacking the data together from all of the receivers leads to the following compact expression:

$$\mathbf{x}(t) = \begin{bmatrix} \mathbf{x}_1(t) \\ \vdots \\ \mathbf{x}_K(t) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_K \end{bmatrix} [\mathbf{B}_1 \cdots \mathbf{B}_K] \begin{bmatrix} \mathbf{d}_1(t) \\ \vdots \\ \mathbf{d}_K(t) \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1(t) \\ \vdots \\ \mathbf{e}_K(t) \end{bmatrix} \tag{1.2}$$

$$= \mathbf{HBd}(t) + \mathbf{e}(t) , \tag{1.3}$$

where the definitions of $\mathbf{x}(t), \mathbf{H}, \mathbf{B}, \mathbf{d}(t)$ and $\mathbf{e}(t)$ should be obvious from context. For the sake of simplicity, in what follows we will drop the explicit dependence of the above equations on time. In some figures we will use the notation $\{n_{R_1}, \ldots, n_{R_K}\} \times n_T$ to describe the configuration of the antennas. Thus a $\{1, 1, 1, 1\} \times 4$ system has $K = 4$ users, each with one antenna, and a base station with 4 antennas, while $\{1, 1, 2, 2\} \times 4$ describes the same case, with the exception that two of the users have two antennas.

1.2.1 Capacity

A fundamental tool for characterizing any communication channel is capacity. In a single-user channel, capacity is the maximum amount of information that can be transmitted as a function of available bandwidth given a constraint on transmitted power. In single-user MIMO channels, it is common to assume that there is a constraint on the total power broadcast by all transmit antennas. For the multi-user MIMO channel, the problem is somewhat more complex. Given a constraint on the total transmitted power, it is possible to allocate
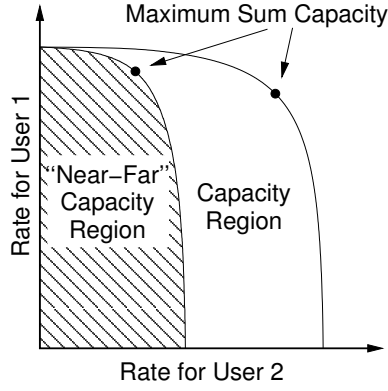
Figure 1.1: An illustration of a multi-user capacity region. The sum capacity may penalize certain users, depending on the shape of the capacity region.

varying fractions of that power to different users in the network, so a single power constraint can yield many different information rates. The result is a "capacity region" like that illustrated in Figure 1.1 for a two-user channel. The maximum capacity for user 1 is achieved when 100% of the power is allocated to user 1; for user 2 the maximum capacity is also obtained when it has all the power. For every possible power distribution in between, there is an achievable information rate, which results in the capacity regions depicted in the illustration. Two regions are shown in Figure 1.1, the bigger one for the case where both users have roughly the same maximum capacity, and the other for a case where they are different (due, for example, to user 2's channel being attenuated relative to user 1). For K users, the capacity region is characterized by a K-dimensional volume.

The maximum achievable throughput of the entire system is characterized by the point on the curve that maximizes the sum of all of the users' information rates, and is referred to as the *sum capacity* of the channel. This point is illustrated in Figure 1.1 by asterisks. Achieving the sum capacity point may not necessarily be the goal of a system designer. One example where this may be the case is when the "near-far" problem occurs, where one user has a strongly attenuated channel compared to other users. As depicted in Figure 1.1, obtaining the sum capacity in such a situation would come at the expense of the user with the attenuated channel.

The sum capacity for a system described by (1.1) has been formulated using the dirty paper coding (DPC) framework (see, for example, [3, 4, 18–20] and [38–41]) for the case of Gaussian noise. The capacity is defined in terms of the achievable rate for each user given the set of covariance matrices for each transmitted data vector $\mathbf{S}_k = \mathcal{E}\{\mathbf{s}_k \mathbf{s}_k^*\}$:

$$R_k = \frac{\log \left| \mathbf{I} + \mathbf{H}_k \left(\sum_{j=1}^{k} \mathbf{S}_k \right) \mathbf{H}_k^* \right|}{\log \left| \mathbf{I} + \mathbf{H}_k \left(\sum_{j=1}^{k-1} \mathbf{S}_k \right) \mathbf{H}_k^* \right|} , \qquad (1.4)$$

assuming that the data for each user is uncorrelated with the others. The sum capacity is then

$$C_S = \max_{\mathbf{S}_k \geq 0; \sum \text{tr}(\mathbf{S}_k) \leq \rho} \sum_{k=1}^{K} R_k \, , \tag{1.5}$$

where ρ is the upper bound on the total transmit power. The input distributions are arbitrary, though the sum capacity can be achieved with Gaussian signals [18, 19]. The capacity region C_R of a given multi-user MIMO system is defined to be the set of all achievable rates $\{R_1, \cdots, R_K\}$ given the power constraint. In general, determining C_R is an unsolved problem, but a solution for the Gaussian case has been reported in [5] building on work in [42].

In the case of users with single receive antennas ($n_{R_j} = 1$ for all j) the sum capacity expression is much simpler:

$$C_S = \max_{\mathbf{D} \in \mathbf{A}} \log |\mathbf{I} + \mathbf{HDH}^*| \, , \tag{1.6}$$

where $\mathbf{A}$ is the set of all $K \times K$ non-negative diagonal matrices $\mathbf{D}$ with $\text{tr}(\mathbf{D}) \leq \rho$. This equation looks much like the capacity of a point-to-point MIMO system with M transmit antennas and K receive antennas, where only the receiver has knowledge of the channel: $\log |\mathbf{I} + (\rho/n_T)\mathbf{HH}^*|$. This comparison makes it easy to see that multi-user sum capacity grows linearly with $\min(M, K)$ under the same conditions as for the single-user case.

1.2.2 Dirty-Paper Coding

As mentioned above, capacity results for the multi-user problem have been achieved using the notion of *dirty-paper coding*, which originates in a 1983 paper [43] by M. H. M. Costa. He studies a channel with Gaussian noise and interference that is known to the transmitter, and makes an analogy to the problem of writing on dirty paper. To describe this idea, let

$$y = s + i + w \, , \tag{1.7}$$

where s is the signal used to transmit a codeword d, i is interference with power Q known deterministically at the transmitter, but unknown to the receiver, $w \sim \mathcal{CN}(0, N)$ is Gaussian noise, and the received data is y. Costa presented the encouraging result that the capacity of this system is the same as if there were no interference present. If the signal has power constraint $|s|^2 \leq \rho$, then the capacity of this system is

$$C = \log \left(1 + \frac{\rho}{N} \right) \tag{1.8}$$

regardless of what Q is. To extend the dirty-paper analogy, the "capacity" of dirty paper is the same as for a sheet without this known "dirt."

This result has been applied to a variety of systems: using what is nominally "analog" spectrum for both analog and digital signals [44], information-embedding applications [45], and in finding the capacity of the MIMO broadcast channel [1–5]. These theoretical results have motivated progress in the development of practical algorithms [46–48] that approach the capacity bound (1.8).

To illustrate dirty-paper principles, we describe a simple technique based on the use of a simple modulo operator. Although this method is very simple, it performs within a few dB of capacity. We define the modulo function f as

$$f_\tau(y) = y - \left\lfloor \frac{y + \tau/2}{\tau} \right\rfloor \tau \; . \tag{1.9}$$

The signal s is created using information about the interference i and the codeword u, as follows:

$$s = f_\tau(u - i) = u - i - \tau k \; , \tag{1.10}$$

where k is any integer. The modulo function reduces the power of the transmitted signal from what it would be if the simple method of $s = u - i$ were used. Applying the modulo function to the received data (1.7) gives

$$\begin{aligned} f_\tau(y) = f_\tau(s + i + w) &= f_\tau(u - i - \tau k + i + w) \\ &= f_\tau(u + w) \; . \end{aligned}$$

The interference has been canceled; there remains only a penalty from applying the modulo function to $u + w$ which may lie outside the interval $(-\tau/2, \tau/2)$, and resulting in $f_\tau(u + w) \neq u + w$. In a practical system τ would be a function of the codeword constellation, chosen for example as described in Section 1.3.3.

To completely overcome this shaping loss, coding must occur over n consecutive samples and the modulo operation is applied with respect to a "good" n-dimensional lattice, rather than with respect to an interval. Finally, as $n \to \infty$ the shaping error disappears and capacity is achieved (See [39] for more information, including how to handle low SNR situations and to find a discussion on what a "good" lattice is). The shaping loss is 1.56 dB when using the simple cubical lattice defined by (1.9) as compared with an infinite-dimensional lattice on a scalar Gaussian interference channel. In Section 1.3.3 we use these DPC ideas to describe and analyze a coding technique for the MIMO downlink channel.

1.2.3 Discussion

We have assumed a data model with a flat-fading or narrow-band channel. However, in many current and next-generation wireless communications applications, this assumption does not hold. Wideband or frequency selective fading channels suffer from inter-symbol interference and a fading characteristic that varies significantly across the frequency band. There are several ways to apply the matrix channel model to this case. In channels where the use of orthogonal frequency division multiplexing (OFDM) is considered, it is possible to implement MIMO processing algorithms separately for each frequency bin, where the channel fading characteristic can be considered to be narrow-band. In what follows, we assume a narrow-band channel model, but note that our discussion can be applied to the wideband case using either OFDM or other common techniques for frequency-selective channels.

One additional property of radio propagation channels that must also be considered in the multi-user MIMO context is how they vary with time, particularly for applications that assume mobility of one or both ends of the wireless link. Two likely applications for multi-user MIMO transmission are wireless local area networks (LANs) and cellular telephony. Wireless LANs are a natural fit for MIMO technology because the rich multipath environment in

the places where they are usually deployed (indoors, office or college campuses, etc.) is an important criterion for achieving high capacity. In this type of channel, user mobility is likely to be very slow, and the channel can be viewed as being quasi-static. Cellular telephone applications are more challenging due to higher user mobility, and the small size and cost constraints of manufacturing mobile devices make the use of multiple antennas problematic. While time-varying channel models have been considered in analyzing simple MIMO systems [49–51], most applications assume quasi-static fading. Further research on techniques for obtaining and tracking channel state information is needed for highly mobile scenarios. Recent research suggests that the prediction horizon for MIMO systems may be much longer than in the SISO case (which has usually proven to be too short to be useful), since multiple antennas reveal more information about the physical structure of the channel [52].

Perhaps the most critical assumption common to all of the recent multi-user MIMO research is the availability of CSI at the transmitter. While single-user MIMO systems benefit from having CSI at the transmitter only when $n_T > n_R$ or at low SNR, a base station transmitting to multiple co-channel users will almost always benefit from CSI. This is because the CSI is not only useful in achieving high SNR at the desired receiver, but also in reducing the interference produced at other points in the network by the desired user's signal. The most common method for obtaining CSI at the transmitter is through the use of training or pilot data in the uplink (*e.g.,* for time-division duplex systems) or via feedback of the receiver's channel estimate found using downlink training data (*e.g.,* for frequency-division duplex transmission). In either case, obtaining CSI at the transmitter is a very challenging and costly problem, but appears justifiable for multi-user channels.

1.3   Single Antenna Receivers

We begin our discussion of multi-user MIMO downlink algorithms with the case most commonly treated in the research literature, namely situations involving users with only one receive antenna: $n_{R_j} = 1$. With only one antenna, the receiver is unable to perform any spatial interference suppression of its own, and the transmitter is responsible for precoding the data in such a way that the interference seen by each user is tolerable. In the discussion that follows, we consider four techniques for solving this problem: channel inversion, regularized channel inversion, sphere encoding and iterative interference balancing, or power control.

1.3.1   Channel Inversion

Channel inversion [15, 16] simply amounts to undoing the effects of the channel via precoding; in other words, we precode the data with the (pseudo-)inverse of the channel prior to transmission, as illustrated in Figure 1.2 for the case where **H** is square. More generally, we define

$$\mathbf{s} = \frac{1}{\sqrt{\gamma}}\mathbf{H}^* \left(\mathbf{HH}^*\right)^{-1}\mathbf{d} \,, \tag{1.11}$$

where it is assumed that $n_T \geq K = n_R$. The scaling factor γ is present to limit the total transmitted power to some predetermined value ρ:

$$\|\mathbf{s}\|^2 = \rho \;\Rightarrow\; \gamma = \frac{1}{\rho}\mathbf{d}^*(\mathbf{HH}^*)^{-1}\mathbf{d} \,. \tag{1.12}$$
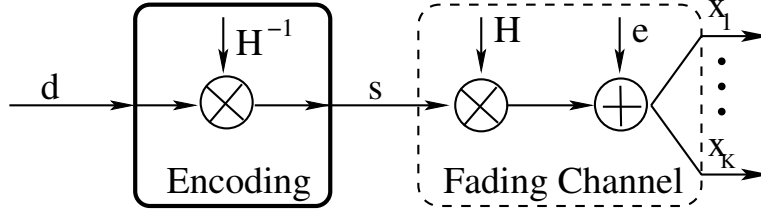
Figure 1.2: Channel Inversion cancels all interference, but requires high power to cancel the small elements of $\mathbf{H}$.

Ideally, all inter-user interference is canceled by this approach, reducing the problem to K separate scalar channels, and each user sees only the desired symbol in additive noise:

$$x_j = \frac{1}{\sqrt{\gamma}} d_j + e_j \; .$$
(1.13)

One issue that may be a problem in practice is the fact that the scaling γ is data-dependent, and will in general change from symbol to symbol. To avoid this problem, γ can be chosen so that the *average* transmit power is ρ, which leads to

$$\gamma = \frac{1}{\rho} \text{trace} \left[(\mathbf{HH}^*)^{-1} \right]$$
(1.14)

if the users' symbols are independent and have average unit power.

Obviously, a more serious problem arises if the channel is ill-conditioned. In such cases, at least one of the singular values of $(\mathbf{HH}^*)^{-1}$ is very large, γ will be large, and the SNR at the receivers will be low. It is interesting to contrast channel inversion with least-squares or "zero-forcing" (ZF) receive beamforming, which applies a dual of the transformation in (1.11) to the receive data. Such beamformers are well-known to cause noise amplification when the channel is nearly rank deficient. Here, on the transmit side, ZF produces signal attenuation instead. In fact, as shown in [6], the problem is very serious, even for what one might consider the "ideal" case, *i.e.,* where the elements of $\mathbf{H}$ are independent, identically distributed Rayleigh random variables. If the elements of $\mathbf{d}$ are modeled as independent zero-mean unit-variance Gaussian random variables, it can be shown [6] that the probability density function of γ is given by

$$p(\gamma) = K \frac{\gamma^{K-1}}{(1+\gamma)^{K+1}} \; ,$$
(1.15)

when $n_T = K = n_R$, and γ has an infinite mean! As a consequence, the capacity of channel inversion does not increase linearly with K, unlike the capacity bound.

1.3.2   Regularized Channel Inversion

When rank-deficient channels are encountered in ZF receive beamforming, a common approach to reducing the effects of noise amplification is to regularize the inverse in the ZF

filter. If the noise is spatially white and an appropriate regularization value is chosen, this approach is equivalent to using a minimum mean-squared error (MMSE) criterion to design the beamformer weights. Applying this principle to the transmit side suggests the following solution:

$$\mathbf{s} = \frac{1}{\sqrt{\gamma}} \mathbf{H}^* \left(\mathbf{H}\mathbf{H}^* + \zeta \mathbf{I} \right)^{-1} \mathbf{d} , \tag{1.16}$$

where ζ is the regularization parameter. The presence of a non-zero value for ζ will mean that the transmit beamformer does not exactly cancel the "mixing" effect of the channel, resulting in some level of inter-user interference. The key is to define a value for ζ that optimally trades off the numerical condition of the matrix inverse (which impacts the normalization required for the power constraint) against the amount of interference that is produced. In [6], it is shown that choosing $\zeta = K/\rho$ approximately maximizes the SINR at each receiver, and unlike standard channel inversion, leads to linear capacity growth with K.
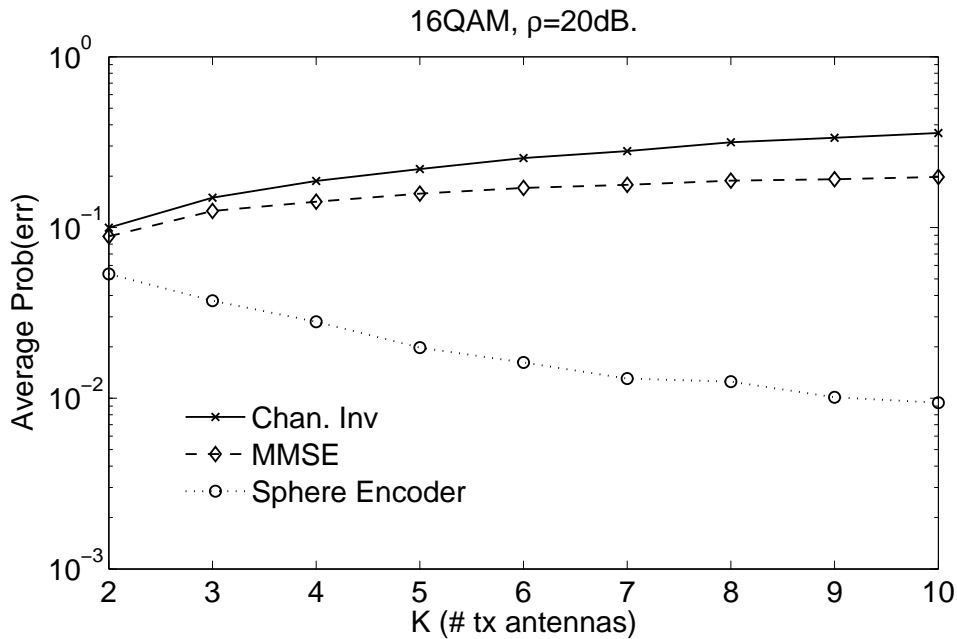

Figure 1.3: Comparing uncoded symbol error rates for standard and regularized channel inversion for $\rho = 20$dB as a function of K. The performance of channel inversion decreases with K, while regularized inversion improves slightly at high transmit power. The performance of the sphere encoder described in Section 1.3.3 is also shown.

Figures 1.3 and 1.4 compare respectively the symbol error rates and capacity of standard and regularized channel inversion. Figure 1.3 shows average error rates as a function of K for $\rho = 20$dB SNR and a 16-QAM signaling (the SNR is defined as ρ since the elements of **e** are assumed to have unit power). The elements of the channel matrices were simulated as independent, unit-variance Rayleigh random variables. Note that the performance of standard

channel inversion, as well as regularized inversion degrades as K increases; the performance of the "sphere encoder," which will be described in Section 1.3.3, improves with K. Figure 1.4 plots capacity as a function of K assuming $n_T = K = n_R$ and $\rho = 10$dB. The plot also shows that there is still a considerable gap between the performance of regularized inversion and the sum capacity of the system.
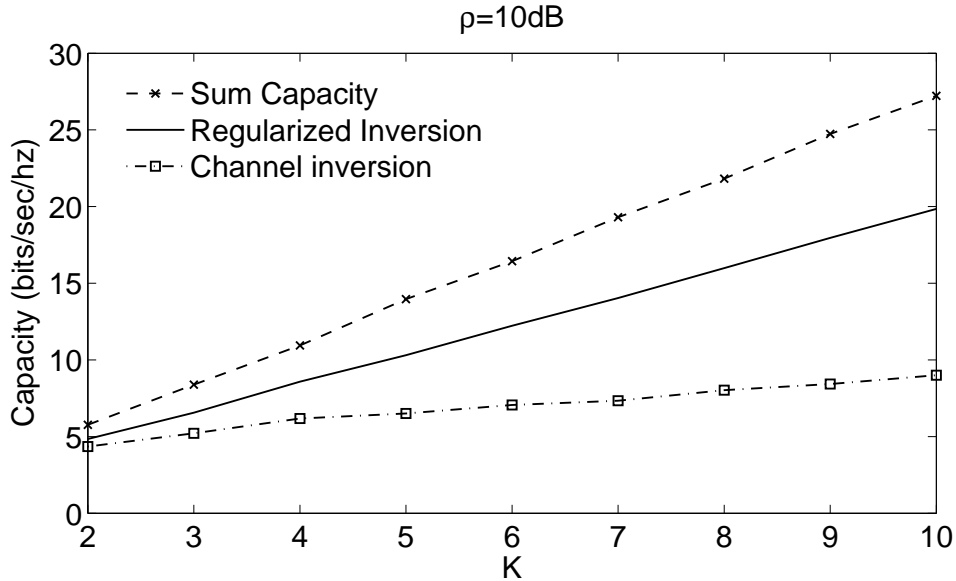

Figure 1.4: Comparison of the sum-capacity (dashed line) as a function of K (where $n_T = K$) for $\rho = 10$dB with the regularized channel inversion sum-rate (solid line) and the standard channel inversion sum-rate (dash-dotted line).

1.3.3   Sphere Encoding

The simulation results of the previous section indicate that channel inversion techniques are not capacity optimal. As mentioned above, dirty-paper coding (DPC) techniques more closely approach (and in some cases achieve) multi-user capacity, and thus may be of interest when capacity is the primary design criterion. DPC is different from other downlink approaches in that the transmitted data is a non-linear function of the information symbols, as well as the interference environment. For this reason, DPC is sometimes referred to as interference-depending coding. Due to their non-linear nature and their need for high-dimensional lattices, DPC techniques are often difficult to implement in practice.

Technically, DPC codes do not constitute beamforming *per se*, but they can be used in conjunction with beamforming as illustrated below. In this section we present a simple DPC technique that fits in well with the channel inversion algorithms already discussed. Figure 1.5 illustrates the approach we will consider, which is referred to as *vector precoding*. As discussed above, channel inversion performs poorly because the scaling factor γ in (1.12) can
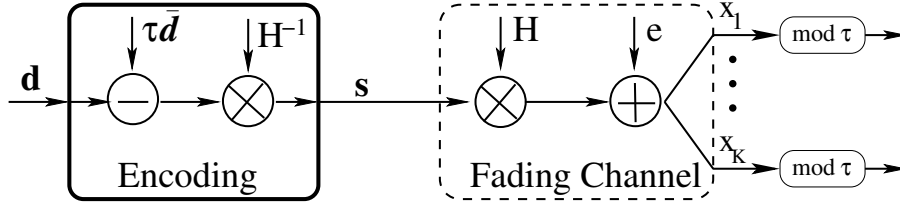
(a) Modulo Vector Precoding.
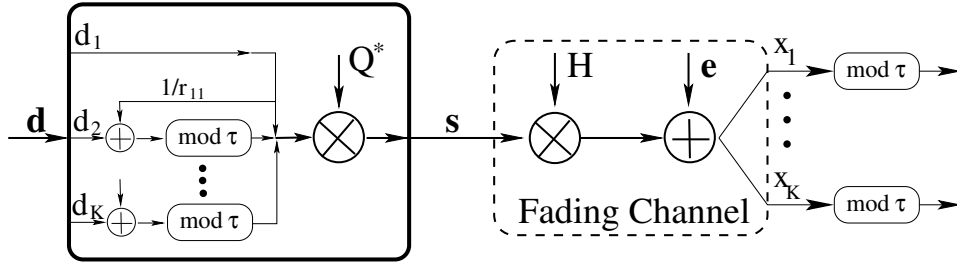
(b) QR−based, Successive Precoding.

Figure 1.5: Part (a) shows the vector precoding technique; a vector chosen to minimize the signal power is added to the data to be transmitted. Part (b) shows that QR-based techniques successively cancel interference from previous users.

be large when the channel is ill-conditioned, and the vector $\mathbf{d}$ happens to (nearly) align itself with a right singular vector of $(\mathbf{HH}^*)^{-1}$ with large singular value. The idea behind the technique proposed in [8, 17] is to "perturb" the symbol vector $\mathbf{d}$ by some value $\bar{\mathbf{d}}$ such that $\mathbf{d} + \bar{\mathbf{d}}$ is directed towards singular vectors of $(\mathbf{HH}^*)^{-1}$ with smaller singular values, and in such a way that the receivers can still decode $\mathbf{d}$ without knowledge of $\bar{\mathbf{d}}$. In particular, [8, 17] constrains $\bar{\mathbf{d}}$ to lie on a (complex) integer lattice:

$$\bar{\mathbf{d}} = \tau(\mathbf{a} + j\mathbf{b}) , \qquad (1.17)$$

where $\mathbf{a}, \mathbf{b}$ are vectors of integers and τ is a real-valued constant, and calculates $\bar{\mathbf{d}}$ based on the following optimization problem:

$$\bar{\mathbf{d}} = \arg\min_{\bar{d}} \ (\mathbf{d} + \bar{\mathbf{d}})^*(\mathbf{HH}^*)^{-1}(\mathbf{d} + \bar{\mathbf{d}})$$
$$\text{s.t. } \bar{\mathbf{d}} = \tau(\mathbf{a} + j\mathbf{b}) . \qquad (1.18)$$

This is an integer-lattice least-squares problem, and can be solved using standard sphere algorithm methods [53–55], and other related techniques [56, 57]. Since it is used on the transmit side for this application, in [8, 10, 17] it is referred to as *sphere encoding* or *sphere precoding*. Using this method, the vector of data at the receivers is given by

$$\mathbf{x} = \frac{1}{\sqrt{\gamma}}\mathbf{d} + \frac{1}{\sqrt{\gamma}}\tau(\mathbf{a} + j\mathbf{b}) + \mathbf{e} , \qquad (1.19)$$

where, as before, γ is chosen to maintain a constant (average) transmit power ρ.

To eliminate the contribution of the vector perturbation, the receivers employ the modulo function (1.9). If τ and γ (or $\mathcal{E}\{\gamma\}$) are known at the receivers, then in the absence of noise,

$$f_\tau \left(\sqrt{\gamma} x_j \right) = f_\tau(d_j + \tau a_j + \tau b_j) = d_j . \tag{1.20}$$

Small values of τ are advantageous because they allow for a denser perturbation lattice, and hence more flexibility in maximizing received SINR. However, τ must be chosen large enough to allow for unambiguous decoding. In [8, 17], it is suggested that τ be chosen as

$$\tau = 2(d_{max} + \Delta/2) , \tag{1.21}$$

where d_{max} is the distance from the origin to the farthest constellation point, and Δ is the maximum distance between any two constellation points.

A simple numerical example is now presented to illustrate the algorithm. For simplicity, we consider the special case of binary pulse-amplitude modulated (PAM) signaling over real-valued channels with $K = 2$ and $\rho = 1$. A near-singular channel matrix is chosen to illustrate the benefit of non-linear precoding:

$$\mathbf{H} = \begin{bmatrix} -0.0521 & 0.17 \\ -0.661 & 1.80 \end{bmatrix} . \tag{1.22}$$

Suppose the data to be transmitted is $\mathbf{d} = [-1, \ 1]^T$, with noise $\mathbf{e} = [0.011, \ 0.001]^T$. Simple channel inversion gives the signal $\mathbf{H}^{-1}\mathbf{d} = [-106, \ -38.4]^T$, which results in $\gamma = 12,700$. Transmitting the normalized signal through the channel results in $\mathbf{x} = [0.00213, \ 0.00987]^T$, which gives decoded PAM symbols $[1, \ 1]$ when using the sign of the elements of $\mathbf{x}$ to decode. In contrast, sphere encoding results in $\bar{\mathbf{d}} = \tau[0, \ 2]^T$, which results in a signal with a more attractive $\gamma = 36.5$. For our 2-PAM constellation, we choose $\tau = 4$, resulting in the received signal $\mathbf{x} = [-0.154, \ -1.82]^T$ and $f_\tau(\mathbf{x}) = [-0.154, \ 0.166]^T$. In this case decoding based on $f_\tau(\mathbf{x})$ returns the correct symbols $[-1, \ 1]^T$.

Figure 1.6 shows a plot of the uncoded symbol error probability of the algorithms discussed thus far for a case with $n_T = 10$, $n_R = 10$, a Rayleigh fading channel and QPSK signaling. "Sphere Encoder" denotes the modulo precoding algorithm described above, and "Reg. Sphere Encoder" refers to the use of vector modulo precoding together with regularized channel inversion. Regularization improves performance, but by a smaller margin than in the case of standard channel inversion. It is clear from the plot that, for SNRs high enough to achieve reliable decoding, modulo precoding offers a significant improvement in performance over channel inversion and regularized inversion. The modulo precoding technique presented here represents perhaps the simplest form of DPC for the multi-user MIMO problem, *i.e.,* one involving a simple cubical lattice. As reported in [8], improved performance can be expected if more complicated, higher-dimensional lattices [39, 58] are employed. In particular, these techniques may improve the low-SNR performance of the modulo precoding techniques, which perform slightly worse than their linear counterparts in Figure 1.6. Finally, we note that a suboptimal but more computationally efficient version of the modulo precoding algorithm has recently been presented in [9].

A full analysis of the algorithm appears difficult, due in part to the difficulty in obtaining a distribution for γ. We focus instead on understanding the performance gains seen in our
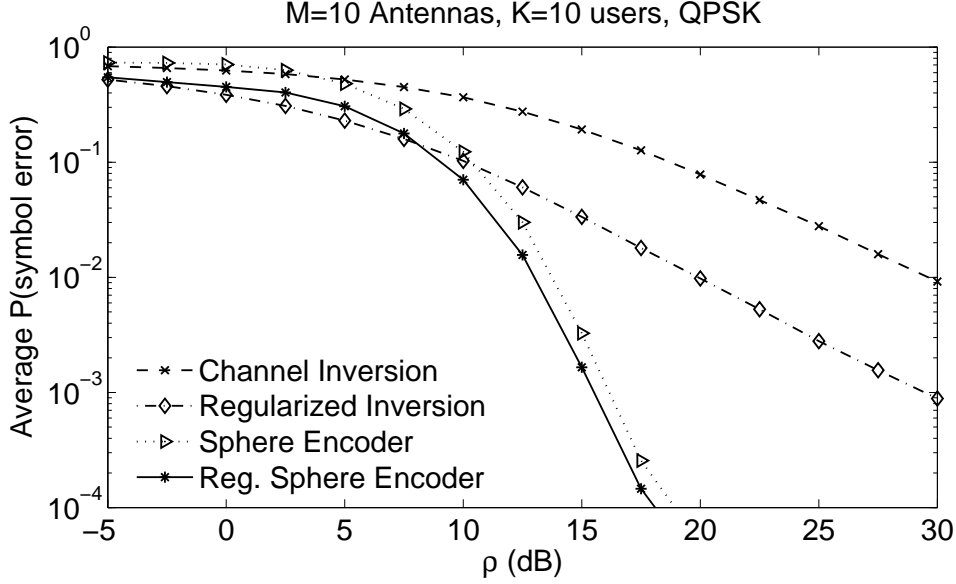
Figure 1.6: Uncoded probability of symbol error for various downlink algorithms as a function of transmit power ρ.

simulations. The precoding process aligns $\mathbf{d} - \bar{\mathbf{d}}$ with the singular values of the inverse channel. Let $\mathbf{UDV}^* = \mathbf{H}^{-1}$ be the singular value decomposition (SVD) of the channel inverse, so that $\boldsymbol{\delta}_{CI} = |\mathbf{V}^*\mathbf{d}|$ is the channel-inversion data vector rotated by the right singular vectors of $\mathbf{H}^{-1}$, and $\boldsymbol{\delta}_{SE} = |\mathbf{V}^* (\mathbf{d} - \bar{\mathbf{d}})|$, is the equivalent for the sphere encoded data. The sphere encoder minimizes the cost function

$$\gamma = \|\boldsymbol{\delta}_{SE}^* \boldsymbol{\sigma}\|^2 \tag{1.23}$$

over $\bar{\mathbf{d}}$, where $\boldsymbol{\sigma}$ is a vector containing the singular values of $\mathbf{H}^{-1}$. For channel inversion we have $\gamma = \|\boldsymbol{\delta}_{CI}^* \boldsymbol{\sigma}\|^2$. In [6] it is shown that $\mathcal{E}\{\gamma\} = \infty$ for plain channel inversion. In contrast, for the sphere encoder, $\mathcal{E}\{\gamma\}$ is shown in [8, 17] to be approximately constant with K. Specifically, $\mathbf{d} - \bar{\mathbf{d}}$ is chosen to orient itself toward each singular vector in inverse proportion to the singular value of the inverse channel matrix:

$$\mathcal{E}\{\sigma_1 \delta_1\} = \ldots = \mathcal{E}\{\sigma_K \delta_K\} . \tag{1.24}$$

We illustrate this in Figure 1.7, where $\delta_k \sigma_k$ are shown averaged over 10000 samples for a case where $K = n_t = 10$ users and a 16-QAM constellation is employed. While basic channel inversion does not modify or perturb the transmitted symbols, the sphere encoding technique attempts to orient the symbol vector towards each singular vector in inverse proportion to the singular values $\boldsymbol{\sigma}$. For comparison we also show results for regularized inversion and regularized vector precoding, both with $\zeta = K/\rho$; for these curves the $\boldsymbol{\sigma}$ is obtained from the regularized inverse.
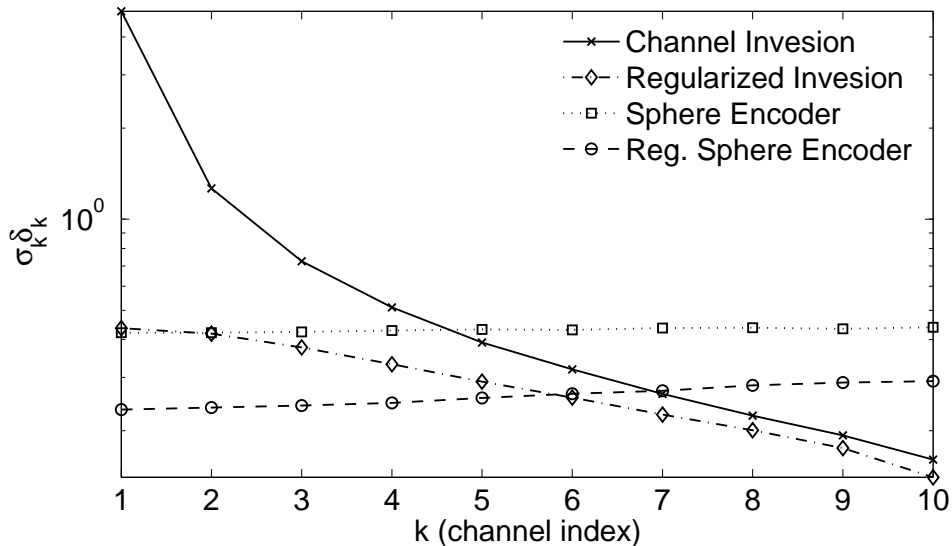
Figure 1.7: The integer offset vector $\bar{\mathbf{d}}$ is chosen such that it orients itself towards each singular vector in inverse proportion to the singular values $\boldsymbol{\sigma}$. Values of $\delta_k \sigma_k$ are shown, averaged over 1000 samples.

1.3.4 Computationally Efficient Precoding

Though the vector precoding technique in the previous section is very powerful, it is somewhat expensive computationally. There are several ways to increase the speed of the integer least-squares search, including successive algorithms based on the QR and V-BLAST decompositions, and on the use of lattice reduction algorithms. We will present a simplified technique which generates the integer offset $\bar{\mathbf{d}}$ by repeated application of a modulo operation inspired by scalar Tomlinson-Harashima Precoding (THP) [59, 60]. The method uses a QR decomposition of the channel matrix $\mathbf{H}$, where the resulting triangular structure leads to the kth user seeing interference only from users $1, \ldots, k-1$. The transmitter compensates for this interference by using its knowledge of $s_1, \ldots, s_{k-1}$ to generate s_k from u_k, for $k = 2, \ldots, K$. Methods based on the QR decomposition have been explored for use with DPC codes in [1]. Similar algorithms have been used for crosstalk cancellation in digital subscriber lines [61] and for CDMA transmission to distributed receivers [62]. The achievable capacity of a greedy form of this scheme is analyzed in [11], where it is shown to be close to the sum-capacity.

Let $\mathbf{H} = \mathbf{R}\mathbf{Q}$, where $\mathbf{R}$ is a lower triangular matrix, and $\mathbf{Q}$ is a unitary matrix; let $\mathbf{D}$ be a diagonal matrix composed of the diagonal entries of $\mathbf{R}$, and $\alpha = \rho/(1+\rho)$. We first generate the signal $\tilde{\mathbf{s}}$, and then form the transmitted signal $\mathbf{s} = \mathbf{Q}^*\tilde{\mathbf{s}}$. Because the matrix $\mathbf{R} - \mathbf{D}$ is

zero on and above the diagonal, a successive technique can be used to generate $\tilde{\mathbf{s}}$:

$$\tilde{s}_1 = d_1$$

$$\tilde{s}_2 = f_\tau \left(d_2 - \alpha \frac{r_{21}}{r_{22}} \tilde{s}_1 \right)$$

$$\vdots$$

$$\tilde{s}_K = f_\tau \left(d_K - \alpha \sum_{i=1}^{K-1} \frac{r_{Ki}}{r_{KK}} \tilde{s}_i \right) , \tag{1.25}$$

where d_k is the kth diagonal entry of D, and r_{ij} is the entry on the ith row and in the jth column of $\mathbf{R}$. We may write this equation in terms of the vector of integers $\bar{\mathbf{d}}$ that the modulo function effectively adds to the signal:

$$\tilde{\mathbf{s}} = \left((1-\alpha)I + \alpha \mathbf{D}^{-1}\mathbf{R} \right)^{-1} \left(\mathbf{d} + \tau\bar{\mathbf{d}} \right) . \tag{1.26}$$

The signal $\mathbf{s} = \mathbf{Q}^*\tilde{\mathbf{s}}$ is formed, normalized, and then sent through the channel. The K users receive

$$\mathbf{x} = \frac{1}{\sqrt{\gamma}} \mathbf{H}\mathbf{s} + \mathbf{e} = \frac{1}{\sqrt{\gamma}} \mathbf{R} \left((1-\alpha)I + \alpha \mathbf{D}^{-1}\mathbf{R} \right)^{-1} \left(\mathbf{d} + \tau\bar{\mathbf{d}} \right) + \mathbf{e}.$$

The parameter α increases the SINR for each user, similar to what the parameter ζ does with regularized inversion in Section 1.3.2. Let y_k be the data received at the kth user; decoding occurs at user k based on [1, 39]

$$y'_k = f_{\tilde{\tau}_k} (\alpha y_k) ,$$

where $\tilde{\tau}_k = \tau \frac{r_{kk}}{\sqrt{\gamma}}$. Each receiver models the received data as

$$y'_k = \alpha \frac{r_{kk}}{\sqrt{\gamma}} d_k + w'_k ,$$

where w'_k combines the additive receiver noise w_k and the interference. We do not analyze the algorithm but simply mention that at high ρ (where $\alpha \to 1$),

$$\mathbf{s} = \mathbf{Q}^*\mathbf{R}^{-1}\mathbf{D} \left(\mathbf{d} + \tau\bar{\mathbf{d}} \right) , \tag{1.27}$$

which yields

$$y'_k = f_{\tilde{\tau}_k} \left(\frac{r_{kk}}{\sqrt{\gamma}} d_k + \tau l_k + w_k \right) . \tag{1.28}$$

When $\alpha = 1$ there is no interference at user k from the other users' signals. The algorithm described here differs slightly from those of [61, 62] in our introduction of the parameter α. However, our use of α is well-known in the DPC literature [1, 39], including in the original paper by Costa [43]. An important characteristic of these QR-based algorithms is that they do not achieve full diversity, although they do provide a significant computational advantage. For $K = N_R$, the complexity is of order K^2. This is much less that for typical integer least-squares algorithms, which have expected complexity of order K^3 as described in [63]. A high peak-to-average power ratio can cause clipping in power amplifiers and accuracy problems

in processors with limited wordlength. We note that the technique described here explicitly limits the peak signal strength; further research is needed on the use of a modulo function to minimize the peak-to-average power ratio.

A V-BLAST-type ordering of the users can also be applied to the QR decomposition [9, 64], with similar complexity in the resulting precoding algorithm. In this case a permutation of the users is obtained as part of the V-BLAST decomposition; the rest of the algorithm is similar to that above for the QR-based techniques. A MMSE V-BLAST decomposition returns $\mathbf{Q}$ and $\mathbf{R}$ matrices that are no longer orthogonal and triangular, respectively. Some interference is allowed at each user in order to increase the overall SINR. For a full description and analysis of the basic V-BLAST technique, see [64]. The MMSE version of this technique performs especially well [65]; however, it also does not achieve the full spatial/multi-user diversity available. This can be easily seen by noting that at least one channel is processed linearly; at high SNR this channel will give an upper bound on performance and result in the same diversity as the linear techniques described in Section 1.3.1.

In contrast to the above techniques, an algorithm for finding the integer offset vector using the Lenstra-Lenstra-Lovász (LLL) algorithm [66] does give full diversity; at high SNR the slope of the error curves for this technique are the same as for vector precoding. The LLL matrix decomposition of the channel results in an integer matrix with unit determinant, and a reduced matrix $\mathbf{B}$:

$$\mathbf{H} = \mathbf{BT}, \quad \text{such that} \quad |\mathbf{T}| = \pm 1 \ . \tag{1.29}$$

Thus $\mathbf{B}$ has the same determinant as $\mathbf{H}$. This decomposition can be used to obtain the Babai estimate of the integer offset:

$$\bar{\mathbf{d}} = \tau \mathbf{T} \left\lfloor \frac{\mathbf{T}^{-1} \mathbf{d}}{\tau} \right\rceil \ . \tag{1.30}$$

MMSE versions of this algorithm exist for the uplink [67], though they have not been explored as much in the downlink setting.

1.3.5 Power Control

As mentioned above, sum rate maximization in "near-far" scenarios may result in one or two strong users taking a dominant share of the available power, potentially leaving weak users with little or no throughput. Consequently, in practice, the dual *power control* problem is often of more interest, *i.e.,* minimizing power output at the transmitter subject to achieving a desired QoS for each user. We illustrate this approach below for the case where QoS is measured in terms of SINR. Assuming linear transmit beamformers and unit-power data symbols and noise, the SINR for user j can be expressed as:

$$\text{SINR}_j \ = \frac{\mathbf{b}_j^* \mathbf{R}_j \mathbf{b}_j}{\sum_{k \neq j} \mathbf{b}_k^* \mathbf{R}_k \mathbf{b}_k + 1} \ , \tag{1.31}$$

where either $\mathbf{R}_j = \mathbf{h}_j \mathbf{h}_j^*$ or $\mathbf{R}_j = \mathcal{E}\{\mathbf{h}_j \mathbf{h}_j^*\}$ depending on the type of CSI available at the transmitter.

Given a desired minimum SINR for each user, which we denote by η_j, the power control

problem can be formulated as follows:

$$\min_{b_1,\cdots,b_K} \sum_{k=1}^{K} \mathbf{b}_k^* \mathbf{b}_k$$

$$\text{s.t. } \frac{\mathbf{b}_j^* \mathbf{R}_j \mathbf{b}_j}{\sum_{k \neq j} \mathbf{b}_k^* \mathbf{R}_k \mathbf{b}_k + 1} \geq \eta_j \ , \ \ j = 1, \cdots, K \ . \tag{1.32}$$

In [23, 24], iterative algorithms are presented that solve this problem when a feasible solution exists (*i.e.,* if the SINR constraints can be met). An alternative formulation of the problem is presented in [25], where (1.32) is recast as a minimization over the matrices $\mathbf{W}_j = \mathbf{b}_j \mathbf{b}_j^*$ rather than the beamformers $\mathbf{b}_j$ directly. It is shown that the constraint that $\mathbf{W}_j$ be rank one can be relaxed, and the resulting optimization problem will still have an optimal rank-one solution. The advantage of this approach is that the problem becomes a semidefinite optimization, for which efficient numerical algorithms exist. While the above discussion has focused on linear beamforming, an approach based on vector precoding would be a natural extension of this work.

1.4   Multiple Antenna Receivers

With only a single antenna, the users in the network can perform no spatial interference suppression of their own, and can only receive data over a single spatial channel. With multiple antennas, these restrictions are removed, provided that the transmitter and receiver can coordinate their spatial processing, and appropriately allocate the available spatial resources. In this chapter, we present several methods that take advantage of the presence of multiple antennas at the receivers for increased throughput, enhanced interference suppression, or both.

1.4.1   Channel Block-Diagonalization

The single-antenna techniques of the previous section could be directly applied in the multiple-antenna receiver case, provided that $n_R \leq n_T$, *i.e.,* the number of transmit antennas is greater than the number of receive antennas summed over all the users. In such cases, each receive antenna is considered to be a separate "user," and each transmitted data stream is decoded independently on each receive antenna as if it were a SISO channel. As mentioned above, while this approach results in a very simple receiver, it overly constrains the problem and will lead to suboptimal performance.

Rather than forcing $\mathbf{HB}$ in (1.2) to be diagonal (or nearly so), an alternative is to make it block-diagonal [7, 26–30]. This removes inter-user interference, but requires that the receiver perform some type of spatial demultiplexing to separate and decode the individual data streams sent to it. To be precise, the goal is to find $\mathbf{B}$ such that

$$\mathbf{HB} = \begin{bmatrix} \mathbf{M}_1 & & \\ & \ddots & \\ & & \mathbf{M}_K \end{bmatrix}, \tag{1.33}$$

where $\mathbf{M}_j$ is $n_{R_j} \times n_{R_j}$ assuming that up to n_{R_j} data streams are transmitted to user j (some of the columns of $\mathbf{M}_j$ could be zero so that $m_j \leq n_{R_j}$). There are several criteria that could

be used to determine $\mathbf{M}_j$. Below, we present an algorithm that is sum-capacity-achieving under the block-diagonal constraint [7].

Define $\tilde{\mathbf{H}}_j$ as the following $(n_R - n_{R_j}) \times n_T$ matrix:

$$\tilde{\mathbf{H}}_j = \left[\, \mathbf{H}_1^T \, \cdots \, \mathbf{H}_{j-1}^T \, \mathbf{H}_{j+1}^T \, \cdots \, \mathbf{H}_K^T \,\right]^T \,. \tag{1.34}$$

If we denote the rank of $\tilde{\mathbf{H}}_j$ as $\tilde{L}_j$, then the nullspace of $\tilde{\mathbf{H}}_j$ has dimension $n_T - \tilde{L}_j \geq n_{R_j}$. The SVD of $\tilde{\mathbf{H}}_j$ is partitioned as follows:

$$\tilde{\mathbf{H}}_j = \tilde{\mathbf{U}}_j \tilde{\boldsymbol{\Sigma}}_j \left[\, \tilde{\mathbf{V}}_j^{(1)} \ \tilde{\mathbf{V}}_j^{(0)} \,\right]^* \,, \tag{1.35}$$

where $\tilde{\mathbf{V}}_j^{(0)}$ holds the $n_T - \tilde{L}_j$ singular vectors in the nullspace of $\tilde{\mathbf{H}}_j$. The columns of $\tilde{\mathbf{V}}_j^{(0)}$ are candidates for user j's beamforming matrix $\mathbf{B}_j$, since they will produce zero interference at the other users. Since $\tilde{\mathbf{V}}_j^{(0)}$ potentially holds more beamformers than the number of data streams that user j can support, an optimal linear combination of these vectors must be found to form $\mathbf{B}_j$, which can have at most n_{R_j} columns. To do this, the following SVD is formed:

$$\mathbf{H}_j \tilde{\mathbf{V}}_j^{(0)} = \left[\mathbf{U}_j^{(1)} \ \mathbf{U}_j^{(0)}\right] \left[\begin{array}{cc} \boldsymbol{\Sigma}_j & 0 \\ 0 & 0 \end{array}\right] \left[\, \mathbf{V}_j^{(1)} \ \mathbf{V}_j^{(0)} \,\right]^* \,, \tag{1.36}$$

where $\boldsymbol{\Sigma}_j$ is $L_j \times L_j$ and $\mathbf{V}_j^{(1)}$ represents the L_j singular vectors with non-zero singular values. The $L_j \leq n_{R_j}$ columns of the product $\tilde{\mathbf{V}}_j^{(0)} \mathbf{V}_j^{(1)}$ represent (to within a power loading factor) the beamformers that maximize the information rate for user j subject to producing zero inter-user interference.

The transmit beamformer matrix will thus have the following form:

$$\mathbf{B} = \left[\, \tilde{\mathbf{V}}_1^{(0)} \mathbf{V}_1^{(1)} \ \cdots \ \tilde{\mathbf{V}}_K^{(0)} \mathbf{V}_K^{(1)} \,\right] \boldsymbol{\Lambda}^{1/2} \,, \tag{1.37}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements scale the power allocated to each "sub-channel." With $\mathbf{B}$ chosen as in (1.37), the capacity of the block-diagonalization (BD) method becomes

$$C_{BD} = \max_{\boldsymbol{\Lambda}} \log_2 \left|\mathbf{I} + \boldsymbol{\Sigma}^2 \boldsymbol{\Lambda}\right| \quad \text{s.t.} \quad \mathrm{Tr}(\boldsymbol{\Lambda}) = \rho \,, \tag{1.38}$$

where

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc} \boldsymbol{\Sigma}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_K \end{array}\right] \,. \tag{1.39}$$

The optimal power loading coefficients in $\boldsymbol{\Lambda}$ are then found using water-filling on the diagonal elements of $\boldsymbol{\Sigma}$. Forcing the inter-user interference to zero also allows for a power control formulation of the above approach. This is done by performing water-filling on each $\boldsymbol{\Sigma}_j$ individually in order to achieve the desired rate for user j, then forming $\boldsymbol{\Lambda}$ from the diagonal matrices that result for each user.

Figure 1.8 illustrates the performance of the BD algorithm and several alternatives for a case involving $n_T = 4$ and $n_R = 4$ with $\rho = 10$dB. The elements of $\mathbf{H}$ were independent

Rayleigh random variables with unit variance, and the cumulative distribution function of the capacity achieved by each method is plotted. The BD algorithm is implemented for three different scenarios: four users with one antenna each ($\{1, 1, 1, 1\} \times 4$), two users with two antennas each ($\{2, 2\} \times 4$), and a single user with 4 antennas (referred to as "1 User" in the figure). "Inversion" refers to channel inversion with equal power distributed to each data stream, and "TDM" refers to the case where no channel information is available and the users are simply time-multiplexed. Note that the difference between channel inversion and BD in the $\{1, 1, 1, 1\} \times 4$ case is due to the fact that BD employs an optimal power allocation via water-filling. The single-user performance is obviously the best, since it does not require the block-diagonal constraint. The improved performance of BD in the $\{2, 2\} \times 4$ case compared with the $\{1, 1, 1, 1\} \times 4$ scenario demonstrates the advantage of relaxing the requirement that the channel be identically diagonalized.
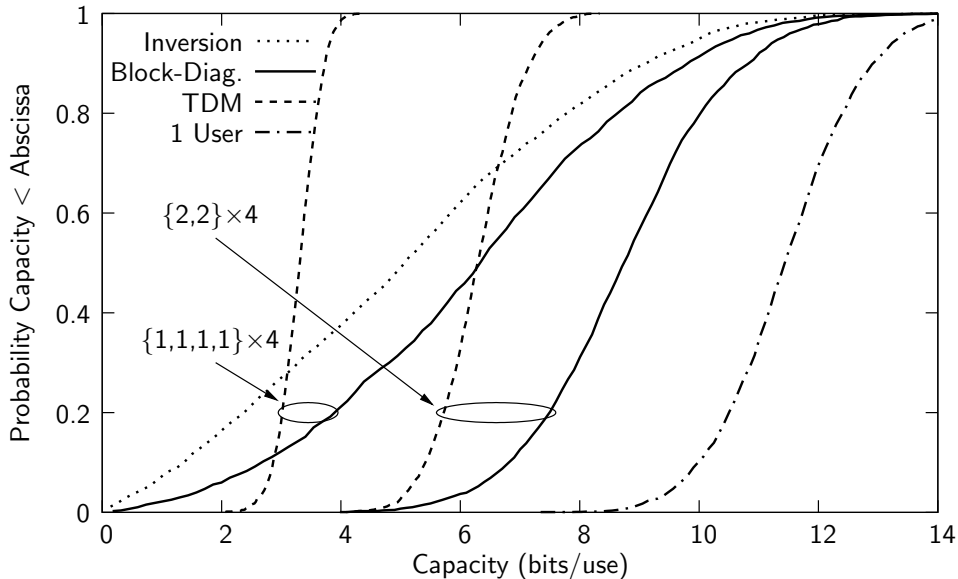

Figure 1.8: Cumulative distributions of the sum capacity for $n_T = n_R = 4$ achieved by several transmit beamforming strategies.

1.4.2 Combined Block Diagonalization and MMSE THP Precoding

Block diagonalization (BD) always outperforms channel inversion, but it is still worse than MMSE THP transmit filtering when the users are equipped with one antenna each. In a scenario where we have multiple users equipped with one or more antennas, the performance of the single-antenna users degrades the overall system performance. In [68] a combination of MMSE THP and block diagonalization is proposed, where MMSE THP is used for the single-antenna users and block diagonalization for the multiple-antenna users. This approach significantly improves the performance of the single-antenna users, and hence also that of the
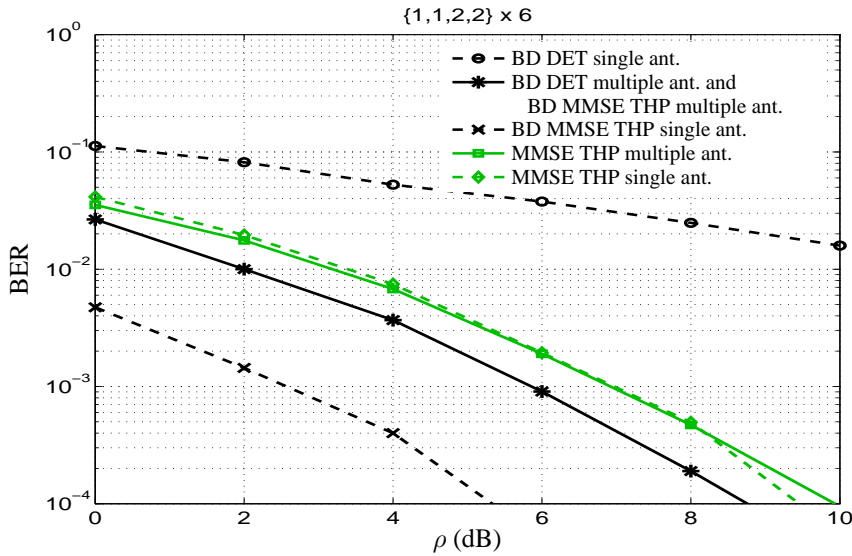
Figure 1.9: Average BER of the single- as well as the multiple-antenna users as a function of the SNR. For the single-antenna users we use either BD or MMSE THP.

overall system.

In the combined BD and MMSE THP approach, the modulation matrices for multiple-antenna users are chosen to lie in the null space of the channel matrices of the other users including those with single antennas. In this way the equivalent channel for the single-antenna users looks as if there are no multiple-antenna users present. This improves diversity for these users, which in turn improves their BER performance. MMSE THP is applied on the network channel corresponding to these single-antenna users. The data transmitted to the multiple-antenna users is also precoded using THP in order to eliminate the multi-user interference which in this case only originates from the single-antenna users.

Figures 1.9 and 1.10 illustrate the performance of this combined technique. In the simulations, the data rate for each user is assumed to be proportional to the number of receive antennas. In Figure 1.9 we compare the performance of the multiple- and single-antenna users in a system with the configuration $\{1, 1, 2, 2\} \times 6$. The BER performance of the single-antenna users is represented using dashed lines. Here we compare the following algorithms: BD with dominant eigenmode transmission (DET), MMSE THP, and BD MMSE THP. In the case of MMSE THP, both single- and multiple-antenna users have similar performance. For BD DET and BD MMSE THP, there is a difference between the performance of the single- and multiple-antenna users. From the results for BD DET and BD MMSE THP, where the only difference is that in the first case we use BD for single-antenna users and in the second case we use MMSE THP, we can see that MMSE THP clearly outperforms BD for single-antenna users. In Fig. 1.10, the overall system performance is depicted comparing MMSE THP, BD, and the combination of BD and MMSE THP. The combination of BD and MMSE
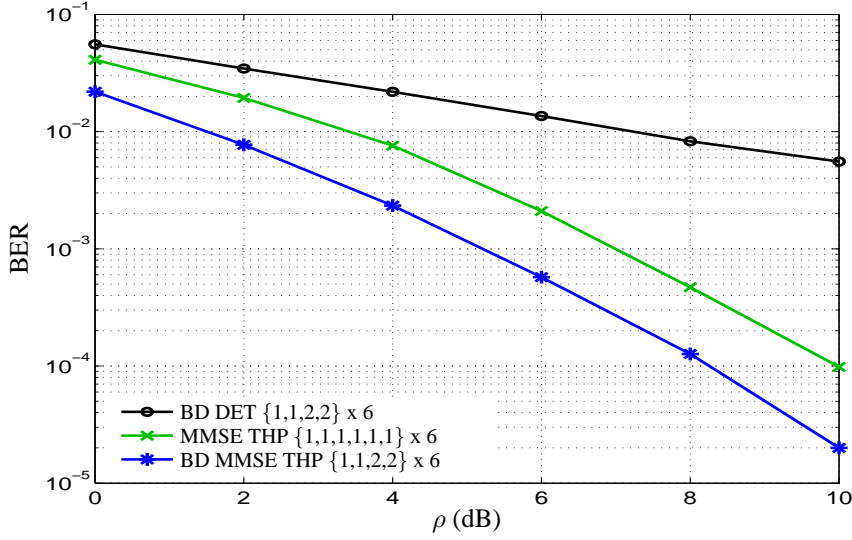
Figure 1.10 Overall BER of all users as a function of the SNR.

THP achieves an SNR gain of about 2 dB with respect to MMSE THP.

1.4.3 Coordinated Tx/Rx Beamforming

Strictly speaking, the BD algorithm does not require $n_T \geq n_R$. However, when there are more than just a couple of users, n_R is usually close to the lower bound on the number of transmit antennas.[1] In this section, we examine methods that have a less stringent constraint on n_T, namely that n_T be no smaller than the total number of data streams to be transmitted. For example, if $m_j = 1$ for all j, then $n_T \geq K$ would be required. Obviously, in a real system where the total number of users serviced by a basestation is very large, spatial multiplexing must be augmented by other multiple access techniques such as time and frequency multiplexing. A key question is how to best group the K users to be spatially multiplexed together into a given time/frequency slot.

To begin, consider the case where $m_j = 1$, and each receiver uses a beamformer $\mathbf{w}_j$ in

---

[1]Technically, the BD approach requires

$$n_T > \max \left\{ \text{rank}(\tilde{\mathbf{H}}_1), \cdots, \text{rank}(\tilde{\mathbf{H}}_K) \right\} \ .$$

Since $\text{rank}(\tilde{\mathbf{H}}_j) \leq n_R - n_{R_j}$, it is clear that n_R can be larger than n_T. For example, two users with 3 antennas each could be accommodated by a transmit array with no more than 4 antennas, and possibly fewer depending on the rank of $\tilde{\mathbf{H}}_1$ and $\tilde{\mathbf{H}}_2$.
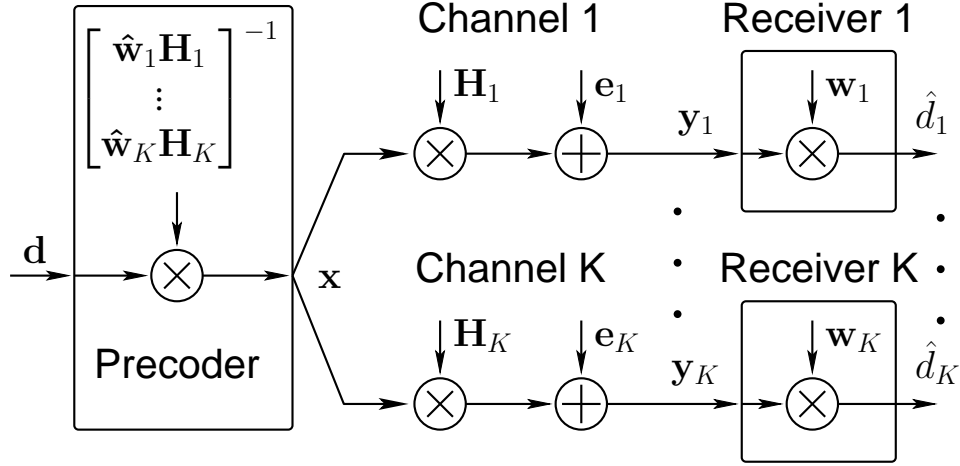
Figure 1.11: An illustration of coordinated transmitter-receiver beamforming, where the transmitter estimates what beamformers the receivers are using, creates a "virtual channel" matrix with one row per user, and uses channel inversion to create the transmit-side beamformers.

decoding the symbol d_j that is sent to it:

$$\bar{x}_j = \mathbf{w}_j^* \mathbf{x}_j = \sum_{k=1}^{K} \mathbf{w}_j^* \mathbf{H}_j \mathbf{b}_k d_k + \mathbf{w}_j^* \mathbf{e}_j \tag{1.40}$$

$$= \sum_{k=1}^{K} \bar{\mathbf{h}}_j^* \mathbf{b}_k d_k + \bar{e}_j \ , \tag{1.41}$$

where $\bar{\mathbf{h}}_j^* = \mathbf{w}_j^* \mathbf{H}_j$ represents the effective channel from the transmit array to the output of the receive beamformer, and $\bar{e}_j = \mathbf{w}_j^* \mathbf{e}_j$ represents the noise at the output of the receive beamformer. If we define $\bar{\mathbf{H}}^* = \begin{bmatrix} \bar{\mathbf{h}}_1 & \cdots & \bar{\mathbf{h}}_K \end{bmatrix}$, then we obtain an equation identical in form to (1.2):

$$\bar{\mathbf{x}} = \bar{\mathbf{H}}\mathbf{B}\mathbf{d} + \bar{\mathbf{e}} \ . \tag{1.42}$$

Each receiver has a single element of $\bar{\mathbf{x}}$ associated with it, so (1.42) has the same dimensions as (1.2) when $n_{R_j} = 1$. The implication is that, if the transmitter somehow has knowledge of $\mathbf{w}_1, \cdots, \mathbf{w}_K$, then it knows $\bar{\mathbf{H}}$, and hence any of the downlink algorithms in Section 1.3 for the single-antenna-per-user case could be used. The coordinated transmit-receive beamforming technique is illustrated in Figure 1.11.

The composite channel $\bar{\mathbf{H}}$ could be estimated directly by the transmitter using uplink training data in a reciprocal time-division duplex (TDD) system, assuming that the receiver will use the conjugate of its transmit weights for downlink reception. However, this approach begs the question of how the receiver chose its beamformer, and whether or not any type of optimal solution is possible. An alternative is to assume the basestation knows what algorithm each receiver uses in computing its own "optimal" receive beamformer. Since the

base generates the interference that each user sees, given CSI it can predict what each user's beamformer will be. For example, suppose it is known that user j employs MMSE receive beamforming. Then

$$\mathbf{w}_j = \left[\mathcal{E} \left\{ \mathbf{x}_j \mathbf{x}_j^* \right\} \right]^{-1} \mathcal{E} \left\{ \mathbf{x}_j d_j^* \right\} \tag{1.43}$$

$$= \left[\sum_{k \neq j} \mathbf{H}_j \mathbf{b}_k \mathbf{b}_k^* \mathbf{H}_j^* + \mathbf{I} \right]^{-1} \mathbf{H}_j \mathbf{b}_j , \tag{1.44}$$

which can be computed at the transmitter. Alternatively, if the receiver uses maximal ratio combining (MRC), then $\mathbf{w}_j = \mathbf{H}_j \mathbf{b}_j$, which is also a function of information known at the transmitter. Whatever the criterion chosen by the receiver, it is likely that the optimal value for $\mathbf{w}_j$ will depend on one or more of the transmit beamformers in $\mathbf{B}$. On the other hand, the choice of $\mathbf{B}$ in (1.42) depends on $\bar{\mathbf{H}}$, which in turn depends on the receive beamformers $\mathbf{w}_j$.

The interdependency of $\mathbf{w}_j$ and $\mathbf{B}$ suggests the following iterative approach:

1. Find an initial value for $\mathbf{w}_1, \cdots, \mathbf{w}_K$. For example, they could be chosen as the principle left singular vectors of the respective channel matrices $\mathbf{H}_j$.

2. Repeat steps 3-4 until convergence.

3. Given $\mathbf{w}_1, \cdots, \mathbf{w}_K$, calculate $\bar{\mathbf{H}}$ and find $\mathbf{B}$ using any of the algorithms discussed above (MMSE, MRC, or other).

4. Given $\mathbf{B}$, recalculate the receive beamformers $\mathbf{w}_1, \cdots, \mathbf{w}_K$ according to their respective algorithms.

Convergence can be said to have occurred, for example, when no appreciable change in the achieved SINR or sum rate is observed from one iteration to the next. Algorithms of this general form have been presented in [7, 32–37]. While analytical results for these approaches are scarce, empirical evidence suggests they have reliable convergence behavior.

In situations where $m_j \geq 1$, solutions similar to those in Section 1.4.1 are possible, where in this case it is the effective channel $\bar{\mathbf{H}}$ that is block-diagonalized. For this case, step 3 in the above iterative algorithm is replaced by either the capacity or the power control formulation of the BD algorithm, and rather than computing the the the $\mathbf{w}_1, \cdots, \mathbf{w}_K$ vectors in step 4 independently, they are taken from the left singular vectors $\mathbf{U}_j^{(1)}$ in equation (1.36) with $\mathbf{H}_j$ replaced by $\bar{\mathbf{H}}_j$ (when $m_j = 1$ this is equivalent to using MRC beamformers). Figure 1.12 plots the cumulative distribution functions of capacity for the coordinated Tx/Rx beamforming algorithm described above. The SNR for this example is 10dB and the channels were all composed of independent, Rayleigh distributed entries. Several base/user geometries were considered: 4×4 (single-user case), $\{2, 2\} \times 4$ with $m_1 = m_2 = 2$, $\{4, 4\} \times 4$ also with $m_1 = m_2 = 2$, $\{2, 2, 2, 2\} \times 4$ and $\{4, 4, 4, 4\} \times 4$. In the latter two scenarios, one sub-channel is allocated to each user, and channel inversion is used to determine $\mathbf{B}$. When $m_j = 2$, the BD algorithm is assumed. As expected, the more total receive antennas that are available, the more flexibility there is in finding a good solution, and the higher the capacity. The 4×4 single-user system outperforms the $\{2, 2\} \times 4$ case since it does not require the block-diagonal constraint. Similarly, the $\{4, 4\} \times 4$ channel achieves higher capacity than the
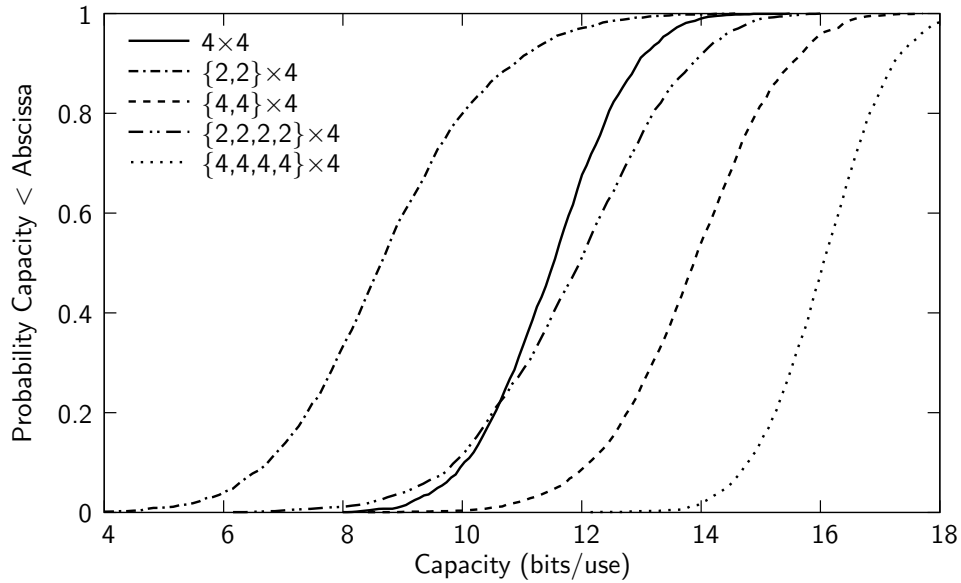
Figure 1.12: Cumulative distributions of the sum capacity for coordinated Tx/Rx beamforming in several scenarios.

$\{2, 2, 2, 2\} \times 4$ system since the block-diagonal constraint is less restrictive than full channel diagonalization.

An obvious extension of this work is to combine the joint transmit-receive beamforming technique with the vector precoding approach of Section 1.3.3. At the transmitter, the nonlinear integer offset would be applied based on the effective channel which would include the transmit beamformers, with the modulo operation at the receivers occurring after the receive beamformers are applied. Similar to the significant diversity gain seen in Section 1.3.3 over linear techniques, the combined technique is anticipated to provide a significant diversity gain over linear joint transmit-receive beamforming.

1.5 Open Problems

1.5.1 Coding and Capacity

Analytic solutions for determining the multi-user capacity region in arbitrary scenarios is the most visible open problem in the area of multi-antenna multi-user coding. The case of Gaussian noise was recently solved [5], but the general problem has not been solved. This and other theoretical results for the broadcast channel have built on dirty-paper coding techniques, which are rather complex. An important area of research is finding simple techniques for approaching the dirty-paper limit. In Section 1.3.3 we used a simple modulo operation to approach capacity; this idea may be extended to higher-dimensional lattices. Though these lattices are difficult to use for encoding and decoding, low-complexity techniques [9, 67]

based on the LLL [66] algorithm exist.

A classic coding approach would be to use a QR-type approach as in Section 1.3.4 but with a more complex dirty-paper coding scheme replacing the simple modulo function. Candidate coding schemes include those based on nested lattices [58], repeat-accumulate codes [40], low-density-parity-check codes [46, 47], and trellis precoding [48]. It remains to be seen if the complexity of these schemes are prohibitive in practical MIMO systems. Though we have focused on the frequency-flat case with CSI assumed known, future wireless cellular systems will include appropriate coding to obtain temporal and frequency diversity as well as the spatial and multi-user diversity mentioned above.

1.5.2 Partial or Imperfect CSI

Almost all of the work on both theoretical and practical coding approaches for the MIMO broadcast channel has assumed that the transmitter and receivers all know the channel exactly [1, 2, 4, 8–10, 69]. Accurate CSI may be easy to obtain when the channel is changing slowly, *e.g.,* as for indoor scenarios, but it is much more difficult in situations where the basestation or users are highly mobile. An analysis of the penalty for using imperfect or outdated feedback of channel information would be of significant benefit to system designers; preliminary simulations indicate that the techniques of Section 1.3.3 are robust to channel estimation error or "stale" CSI. The sum-capacity when only the transmitter or when no one knows the channel would also provide insight for practical coding schemes. A related area of open research is analysis of a system where the transmitter and/or receiver know only the statistics of the channel coefficients. References to several papers that have addressed this problem can be found in [14].

It has been proposed [52] that MIMO channel prediction has the ability to lengthen the time between training intervals over that obtained by SISO prediction techniques. This is especially intriguing for multi-user scenarios where the CSI burden at the transmitter is considerably higher than in the single-user case. Other CSI-related issues that require additional research include: algorithms that take the statistics of the channel estimation error into account, channel feedback methods that consume minimal bandwidth, and analysis of the trade-offs between the amount of CSI fed back to the transmitter and the gain available from using the CSI. A comparison of the resources needed for channel estimation in a time-division duplex versus a frequency-division duplex system is given in [70].

1.5.3 Scheduling

In systems with a single base-station antenna, it is well known [71] that transmitting to the user with the strongest channel at any given time achieves the sum-rate capacity. The resulting "multi-user diversity" is expected to be present as well in the multi-antenna case. The idea is that when a large number of users are sharing a network with rapidly time-varying channels, a base station may use intelligent scheduling algorithms to improve capacity by transmitting to a subset of users. In situations where $K \gg n_T$, one obvious possibility is to transmit to a set of n_T users at each symbol time. The n_T users with strongest average channel strength could be selected, or some orthogonality criterion could be used [72].

In situations where minimal channel variation is to be found, "opportunistic beamforming" has been used to create artificial channel variation in situations where it might not oth-

erwise exist [73]. *Proportional fair scheduling* is also investigated in [73] as a way to ensure that users are treated fairly over some time interval. Some work has been done in extending the opportunistic beamforming idea to the case where base station and mobile stations all have multiple antennas [74], but this has not been studied extensively. Multiple random beams are used in [75] with limited feedback to communicate with many users. Other recent studies in the area of scheduling include [76–90].

1.5.4 Resource Allocation

Related to the scheduling problem are numerous issues that arise when considering practical network implementations. For example, a typical scenario involves more users than transmit antennas; if SDMA is used to supplement existing TDMA and FDMA implementations, it is important to consider how the users in the network will be grouped together. In particular, since the different spatial channels are non-orthogonal, it is critical that only "spatially-compatible" users be chosen to be time- or frequency-coincident [91]. Efficient methods are needed to determine how to optimally determine which users in a network should be spatially multiplexed.

Downlink processing is only one aspect of the multi-user MIMO problem. The uplink MIMO MAC has received significant attention in recent years (see [3, 4, 14] for a discussion of the MIMO MAC and some of the duality relationships it shares with the MIMO broadcast channel). We have focused solely on the cellular network architecture with a base terminal and users that communicate data to the base. *Ad-hoc* networks composed of multiple antenna nodes are of increasing interest, especially in military applications. An important question is how MIMO nodes could be exploited in message relaying, where signal "hops" are required to connect widely separated network nodes. When considering all of these ideas together in the context of a network where all users have arrays, complexity grows very quickly; it is unlikely that globally optimal solutions can be easily found. The ability of heuristic algorithms for scheduling and relaying to achieve the available capacity remains an important area for future research. Some preliminary studies on the application of MIMO techniques to ad-hoc networks have been conducted (*e.g.,* see [92–95]).

Though the general area of multi-user communications has been well-studied, the addition of multiple antennas in a wireless network opens up many new areas of research which have not yet been addressed. Space does not permit a complete treatment of this topic; in addition to the references cited above, there have been a number of other important results that have not been mentioned, including methods based on estimating physical channel parameters (*e.g.,* directions of arrival, *etc.*) [96–100], "multi-cell" or multi-basestation MIMO [101, 102], and others [103–107].

1.6 Summary

A brief overview of coding techniques for the multi-user MIMO downlink has been given in this chapter. We began with a description of techniques for the special case where each user in the network has a single antenna. The capacity of channel inversion was shown to approach a constant as the number of antennas grows, a fact which indicates its inefficiency. A combination of regularized channel inversion and integer perturbation of the data to be

transmitted is shown to operate near capacity when combined with an appropriate channel code. Linear techniques for multiple-antenna receivers were also presented, including block diagonalization of the channel and joint transmit/receive beamforming.

Despite the progress in this area, many open problems remain. Efficient techniques for multi-user scheduling, acquisition of CSI at the base station, and coding must be developed. Future wireless cellular systems may include channel tracking techniques at the base station, and use of this CSI with appropriate coding to obtain temporal, frequency, spatial, and multi-user diversity. The complexity of this system is a serious challenge for researchers, since it grows rapidly with the number of antennas, users, bandwidth and code length.

Acknowledgments

References

[1] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 43, pp. 1691–1706, July 2003.

[2] W. Yu and J. Cioffi, "Sum capacity of a Gaussian vector broadcast channel," in *Proceedings IEEE International Symposium on Information Theory*, July 2002, p. 498.

[3] P. Viswanath and D. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, August 2003.

[4] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, August 2003.

[5] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian mimo broadcast channel," in *Proceedings Conf. on Information Sciences and Systems (CISS)*, Princeton, NJ, March 2004.

[6] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multi-Antenna Multi-User Communication—Part I: Channel Inversion and Regularization," *Submitted to IEEE Trans. Comm.*, Sept. 2003, available from http://mars.bell-labs.com/.

[7] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-Forcing Methods for Downlink Spatial Multiplexing in Multi-User MIMO Channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, February 2004.

[8] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multi-Antenna Multi-User Communication—Part II: Perturbation," *Submitted to IEEE Trans. Comm.*, Sept 2003, available from http://mars.bell-labs.com/.

[9] C. Windpassinger, R. F. H. Fischer, and J. B. Huber, "Lattice-reduction-aided broadcast precoding," in *Proceedings 5th ITG Conference on Source and Channel Coding*, Erlangen, Germany, Jan. 2004.

[10] S. Shi and M. Schubert, "Precoding and power loading for multi-antenna broadcast channels," in *Proceedings Conf. on Information Sciences and Systems*, Princeton, NJ, March 2004.

[11] Z. Tu and R. S. Blum, "Multiuser Diversity for a Dirty Paper Approach," *IEEE Communications Letters*, vol. 7, no. 8, pp. 370–372, August 2003.

[12] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich Scattering Wireless Channel," in *Proc. IEEE ISSSE Conf.*, Pisa, Italy, September 1998.

[13] G. Golden, G. Foschini, R. Valenzuela, and P. Wolniansky, "Detection Algorithm and Initial Laboratory Results using the V-BLAST Space-Time Communications Architecture," *Electronics Letters*, vol. 35, no. 1, pp. 14–15, Jan. 1999.

[14] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity Limits of MIMO Channels," *IEEE J. Sel. Areas in Commun.*, vol. 21, no. 5, pp. 684–702, June 2003.

[15] D. Gerlach and A. Paulraj, "Adaptive Transmitting Antenna Arrays with Feedback," *IEEE Sig. Proc. Letters*, vol. 1, no. 10, pp. 150–152, October 1994.

[16] T. Haustein, C. von Helmolt, E. Jorwieck, V. Jungnickel, and V. Pohl, "Performance of MIMO systems with Channel Inversion," in *Proc. IEEE Vehicular Technology Conf.*, Birmingham, AL, May 2002.

[17] B. M. Hochwald and C. B. Peel, "Vector Precoding for the Multi-Antenna, Multi-User Channel," in *Proc. Allerton Conference on Communication, Control, and Computing*, October 2003.

[18] W. Yu and J. Cioffi, "Trellis precoding for the broadcast channel," in *Proceedings of 2001 IEEE GlobeCom*, November 2001, pp. 1344–1348.

[19] ——, "Sum Capacity of a Gaussian Vector Broadcast Channel," in *Proc. ISIT*, Lausanne, Switzerland, 2002, p. 498.

[20] G. Caire and S. Shamai, "On the Achievable Throughput of a Multi-Antenna Gaussian Broadcast Channel," *IEEE Trans. Info. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[21] C. B. Peel, "Studies in Multiple-Antenna Wireless Communications," Ph.D. dissertation, Brigham Young Universty, Provo UT (USA), January 2004, http://contentdm.lib.byu.edu/ETD/image/etd331.pdf.

[22] S. Shi and M. Schubert, "Precoding and Power Loading for Multi-Antenna Broadcast Channels," in *Proc. Conf. on Information Sciences & Systems*, Princeton, NJ, March 2004.

[23] F. Rashid-Farrokhi, K. R. Liu, and L. Tassiulas, "Transmit Beamforming and Power Control for Cellular Wireless Systems," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1437–1450, October 1998.

[24] E. Visotsky and U. Madhow, "Optimum Beamforming Using Transmit Antenna Arrays," in *Proceedings of the IEEE Vehicular Technology Conference*, vol. 1.   Houston, TX: IEEE, May 16-20 1999, pp. 851–856.

[25] M. Bengtsson and B. Ottersten, "Optimal and Suboptimal Beamforming," in *Handbook of Antennas in Wireless Communications*, L. Godara, Ed.   CRC Press, August 2001.

[26] R. L.-U. Choi and R. D. Murch, "A Downlink Decomposition Transmit Pre-processing Technique for Multi-user MIMO Systems," in *Proceedings of IST Mobile & Wireless Telecommunications Summit*, June 2002.

[27] A. Bourdoux and N. Khaled, "Joint Tx-Rx Optimization for MIMO-SDMA Based on a Null-Space Constraint," in *Proc. IEEE Vehic. Tech. Conf.*, September 2002, pp. 171–174.

[28] Q. H. Spencer and M. Haardt, "Capacity and Downlink Transmission Algorithms for a Multi-user MIMO Channel," in *Conference Record of the 36th Asilomar Conference on Signals, Systems and Computers*.   IEEE, November 2002.

[29] M. Rim, "Multi-user downlink beamforming with multiple transmit and receive antennas," *Electronics Letters*, vol. 38, no. 25, pp. 1725–1726, 5th December 2002.

[30] R. Choi and R. Murch, "A Transmit Preprocessing Technique for Multiuser MIMO Systems Using a Decomposition Approach," *IEEE Trans. Wireless Comm.*, vol. 3, no. 1, pp. 20–24, January 2004.

[31] Q. H. Spencer, "Transmission Strategies for Wireless Multi-user, Multiple-Input, Multiple-Output Communication Channels," Ph.D. dissertation, Brigham Young University, March 2004, http://contentdm.lib.byu.edu/ETD/image/etd378.pdf.

[32] J.-H. Chang, L. Tassiulas, and F. Rashid-Farrokhi, "Joint Transmitter Receiver Diversity for Efficient Space Division Multiaccess," *IEEE Transactions on Wireless Communications*, vol. 1, no. 1, pp. 16–27, January 2002.

[33] Z. Pan, K.-K. Wong, and T. Ng, "MIMO Antenna System for Multi-User Multi-Stream Orthogonal Space Division Multiplexing," in *Proceedings of the IEEE International Conference on Communications*, vol. 5.   Anchorage, Alaska: IEEE, May 2003, pp. 3220–3224.

[34] K.-K. Wong, R. Murch, and K. B. Letaief, "A Joint-Channel Diagonalization for Multiuser MIMO Antenna Systems," *IEEE Trans. Wireless Comm.*, vol. 2, no. 4, pp. 773–786, July 2003.

[35] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Fast Power Minimization with QoS Constraints in Multi-User MIMO Downlinks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*   IEEE, April 2003.

[36] R. L.-U. Choi, M. T. Ivrlač, R. D. Murch, and J. A. Nossek, "Joint Transmit and Receive Multiuser MIMO Decomposition Approach for the Downlink of Multi-user MIMO Systems," in *Proceedings of the IEEE 58th Vehicular Technology Conference.*   Orlando, FL: IEEE, October 6-9 2003.

[37] Q. H. Spencer and A. L. Swindlehurst, "A Hybrid Approach to Spatial Multiplexing in Multi-User MIMO Downlinks," 2004, *EURASIP Journal on Wireless Communications and Networking*, to appear.

[38] S. Vishwanath, N. Jindal, and A. Goldsmith, "On the capacity of multiple input multiple output broadcast channels," in *Proceedings of the IEEE Int. Conf. on Communications*, New York, NY, April 2002, see also http://wsl.stanford.edu/Andrea_publications.html for journal version.

[39] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice strategies for cancelling known interference," in *Proceedings International Symposium on Information Theory and its Applications*, Honolulu, Hawaii, Nov. 2000, pp. 681–684.

[40] U. Erez and S. ten Brink, "Approaching the dirty paper limit for canceling known interference," in *Proc. Allerton Conference on Communication, Control, and Computing*, October 2003.

[41] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink Capacity Evaluation of Cellular Networks with Known Interference Cancellation," *IEEE J. Select. Areas Commun.*, vol. 21, no. 5, pp. 802–811, June 2003.

[42] S. Vishwanath, G. Kramer, S. S. (Shitz), S. Jafar, and A. Goldsmith, "Capacity bounds for Gaussian vector broadcast channels," in *Proceedings DIMACS Workshop on Signal Processing for Wireless Transmission*, October 2002.

[43] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Information Theory*, vol. 29, pp. 439–441, May 1983.

[44] S. S. Pradhan and K. Ramchandran, "Enhancing analog image transmission systems using digital side information: a new wavelet-based image coding paradigm," in *Proceedings of the Data Compression Conference*, Snowbird, Utah, March 2001, pp. 63–72.

[45] J. J. Eggers, J. K. Su, and B. Girod, "A blind watermarking scheme based on structured codebooks," in *Proceedings of IEE Secure Images and Image Authentication*, vol. 4, London, UK, April 2000, pp. 1–6.

[46] G. Caire and S. Shamai, "LDPC coding for interference mitigation at the transmitter," in *Proceedings 40th annual Allerton Conference on Communication, Control, and Computing*, October 2002.

[47] ——, "Writing on dirty tape with LDPC codes," in *Multiantenna Channels: Capacity, Coding and Signal Processing*, ser. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 62, 2002, pp. 123–140.

[48] T. Philosof, U. Erez, and R. Zamir, "Precoding for interference cancellation at low snr," in *22nd Convention of IEEE Israel Section*, Tel-Aviv University, December 2002.

[49] C. B. Peel and A. L. Swindlehurst, "Effective SNR for Space-Time Modulation over a Time-Varying Rician Channel," *IEEE Transactions on Communications*, vol. 52, no. 1, pp. 17–23, Jan 2004.

[50] ——, "Performance of space-time modulation for a general time-varying Rician channel model," *IEEE Transactions on Wireless Communications, Accepted for publication*, vol. 3, no. 3, pp. 1003–1012, May 2004.

[51] ——, "Capacity-optimal training for space-time modulation in time varying fading." in *The 2003 International Conference on Communications*, vol. 4, May 2003, pp. 2698–2702.

[52] T. Svantesson and A. L. Swindlehurst, "A performance bound for prediction of a multipath mimo channel," in *Proc. 37th Asilomar Conference on Signals, Systems, and Computers, Session: Array Processing for Wireless Communications*, Pacific Grove, CA, November 2003.

[53] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short lengths in a lattice, including a complexity analysis," *Mathematics of Computation*, vol. 44, pp. 463–471, April 1985.

[54] M. O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice code decoder for space-time codes," *IEEE Communications Letters*, vol. 4, no. 5, pp. 161–163, May 2000.

[55] A. M. Chan and I. Lee, "A new reduced-complexity sphere decoder for multiple antenna systems," in *Proceedings International Conference on Communications*, vol. 1, New York, April 2002, pp. 460–464.

[56] R. Kannan, "Improved algorithms for integer programming and related lattice problems," in *Proceedings of the ACM Symposium on the Theory of Computing*, Boston, April 1983, pp. 193–206.

[57] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point searches in lattices," *IEEE Trans. Information Theory*, vol. 48, no. 8, pp. 2201–2214, August 2002.

[58] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Information Theory*, vol. 48, no. 6, pp. 1250–1276, June 2002.

[59] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electronics Letters*, vol. 7, pp. 138–139, March 1971.

[60] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Trans. Communication*, vol. 20, no. 4, pp. 774–780, August 1972.

[61] G. Ginis and J. Cioffi, "A multi-user precoding scheme achieving crosstalk cancellation with application to DSL systems," in *Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2000, pp. 1627–1631.

[62] R. Fischer, C. Windpassinger, A. Lampe, and J. Huber, "MIMO precoding for decentralized receivers," in *Proceedings IEEE International Symposium on Information Theory*, Lausanne, Switzerland, June/July 2002, p. 496.

[63] B. Hassibi and H. Vikalo, "On the expected complexity of integer least-squares problems," in *Proceedings ICASSP*, Orlando, 2002.

[64] C. Windpassinger, T. Vencel, and R. F. H. Fischer, "Precoding and loading for BLAST-like systems," in *Proceedings of the IEEE International Conference on Communications*, vol. 5, Anchorage, Alaska, May 2003, pp. 3061–3065.

[65] M. Joham, J. Brehmer, and W. Utschick, "MMSE approaches to multiuser spatio-temporal Tomlinson-Harashima precoding," in *Proc. 5th International ITG Conference on Source and Channel Coding (ITG SCC'04)*, January 2004, pp. 387–394.

[66] A. K. Lenstra, H. W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, pp. 515–534, 1982.

[67] D. Wübben, R. Böhnke, V. Kühn, and K.-D. Kammeyer, "Mmse-based lattice-reduction for near-ml detection of mimo systems," in *ITG Workshop on Smart Antennas*, Munich, Germany, March 2004.

[68] V. Stankovic, M. Haardt, and M. Fuchs, "Combination of block diagonalization and THP transmit filtering for downlink beamforming in multi-user MIMO systems," in *Proc. European Conference on Wireless Technology (ECWT 2004)*, Amsterdam, The Netherlands, Oct. 2004, to be published.

[69] P. Viswanath and D. Tse, "Sum capacity of the multi-antenna broadcast channel," in *Proceedings IEEE International Symposium on Information Theory*, July 2002, p. 497.

[70] T. L. Marzetta and B. M. Hochwald, "Fast transfer of channel state information in wireless systems," *Submitted to IEEE Transactions on Signal Processing*, 2004.

[71] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedigns of the IEEE Int. Conf. on Communications*, June 1995.

[72] M. J. Lopez, "Multiplexing, scheduling, and multicasting strategies for antenna arrays in wireless networks," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[73] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic Beamforming Using Dumb Antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[74] R. W. Heath, M. Airy, and A. J. Paulraj, "Multiuser Diversity for MIMO Wireless Systems With Linear Receivers," in *Conference Record of the 35th Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, CA: IEEE, November 4-7 2001, pp. 1194–1199.

[75] M. Sharif and B. Hassibi, "On the capacity of mimo broadcast channel with partial side information," in *Proc. of the Thrity-Seventh Asilomar Conference on Signals, Systems and Computers*, November 2003, pp. 958–962.

[76] R. H. Jr, M. Airy, and A. Paulraj, "Multiuser Diversity for MIMO Wireless Systems with Linear Receivers," in *35th Asilomar Conf. on Signals, Systems and Comp.*, Asilomar, CA., November 2001, pp. 1194–1199.

[77] H. Viswanathan and K. Kumaran, "Rate Scheduling in Multiple Antenna Downlink Wireless Systems," in *Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, October 3-5 2001.

[78] I. Koutsopoulos and L. Tassiulas, "Adaptive Resource Allocation in SDMA-Based Wireless Broadband Networks With OFDM Signaling," in *Proceedings of IEEE INFOCOM 2002*, vol. 3, 2002, pp. 1376–1385.

[79] V. Lau, Y. Liu, and T.-A. Chen, "Optimal Multi-User Space Time Scheduling for Wireless Communications," in *Proc. IEEE Vehic. Tech. Conf.*, September 2002, pp. 1939–1942.

[80] R. Gozali, R. Buehrer, and B. Woerner, "On the Performance of Scheduling over Space-Time Architectures," in *Proc. Vehicular Technology Conf.*, September 2002, pp. 415–419.

[81] M. Demirkol and M. Ingram, "Stream Control in Networks with Interfering MIMO Links," in *Proc. IEEE Wireless Comm. & Networking Conf.*, March 2003, pp. 343–348.

[82] B. Farhang-Boroujeny and Q. Spencer, "Layering Techniques for Space-Time Communication in Multi-User Networks," in *Proceedings of the IEEE 58th Vehicular Technology Conference*. Orlando, FL: IEEE, October 6-9 2003.

[83] D. Mazzarese and W. Krzymien, "High Throughput Downlink Cellular Packet Data Access with Multiple Antennas and Multiuser Diversity," in *Proc. IEEE Vehic. Tech. Conf.*, vol. 2, 2003, pp. 1079–1083.

[84] J. Chung, C.-S. Hwang, K. Kim, and Y. Kim, "A Random Beamforming Technique in MIMO Systems Exploiting Multiuser Diversity," *IEEE J. Sel. Areas in Commun.*, vol. 21, no. 5, pp. 848–855, June 2003.

[85] H. Boche and M. Wiczanowski, "Stability region of arrival rates and optimal scheduling for mimo-mac-a cross-layer approach," in *Proc. 2004 International Zurich Seminar on Communications*, Feb 2004, pp. 18–21.

[86] Z. Tu and R. Blum, "Multiuser Diversity for a Dirty Paper Approach," *IEEE Communications Lett.*, vol. 7, no. 8, pp. 370–372, August 2003.

[87] L. Dong, T. Li, and Y.-F. Huang, "Opportunistic Transmission Scheduling for Multiuser MIMO Systems," in *Proc. ICASSP*, vol. 5, April 2003, pp. V65–V68.

[88] O.-S. Shin and K. Bok, "Antenna-Assisted Round Robin Scheduling for MIMO Cellular Systems," *IEEE Communications Lett.*, vol. 7, no. 3, pp. 109–111, March 2003.

[89] D. Aktas and H. Gamal, "Multiuser scheduling for MIMO Wireless Systems," in *Proc. IEEE Vehic. Tech. Conf.*, October 2003, pp. 1743–1747.

[90] K.-K. Wong, "Adaptive Space-Division-Multiplexing and Bit-and-Power Allocation in Multiuser MIMO Flat Fading Broadcast Channels," in *Proceedings of the IEEE 58th Vehicular Technology Conference*. Orlando, FL: IEEE, October 6-9 2003.

[91] Q. H. Spencer and A. L. Swindlehurst, "Channel Allocation in Multi-user MIMO Wireless Communications Systems," in *Proceedings of the IEEE International Conference on Communications*. Paris: IEEE, June 2004.

[92] M. Dohler, A. Gkelias, and H. Aghvami, "A Resource Allocation Strategy for Distributed MIMO Multi-Hop Communication Systems," *IEEE Communications Lett.*, vol. 8, no. 2, pp. 99–101, February 2004.

[93] M. Dohler, J. Dominquez, and H. Aghvami, "Link Capacity Analysis for Virtual Antenna Arrays," in *Proc. IEEE Vehic. Tech. Conf.*, September 2002, pp. 440–443.

[94] S. Bellofiore, J. Foutz, R. Govindarajula, I. Bahceci, C. Balanis, A. Spanias, J. Capone, and T. Duman, "Smart Antenna System Analysis, Integration and Performance for Mobile Ad-Hoc Networks (MANETs)," *IEEE Trans. on Ant. and Prop.*, vol. 50, pp. 571–581, May 2002.

[95] Y. Chang and Y. Hua, "Application of Space-Time Linear Block Codes to Parallel Wireless Relays in Mobile Ad Hoc Networks," in *Proc. 37th Asilomar Conf. on Signals, Systems, and Computers*, 2003, pp. 1002–1006.

[96] P. Zetterberg and B. Ottersten, "The Spectrum Efficiency of a Base Station Antenna Array System for Spatially Selective Transmission," *IEEE Trans. on Vehic. Tech.*, vol. VT-44, no. 3, pp. 651–660, August 1995.

[97] J. Fonollosa, J. Goldberg, and G. Vazquez, "Downlink Beamforming in Cellular Mobile Communications," in *Proc. SPAWC Workshop*, Paris, France, 1997, pp. 197–200.

[98] C. Farsakh and J. A. Nossek, "Spatial Covariance Based Downlink Beamforming in an SDMA Mobile Radio System," *IEEE Transactions on Communications*, vol. 46, no. 11, pp. 1497–1506, November 1998.

[99] T. Asté, P. Forster, L. Féty, and S. Mayrargue, "Downlink Beamforming Avoiding DOA Estimation for Cellular Mobile Communications," in *Proc. ICASSP*, vol. 6, Seattle, WA, 1998, pp. 3313–3316.

[100] K. Hugl, J. Laurila, and E. Bonek, "Downlink Beamforming for Frequency Division Duplex Systems," in *Proc. IEEE Vehic. Tech. Conf.*, vol. 3, 1999, pp. 1449–1453.

[101] S. Shamai and B. Zaidel, "Enhancing the Cellular Downlink Capacity via Co-Processing at the Transmitting End," in *Proc. Vehicular Technology Conf.*, May 2001, pp. 1745–1749.

[102] G. J. F. S. A. Jafar and A. Goldsmith, "PhantomNet : Exploring Optimal Multicellular Multiple-Antenna Systems," in *Proc. IEEE Vehicular Technology Conf.*, Vancouver, Canada, September 2002, pp. 24–28.

[103] K.-K. Wong, R. Murch, and K. Letaief, "Performance Enhancement of Multiuser MIMO Wireless Communication Systems," *IEEE Trans. Commun.*, vol. 50, no. 12, pp. 1960–1970, December 2002.

[104] M. Schubert and H. Boche, "Joint 'Dirty Paper' Pre-coding and Downlink Beamforming," in *Proceedings of the IEEE 7th International Symposium on Spread Spectrum Techniques and Applications*, vol. 2.   Prague, Czech Republic: IEEE, September 2-5 2002, pp. 536–540.

[105] H. Boche, M. Scbubert, and E. A. Jorswieck, "Trace Balancing for Multiuser MIMO Downlink Transmission," in *Conference Record of the 36th Asilomar Conference on Signals, Systems and Computers*.   Pacific Grove, CA: IEEE, November 2002, pp. 1379–1383.

[106] M. T. Ivrlač, R. L.-U. Choi, R. D. Murch, and J. A. Nossek, "Effective Use of Long-term Transmit Channel State Information in Multi-user MIMO Communication Systems," in *Proceedings of the IEEE 58th Vehicular Technology Conference*.   Orlando, FL: IEEE, October 6-9 2003.

[107] M. Schubert and H. Boche, "Solution of the Multiuser Downlink Beamforming Problem With Individual SINR Constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, January 2004.

[108] C. Peel, "On Dirty-Paper Coding," *IEEE Sig. Proc. Magazine*, vol. 20, no. 3, pp. 112–113, May 2003.

[109] C. Peel, Q. Spencer, A. Swindlehurst, and M. Haardt, "Downlink Transmit Beamforming in Multi-User MIMO Systems," in *Proc. IEEE Workshop on Sensor Array & Multichannel Sig. Proc.*, Stiges, Spain, 2004.

[110] Q. Spencer, C. Peel, A. Swindlehurst, and M. Haardt, "An Introduction to the Multi-User MIMO Downlink," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 60–67, Oct. 2004.