Chapter 2
Single-user MIMO

In this chapter we study the single-user MIMO channel which is used to model the communication link between a base station and a user. We explore in more detail the fundamental results that were briefly described in the previous chapter. We derive the open-loop and closed-loop MIMO channel capacities and describe techniques for achieving capacity including architectures known as V-BLAST and D-BLAST. We also describe classes of suboptimal techniques such as linear receivers, space-time coding for transmitting a single data stream from multiple antennas, and precoding when there is limited knowledge of CSI at the transmitter.

2.1 Channel model

Figure 2.1 shows the baseband model for a single-user (M, N) MIMO link with M transmit and N receive antennas. A stream of data bits is communicated over the channel. We let $d^{(i)} \in \{+1, -1\}$ represent the data bit with index $i = 0, 1, \ldots$. The data stream is processed to create a sequence of transmitted data symbols. We let $s_m^{(t)} \in \mathbb{C}$ denote the complex baseband signal transmitted from antenna m during period t. For a given symbol period t, the channel between the mth $(m = 1, \ldots, M)$ transmit antenna and the jth $(j = 1, \ldots, N)$ receive antenna is characterized by a scalar value $h_{j,m}^{(t)} \in \mathbb{C}$ which represents the complex amplitude of the narrowband, frequency-nonselective channel. Because each receive antenna is exposed to all transmit antennas, the baseband signal received at antenna j during time t can be written as a linear combination of the transmitted signals:

H. Huang et al., *MIMO Communication for Cellular Networks*,
DOI 10.1007/978-0-387-77523-4_2, © Springer Science+Business Media, LLC 2012

$$x_j^{(t)} = \sum_{m=1}^{M} h_{j,m}^{(t)} s_m^{(t)} + n_j^{(t)}, \tag{2.1}$$

where $n_j^{(t)}$ is complex additive noise. By stacking the received signals from all N antennas in a tall vector, we can write:

$$\begin{bmatrix} x_1^{(t)} \\ \vdots \\ x_N^{(t)} \end{bmatrix} = \begin{bmatrix} h_{1,1}^{(t)} & \cdots & h_{1,M}^{(t)} \\ \vdots & \ddots & \vdots \\ h_{N,1}^{(t)} & \cdots & h_{N,M}^{(t)} \end{bmatrix} \begin{bmatrix} s_1^{(t)} \\ \vdots \\ s_M^{(t)} \end{bmatrix} + \begin{bmatrix} n_1^{(t)} \\ \vdots \\ n_N^{(t)} \end{bmatrix} \tag{2.2}$$

which can be written in the more compact form

$$\mathbf{x}^{(t)} = \mathbf{H}^{(t)} \mathbf{s}^{(t)} + \mathbf{n}^{(t)}, \tag{2.3}$$

where $\mathbf{x}^{(t)} \in \mathbb{C}^{N\times 1}$, $\mathbf{H}^{(t)} \in \mathbb{C}^{N\times M}$, $\mathbf{s}^{(t)} \in \mathbb{C}^{M\times 1}$, and $\mathbf{n}^{(t)} \in \mathbb{C}^{N\times 1}$. For simplicity, we will typically drop the time index t. The noise vector $\mathbf{n}$ is assumed to be zero-mean, spatially white (ZMSW), circularly symmetric, additive complex Gaussian, with each component having variance σ^2: $\mathbb{E}(\mathbf{n}\mathbf{n}^H) = \sigma^2 \mathbf{I}_N$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix. (When the noise is spatially colored, i.e., when the covariance of $\mathbf{n}$ is not a multiple of the identity matrix, we can suppose that the receiver whitens the noise first, by multiplying the received signal vector by the inverse square root of the noise covariance.) The components of the signal vectors $\mathbf{s}^{(t)}$, $t = 1, 2, \ldots$ are the encoded symbols obtained by processing an information bit stream $d^{(1)}, d^{(2)}, \ldots$ which we denote by $\{d^{(i)}\}$. The signal vector is modeled as a stationary random process with zero mean $\mathbb{E}(\mathbf{s}) = \mathbf{0}_M$ and covariance $\mathbf{Q} := \mathbb{E}(\mathbf{s}\mathbf{s})^H$. The signal is subject to the power constraint $\operatorname{tr} \mathbf{Q} = \mathbb{E}(\|\mathbf{s}\|^2) = P$.

The realization $\mathbf{H}^{(t)}$ is drawn from a stationary, ergodic random process to model the fading of the wireless channel. Due to the movement of the transmitter, receiver, and local scatterers, the signal transmitted from antenna m and received by antenna j experiences multipath fading caused by varying path lengths to the scatterers. As a result of the central limit theorem, the complex amplitude of the combined multipath signals can be modeled as a complex Gaussian random variable. If the spacing between the M transmit antennas is sufficiently large relative to the channel angle spread (which is determined by the height of the antennas relative to the height of the local scatterers), then the M channel coefficients $h_{j,1}^{(t)}, \ldots, h_{j,M}^{(t)}$ for receive antenna j will be uncorrelated. Likewise, if the spacing between the N receive antennas is sufficiently large, then the N channel coefficients $h_{1,m}^{(t)}, \ldots, h_{N,m}^{(t)}$ for

transmit antenna m will be uncorrelated. A channel in which the coefficients of $\mathbf{H}^{(t)}$ are uncorrelated (or weakly correlated) is said to be *spatially rich*.

Typically, we will assume in this book that for a given symbol interval t, the elements of $\mathbf{H}^{(t)}$ are not only spatially rich but also independent and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance. Because the amplitude of each element has a Rayleigh distribution, this channel distribution is known as an *i.i.d. Rayleigh* distribution. As a result of the channel normalization, the average received signal power is P, and the signal-to-noise ratio (SNR), defined as the ratio of the received signal power and noise power, is P/σ^2.

With regard to the time evolution of the channel realizations, we define two types of channel model.

1. **Fast-fading**: the channel changes fast enough between symbol periods that each coding block effectively spans the entire distribution of the random process (i.e., ergodicity holds).
2. **Block-fading**: the channel is fixed for the duration of a coding block, but it changes from one block to another.

In practice, coding block lengths are on the order of a millisecond, so users with low mobility (stationary or pedestrian users) experience slowly fading channels consistent with the block-fading model. In this book, we focus mainly on the block-fading model. Further, we usually assume an i.i.d. Rayleigh fading model for the channel.

In the rest of this section, we briefly describe more general channel models that account for propagation environments that are not spatially rich and therefore induce correlated fading across transmitter and receiver antenna pairs.

2.1.1 Analytical channel models

Analytical channel models attempt to describe the end-to-end transfer functions between the transmitting and receiving antenna arrays by accounting for physical propagation and antenna array characteristics [6]. Most analytical channel models capture the various propagation mechanisms through the correlations of the random channel coefficients. Below we describe the most well-known correlation-based analytical MIMO channel models.

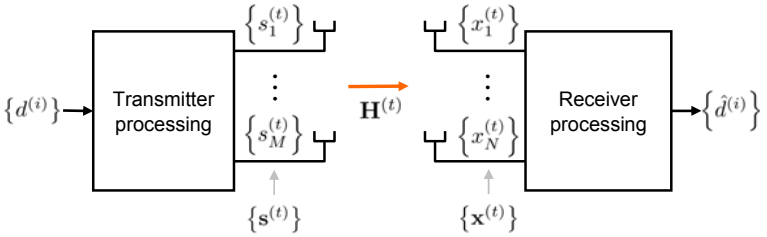

Fig. 2.1 The (M, N) single-user MIMO link. A stream of data bits $\left\{d^{(i)}\right\}$ is processed to form a stream of encoded symbol vectors $\left\{\mathbf{s}^{(t)}\right\}$. The signal is transmitted from M antennas over the channel $\mathbf{H}$ and is received by N antennas. The received signal $\left\{\mathbf{x}^{(t)}\right\}$ is processed to provide estimates of the data stream bits $\left\{\hat{d}^{(i)}\right\}$.

2.1.1.1 Kronecker MIMO channel model

The Kronecker MIMO channel model is probably the best-known correlation-based model and stems from early efforts in the community [7–11] to find models which correspond to a given pair of transmission and receiver correlation matrices (denoted for simplicity as $\mathbf{R}_T = \mathbb{E}\left(\mathbf{H}^H\mathbf{H}\right)$ and $\mathbf{R}_R = \mathbb{E}\left(\mathbf{H}\mathbf{H}^H\right)$, respectively). It hinges on the assumption that these two correlation matrices are *separable*. Mathematically, this assumption is expressed as

$$\mathbf{R}_{\mathbf{H}} = \mathbf{R}_T \otimes \mathbf{R}_R, \tag{2.4}$$

where $\mathbf{R}_{\mathbf{H}}$ is defined as $\mathbf{R}_{\mathbf{H}} = \mathbb{E}\left(\mathrm{vec}\left(\mathbf{H}\right)\mathrm{vec}\left(\mathbf{H}\right)^H\right)$, where the vec operator stacks the columns of the operand matrix vertically, and $\otimes$ denotes the Kronecker product between two matrices. It can be shown that, in this case, the channel matrix can be expressed as

$$\mathbf{H} = \mathbf{R}_R^{1/2}\mathbf{H}_{\text{i.i.d.}}\mathbf{R}_T^{1/2} \tag{2.5}$$

where $\mathbf{H}_{\text{i.i.d.}}$ is a $N \times M$ matrix of i.i.d. circularly symmetric complex Gaussian random variables of zero mean and unit variance. The assumption of separable transmit/receive correlations of course limits the generality of this model, as it is unable to capture any coupling between direction of departure and direction of arrival spectra. Examples of simple channel models that are not captured by the Kronecker channel model are the so-called "keyhole channel," as well as the single- and double-bounce models [12] described below.

However, the Kronecker model has been very popular due to its successful role in quantifying MIMO capacity of correlated channels as well as to its modularity, which allows separate transmitter and receiver array optimization.

2.1.1.2 Single-bounce analytical MIMO channel model

In this case we assume that the signal transmitted from each transmitter bounces once off each of a set of (say V) scatterers before it reaches any receiver antenna. In this single-bounce case, the MIMO channel can be modeled as

$$\mathbf{H}_{\mathrm{SB}}(N, M, V) = \mathbf{\Phi}_{\mathrm{R}}(N, V)\, \mathbf{H}_{\mathrm{i.i.d.}}(V, V)\, \mathbf{\Phi}_{\mathrm{T}}^{H}(M, V), \tag{2.6}$$

where $\mathbf{\Phi}_R$ and $\mathbf{\Phi}_T$ are matrices that define the electrical path lengths from the V scatterers and the N, M antenna elements, respectively. In fact, it turns out that these matrices are the matrix square roots of the correlation matrices on each side of transmission, i.e.

$$\mathbf{H}_{\mathrm{SB}}(N, M, V) = \mathbf{R}_{\mathrm{R}}^{1/2}(N, N)\, \mathbf{H}_{\mathrm{i.i.d.}}(N, M)\, \mathbf{R}_{\mathrm{T}}^{1/2}(M, M). \tag{2.7}$$

In the above, the first two arguments within the parentheses denote the matrix dimensions; the third argument, when present, denotes the assumed number of scatterers. The model in (2.7) has been used successfully to characterize many practical cases where correlation among antenna elements on each side of the link is present (e.g. due to their proximity or a limited angle spread), despite the fundamental richness of the in-between propagation environment. In other words, it models local correlation well. It should be noted of course that when V is smaller than $\min(M, N)$, the channel in (2.7) will suffer severe degradation in its richness, as it will lose rank (notice that such a phenomenon cannot be captured by the Kronecker model in (2.5)). To maximize richness, V should be greater than or equal to NM. The middle ground between $\min(M, N)$ and NM provides intermediate levels of richness (see [12] for some simulated results). It should also be noted that smaller scale effects, such as those due to mutual coupling, are not captured in this model (see [13, 14]).

2.1.1.3 Double-bounce and keyhole MIMO analytical channel models

In some cases, channel richness is compromised by the fact that some waves follow common paths, thus limiting the independence between some signals; as noted in [8, 15], this could be due either to a separation in free space or to some sort of wave-guiding effect. A model that captures these effects is the so-called double-bounce model, which is an extension of the single-bounce model that includes a second ring of scatterers and is described by the following equation:

$$\mathbf{H}_{\mathrm{DB}}\left(N, M, V\right) = \mathbf{\Phi}_{\mathrm{R}}(N, V_1)\mathbf{H}_{\text{i.i.d.}}\left(V_1, V_1\right)\mathbf{X}\left(V_1, V_2\right)\mathbf{\Phi}_{\mathrm{T}}(M, V_2)^{H}. \quad (2.8)$$

where V_1 and V_2 denote the number of scatterers in the first and second ring, respectively. A special case of the model in (2.8), where the resulting matrix has only a single nonzero eigenvalue, is the so-called *keyhole* or *pinhole* channel [8, 15].

2.1.1.4 The Weichselberger MIMO analytical channel model

This model attempts to relax the separability between transmitter and receiver correlations by exploiting the eigenvalue decomposition of the corresponding correlation matrices, shown below:

$$\begin{aligned} \mathbf{R}_T &= \mathbf{U}_T\mathbf{\Lambda}_T\mathbf{U}_T^H \\ \mathbf{R}_R &= \mathbf{U}_R\mathbf{\Lambda}_R\mathbf{U}_R^H \end{aligned}, \quad (2.9)$$

where $\mathbf{U}_T$, $\mathbf{U}_R$ are unitary and $\mathbf{\Lambda}_T$, $\mathbf{\Lambda}_R$ are diagonal matrices. The Weichselberger MIMO channel model is given by the following expression:

$$\mathbf{H} = \mathbf{U}_R\left(\mathbf{\Omega} \bullet \mathbf{H}_{\text{i.i.d.}}\right)\mathbf{U}_T^H, \quad (2.10)$$

where $\mathbf{\Omega}$ is a $N \times M$ coupling matrix that determines the average power coupling between the transmit and receive eigenmodes and $\bullet$ denotes the Schur-Hadamard product (element-wise multiplication). In fact, the Kronecker model is a special case of the Weichselberger model where the coupling matrix $\mathbf{\Omega}$ has rank 1. Other classes of random analytical MIMO channel models include propagation-based versions, such as:

- The finite scatterer model, which assumes a finite number of scatterers and models the angles of departure and arrival, scattering coefficient and delay for each scatterer [16]. This model allows incorporation of both single-bounce and double-bounce scattering.
- The maximum entropy model [17], which attempts to incorporate properties of the propagation environment and system parameters via the maximum entropy principle so as to maximize the model's match to the a priori known information about the link.
- The virtual channel model which exploits the so-called "deconstructed" MIMO channel representation proposed in [18], capturing the "inner" propagation environment between virtual transmission and reception scatterers.

2.1.1.5 The Ricean MIMO channel model

Similarly to the case of scalar channels, when a line of sight (LOS) exists between the transmitter and receiver, the channel is modeled as the sum of a random part representing the non-LOS component and a deterministic part that represents the LOS component. The well-known scalar Ricean channel can be extended to the MIMO case as follows:

$$\mathbf{H} = \frac{\mathbf{H}_{\mathrm{R}} + \sqrt{K}\mathbf{H}_{\mathrm{D}}}{\sqrt{1+K}}, \tag{2.11}$$

where $K \geq 0$ is the Rice factor (also called the "K factor"), $\mathbf{H}_{\mathrm{D}}$ denotes the LOS deterministic channel matrix and $\mathbf{H}_{\mathrm{R}}$ denotes the random channel matrix that can be modeled according to any of the MIMO channel models presented above.

2.1.2 Physical channel models

In contrast to the analytical channel models, physical channel models focus on the properties of the physical environment between transmitter and receiver array. Two classes of physical channel model are briefly described below:

- **Ray-tracing models**
  Ray-tracing (RT) models (see [6, 19]) are widely considered to be among the most reliable deterministic channel models for wireless communica-

tions: they rely on the theory of geometrical optics to predict the way the electromagnetic waves will reach the receiver after they interact with the environment's obstacles (which cause reflection, absorption, diffraction, and so on). The Achilles' heal of the RT approach is the need to know in advance the physical obstacles of the propagation environment between the transmitter and receiver. MIMO extensions of the RT approach have been proposed in [20, 21]. In the MIMO case, the pattern and polarization of each antenna must be taken into account; however, this can be done in a modular fashion, making the technique applicable to any known antenna array configuration.

- **Geometry-based stochastic physical models**
  Contrary to the deterministic nature of the RT approach described above, which exploits the propagation environment's geometry in a deterministic fashion, geometry-based stochastic physical models (GSCM) model the scatterer locations in stochastic (random) terms, i.e. via their statistical distributions. Beyond Lee's original model of deterministic scatterer locations on a circle around the mobile [22], various random scatterer distributions (including scatterer clustering) have since been proposed in, for example, [23–26]. In the single-bounce approach each transmit/receive path is broken into two sub-paths: transmitter-to-scatterer and scatterer-to-receiver (described by their direction of departure, direction of arrival, and path distance); the scatterer itself is modeled typically via the introduction of a random phase shift. Multiple-bounce scattering has also been proposed in order to address more complicated propagation environments (see [27–29]). As mentioned above, MIMO versions of these models are derived by considering the specific configuration and characteristics of the antenna arrays on each side of the link.

2.1.3 Other extensions

The models mentioned above typically assume narrow-band propagation; in other words, they are frequency-flat. Several wideband extensions have been proposed in the literature to capture broadband communication links (see e.g. references in [6]); these are especially relevant in view of the emerging LTE/LTE Advanced and WiFi/WiMAX type systems, which typically use OFDM and operate in bandwidths on the order of several tens of MHz. Also,

it was assumed that the propagation channel is time-invariant; time-varying extensions have also been proposed [30, 31]; these are especially relevant in cases of high user mobility. Mutual coupling between antenna elements (in particular its role in affecting the channel's spatial correlation properties) in the above representative classes of MIMO channel models have been proposed in [13, 14]. Finally, it should be mentioned that large collaborative efforts have been undertaken over the last decade or so in order to propose MIMO channel models that are fit for current and emerging wireless standards. These include the COST259 [32], COST273 [29], IEEE 802.11n [33], Hiperlan2 [34], Stanford University Interim (SUI) [26] and IEEE 802.16 [35]. The Spatial Channel Model described in [36] is used as the basis for standards-related simulations for 3GPP and 3GPP2.

2.2 Single-user MIMO capacity

In this section, expanding on the brief discussion on SU-MIMO capacity in Chapter 1, we derive the capacity for the single-user MIMO channel. We first derive the open-loop and closed-loop MIMO capacity for a fixed channel realization, and then we study the performance of the capacity averaged over random channel realizations.

2.2.1 Capacity for fixed channels

To begin with, we will focus on the case where the channel matrix $\mathbf{H}^{(t)}$ in (2.3) equals some fixed matrix $\mathbf{H} \in \mathbb{C}^{N\times M}$ for all t, i.e., the channel is time-invariant. We will further assume that $\mathbf{H}$ is known exactly to both the transmitter and receiver. After dropping the time index t in (2.3) for convenience, the input-output relationship of the channel reduces to

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{n}. \tag{2.12}$$

The channel input $\mathbf{s}$ is subject to an average power constraint of P. The additive noise vector $\mathbf{n}$ has a circularly symmetric complex Gaussian distribution with zero mean and covariance $\sigma^2\mathbf{I}_N$.

By Shannon's channel coding theorem [37], the capacity $C\left(\mathbf{H}, P/\sigma^2\right)$ of the above channel, defined as the maximum data rate at which the decoding error probability at the receiver can be made arbitrarily small with sufficiently long codewords, is given by the maximum mutual information $I(\mathbf{s};\mathbf{x})$ between the input $\mathbf{s}$ and the output $\mathbf{x}$, over all possible distributions for $\mathbf{s}$ that satisfy the power constraint $\mathrm{tr}\left(E\left[\mathbf{s}\mathbf{s}^H\right]\right) \leq P$. There is no loss of optimality here in restricting $\mathbf{s}$ to have zero mean, since $\mathbf{s} - E[\mathbf{s}]$ automatically satisfies the power constraint if $\mathbf{s}$ does, and also yields the same mutual information with $\mathbf{x}$ as $\mathbf{s}$. Now,

$$I(\mathbf{s};\mathbf{x}) = h(\mathbf{x}) - h\left(\mathbf{x} \mid \mathbf{s}\right) \tag{2.13}$$

$$= h(\mathbf{x}) - h(\mathbf{n}), \tag{2.14}$$

where $h(\mathbf{z})$ denotes the differential entropy of the random vector $\mathbf{z}$, and $h\left(\mathbf{z} \mid \mathbf{y}\right)$ the conditional differential entropy of $\mathbf{z}$ given $\mathbf{y}$.

The differential entropies can be evaluated using the following important result about differential entropy (see, e.g., [38] for a proof): *If $\mathbf{z}$ is any zero-mean complex random vector with covariance $E\left[\mathbf{z}\mathbf{z}^H\right] = \mathbf{R}_z$, then $h(\mathbf{z}) \leq \log|\pi e\mathbf{R}_z|$, with equality holding if and only if $\mathbf{z}$ has a circularly symmetric complex Gaussian distribution.*

Thus in (2.14), we have $h(\mathbf{n}) = \log\left|\pi e\sigma^2\mathbf{I}_N\right|$. Maximizing $I(\mathbf{s};\mathbf{x})$ therefore amounts to maximizing $h(\mathbf{x})$. Further, for any zero-mean $\mathbf{s}$ with covariance $E\left[\mathbf{s}\mathbf{s}^H\right] = \mathbf{R}_s$, the channel output $\mathbf{x}$ is also zero-mean and has the covariance $\sigma^2\mathbf{I}_N + \mathbf{H}\mathbf{R}_s\mathbf{H}^H$. Consequently,

$$h(\mathbf{x}) \leq \log\left|\pi e\left(\sigma^2\mathbf{I}_N + \mathbf{H}\mathbf{R}_s\mathbf{H}^H\right)\right|, \tag{2.15}$$

with equality if and only if $\mathbf{x}$ is circularly symmetric complex Gaussian. The latter condition holds when the input $\mathbf{s}$ is itself circularly symmetric complex Gaussian. We can therefore conclude that, among all zero-mean input distributions with a given covariance $\mathbf{R}_s$, the one that maximizes $I(\mathbf{s};\mathbf{x})$ is circularly symmetric complex Gaussian. Further, the corresponding mutual information is

$$I(\mathbf{s};\mathbf{x}) = \log\left|\pi e\left(\sigma^2\mathbf{I}_N + \mathbf{H}\mathbf{R}_s\mathbf{H}^H\right)\right| - \log\left|\pi e\sigma^2\mathbf{I}_N\right| \tag{2.16}$$

$$= \log\frac{\left|\sigma^2\mathbf{I}_N + \mathbf{H}\mathbf{R}_s\mathbf{H}^H\right|}{\left|\sigma^2\mathbf{I}_N\right|} \tag{2.17}$$

$$= \log\left|\mathbf{I}_N + \left(1/\sigma^2\right)\mathbf{H}\mathbf{R}_s\mathbf{H}^H\right|. \tag{2.18}$$

Therefore the problem of determining the capacity $C\left(\mathbf{H}, P/\sigma^2\right)$ of the channel in (2.12) is reduced to that of finding the input covariance $\mathbf{R}_s$ that maximizes the RHS of (2.18), subject to the constraint $\mathrm{tr}\left(\mathbf{R}_s\right) \leq P$:

$$C\left(\mathbf{H}, P/\sigma^2\right) = \max_{\substack{\mathbf{R}_s \succeq 0 \\ \mathrm{tr}(\mathbf{R}_s) \leq P}} \log\left|\mathbf{I}_N + \left(1/\sigma^2\right)\mathbf{H}\mathbf{R}_s\mathbf{H}^H\right|. \tag{2.19}$$

Clearly, in (2.19), any specific choice of the input covariance $\mathbf{R}_s$ satisfying $\mathrm{tr}\left(\mathbf{R}_s\right) \leq P$ will yield an achievable data rate, i.e., a lower bound on the channel capacity. One such choice that is often of interest is $\mathbf{R}_s = (P/M)\,\mathbf{I}_M$, which corresponds to an *isotropic* input, i.e., sending independent data streams at the same power from each of the transmit antennas. The corresponding achievable rate, which we will loosely term the "open-loop capacity" of the channel and denote by $C^{\mathrm{OL}}\left(\mathbf{H}, P/\sigma^2\right)$, is given by

$$C^{\mathrm{OL}}\left(\mathbf{H}, P/\sigma^2\right) = \log\left|\mathbf{I}_N + \frac{P}{M\sigma^2}\mathbf{H}\mathbf{H}^H\right|. \tag{2.20}$$

In order to motivate the choice of an isotropic input and the concept of open-loop capacity, one can consider a situation where the transmitter has no knowledge of the channel matrix $\mathbf{H}$ (but the receiver still knows it perfectly). The isotropy of the additive noise $\mathbf{n}$ then suggests that the transmitter should employ an isotropic input, hedging against its ignorance of the channel by signaling with equal power in M orthogonal directions. More rigorous justifications can be given for the optimality of an isotropic input in the context of an ergodic channel model with spatially white noise [38].

2.2.1.1 Optimal input covariance

We will now sketch the derivation of the optimal input covariance $\mathbf{R}_s$ in (2.19). The key idea here is to show that the MIMO channel can be decomposed into several single-input single-output (SISO) channels that operate in parallel without interfering with each other, and must share the total available transmit power of P. The optimal power allocation between these SISO channels can then be obtained by a procedure commonly referred to as "waterfilling" (the reason for the name will soon become clear). While the derivation is of secondary importance for this book, we will provide this rough proof because it reveals the important concept of *spatial modes*.

The decomposition of the MIMO channel into non-interfering SISO channels is based on the *singular value decomposition* (SVD) of the $N \times M$ channel matrix $\mathbf{H}$. This decomposition allows us to express $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H, \tag{2.21}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $N \times N$ and $M \times M$ unitary matrices, respectively (so $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}_N$, $\mathbf{V}\mathbf{V}^H = \mathbf{V}^H\mathbf{V} = \mathbf{I}_M$) and $\boldsymbol{\Sigma}$ is an $N \times M$ diagonal matrix. Each element of diag $(\boldsymbol{\Sigma})$ is a *singular value* of $\mathbf{H}$, i.e., the positive square root of an eigenvalue of either $\mathbf{H}\mathbf{H}^H$ (if $N \leq M$) or $\mathbf{H}^H\mathbf{H}$ (if $N \geq M$). Moreover, the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{H}\mathbf{H}^H$, and the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{H}^H\mathbf{H}$.

Using (2.21) in (2.12), we get

$$\mathbf{x} = \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H\right)\mathbf{s} + \mathbf{n} \Rightarrow \mathbf{U}^H\mathbf{x} = \left(\mathbf{U}^H\mathbf{U}\right)\boldsymbol{\Sigma}\left(\mathbf{V}^H\mathbf{s}\right) + \mathbf{U}^H\mathbf{n} \Rightarrow \mathbf{x}' = \boldsymbol{\Sigma}\mathbf{s}' + \mathbf{n}', \tag{2.22}$$

where $\mathbf{x}' = \mathbf{U}^H\mathbf{x}$, $\mathbf{s}' = \mathbf{V}^H\mathbf{s}$, and $\mathbf{n}' = \mathbf{U}^H\mathbf{n}$. Note that $\mathbf{n}'$ has the same distribution as $\mathbf{n}$, since it is obtained by a unitary linear transformation of a zero-mean circularly symmetric complex Gaussian vector whose covariance is a multiple of the identity matrix. So the components of $\mathbf{n}'$ are all independent, circularly symmetric, complex Gaussian random variables of mean 0 and variance σ^2. Note also that the signal terms of (2.22) are *uncoupled*, due to the diagonal structure of $\boldsymbol{\Sigma}$.

Let us assume now that rank$(\mathbf{H}) = r$ (where $r \leq \min(M, N)$). The matrix $\boldsymbol{\Sigma}$ will then have r positive diagonal elements, which we will denote by $\lambda_i, i = 1, \ldots, r$. These are the singular values of $\mathbf{H}$, and $\lambda_i^2, i = 1, \ldots, r$ are the eigenvalues of $\mathbf{H}\mathbf{H}^H$. We will assume further that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. So (2.22) can equivalently be written as:

$$x_i' = \lambda_i s_i' + n_i', \;\; i = 1, \ldots, r. \tag{2.23}$$

(If $r < N$ there are also $N - r$ equations of the type $x_i' = n_i', \; i = r+1, \ldots, N$, which contain no input signal information, and can therefore be neglected.) Note that (2.23) describes an ensemble of r parallel, non-interfering SISO channels, with gains $\lambda_1, \lambda_2, \ldots, \lambda_r$ and noise variance σ^2. As a result, we can depict the equivalent signal model as shown in Figure 2.2.

Assuming now that the transmitter allocates power $P_i = E\left|s_i'\right|^2$ to the i^{th} channel in (2.23), the SNR on the i^{th} SISO channel is $\rho_i = \lambda_i^2 P_i/\sigma^2$, and the rate achievable over it is $R_i = \log_2(1 + \rho_i)$. The overall rate achieved

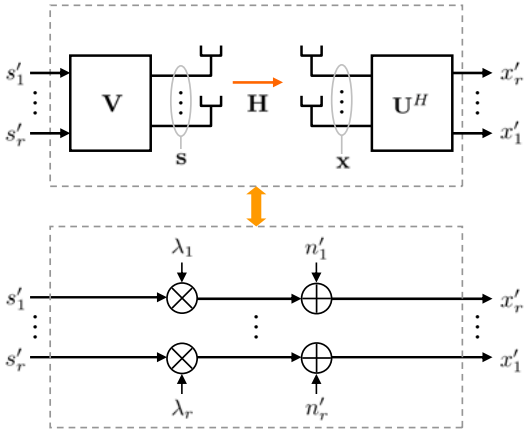

Fig. 2.2 Decomposition of the MIMO channel into r constituent SISO channels, where r is the rank of $\mathbf{H}$.

is $\sum_i R_i$, i.e., the sum of the rates over the individual SISO channels. The capacity of the MIMO channel is then obtained by maximizing the overall rate over all power allocations, subject to the total transmit power constraint:

$$C^{\mathrm{CL}}(\mathbf{H}, P/\sigma^2) = \max_{\substack{P_1,P_2,\ldots,P_r \\ \sum_i P_i = P}} \sum_{i=1}^{r} \log_2\left(1 + \frac{P_i \lambda_i^2}{\sigma^2}\right). \tag{2.24}$$

The objective function in (2.24) is concave in the variables P_i and can be maximized using Lagrangian methods (see [38]), yielding the following solution:

$$P_i^{\mathrm{Opt}} = \left(\mu - \frac{\sigma^2}{\lambda_i^2}\right)^+, \text{with } \mu \text{ chosen such that } \sum_{i=1}^{r} P_i^{\mathrm{Opt}} = P. \tag{2.25}$$

Here $(a)^+ = \max(a, 0)$. The capacity of the channel is then given by [38]

$$C^{\mathrm{CL}}(\mathbf{H}, P/\sigma^2) = \sum_{i=1}^{r} \left[\log_2\left(\frac{\lambda_i^2 \mu}{\sigma^2}\right)\right]^+. \tag{2.26}$$

The covariance matrix of the transmitted signal $\mathbf{s}$ is given by:

$$\mathbf{R}_{\mathbf{s}} = \mathbf{V}\left[\mathrm{diag}\left(P_1, \cdots, P_M\right)\right]\mathbf{V}^H. \tag{2.27}$$

The optimal power allocation between the eigenmodes of the channel, given in (2.25), can be computed by a procedure known as "waterfilling" [37]. Water is poured in a two-dimensional container whose base consists of r steps, where the "height" of each step is σ^2/λ_j^2. If we let μ in (2.25) be the water level, the optimal power allocated to the jth eigenmode P_j^{Opt} is the difference between the water level and the height of the jth step (provided the water level is higher than that step), where μ is set so the total allocated power is P.

Figure 2.3 illustrates the waterfilling algorithm for a MIMO channel with three eigenmodes: $\lambda_1 > \lambda_2 > \lambda_3$. If SNR is very low, the water level, as indicated by the horizontal dotted line, covers only the first step. This indicates that all the power P is allocated to the dominant eigenmode ($P_1^{\text{Opt}} = P$) and no power is allocated to the others ($P_2^{\text{Opt}} = P_3^{\text{Opt}} = 0$). As the SNR increases, the other eigenmodes will be activated. For very high SNR, all three modes are activated, and the difference in height between the steps is insignificant compared to the water level. In this case, the power is allocated approximately equally among the three eigenmodes: $P_1^{\text{Opt}} \approx P_2^{\text{Opt}} \approx P_3^{\text{Opt}}$.

2.2.2 Performance gains

Having established the capacity of open- and closed-loop MIMO channels, we now discuss the performance gains of MIMO relative to conventional single-antenna techniques. In this section, we study in more detail the performance gains mentioned briefly in Chapter 1, namely that the MIMO gains in the low-SNR regime come about through antenna combining, and that the gains in the high-SNR regime come from spatial multiplexing. We also consider the capacity gains as the number of transmit and receive antennas increases without bound.

Under a block-fading channel model, the channel realization is random from block to block, and the capacity for each realization is a random variable. A useful performance measure is the *average capacity* obtained by taking the expectation of the capacity with respect to the distribution of $\mathbf{H}$. The average open-loop and closed-loop capacities are defined respectively as

$$\bar{C}^{\text{OL}}(M, N, P/\sigma^2) := \mathbb{E}_{\mathbf{H}} C^{\text{OL}}(\mathbf{H}, P/\sigma^2) \tag{2.28}$$

$$\bar{C}^{\text{CL}}(M, N, P/\sigma^2) := \mathbb{E}_{\mathbf{H}} C^{\text{CL}}(\mathbf{H}, P/\sigma^2). \tag{2.29}$$

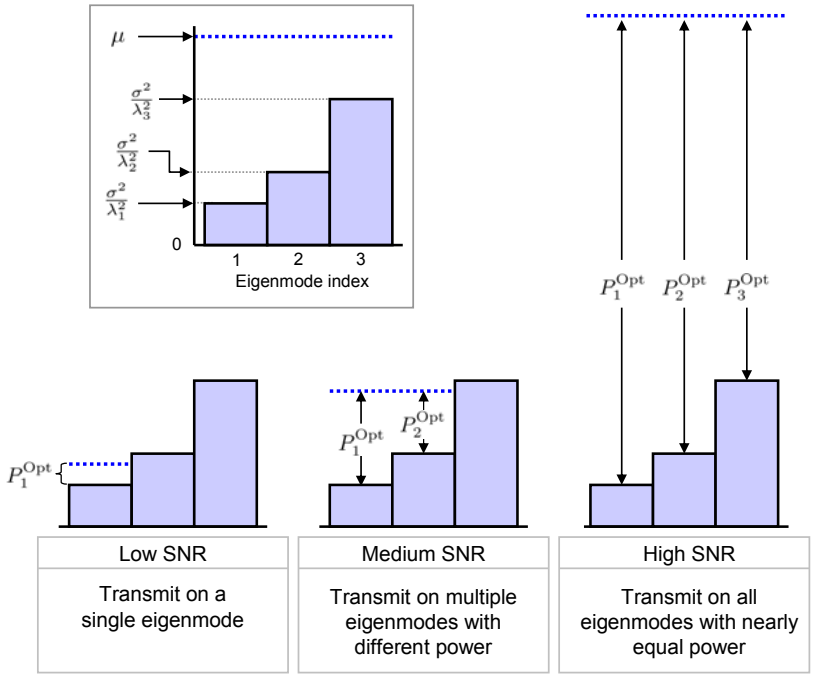

Fig. 2.3 The waterfilling algorithm determines the optimal allocation of power among parallel Gaussian channels that result from the decomposition of a MIMO channel. The channel has three eigenmodes, and the power allocation is shown for low, medium, and high values of SNR (P/σ^2).

(In the context of fast fading channels, the average open-loop capacity as defined here can also be interpreted as the "ergodic capacity" of the channel [39].) We will typically assume an i.i.d. Rayleigh distribution for the components of $\mathbf{H}$.

2.2.2.1 Low SNR

In the low-SNR regime where P approaches zero, the open-loop capacity for a fixed channel $\mathbf{H}$ with rank $r \leq \min(M, N)$ can be approximated as

$$C^{\mathrm{OL}}(\mathbf{H}, P/\sigma^2) = \log_2 \det\left(\mathbf{I}_N + \frac{P}{M\sigma^2}\mathbf{H}\mathbf{H}^H\right) \tag{2.30}$$

$$= \log_2 \prod_{i=1}^{r} \left(1 + \frac{P}{M\sigma^2}\lambda_i^2(\mathbf{H})\right) \tag{2.31}$$

$$\approx \log_2 \left(1 + \sum_{i=1}^{r} \frac{P}{M\sigma^2}\lambda_i^2(\mathbf{H})\right) \tag{2.32}$$

$$= \log_2 \left[1 + \frac{P}{M\sigma^2} \operatorname{tr}\left(\mathbf{H}\mathbf{H}^H\right)\right] \tag{2.33}$$

$$\approx \frac{P}{M\sigma^2} \operatorname{tr}\left(\mathbf{H}\mathbf{H}^H\right) \log_2 \mathrm{e}, \tag{2.34}$$

where $\lambda_i^2(\mathbf{H})$ are the eigenvalues of $\mathbf{H}\mathbf{H}^H$ (and $\lambda_i(\mathbf{H})$ are the singular values of $\mathbf{H}$). Equation (2.32) follows from the dominance of the linear terms for P approaching zero, and (2.34) follows from the approximation

$$\log_2(1 + x) \approx x \log_2 \mathrm{e} \tag{2.35}$$

for x approaching zero. Therefore the average open-loop capacity for i.i.d. Rayleigh channels at low SNR is:

$$\bar{C}^{\mathrm{OL}}(M, N, P/\sigma^2) \approx \frac{P}{M\sigma^2}\mathbb{E}\left[\operatorname{tr}(\mathbf{H}\mathbf{H})^H\right] \log_2 \mathrm{e} \tag{2.36}$$

$$= \frac{P}{M\sigma^2}\mathbb{E}\left[\sum_{m=1}^{M}\sum_{n=1}^{N} |h_{n,m}|^2\right] \log_2 \mathrm{e} \tag{2.37}$$

$$= N\frac{P}{\sigma^2} \log_2 \mathrm{e}, \tag{2.38}$$

where (2.38) follows from $\mathbb{E}\left[\sum_{m=1}^{M}\sum_{n=1}^{N} |h_{n,m}|^2\right] = MN$ for i.i.d. Rayleigh channels. Hence at low SNR, the average open-loop capacity scales linearly with the number of receive antennas N:

$$\lim_{P/\sigma^2 \to 0} \frac{\bar{C}^{(\mathrm{OL})}(M, N, P/\sigma^2)}{P/\sigma^2} = N \log_2 \mathrm{e}. \tag{2.39}$$

In the low-SNR regime, multiple transmit antennas do not improve the capacity, and the capacity of any (M, N) channel with $M \geq 1$ is asymptotically equivalent.

For the closed-loop capacity, waterfilling at asymptotically low SNR puts all the power P into the single best eigenmode. (With i.i.d. Rayleigh channels, the singular values will be unique with probability 1.) The average capacity is therefore

$$\bar{C}^{(\mathrm{CL})}(M,N,P/\sigma^2) = \mathbb{E}\left[\max_{\sum P_i \le P} \sum_{i=1}^{r} \log_2\left(1+\frac{P_i}{\sigma^2}\lambda_i^2(\mathbf{H})\right)\right] \tag{2.40}$$

$$\approx \mathbb{E}\left[\max_{\sum P_i \le P} \sum_{i=1}^{r} \frac{P_i}{\sigma^2}\lambda_i^2(\mathbf{H})\right]\log_2 \mathrm{e} \tag{2.41}$$

$$\approx \frac{P}{\sigma^2}\mathbb{E}\left(\lambda_{\max}^2(\mathbf{H})\right)\log_2 \mathrm{e}, \tag{2.42}$$

where (2.41) follows from (2.35), and $\lambda_{\max}^2(\mathbf{H})$ is the maximum eigenvalue value of $\mathbf{H}\mathbf{H}^H$. Hence

$$\lim_{P/\sigma^2\to 0} \frac{\bar{C}^{(\mathrm{CL})}(M,N,P/\sigma^2)}{P/\sigma^2} = \mathbb{E}\left(\lambda_{\max}^2(\mathbf{H})\right)\log_2 \mathrm{e}. \tag{2.43}$$

Because $\mathbb{E}\left(\lambda_{\max}^2(\mathbf{H})\right) \ge \max(M,N)$ for i.i.d. Rayleigh channels, closed-loop MIMO capacity at low SNR benefits from combining at either the transmitter or receiver.

2.2.2.2 High SNR

From (2.20), the average capacity for i.i.d. Rayleigh channels in the limit of high SNR can be written as:

$$\bar{C}^{(\mathrm{OL})}(M,N,P/\sigma^2) = \mathbb{E}\left[\sum_{i=1}^{\min(M,N)} \log_2\left(1+\frac{P}{M\sigma^2}\lambda_i^2(\mathbf{H})\right)\right]$$

$$\approx \min(M,N)\log_2\left(\frac{P}{M\sigma^2}\right) + \sum_{i=1}^{\min(M,N)} \mathbb{E}\left(\log_2 \lambda_i^2(\mathbf{H})\right), \tag{2.44}$$

where (2.44) derives from the following approximation for large x:

$$\log_2(1+x) \approx \log_2(x). \tag{2.45}$$

Because $\mathbb{E}\left(\log_2 \lambda_i^2(\mathbf{H})\right) > -\infty$ for all i, it follows that

$$\lim_{P/\sigma^2\to\infty} \frac{\bar{C}^{\mathrm{OL}}(M,N,P/\sigma^2)}{\log_2 P/\sigma^2} = \min(M,N). \tag{2.46}$$

Therefore open-loop MIMO achieves a multiplexing gain of $\min(M,N)$ at high SNR.

For the closed-loop capacity at asymptotically high SNR, waterfilling puts equal power in each of the $\min(M,N)$ eigenmodes. Therefore the average

capacity is

$$\begin{aligned}\bar{C}^{(\mathrm{CL})}(M,N,P/\sigma^2) &= \mathbb{E}\left[\max_{\sum P_i \le P} \sum_{i=1}^{\min(M,N)} \log_2\left(1+\frac{P_i}{\sigma^2}\lambda_i^2(\mathbf{H})\right)\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{\min(M,N)} \log_2\left(1+\frac{P/\sigma^2}{\min(M,N)}\lambda_i^2(\mathbf{H})\right)\right] \\ &\approx \min(M,N)\log_2\left(\frac{P/\sigma^2}{\min(M,N)}\right) + \sum_{i=1}^{\min(M,N)} \mathbb{E}\left(\log_2 \lambda_i^2(\mathbf{H})\right), \qquad (2.47)\end{aligned}$$

where (2.47) follows from (2.45). Hence

$$\lim_{P/\sigma^2\to\infty} \frac{\bar{C}^{\mathrm{CL}}(M,N,P/\sigma^2)}{\log_2 P/\sigma^2} = \min(M,N), \qquad (2.48)$$

and closed-loop MIMO achieves the same multiplexing gain as open-loop MIMO (2.46) despite the advantage of CSIT. For both open- and closed-loop MIMO, the multiplexing gain at high SNR requires multiple antennas at both the transmitter and receiver.

2.2.2.3 Large number of antennas

When M and N go to infinity with the ratio M/N converging to α, and the SNR remains fixed at P, the open-loop capacity per transmit antenna converges almost surely to a constant [40]:

$$\begin{aligned}\lim_{\substack{M,N\to\infty \\ M/N\to\alpha}} \frac{\bar{C}^{\mathrm{OL}}(M,N,P/\sigma^2)}{M} &= \log\left[1+S\left(\alpha,\frac{P/\sigma^2}{\alpha}\right)\right] + \frac{1}{P/\sigma^2} S\left(\alpha,\frac{P/\sigma^2}{\alpha}\right) \\ &\quad - \frac{1}{\alpha} + \frac{1}{\alpha}\log\left[1+P/\sigma^2 - \frac{P/\sigma^2}{\alpha} + S\left(\alpha,\frac{P/\sigma^2}{\alpha}\right)\right], \qquad (2.49)\end{aligned}$$

where

$$S(\alpha,\rho) = \frac{1}{2}\left[\rho - \rho\alpha - 1 + \sqrt{(\rho-\rho\alpha-1)^2+4\rho}\right].$$

Thus the capacity grows linearly with the number of antennas. The quantity $S(\alpha,\rho)$ can be interpreted as the asymptotic SINR at the output of a linear MMSE receiver for the signal from each of the M transmit antennas.

For the open-loop $(M,1)$ MISO channel with i.i.d. Rayleigh distribution, the transmit power is distributed among the M antennas, and the average

capacity is given by

$$\bar{C}^{\mathrm{OL}}(M,1,P/\sigma^2) = \mathbb{E}\left[\log_2\left(1+\frac{P}{M\sigma^2}Z,\right)\right], \tag{2.50}$$

where Z is a chi-square random variable with $2M$ degrees of freedom. Asymptotically, as $M \to \infty$, the open-loop MISO capacity converges as a result of the law of large numbers:

$$\lim_{M\to\infty} \bar{C}^{\mathrm{OL}}(M,1,P/\sigma^2) = \log_2\left(1+P/\sigma^2\right). \tag{2.51}$$

Therefore the result of transmit diversity (diversity is the only phenomenon taking place in multi-antenna transmission with single-antenna reception) is to remove the effect of fading when enough transmit antennas are available.

2.2.3 Performance comparisons

The CDF of the open-loop (4,1) MISO and (1,4) SIMO capacities are shown in Figure 2.4 for i.i.d. Rayleigh channel realizations with SNR $P/\sigma^2 = 10$. The circles indicate the average capacities. For MISO channels, the average capacity increases as M increases, and the CDF becomes steeper, indicating there is less variation in the capacity as a result of diversity gain.

On the other hand, receiver combining for the SIMO channel results in both diversity gain and combining gain. Increasing the number of receive antennas results in a steeper CDF due to diversity and a shift with respect to the open-loop MISO curve due to combining gains. The performance of a $(1,N)$ SIMO channel is equivalent to that of a $(N,1)$ closed-loop MISO channel.

Figure 2.5 shows the average capacity of various link configurations versus SNR. The $(4,1)$ OL MISO capacity yields a small improvement over the SISO performance. The $(4,1)$ CL, and $(1,4)$ performance is better as a result of coherent transmitter or combining gain, but the slope of the capacity curve with respect to $\log_2 P/\sigma^2$ is the same as SISO's. At high SNR, the open-loop and closed-loop MIMO techniques achieve a multiplexing gain of 4, indicated by the slope of the capacity. For asymptotically high SNR, the open-loop and closed-loop capacities are equivalent, and there is already negligible difference for SNRs greater than 20 dB. For MIMO, every doubling (3 dB increase) in

SNR results in 4 bps/Hz of additional capacity. For SIMO or MISO, every doubling results in only 1 bps/Hz of additional capacity.

Figure 2.6 shows the average capacity of the same link configurations for a lower range of SNRs. At very low SNRs, the optimal transmission strategy benefits from diversity and combining but not from multiplexing. Compared to $(1, 4)$ SIMO, additional transmit antennas under $(4, 4)$ OL MIMO do not provide any benefit. Using knowledge of the channel at the transmitter, $(4, 4)$ CL MIMO achieves additional capacity by steering power in the direction of the channel's dominant eigenmode. To better visualize the relative gains due to multiple antennas compared to SISO, Figure 2.7 shows the ratios of the MIMO, SIMO, and MISO average capacities versus the SISO average capacity as a function of SNR. For $(4, 4)$ OL MIMO, the ratio is 4 for both low and high SNRs but dips below 4 in between.

For CL MIMO, the number of transmitted streams as determined by waterfilling depends on the SNR. Figure 2.8 shows the average number of transmitted streams (average number of eigenmodes with nonzero power) for different antenna configurations as a function of SNR. For SNRs below -15 dB, capacity is achieved by transmitting a single stream for all cases. As the SNR increases, the probability of transmitting multiple streams increases. For (2,2) and (4,4) multiplexing the maximum number of streams $\min(M, N)$ occurs with probability 1 for SNRs of at least 30 dB. For (2,4), full multiplexing with probability 1 occurs for SNRs of at least 10 dB.

Figure 2.9 shows the average capacity versus the number of antennas M for (M, M) MIMO and $(1, M)$ SIMO. Because the MIMO capacity is roughly $M \log_2(P/\sigma^2)$ for high SNR, the slope of the curve versus M depends on the SNR.

2.3 Transceiver techniques

The previous section describes the theoretical capacity of MIMO links but only hints at the transceiver (transmitter and receiver) structure required to achieve those rates. In this section we discuss transceiver implementation for achieving open- and closed-loop capacity and a number of relevant suboptimal techniques, including linear receivers and space-time coding.

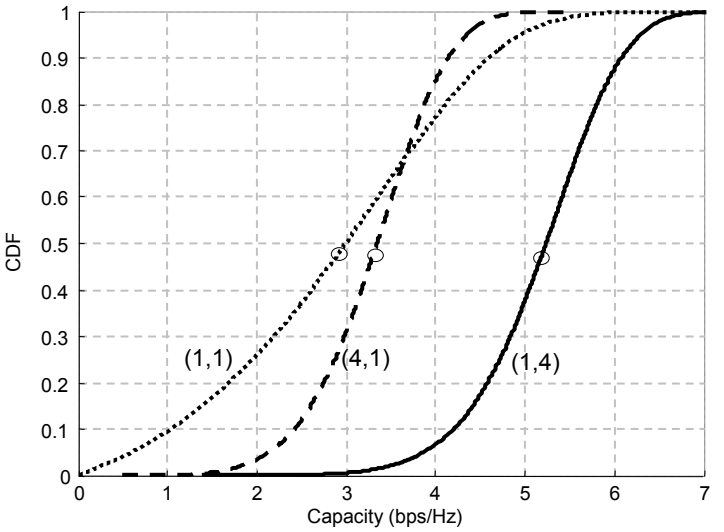

Fig. 2.4 CDF of capacity for SISO, open-loop MISO and SIMO channels for i.i.d. Rayleigh channels. The MISO channel increases reliability by providing diversity gain. The SIMO channel provides both diversity and combining gain.

2.3.1 Linear receivers

Let us consider the received signal (2.3) for the (M, N) MIMO channel where the data stream from the mth transmit antenna is highlighted:

$$\mathbf{x} = \mathbf{h}_m s_m + \sum_{j \neq m} \mathbf{h}_j s_j + \mathbf{n}, \tag{2.52}$$

where $\mathbf{h}_m$ is the m^{th} column of the channel matrix $\mathbf{H}$. We assume that the power of the mth stream is $P_m := \mathbb{E}\left[|s_m|^2\right]$ and that the noise vector $\mathbf{n}$ is ZMSW Gaussian with covariance $\sigma^2 \mathbf{I}_M$.

We are interested in the class of *linear* receivers which computes a decision statistic r_m for the mth data stream by correlating the received signal $\mathbf{x}$ with an appropriately chosen vector $\mathbf{w}_m$:

$$r_m = \mathbf{w}_m^H \mathbf{x} \tag{2.53}$$

$$= (\mathbf{w}_m^H \mathbf{h}_m) s_m + \sum_{j \neq m} (\mathbf{w}_m^H \mathbf{h}_j) s_j + \mathbf{w}_m^H \mathbf{n}. \tag{2.54}$$

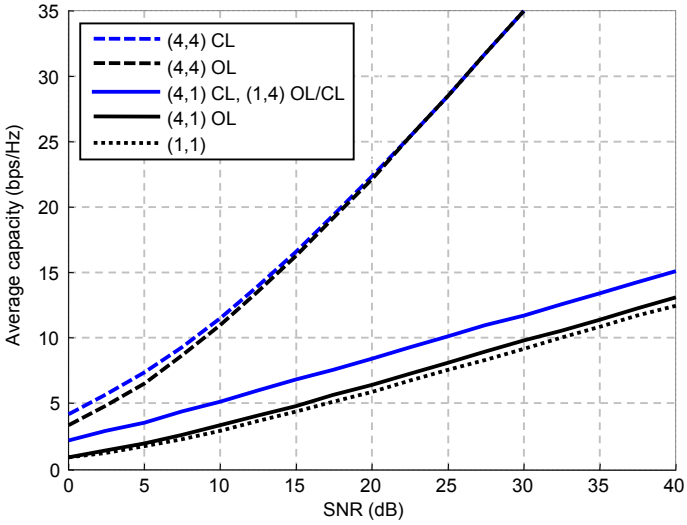

Fig. 2.5 Average capacity versus SNR for i.i.d. Rayleigh channels. At high SNR, (4,4) OL and CL MIMO provide a multiplexing gain of 4.

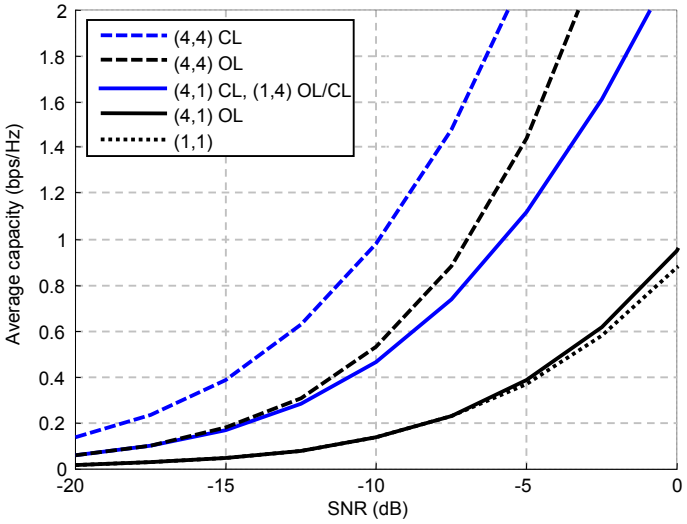

Fig. 2.6 Average capacity versus SNR for i.i.d. Rayleigh channels. At low SNR, (4,4)OL MIMO, (4,1)CL MISO, and (1,4)OL/CL SIMO provide a power gain of 4 as a result of combining.

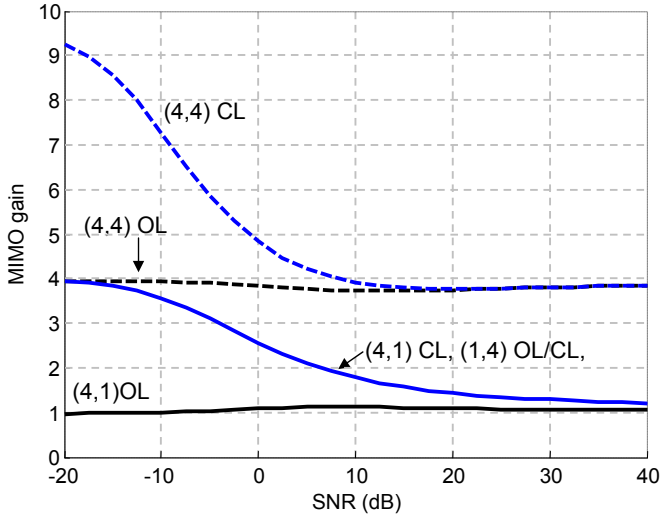

Fig. 2.7 Ratio of MIMO, SIMO, and MISO average rates versus the SISO average rate as a function of SNR for i.i.d. Rayleigh channels.

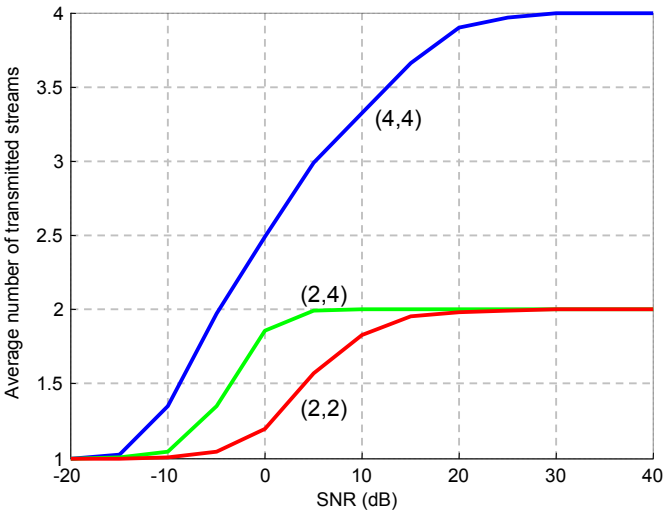

Fig. 2.8 Average number of transmitted streams for CL MIMO with i.i.d. Rayleigh channels. For lower SNRs, only a single stream is transmitted. Spatial multiplexing occurs at higher SNRs.

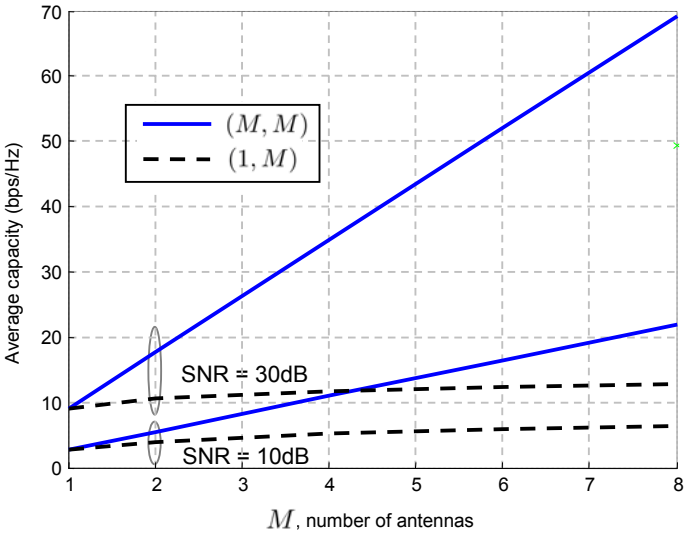

Fig. 2.9 Average capacity versus number of antennas M for i.i.d. Rayleigh channels. The slope of the MIMO capacity depends on the SNR.

The *output signal-to-noise ratio* (SNR), defined as the ratio of the receiver output power of the desired stream to the receiver output power of the thermal noise, is given by

$$\frac{\mathbb{E}\left[|\mathbf{w}_m^H \mathbf{h}_m s_m|^2\right]}{\mathbb{E}\left[|\mathbf{w}_m^H \mathbf{n}\right]} = \frac{|\mathbf{w}_m^H \mathbf{h}_m|^2 P_m}{||\mathbf{w}_m||^2 \sigma^2}. \tag{2.55}$$

In contrast to the SNR of the received signal P/σ^2 defined in Section 2.1, we emphasize that the output SNR (2.55) is defined at the output of the receiver processing. The *output signal-to-interference-plus-noise ratio* (SINR) is defined as the ratio of the receiver output power of the desired stream to the sum of the receiver output power of the thermal noise and interference from the other streams. Because the thermal noise and data streams are uncorrelated, the SINR is given by:

$$\frac{\mathbb{E}\left[\left|\mathbf{w}_m^H \mathbf{h}_m s_m\right|^2\right]}{\mathbb{E}\left[\left|\mathbf{w}_m^H \left(\sum_{j \neq m} \mathbf{h}_j s_j + \mathbf{n}\right)\right|^2\right]} = \frac{\left|\mathbf{w}_m^H \mathbf{h}_m\right|^2 P_m}{\sum_{j \neq m} \left|\mathbf{w}_m^H \mathbf{h}_j\right|^2 P_j + \left\|\mathbf{w}_m\right\|^2 \sigma^2}. \tag{2.56}$$

We consider two linear receivers: the *matched filter* (MF) receiver and the *minimum mean-squared error* (MMSE) receiver. The MF is defined as the correlator matched to the desired stream's channel:

$$\mathbf{w}_m = \mathbf{h}_m. \tag{2.57}$$

Because the noise vector is Gaussian, this receiver maximizes the output SNR [41], but it is oblivious to the interference from the other data streams. It requires knowledge of the desired stream's channel but no knowledge of the other streams'. From (2.55), the output SNR is $\|\mathbf{h}_m\|^2 P_m/\sigma^2$, and the output SINR is

$$\Gamma_{\mathrm{MF},m} = \frac{\|\mathbf{h}_m\|^4 P_m}{\|\mathbf{h}_m\|^2 \sigma^2 + \sum_{j\neq m} |\mathbf{h}_m^* \mathbf{h}_j|^2 P_j}. \tag{2.58}$$

The MF receiver is also known as the *maximal ratio combiner* (MRC) because it weights and combines the received signal components to maximize the output SNR.

The MMSE receiver is a more sophisticated linear receiver that accounts for the presence of interference by minimizing the mean-squared error between the receiver output and the desired data stream s_m:

$$\begin{aligned}
\mathbf{w}_m &= \arg\min_{\mathbf{w}} \mathbb{E}\left[\left|\mathbf{w}^H\mathbf{x} - s_m\right|^2\right] \\
&= \arg\min_{\mathbf{w}} \mathbf{w}^H\left(\mathbf{H}\mathbf{P}\mathbf{H}^H + \sigma^2\mathbf{I}_N\right)\mathbf{w} - 2\mathbf{w}^H\mathbf{h}_m P_m + P_m \\
&= \left(\mathbf{H}\mathbf{P}\mathbf{H}^H + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{h}_m P_m,
\end{aligned}$$

where $\mathbf{P} := \mathrm{diag}(P_1, \ldots, P_M)$ is the diagonal matrix of powers. Using the matrix inversion lemma [42] for invertible $\mathbf{A}$:

$$\left(\mathbf{A} + \mathbf{b}\mathbf{b}^H\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^H\mathbf{A}^{-1}}{1 + +\mathbf{b}^H\mathbf{A}^{-1}\mathbf{b}}, \tag{2.59}$$

and defining $\mathbf{X} := \sum_{j\neq m} \mathbf{h}_j\mathbf{h}_j^H P_j + \sigma^2\mathbf{I}_N$, we can write the MMSE receiver as

$$\begin{aligned}
\mathbf{w}_m &= \left(\mathbf{H}\mathbf{P}\mathbf{H}^H + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{h}_m P_m \\
&= \left(\mathbf{X} + \mathbf{h}_m\mathbf{h}_m^H P_m\right)^{-1}\mathbf{h}_m P_m \\
&= \mathbf{X}^{-1}\mathbf{h}_m P_m - \frac{\mathbf{X}^{-1}\mathbf{h}_m\mathbf{h}_m^H\mathbf{X}^{-1}\mathbf{h}_m P_m^2}{1 + \mathbf{h}_m^H\mathbf{X}^{-1}\mathbf{h}_m P_m}
\end{aligned}$$

$$= \frac{\mathbf{X}^{-1}\mathbf{h}_m P_m}{1 + \mathbf{h}_m^H \mathbf{X}^{-1}\mathbf{h}_m P_m}. \tag{2.60}$$

Using (2.56) and (2.60), the SINR of the mth stream at the MMSE receiver output is

$$\begin{aligned} \Gamma_{\mathrm{MMSE},m} &= \frac{\mathbf{w}_m^H \mathbf{h}_m P_m \mathbf{h}_m^H \mathbf{w}_m}{\mathbf{w}_m^H \mathbf{X} \mathbf{w}_m} \\ &= \frac{P_m}{\sigma^2} \mathbf{h}_m^H \left(\mathbf{I}_N + \sum_{j \neq m} \frac{P_j}{\sigma^2} \mathbf{h}_j \mathbf{h}_j^H \right)^{-1} \mathbf{h}_m. \end{aligned} \tag{2.61}$$

The MMSE receiver is the linear receiver which maximizes the SINR [43], and in this sense, it is often said to be the optimal linear receiver. We note that the MMSE receiver for a particular data stream can also be obtained by whitening the total noise plus interference affecting that stream, and then computing the matched filter for the equivalent channel after whitening.

As given in (2.60), the MMSE receiver requires knowledge of all channels. This requirement is reasonable for many situations where pilot signals are transmitted from each of the antennas. (See Section 2.7 for a discussion on acquiring channel estimates.) If the channel estimates are unreliable or unavailable, blind receivers techniques could be used [44].

Now suppose that M and N both go to infinity and $M/N \to \alpha$. In this large-system limit, it can be shown that [43]

$$\Gamma_{\mathrm{MF},m} \to \frac{P/\sigma^2}{\alpha(1 + P/\sigma^2)} \tag{2.62}$$

and

$$\Gamma_{\mathrm{MMSE},m} \to \frac{1}{2} \left[\left(\frac{P}{\sigma^2 \alpha} - \frac{P}{\sigma^2} - 1 \right) + \sqrt{\left(\frac{P}{\sigma^2 \alpha} - \frac{P}{\sigma^2} - 1 \right)^2 + \frac{4P}{\sigma^2 \alpha}} \right]. \tag{2.63}$$

Figure 2.10 shows the mean SINR (averaged over transmit antennas as well as i.i.d. Rayleigh channel realizations) versus average SNR P/σ^2 for both the MF and MMSE linear receivers. In each case, the solid lines show the results for $M = 4$ and $N = 4$ (as in (2.58) and (2.61)), while the dashed lines show the asymptotic results for $\alpha = 1$ (as in (2.62) and (2.63)). It can be observed that, in contrast to the interference-aware MMSE receiver, the SINR attainable with the interference-oblivious MF receiver saturates as the SNR is increased, indicating that it is interference-limited. (In fact, it can be

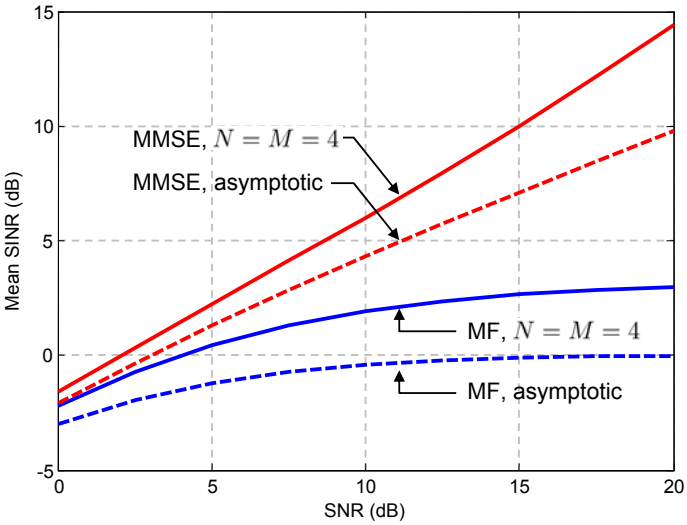

Fig. 2.10 Mean SINR per transmit antenna for MF and MMSE receivers. The asymptotic results assume that M and N both go to infinity with $M/N \to 1$.

shown [39] that the linear MMSE receiver attains the optimal multiplexing gain of $\min(M, N)$, i.e., the rate achievable with it as a function of SNR exhibits the same slope at high SNR as the capacity of the MIMO channel.) The trends exhibited by the asymptotic cases are already apparent for a link with relatively few antennas.

2.3.2 MMSE-SIC

The performance of the MMSE receiver could be improved by following it with a nonlinear *successive interference cancellation* (SIC) stage, shown in Figure 2.11. Suppose that we detect the data symbol s_1 from the first antenna. Its SINR is $\Gamma_{\mathrm{MMSE},1}$, given by (2.61). Assuming s_1 is detected correctly and assuming the receiver has ideal knowledge of $\mathbf{h}_1$, it can be cancelled from the received signal $\mathbf{x}$, yielding:

$$\mathbf{x} - \mathbf{h}_1 s_1 = \sum_{j=2}^{M} \mathbf{h}_j s_j + \mathbf{n}. \tag{2.64}$$

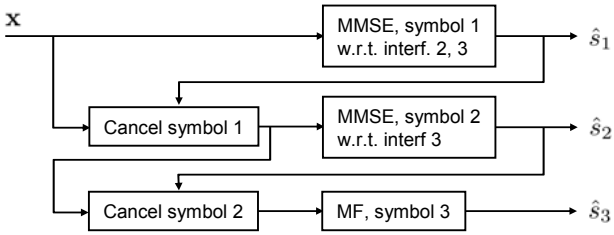

Fig. 2.11 MMSE-SIC detector for $M = 3$ symbols. Symbols are detected and cancelled in order, yielding estimates $\hat{s}_1, \hat{s}_2, \hat{s}_3$.

Given (2.64), data symbol s_2 can be detected using an MMSE receiver. Because there is no contribution from s_1 in (2.64), its SINR is:

$$\Gamma_2 = \frac{P_2}{\sigma^2}\mathbf{h}_2^H \left(\sum_{j=3}^{M} \mathbf{I}_N + \frac{P_j}{\sigma^2}\mathbf{h}_j\mathbf{h}_j^H \right)^{-1} \mathbf{h}_2.$$

If we successively detect the data symbols in order $s_3, s_4, \ldots, M$ and cancel their contributions, the mth stream ($m = 1, \ldots, M-1$) experiences interference from streams $m+1, m+2, \ldots, M$. The SINR for the mth stream is therefore

$$\Gamma_m = \frac{P_m}{\sigma^2}\mathbf{h}_m^H \left(\sum_{j=m+1}^{M} \mathbf{I}_N + \frac{P_j}{\sigma^2}\mathbf{h}_j\mathbf{h}_j^H \right)^{-1} \mathbf{h}_m. \tag{2.65}$$

The Mth stream is detected in the presence of only Gaussian noise. Using a matched filter, its SNR is

$$\Gamma_M = \frac{P_M}{\sigma^2} \left\| \mathbf{h}_M \right\|^2. \tag{2.66}$$

In the discussion of the MMSE-SIC detector (and the MF and MMSE detectors), we have focused on the detection of the data symbols s_m without any regard to channel encoding. In general, these data symbols are the output of a channel encoder, and we show in the next section how the MMSE-SIC detector in conjunction with a channel decoder can be used to achieve the open-loop MIMO capacity.

2.3.3 V-BLAST

The open-loop capacity for a fixed (M, N) MIMO link can be achieved using the *vertical BLAST* (V-BLAST) architecture, which uses an MMSE-SIC receiver structure where the interference cancellation is performed with respect to the decoded data streams. In the MIMO literature, the term "V-BLAST" is used to describe a variety of transceiver architectures and can be applied to block- or fast-fading channels. However, we use the term to refer specifically to the transmitter architecture shown in Figure 2.12 where the information data stream is multiplexed into M lower-rate streams that are independently encoded. This transmit architecture is sometimes referred to as *per-antenna rate control* (PARC) because the rate of each antenna's data stream is adjusted based on the channel realization $\mathbf{H}$.

From (2.3), the encoded transmitted signal vector for symbol time t is denoted as $\mathbf{s}^{(t)}$, and we let $\left\{\mathbf{s}^{(t)}\right\}$ denote the stream of vectors associated with a coding block. Similarly we let $\left\{\mathbf{x}^{(t)}\right\}$ denote the corresponding block of received signals. We assume that the channel is stationary for the duration of the coding block and that each stream has power $P_m = P/M$, $m = 1, \ldots, M$. Applying an MMSE detector to the received signal vectors $\left\{\mathbf{x}^{(t)}\right\}$, the output SINR is (2.61):

$$\Gamma_1 = \frac{P}{M\sigma^2}\mathbf{h}_1^H \left(\mathbf{I}_N + \sum_{j=2}^{M} \frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right)^{-1} \mathbf{h}_1. \tag{2.67}$$

If the data stream for the first antenna $\left\{s_1^{(t)}\right\}$ is encoded using a capacity-achieving code corresponding to rate $\log_2(1 + \Gamma_1)$, then it can be decoded without error. Using ideal knowledge of $\mathbf{h}_1$, its contribution to the received signal $\left\{\mathbf{x}^{(t)}\right\}$ can be cancelled. In general, if the mth data stream is encoded with rate $\log_2(1 + \Gamma_m)$, where from (2.65),

$$\Gamma_m = \frac{P}{M\sigma^2}\mathbf{h}_m^H \left(\mathbf{I}_N + \sum_{j=m+1}^{M} \frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right)^{-1} \mathbf{h}_m, \tag{2.68}$$

then it can be decoded and cancelled from the received signal so that data streams $m + 1, m + 2, \ldots, M$ do not experience interference from it. Using (2.68) and the matrix identities [42]

$$\frac{\det(\mathbf{A})}{\det(\mathbf{B})} = \det(\mathbf{B}^{-1}\mathbf{A}) \text{ and}$$
$$\det(\mathbf{I} + \mathbf{AB}) = \det(\mathbf{I} + \mathbf{BA}),$$

the rate achievable by stream m can be written as

$$\begin{aligned}\log_2(1+\Gamma_m) &= \log_2\left[1 + \frac{P}{M\sigma^2}\mathbf{h}_m^H\left(\mathbf{I}_N + \sum_{j=m+1}^{M}\frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right)^{-1}\mathbf{h}_m\right] \\ &= \log_2\det\left[\mathbf{I}_N + \frac{P}{M\sigma^2}\left(\mathbf{I}_N + \sum_{j=m+1}^{M}\frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right)^{-1}\mathbf{h}_m\mathbf{h}_m^H\right] \\ &= \log_2\frac{\det\left(\mathbf{I}_N + \sum_{j=m}^{M}\frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right)}{\det\left(\mathbf{I}_N + \sum_{j=m+1}^{M}\frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right)}. \end{aligned} \tag{2.69}$$

If we add the achievable rates for all M streams, then from (2.69), all the terms are cancelled except for the numerator of the rate for stream 1. The achievable sum rate is therefore

$$\begin{aligned}\sum_{m=1}^{M}\log_2(1+\Gamma_m) &= \log_2\det\left(\mathbf{I}_N + \sum_{j=1}^{M}\frac{P}{M\sigma^2}\mathbf{h}_j\mathbf{h}_j^H\right) \\ &= \log_2\det\left(\mathbf{I}_N + \frac{P}{M\sigma^2}\mathbf{H}\mathbf{H}^H\right). \end{aligned} \tag{2.70}$$

Noting the equivalence between (2.70) and (2.20), we conclude that the PARC strategy with an MMSE-SIC receiver achieves the open-loop MIMO capacity for block-fading channels.

For a fast-fading channel with i.i.d. Rayleigh distribution, using the statistics of $\mathbf{H}$, we can set the rate of stream m to be

$$R_m = \mathbb{E}_{\mathbf{H}}\left[\log_2\left(1+\Gamma_m\right)\right], \tag{2.71}$$

where Γ_m is from (2.68). Then, from (2.70), the achievable sum rate is

$$\sum_{m=1}^{M}\left[\mathbb{E}_{\mathbf{H}}\log_2(1+\Gamma_m)\right] = \mathbb{E}_{\mathbf{H}}\left[\log_2\det\left(\mathbf{I}_N + \frac{P}{M\sigma^2}\mathbf{H}\mathbf{H}^H\right)\right]. \tag{2.72}$$

Therefore the V-BLAST architecture also achieves the ergodic capacity for fast-fading channels. We emphasize that for a block-fading channel, the rates

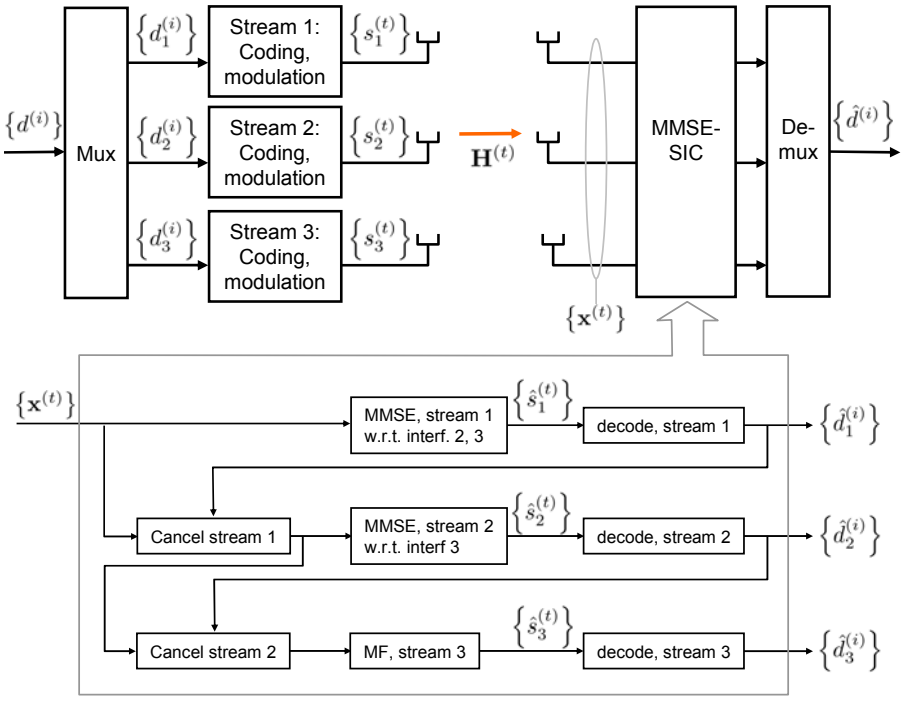

Fig. 2.12 Transceiver for achieving OL-MIMO capacity using the V-BLAST transmit architecture and an MMSE-SIC receiver.

are set as a function of the realization $\mathbf{H}$, while for a fast-fading channel, the rates are based on the statistics of $\mathbf{H}$.

Even though we have assumed that the data streams are decoded and cancelled in order from $m = 1$ to M, we note that the sum rate computed via (2.70) is independent of the order. Therefore, the OL MIMO capacity can be achieved for any ordering, as long as each antenna's stream is encoded with the appropriate rate as determined by the particular ordering.

The optimality of V-BLAST and PARC with MMSE-SIC was shown in [45] and is based on the sum-rate optimality of the MMSE-SIC for the multiple access channel [46]. This topic will be revisited in Chapter 3. While the PARC strategy assumes equal power on each stream, the throughput can be increased by optimizing the power distribution among the streams [45] using knowledge of $\mathbf{H}$ at the transmitter.

2.3.4 D-BLAST

In contrast to V-BLAST, where data streams for each antenna are encoded independently, *Diagonal BLAST* (D-BLAST) is an alternative technique for achieving the open-loop capacity that transmits the symbols for each coding block from all M antennas. Figure 2.13 shows the D-BLAST transceiver architecture. The information stream is encoded as U blocks of ML symbols. In the context of D-BLAST, each coding block is known as a *layer*. Layer $u = 1, \ldots, U$ consists of two subblocks of L symbols: $\left\{b_1^{(u)}\right\}$ and $\left\{b_2^{(u)}\right\}$. The blocks are transmitted in a staggered fashion so that L symbols $\left\{b_2^{(u)}\right\}$ are transmitted from antenna 2, followed by L symbols $\left\{b_1^{(u)}\right\}$ transmitted from antenna 1. The layer u transmission on antenna 1 occurs at the same time as the layer $u+1$ transmission on antenna 2. During the first L symbol periods, symbols $\left\{b_2^{(1)}\right\}$ are transmitted from antenna 2, and nothing is transmitted from antenna 1. During the last L symbol periods, $\left\{b_1^{(U)}\right\}$ are transmitted from antenna 1, and nothing is transmitted from antenna 2.

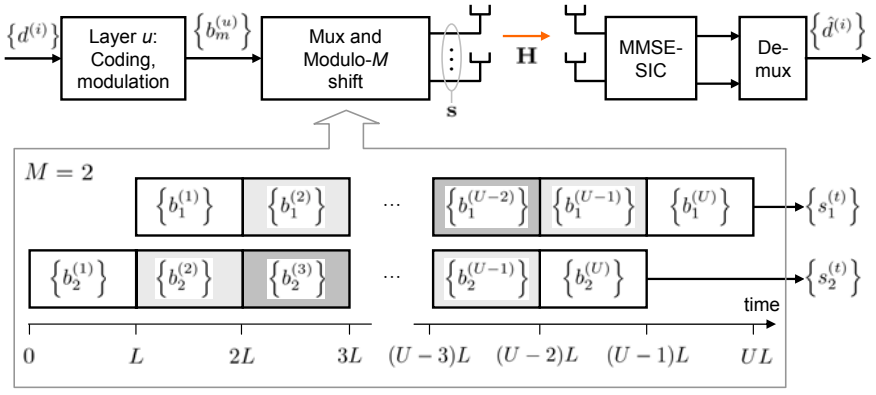

Fig. 2.13 The D-BLAST transmitter cyclicly shifts the association of each stream with all M antennas. The modulo-M shift is illustrated for $M = 2$ antennas and U layers.

To decode layer 1, symbols $\left\{b_2^{(1)}\right\}$ are detected using a matched filter in the presence of thermal noise. The output SNR is

$$\Gamma_2 = \frac{P}{2\sigma^2}|\mathbf{h}_2|^2. \tag{2.73}$$

The symbols $\left\{b_1^{(1)}\right\}$ are detected using an MMSE receiver in the presence of interference from antenna 2 from symbols $\left\{b_2^{(2)}\right\}$. The output SINR is

$$\Gamma_1 = \frac{P}{2\sigma^2}\mathbf{h}_1^H\left(\mathbf{I}_N + \frac{P}{2\sigma^2}\mathbf{h}_2\mathbf{h}_2^H\right)^{-1}\mathbf{h}_1. \tag{2.74}$$

If the $2L$ symbols of layer 1 are encoded using a compound code at a rate $R < \log_2(1+\Gamma_1) + \log_2(1+\Gamma_2)$, then these symbols can be reliably decoded and canceled from the received signal stream.

The decoding of layer u follows the same procedure. Symbols $\left\{b_2^{(u)}\right\}$ are detected using a matched filter because symbols from the previous layer transmitted on antenna 1 have been cancelled. Then symbols $\left\{b_1^{(u)}\right\}$ are detected using an MMSE in the presence of interference from $\left\{b_2^{(u+1)}\right\}$. If U layers are transmitted, the achievable rate is

$$\frac{U}{U+1}\left[\log_2(1+\Gamma_1) + \log_2(1+\Gamma_2)\right], \tag{2.75}$$

where the fraction is due to the empty frames during the first and last layers. This overhead vanishes as U increases.

The procedure described for the $M = 2$ antenna case can be generalized for more antennas, so that the symbols for a single layer are staggered over ML symbol periods and transmitted over all M antennas. The symbols from antenna $m = 1, ..., M$ are detected in the presence of interference from antennas $m+1, \ldots, M$. The SINR achieved for detecting symbols from antenna m is (2.68), and the overall achievable rate (for large U) is the open-loop MIMO capacity (2.70).

In practice, because the transmitter does not have knowledge of the channel, it does not know at what rate to encode the information. The receiver could estimate the channel $\mathbf{H}$, determine the channel capacity, and feed back this information to the transmitter. The D-BLAST encoder needs to know only the MIMO capacity whereas the V-BLAST encoder needs to know the achievable rates of each stream. D-BLAST would therefore require less feedback. However, due to the difficulty in implementing efficient compound codes, the V-BLAST architecture is more commonly implemented.

2.3.5 Closed-loop MIMO

If the channel $\mathbf{H}$ is known at the transmitter, one can achieve capacity by transmitting on the eigenmodes of the channel, as discussed in Section 2.2.1.1. The corresponding transceiver structure is shown in Figure 2.14. The information bit stream is first multiplexed into $\min(M, N)$ lower-rate streams, and the streams are encoded independently according to the rates determined by waterfilling. Given the SVD of the MIMO channel $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$, the $\min(M, N)$ streams are precoded using the first $\min(M, N)$ columns of the $M \times M$ unitary matrix $\mathbf{V}$. At the receiver, the $N \times N$ linear transformation $\mathbf{U}^H$ is applied, and the elements of the first $\min(M, N)$ rows are demodulated and decoded. The information bits for the $\min(M, N)$ streams are demultiplexed to create an estimate for the original bit stream.

For the special case of the $(M, 1)$ MISO channel, the data stream is precoded with the unit-normalized complex conjugate of the channel vector $\mathbf{h} \in \mathbb{C}^{1\times M}$: $(\mathbf{v} = \mathbf{h}^H / \|\mathbf{h}\|)$. This weighting is sometimes known as *maximal ratio transmission* (MRT), and it is the dual of MRC receiver for the $(1, N)$ SIMO channel.

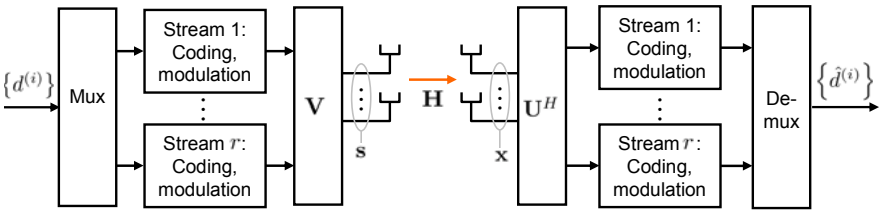

Fig. 2.14 Capacity-achieving transceiver for CL MIMO based on the SVD of $\mathbf{H}$. The power allocated to each stream is determined by waterfilling.

2.3.6 Space-time coding

If the channel is known at the transmitter, the SVD-based strategy described in Section 2.3.5 achieves the closed-loop capacity for any (M, N) link. If the channel is not known at the transmitter, the strategies for achieving open-

loop capacity described in Section 2.3.5 apply only when $M \leq N$. *Space-time coding* is a class of techniques for achieving diversity gains in MISO channels when the channel state information is not known at the transmitter [47] [48]. Multiple receive antennas could be used to achieve combining and additional diversity gains. We will outline the basic principles of *space-time block codes* (STBCs), which are illustrative and representative of what space-time coding can achieve in MISO channels. The space-time block-coding transmission architecture is shown in Figure 2.15.

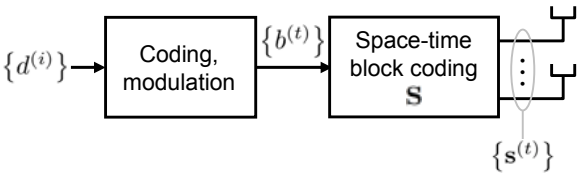

Fig. 2.15 In a space-time block-coding transmission architecture, the coded symbols $\{b^{(t)}\}$ are mapped to the transmitted symbol vector using, for example, (7.2)

.

In this architecture, a data stream is encoded using an outer channel encoder, and a space-time block encoder maps a block of Q encoded symbols $b_1, \ldots, b_Q$ onto the M antennas over L symbol periods. This mapping is represented by an $M \times L$ matrix $\mathbf{S}$, where the (m, l)th element of $\mathbf{S}$ ($m = 1, \ldots, M$, $l = 1, \ldots, L$) is the symbol transmitted from antenna m during symbol period l. In general, each element of $\mathbf{S}$ is a linear combination of $b_1, \ldots, b_Q$ and of the respective complex conjugates of these symbols $b_1^*, \ldots, b_Q^*$. The parameters L and $R := Q/L$ are known respectively as the *code delay* and *code rate* of space-time block code $\mathbf{S}$. Typically, the mapping parameters are chosen so that $L \geq M$ and $R \leq 1$. The performance of a space-time code can be measured by its *diversity order* which we define as the magnitude of the slope of the average symbol error rate at the receiver versus SNR (in a log-log scale).

For an $(M, 1)$ channel, the optimal (maximum) diversity order is M and can be achieved if $\mathbf{S}\mathbf{S}^H$ is proportional to the identity matrix $\mathbf{I}_M$ [49]. It is also desirable for a code to be *full rate* (i.e., $R = 1$ and $Q = L$) and *delay optimal* (i.e., $L = M$) so that the code is time efficient. A space-time block

code is *ideal* if it is full rate and delay optimal (so that $L = M = Q$) and if it achieves maximum diversity order.

For the case of $M = 2$ transmit antennas, a very popular and remarkably efficient space-time block code (the one that really defined the class of space-time block codes) is the *Alamouti space-time block code* [50]. Given a sequence of encoded symbols $\left\{b^{(t)}\right\}$ $(t = 0, 1, \ldots)$, each pair of symbols on successive time intervals $b^{(2j)}$ and $b^{(2j+1)}$ $(j = 0, 1, \ldots)$ is transmitted over the two antennas on intervals $2j$ and $2j + 1$ as follows:

$$\mathbf{s}^{(2j)} = \begin{bmatrix} b^{(2j)} \\ b^{(2j+1)} \end{bmatrix} \text{ and } \quad \mathbf{s}^{(2j+1)} = \begin{bmatrix} -b^{*(2j+1)} \\ b^{*(2j)} \end{bmatrix}.$$

This code is ideal because it achieves maximum diversity order $M = 2$ with $L = M = Q = 2$. Moreover, it quite remarkably achieves the open-loop capacity for the (2,1) MISO channel for any SNR if an outer capacity-achieving scalar code is used [51] [52]. It also achieves the optimal diversity/ multiplexing tradeoff (see e.g. [53]). For a $(2, N)$ channel with $N > 1$, the Alamouti STBC with maximal ratio combining in general does not achieve the capacity. Not surprisingly, due to its remarkable properties, the Alamouti code has been used in several wireless standards.

To date, no ideal space-time block codes have been found for $M > 2$. However, quasi-orthogonal STBCs have also been proposed that approach the open-loop capacity in the (4,1) case [54] [55]. While generalizations of the quasi-orthogonal concept (and other space-time coding techniques) to arbitrary numbers of antennas have been suggested [56], it should be emphasized that the marginal diversity gains for open-loop MISO techniques diminish as the number of antennas increases. This fact, coupled with the additional overhead required to the channels, reduces the incentive for using too many $(M > 4)$ antennas for space-time coding.

Besides STBCs, other classes of space-time codes include space-time trellis codes [47], linear dispersion codes [57], layered turbo codes [58], and lattice space-time codes [59].

2.3.7 Codebook precoding

If CSIT is ideally known, precoding with waterfilling achieves the closed-loop MIMO capacity. In many practical cases, it is not possible to obtain reliable

CSIT (see Section 4.4). Isotropic transmission is suboptimal, and as we saw in Section 2.2.3, the performance gap between OL-MIMO and CL-MIMO is significant at lower SNRs.

Another suboptimal alternative is to use precoding matrices that are chosen from a finite discrete set known as a *codebook*. (The precoding matrices are sometimes known as *codewords*, but they are not to be confused with the codewords associated with channel encoding.) The codebook is known by both the transmitter and receiver. Under codebook precoding, typically the receiver estimates the channel $\mathbf{H}$ and sends information back to the transmitter to indicate its preferred codeword. Using B feedback bits, the receiver can index up to 2^B codewords in the codebook. A block diagram is shown in Figure 2.16, where the codebook $\mathcal{B}$ consists of 2^B precoding matrices $\mathbf{G}_1, \ldots, \mathbf{G}_{2^B}$. This precoding technique is sometimes known as *limited feedback precoding*.

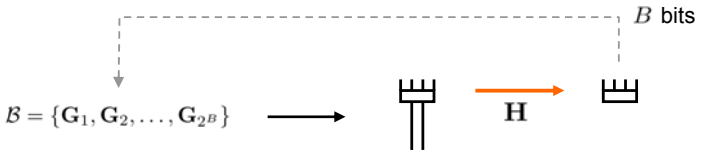

Fig. 2.16 Block diagram for codebook precoding. The user estimates the channel $\mathbf{H}$ and feeds back B bits to indicate its perfered precoding vector. The codebook $\mathcal{B}$ consists of 2^B codeword matrices and is known by the transmitter and the user.

In general, the transmitter can send up to $\min(M, N)$ data streams. If we let $J \leq \min(M, N)$ denote the number of streams, and $\mathbf{u} \in \mathbb{C}^J$ be the vector of data symbols, then the precoding matrix $\mathbf{G} \in \mathcal{B}$ is size $M \times J$, and the transmitted signal is given by $\mathbf{s} = \mathbf{G}\mathbf{u}$. From 2.3, the received signal is

$$\mathbf{x} = \mathbf{H}\mathbf{G}\mathbf{u} + \mathbf{n}. \tag{2.76}$$

When $J = 1$ and the input covariance has rank 1, precoding is often known as *beamforming*.

If the MIMO channel is changing sufficiently slowly, the mobile feedback could be aggregated over multiple feedback intervals so that the aggregated bits index a larger codebook. In general, a larger codebook implies more accurate knowledge of the MIMO channel at the transmitter, resulting in im-

proved throughput. By aggregating the feedback bits over multiple intervals, the codewords can be arranged in a hierarchical tree structure so that the feedback on a given interval is an index of codewords that are the "children nodes" of a codeword indexed by previous feedback [60]. Temporal correlation of the channel can also be exploited by adapting codebooks over time [61] or by tracking the eigenmodes of the channel [62] [63] [64].

2.3.7.1 Single-antenna receiver, $N = 1$

Let us consider the problem of designing a codebook $\mathcal{B}$ for the case of a single-antenna receiver $N = 1$. In this case, the codebook consists of 2^B beamforming vectors: $\mathcal{B} = \{\mathbf{g}_1, \ldots, \mathbf{g}_{2^B}\}$, with $\mathbf{g}_b \in \mathbb{C}^{M\times 1}$. Assuming that the channel $\mathbf{h} \in \mathbb{C}^{1\times M}$ can be estimated ideally, the user chooses the codeword in $\mathcal{B}$ which maximizes its rate:

$$\max_{\mathbf{g}\in\mathcal{B}} \log_2\left(1 + |\mathbf{hg}|^2 \frac{P}{\sigma^2}\right) = \arg\max_{\mathbf{g}\in\mathcal{B}} |\mathbf{hg}|. \tag{2.77}$$

If the channel realizations are drawn from a finite, discrete distribution of 2^B M-dimensional vectors, one would design the codebook to consist of these vectors. The rate-maximizing codeword would be the (normalized) channel vector which corresponds to the maximal ratio transmitter (MRT). Assuming the channels could be estimated without error and the B feedback bits from each user could be received without error, the transmitter would achieve ideal CSIT. In practice, because the channel realizations are drawn from a continuous distribution, the codewords should be designed to optimally span the distribution, as determined by the channel correlation and desired performance metric.

At one extreme, the antennas are spatially uncorrelated, and the MISO channel coefficients each have an i.i.d. Rayleigh distribution. In this case the normalized realization $\mathbf{h}/||\mathbf{h}||$ is distributed uniformly on an M-dimensional unit hypersphere. The optimal rate maximizing strategy is to distribute the 2^B codewords as uniformly as possible on the surface of the hypersphere [65]. This problem is known as the Grassmannian line packing problem: design the codebook $\mathcal{B}$ to maximize the minimum distance between any two codewords

$$\sqrt{1 - \max_{i\neq j} |\mathbf{g}_i^H \mathbf{g}_j|^2}. \tag{2.78}$$

At the other extreme, the antennas are totally correlated, for example in a line-of-sight channel with zero angle spread. Figure 2.17 shows a linear array with M elements lying on the x-axis with uniform spacing d and a user with direction θ with respect to the x-axis. Let us consider the channel response $\mathbf{h}$ measured by a user lying in the general direction $\theta \in [0°, 360°)$. If the channel coefficient of the first element is $h_1 = \alpha \exp(j\gamma)$, then the coefficient at the mth element$(m = 1, ..., M)$ is

$$h_m(\theta) = \alpha \exp\left(\frac{2\pi j d}{\lambda}(m-1)\cos\theta + j\gamma\right), \tag{2.79}$$

where λ is the carrier wavelength. We can use MRT to create a beamforming vector $\mathbf{g}(\theta^*)$ in the direction θ^* by matching the phase of the beamforming weight $g_m(\theta^*)$ to the phase of the channel coefficient $h_m(\theta^*)$, modulo the phase offset γ. With $d = \lambda/2$, the resulting MRT beamforming vector is

$$\mathbf{g}(\theta^*) = \frac{1}{\sqrt{M}} \begin{bmatrix} 1 \\ \exp(\pi j \cos\theta^*) \\ \vdots \\ \exp(\pi j(M-1)\cos\theta^*) \end{bmatrix}. \tag{2.80}$$

Using this beamforming vector, the SNR of a user lying in the direction θ is

$$|\mathbf{h}^H(\theta)\mathbf{g}(\theta^*)|^2 P/\sigma^2. \tag{2.81}$$

The MRT beamforming vector creates a *directional beam* (pointing in the direction θ^*) in the sense that the transmitted signal is co-phased to maximize the SNR of a user lying direction $\theta = \theta^*$. Figure 2.18 shows the MRT beam response for a linear array with $M = 4$ elements and a desired direction $\theta^* = 105°$. (The elements themselves are directional and pointing in the direction $\theta = 90°$, as described in Section 6.4.3. Otherwise, there would also be a response peak in the direction $\theta^* + 180°$.) Codewords could be designed to form directional beams spanning a desired range. For example, if users lie in a 120-degree sector $\theta \in [30°, 150°]$, we could choose to span this range using four MRT beams with directions $\{45°, 75°, 105°, 135°\}$. A user could determine its best codeword from (2.77) and indicate its preference with only $B = 2$ bits.

More general design techniques known as robust minimum variance beamforming can be used to design beamforming vectors for arbitrary antenna array configurations that are robust enough to withstand mismatch between

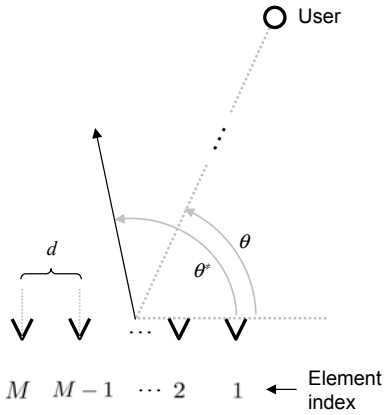

Fig. 2.17 An M-element linear array with inter-element spacing d. The direction of the user is θ, and a beam is pointed in the direction $\theta^* = 105°$.

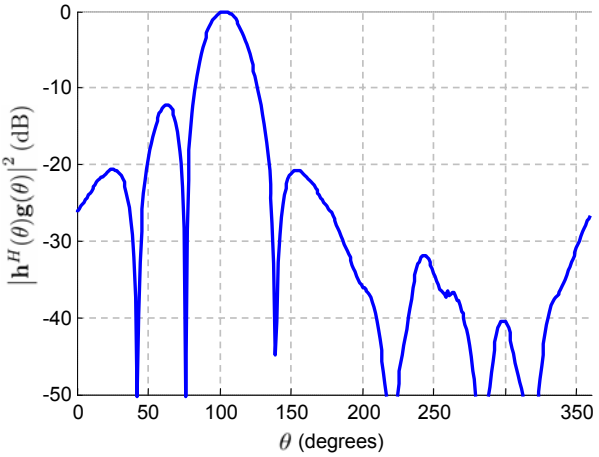

Fig. 2.18 The directional response as a function of the user direction θ for a MRT beamforming vector (2.80) pointing in the direction $\theta^* = 105°$.

measured and actual channel state information $\mathbf{h}(\theta)$ [66]. The design of equal-gain beamformers with limited feedback [67] is also relevant for antenna arrays where the amplitudes of the channel coefficients are highly correlated.

For intermediate situations where the spatial channels are neither fully correlated nor totally uncorrelated, systematic codebook designs have been proposed in [68]. Codebooks can also be designed implicitly using a training sequence of channel realizations drawn from a given spatial correlation function. This technique, based on the Lloyd-Max algorithm [69], is effective for creating specifically tailored codebooks for arbitrary spatial correlations. The training sequence $\{\mathbf{H}_j\}_{j=1}^{N_{TS}}$ is of size N_{TS}, and its elements $\mathbf{H}_j$ are realizations of MIMO channels drawn from a given spatial correlation function. If we let $\mu(\mathbf{H}_j, \mathbf{g}_i)$ be a performance metric for a given channel realization and codebook vector, the algorithm iteratively maximizes the average performance metric

$$\max_{\mathcal{B}} \frac{1}{N_{TS}} \sum_{i=1}^{2^B} \sum_{\mathbf{H}_j \in \mathcal{R}_i} \mu(\mathbf{H}_j, \mathbf{g}_i), \tag{2.82}$$

where $\mathcal{R}_i$ is the partitioned region of the training sequence associated with codeword $\mathbf{g}_i$. The size of the training sequence needs to scale at least linearly with the number of desired codewords to achieve good performance [69], hence the complexity of codebook design scales at least exponentially with the number of feedback bits B. However, because the codebook generation can be performed offline as long as the correlations are known beforehand, the complexity of the algorithm is not an issue. We also note that the algorithm converges to a maximum that is not guaranteed to be global. Nevertheless, it provides a practical way for codebook design even when the statistics of the source are not known or difficult to characterize.

2.3.7.2 Multi-antenna receiver, $N > 1$

If the receiver has multiple antennas ($N > 1$), the beamforming techniques discussed for the case of single-antenna receivers could be used, and the received signal could be coherently combined across the N antennas. For i.i.d. Rayleigh channels with $M > 1$ and $N > 1$, the Grassmannian solution has been shown to maximize the beamforming rate [70] [71] [72].

In order to exploit the potential of spatial multiplexing, precoding matrices with rank $J > 1$ (and $J \leq \min(M, N)$) could be used. It is common to use *multidimensional eigenbeamforming*, where the columns of the precoding

matrix $\mathbf{G} \in \mathbb{C}^{M \times J}$ are orthogonal such that $\mathbf{G}^H\mathbf{G}$ is a diagonal matrix. In doing so, the J streams are transmitted on mutually orthogonal subspaces, as is the case when precoding to achieve closed-loop MIMO capacity. We assume the symbols of the data vector $\mathbf{u} \in \mathbb{C}^J$ are independent and normalized such that $\mathbb{E}\left(\mathbf{u}\mathbf{u}^H\right) = \mathbf{I}_J$. Because the transmit power is $\mathrm{tr}\,\mathbb{E}\left(\mathbf{G}\mathbf{u}\mathbf{u}^H\mathbf{G}^H\right) = P$, we have that $\mathbf{G}^H\mathbf{G} = \mathrm{diag}(P_1, \ldots, P_J)$, where P_j $(j = 1, ..., J)$ is the power allocated to stream j, and $\sum_{j=1}^{J} P_j = P$.

Compared to transmitting with equal power on each stream, non-uniform power allocation requires more feedback and may not result in significant performance gains, especially if the channel is spatially uncorrelated. As described in Chapter 7, spatial multiplexing in 3GPP standards is achieved using codebook-based precoding with equal power on each stream.

A special case of multidimensional eigenbeamforming is to use antenna subset selection, where the columns of the precoder $\mathbf{G}$ are uniquely drawn from the columns of the $M \times M$ identity matrix and appropriately normalized [73]. In doing so, $J \leq \min(M, N)$ streams are uniquely associated with a subset of J transmit antennas. The case of $J = M$ corresponds to the V-BLAST transmission.

2.4 Practical considerations

2.4.1 CSI estimation

In deriving the MIMO capacity and capacity-achieving techniques, we have assumed that the CSI is known perfectly at the receiver and, when necessary for closed-loop MIMO, at the transmitter. In practice, estimates of the CSI at the receiver can be obtained from training signals (also known as pilot or reference signals) sent over time or frequency resources that are orthogonal to the data signals' resources. For M transmit antennas, the optimal training set consists of M mutually orthogonal signals, one assigned for each antenna and with equal power [74]. The reliability of the CSI estimates and the resulting rate performance depend on the fraction of resources devoted to the training signals and the rate of channel variation. As the channel varies more rapidly, additional training resources are required to achieve the same reliability in the channel estimates. (Reference signals are described in more detail in Section 4.4.)

To achieve closed-loop MIMO capacity, the CSI at the transmitter is assumed to be known ideally. If the CSIT is unreliable (see Section 4.4 for acquiring CSIT), the performance will be degraded. In high SNR channels, isotropic transmission should be used as an alternative because it requires no CSIT. In low SNR channels, precoding with limited feedback could be used to provide performance that is more robust to unreliable CSIT.

2.4.2 Spatial richness

The numerical results in this chapter assume that the MIMO channel coefficients are i.i.d. Rayleigh. As mentioned in Section 2.1, the spatial correlation between antennas depends on their spacing relative to the height of the surrounding scatterers. For a base station antenna that is high above the clutter (for example in rural or suburban deployments), a common rule of thumb is that spatial decorrelation could be achieved if the separation is at least 10 wavelengths [75]. On the other hand, if the base antenna is surrounded by scatterers of the same height (for example in rooftop urban deployments), decorrelation could be achieved with separation of only a few wavelengths. Decorrelation between pairs of base station antenna elements can also be achieved using cross-polarized antennas (Section 5.5.1). Mobile terminals are typically assumed to be surrounded by scatterers, and half-wavelength separation is considered sufficient for decorrelation [76].

Full multiplexing gain can be achieved over a SU-MIMO channel if the antenna array elements at both the transmitter and receiver are uncorrelated. As the antennas at either the transmitter or receiver become more correlated, the average capacity of the channel decreases. If the antennas at either end become fully correlated, then the channel cannot support multiplexing, and the multiplexing gain is 1. General characterizations of the capacity as a function of antenna correlation are given in [77].

2.5 Summary

Multiple-antenna techniques can be used to improve the throughput and reliability of wireless communication. In this chapter, we discussed the single-

user (M, N) MIMO link where the transmitter is equipped with M antennas and the receiver is equipped with N antennas.

- The open-loop and closed-loop capacity measure the maximum rate of arbitrarily reliable communication for the case where the channel state information (CSI) is respectively known and not known at the transmitter. CSI at the receiver is always assumed.
- The MIMO capacity in a spatially rich channel scales linearly with the number of antennas. At high SNRs, a multiplexing gain of $\min(M, N)$ is achieved by transmitting multiple streams simultaneously from multiple antennas. At low SNRs, a power gain of N is achieved through receiver combining. CSIT for closed-loop MIMO allows more efficient power distribution, resulting in a higher capacity.
- Closed-loop MIMO capacity can be achieved using linear precoding and linear combining at the transmitter and receiver, respectively, where the transformations are based on the singular-value decomposition of the channel matrix $\mathbf{H}$. Waterfilling is used to determine the optimal power allocation for each of the streams.
- Open-loop MIMO capacity for an (M, N) link (with $M \leq N$) can be achieved using isotropic transmission and an MMSE-SIC receiver. The V-BLAST transmit architecture achieves capacity by sending independent streams with appropriate rate assignment on each antenna. The D-BLAST transmit architecture achieves capacity by cyclically shifting the association of streams with transmit antennas.
- Space-time coding provides diversity gain for open-loop MISO channels. The Alamouti space-time block code achieves the capacity of a (2,1) MISO channel, but otherwise space-time coding cannot achieve the capacity of general MIMO antenna configurations. Space-time coding does not provide multiplexing gain and therefore provides only modest throughput gains as a result of diversity.
- In FDD systems where CSI at the transmitter is not readily available, feedback from the receiver can be used to index a fixed set of precoding matrices, providing suboptimal performance compared to the closed-loop MIMO capacity.

http://www.springer.com/978-0-387-77521-0

MIMO Communication for Cellular Networks
Huang, H.; Papadias, C.B.; Venkatesan, S.
2012, XVI, 316 p., Hardcover
ISBN: 978-0-387-77521-0