1 Introduction

The practical demonstration of the vertical Bell laboratories layered space-time architecture (V-BLAST) multiantenna wireless system by Bell Labs [1], and the theoretical prediction of very high wireless channel capacities in rich scattering environments by Telatar in [2] and Foschini and Gans in [3] in the late 1990s opened up immense possibilities and created wide interest in multiantenna wireless communications. Since then, multiantenna wireless systems, more commonly referred to as multiple-input multiple-output (MIMO) systems, have become increasingly popular. The basic premise of the popularity of MIMO is its theoretically predicted capacity gains over single-input single-output (SISO) channel capacities. In addition, MIMO systems promise other advantages like increased link reliability and power efficiency. Realizing these advantages in practice requires careful exploitation of large spatial dimensions.

Significant advances in the field of MIMO theory and practice have been made as a result of the extensive research and development efforts carried out in both academia and industry [4]–[7]. A vast body of knowledge on MIMO techniques including space-time coding, detection, channel estimation, precoding, MIMO orthogonal frequency division multiplexing (MIMO-OFDM), and MIMO channel sounding/modeling has emerged and enriched the field. It can be safely argued that MIMO systems using 2 to 4 antennas constitute a fairly mature area now. Technological issues in such small systems are fairly well understood and practical implementations of these systems have become quite common. Indeed, MIMO techniques have found their way into major wireless standards like long-term evolution (LTE) and WiFi (IEEE 802.11n/ac), leading to the successful commercial exploitation of MIMO technology.

At this point in time, when numerous papers and several books on MIMO have already been written and MIMO implementations are increasingly being embedded in wireless products, a natural question that arises is "What is next in MIMO?" One can pose this question a little differently: "Have we exploited the full potential of MIMO?" In addressing this question, one can realize that the main potential of MIMO, which is the feasibility of achieving "very" high spectral efficiencies/sum-rates, has not yet been well exploited in practice. Although the early days of MIMO technology witnessed the practical demonstration by Bell Labs of spectral efficiencies as high as 24 bps/Hz using 8 transmit and 12 receive antennas (the V-BLAST system) in fixed indoor environments, a majority of the

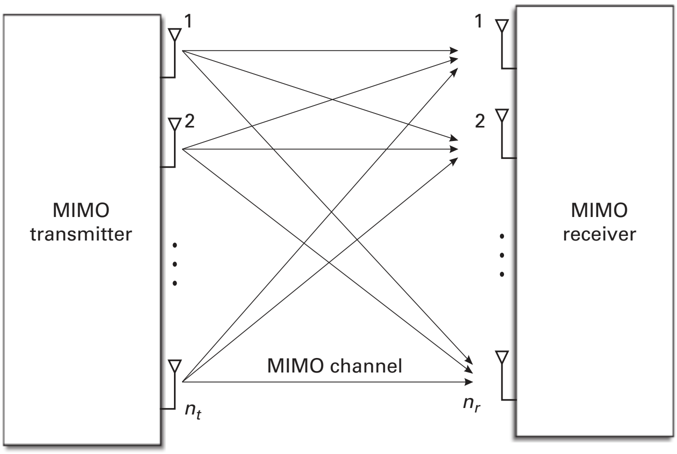

Figure 1.1 Point-to-point MIMO system.

subsequent research and development activities in MIMO seem to have focused on systems with fewer antennas and much lower spectral efficiencies. Except in a few cases (e.g., DoCoMo's 12×12 MIMO system [8]), MIMO configurations with 2–4 antennas and lower than 15 bps/Hz spectral efficiency have dominated MIMO research and development efforts since 2000. However, "large MIMO systems" (MIMO systems with a large number of antennas) are now a practical proposition and, hence, interest in them is growing [9]–[13].

The term large MIMO systems refers to systems in which large numbers (tens to hundreds) of antennas are employed in communication terminals. For example, in a point-to-point MIMO wireless link (Fig. 1.1), both the transmitter and receiver sides can be provided with a large number of antennas to achieve increased data rates without increasing bandwidth (i.e., they achieve very high spectral efficiencies). High-speed wireless backhaul connectivity between base stations (BSs) can adopt such a point-to-point MIMO configuration. In point-to-multipoint MIMO communication (e.g., multiuser MIMO downlink), the BS can be provided with a large number of transmit antennas for multiuser precoding and the user terminal can have one or more receive antennas (Fig. 1.2) so that increased sum-rates can be achieved. Likewise, in multipoint-to-point MIMO communication (e.g., multiuser MIMO uplink), each user terminal can transmit using one or more transmit antennas and the BS can receive through a large number of receive antennas and perform multiuser detection (MUD) using spatial signatures of all the users.

1.1 Multiantenna wireless channels

Multiantenna wireless channels are a broad category of channels that include point-to-point and multiuser channels. One of the defining characteristics of a

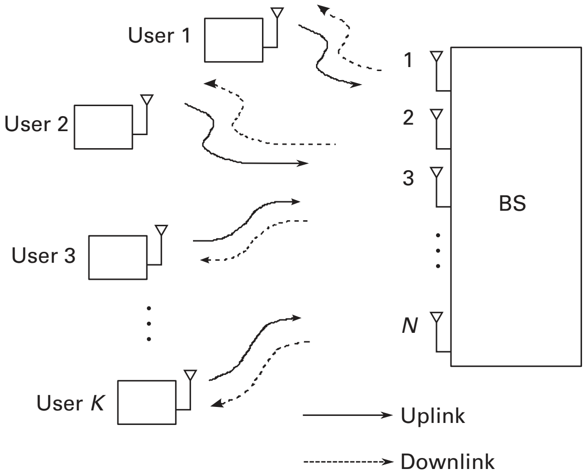

Figure 1.2 Multiuser MIMO system.

wireless channel is the variation of the channel strength over time and over frequency [6],[14]. These variations are typically classified into two types: large scale fading and small scale fading. Large scale fading is due to path loss as a function of distance and shadowing by large objects like buildings, bridges, trees, etc., and is typically frequency independent. Small scale fading, on the other hand, is due to the constructive and destructive interference of the multiple signal paths between the transmitter and receiver. Small scale fading happens at the spatial scale of the order of the carrier wavelength, and is frequency dependent. The channel is then classified as frequency-selective or frequency-flat. When the signaling bandwidth is larger than the coherence bandwidth of the channel (which has an inverse relation with the maximum delay spread of the channel), the channel is frequency-selective [14]. On the other hand, in frequency-flat channels, the signaling bandwidth is much smaller than the coherence bandwidth of the channel. Even when the channel is frequency-selective, techniques like OFDM can convert the channel into multiple frequency-flat channels on which the techniques designed for frequency-flat fading can be employed.

In terms of time variation, wireless channels are further classified as slowly fading or fast fading, depending on the fade rate relative to the signaling rate. If the fade remains constant over the signaling duration, the fading is termed slow (or time-flat) fading, whereas if the fade varies within the signaling duration, it is termed fast (or time-selective) fading. The carrier wavelength and velocity of the communication terminal determine the amount of time selectivity (or Doppler spread) in the channel [14].

Most multiantenna wireless channels with n_t transmit and n_r receive antennas (Fig. 1.1) are modeled as a linear channel with an equivalent baseband channel matrix $\mathbf{H}_c \in \mathbb{C}^{n_r \times n_t}$. The (i, j)th entry of $\mathbf{H}_c$ represents the channel gain from the jth transmit antenna to the ith receive antenna. The channel gains are also referred to as the "channel state information (CSI)." The availability of knowledge of these gains at the receiver and transmitter is an important factor which decides the performance of the communication system. CSI at the receiver (CSIR) refers to the scenario where the receiver has knowledge of the channel gains. Likewise, CSI at the transmitter (CSIT) refers to the scenario where the transmitter has knowledge of the channel gains. In fast fading channels, accurate estimation of the channel gains can become an issue, in which case non-coherent or blind techniques can be considered. In addition, obtaining CSIT through feedback can become ineffective in fast fading. However, in applications where the channel is not varying fast, it is generally possible to estimate the channel gains accurately through pilot-assisted transmission. Also, CSIT based on measured CSI fed back from the receiver is effective in such slow fading channels.

The channel gains can be independent or correlated, depending on various factors including the spacing between antenna elements, the amount of scattering in the environment, pin-hole effects, etc. Mathematical models that characterize the spatial correlation in MIMO channels are used in the performance evaluation of MIMO systems. Spatial correlation at the transmit and/or receive side can affect the rank structure of the MIMO channel resulting in degraded MIMO capacity. The structure of scattering in the propagation environment can also affect the capacity. In addition, transmit correlation in MIMO fading can be exploited by using non-isotropic inputs (precoding) based on knowledge of the channel correlation matrices.

1.2 MIMO system model

A widely studied point-to-point MIMO system model is one which is assumed to be frequency-flat and slow fading. The channel gains are assumed to remain constant over the signaling interval. With these assumptions, the equivalent complex baseband MIMO system model can be written as

$$\mathbf{y}_c = \mathbf{H}_c \mathbf{x}_c + \mathbf{n}_c, \tag{1.1}$$

where $\mathbf{x}_c \in \mathbb{C}^{n_t}$ is the transmitted vector, $\mathbf{y}_c \in \mathbb{C}^{n_r}$ is the received vector, and $\mathbf{n}_c \in \mathbb{C}^{n_r}$ is the additive white Gaussian noise (AWGN) vector. The jth entry of $\mathbf{x}_c$ is the symbol transmitted from the jth transmit antenna, $j = 1, \ldots, n_t$. In a typical communication system, information bits (e.g., the output of some source coder which performs voice or image compression, followed by a channel coder) are grouped into messages, and each message then corresponds to an n_t-dimensional complex vector, whose jth component is transmitted from the jth transmit antenna. In practice, these vectors belong to some codebook $\mathcal{X}$. The transmitter groups $R = \log_2 |\mathcal{X}|$ bits into a message, which is then used to index the codebook. R is often referred to as the rate of the codebook or simply as the rate of transmission. Alternatively, the $\mathbf{x}_c$ vector could be a pilot symbol vector (known to the receiver) during the training phase in a pilot-aided channel estimation scheme. The ith entry of $\mathbf{y}_c$ is the signal received at the ith receive antenna, $i = 1, \ldots, n_r$. Assuming a rich scattering environment, the entries of the channel matrix $\mathbf{H}_c$ are often modeled as independent and identically distributed (iid) $\mathcal{CN}(0, 1)$. Since the transmitter is power constrained, we have $\mathbb{E}[\mathrm{tr}(\mathbf{x}_c\mathbf{x}_c^H)] = P$, where P is the total power available at the transmitter. Also, $\mathbb{E}[\mathbf{n}_c\mathbf{n}_c^H] = \sigma^2\mathbf{I}_{n_r}$, where σ^2 is the noise variance at each receive antenna. The average received signal-to-noise ratio (SNR) at each receive antenna is given by $\gamma = P/\sigma^2$.

The MIMO signal detection problem at the receiver can be stated as: given $\mathbf{y}_c$ and knowledge of $\mathbf{H}_c$, determine $\widehat{\mathbf{x}}_c$, an estimate of the transmitted symbol vector $\mathbf{x}_c$. Likewise, the MIMO channel estimation problem in a training based scheme can be stated as: given knowledge of the transmitted pilot symbol vector $\mathbf{x}_c$, determine $\widehat{\mathbf{H}}_c$, an estimate of the channel gain matrix $\mathbf{H}_c$.

1.3 MIMO communication with CSIR-only

In communication channels, error probability is one of the key performance indicators. Most communication schemes employ channel coding schemes to increase robustness against errors. To achieve an arbitrarily low probability of error, the rate of transmission R must be strictly below the MIMO channel capacity. The MIMO channel capacity is dependent on $\mathbf{H}_c$ and the transmit covariance matrix $\mathbf{K}_x \stackrel{\triangle}{=} \mathbb{E}[\mathbf{x}_c\mathbf{x}_c^H]$, $\mathrm{tr}\{\mathbf{K}_x\} = P$, and is given by

$$C_{MIMO}(\gamma, \mathbf{H}_c, \mathbf{K}_x) = \log_2 \det\left(\mathbf{I}_{n_r} + \frac{1}{\sigma^2}\mathbf{H}_c\mathbf{K}_x\mathbf{H}_c^H\right). \tag{1.2}$$

In the case of availability of CSIR only (i.e., no CSIT), since the transmitter has no knowledge of the channel gains, it cannot adapt its transmission scheme with respect to the channel gains. Therefore, for a fixed γ and rate R, the transmitter uses a fixed codebook, which does not change with changing channel gains. The transmitter codebook selection is very much dependent on whether the channel is slow fading or fast fading.

1.3.1 Slow fading channels

In slow fading channels, where the channel does not change during the length of the codeword, if the channel is such that $C_{MIMO}(\gamma, \mathbf{H}_c, \mathbf{K}_x) < R$, then no detector can recover the transmitted codeword correctly, and the channel is said to be in “outage.” Hence, for slow fading channels with CSIR only, outage cannot be avoided and it is impossible to achieve an arbitrary low probability of error. In such scenarios, an appropriate performance indicator of any encoding–decoding scheme is the codeword error probability or codeword error rate. For codewords of large length, the theoretical limit for the codeword error rate of any encoding–decoding scheme is the channel outage probability, which is defined as

$$P_{outage}(\gamma, R) = \min_{\mathbf{K}_x \,|\mathrm{tr}\{\mathbf{K}_x\}=P} p\big(C_{MIMO}(\gamma, \mathbf{H}_c, \mathbf{K}_x) < R\big). \tag{1.3}$$

Any practical encoding–decoding scheme would have a codeword error rate more than the channel outage probability given in (1.3). Therefore, it is important to design transmit schemes and corresponding receivers which can perform very close to the channel outage probability for all values of γ and R. For slow fading channels, there are two important parameters, namely, diversity gain and multiplexing gain. Diversity gain is a measure of reliability, whereas multiplexing gain is a measure of the degrees of freedom in the MIMO channel. These two parameters are usually related by the so called diversity–multiplexing gain tradeoff [15]. The maximum diversity gain achievable is $n_r n_t$ and the maximum multiplexing gain achievable is $\min(n_r, n_t)$. When the rate of transmission R is fixed, the limiting value (as $\gamma \to \infty$) of the negative of the slope of $\log(P_{outage}(\gamma, R))$ wrt $\log \gamma$ can be no more than $n_r n_t$. For a given scheme, we can therefore define the diversity order achievable (with fixed R) as

$$d = - \lim_{\gamma \to \infty} \frac{\log(P_e(\gamma))}{\log \gamma}, \tag{1.4}$$

where $P_e(\gamma)$ is the codeword error rate of the scheme. For simple MIMO schemes like V-BLAST [16], it can be shown that the maximum diversity order achievable is only n_r. This is because symbols transmitted from the antennas in V-BLAST are independent, and each such symbol reaches the receiver only through n_r different paths.

Space-time block coding is a well-known technique which can achieve the full diversity gain of $n_r n_t$ [5]. To achieve full diversity, symbols are coded across both space and time. Orthogonal space-time block codes (STBC) allow simple decoding achieving full diversity [17],[18]. However, they make sacrifices regarding the multiplexing rate, and are therefore not suited for systems with high target spectral efficiencies. Subsequent to orthogonal STBCs, several high-rate and high-diversity STBCs were proposed. One such class of STBCs is non-orthogonal STBCs (NO-STBCs) from cyclic division algebras (CDAs) [19],[20]. STBCs from CDA can achieve the full diversity of $n_r n_t$ without sacrificing rate.

1.3.2 Fast fading channels

In fast fading channels, the channel fade changes multiple times in the duration of the codeword. By spreading portions of the codeword across multiple fades, the reliability of codeword reception can be improved. In such a scenario, if the MIMO channel is ergodic, in the limit of infinitely long codewords, it is possible to achieve error-free communication if the rate of transmission R satisfies

$$R \leq C_{ergodic}(\gamma) \triangleq \max_{\mathbf{K}_x \,|\mathrm{tr}\{\mathbf{K}_x\}=P} \mathbb{E}_{\mathbf{H}_c}\big[C_{MIMO}(\gamma, \mathbf{H}_c, \mathbf{K}_x)\big]. \tag{1.5}$$

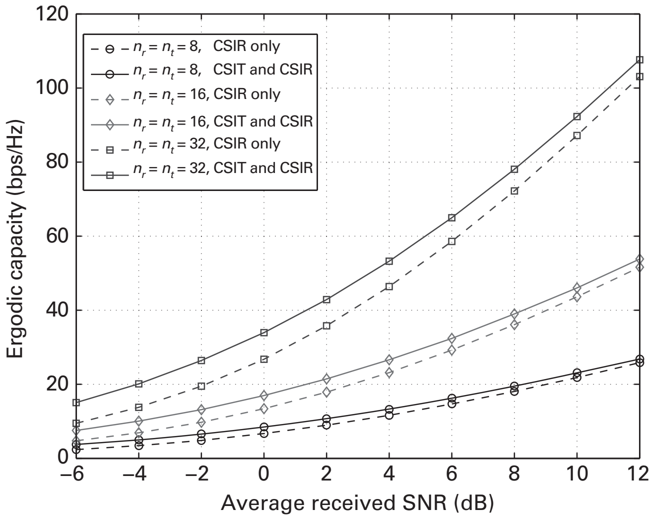

Figure 1.3 Ergodic MIMO capacity for increasing $n_t = n_r$ with (i) CSIR only, and (ii) CSIT and CSIR.

$C_{ergodic}(\gamma)$ is often referred to as the ergodic MIMO capacity, and is achieved with $\mathbf{K}_x = (P/n_t)\mathbf{I}_{n_t}$. This transmit architecture is also known as the V-BLAST scheme, where the symbol streams transmitted from each transmit antenna are uncorrelated. The ergodic MIMO capacity is therefore given by

$$C_{ergodic}(\gamma) = \mathbb{E}_{\mathbf{H}_c}\Big[\log_2 \det\Big(\mathbf{I}_{n_r} + \frac{\gamma}{n_t}\mathbf{H}_c\mathbf{H}_c^H\Big)\Big]. \tag{1.6}$$

The ergodic MIMO capacity increases linearly with increasing $n_t = n_r$ [2]. In Fig. 1.3, the ergodic MIMO capacity is plotted as a function of the average received SNR, γ, for the case with CSIR only as well as for the case with CSIT and CSIR, for different values of $n_t = n_r$. It can be observed that, for a given SNR, the ergodic MIMO capacity increases linearly with $n_t = n_r$. For example, at an SNR of 6.8 dB, the ergodic capacity is 16, 32, and 64 bps/Hz for $n_t = n_r = 8, 16, 32$, respectively. This implies that, at an SNR of 6.8 dB, an $n_t = n_r$ MIMO system would have an ergodic capacity of roughly $2n_t$ bps/Hz.

1.4 MIMO communication with CSIT and CSIR

If MIMO systems are operating in time division duplex (TDD) mode or if MIMO channels are slowly varying, it is possible for both the transmitter as well as the receiver to acquire the CSI. When both CSIT and CSIR are available, the ergodic MIMO capacity is known to be achieved with independent Gaussian inputs beamformed along the right singular vectors of the channel matrix. This transforms the MIMO channel into a set of parallel non-interfering $n = \min(n_t, n_r)$ subchannels. Capacity is then achieved by waterfilling power allocation among these n subchannels [2]. Note that the optimal power allocation is isotropic for the case of CSIR-only.

When CSIT is available, it is possible to use the available power judiciously by allocating more power to the subchannel with higher channel gain. At low SNRs, the availability of CSIT in addition to CSIR has an even higher impact on the ergodic capacity when compared to the CSIR-only scenario. This is because, at low SNRs, capacity is known to increase almost linearly with SNR, and therefore with CSIT the transmitter allocates all available power to the subchannel with the highest channel gain. In contrast to this, with CSIR-only, the available power is equally divided among the subchannels, resulting in a lower achievable capacity when compared to the scenario with CSIT. At high SNRs, waterfilling power allocation distributes roughly equal power to all the subchannels. Therefore, power allocation at high SNRs is almost the same for scenarios with CSIR-only as well as those with CSIR and CSIT. This implies that at high SNRs, both CSIR-only and the CSIR and CSIT scenarios have roughly the same ergodic capacity. This can be seen in Fig. 1.3, where the "CSIT and CSIR" ergodic capacity is plotted, in addition to the "CSIR only" capacity. It can be observed that indeed, for a given $n_t = n_r$ and SNR, γ, the ergodic capacity with "CSIT and CSIR" is more than the ergodic capacity with "CSIR only." Also, the gap between the ergodic capacity of the two scenarios reduces with increasing SNR. Another important fact, which is not highlighted in Fig. 1.3 is that, at low SNRs, the ergodic capacity with "CSIT and CSIR" is more than $n \log_2(1+\gamma)$, which is the capacity of n parallel, independent SISO non-faded AWGN channels [6].

In slow fading channels, the codewords transmitted are subject to block fading (i.e., the channel remains almost the same for the duration of the transmitted codeword). As pointed out earlier, in such block fading scenarios, if the capacity of the channel is below the rate of transmission, there will always be a codeword error (outage) irrespective of the coding scheme used. With the availability of CSIT, however, it is possible to theoretically achieve zero outage probability by adapting the transmitted codewords (i.e., codeword rate and transmit power) for a given long-term average power constraint [21]. This leads to a variable rate transmission scheme, and also a large peak to average requirement on the transmit radio frequency (RF) amplifiers, which are undesirable in many applications. Therefore, in such applications, it is obvious that outages cannot be avoided. Hence, it is important that encoding and decoding schemes are devised to achieve high diversity and multiplexing gains. The maximum diversity gain is $n_t n_r$ and the maximum multiplexing gain is $\min(n_t, n_r)$. CSIT can be used to encode the information symbols into transmit vectors, a process commonly called "precoding." Several precoding schemes are known in the literature. Most known precoding schemes (or precoders for short) achieve either (i) high rate or high diversity at low complexity (e.g., linear precoders and non-linear precoders based on Tomlinson–Harashima precoding) or (ii) both high rate and high diversity but at high complexity (e.g., precoders based on lattice reduction techniques and vector perturbation).

Table 1.1. *Reliability and capacity of SISO, SIMO, and MIMO channels*

Number of antennas	Error probability (P_e)	Capacity (C), bps/Hz
SISO ($n_t = n_r = 1$)		
non-fading	$P_e \propto e^{-\gamma}$	$C = \log_2(1+\gamma)$
fading	$P_e \propto \gamma^{-1}$	$C = \log_2(1+\gamma)$
SIMO ($n_t = 1, n_r > 1$)		
fading	$P_e \propto \gamma^{-n_r}$	$C = \log_2(1+\gamma)$
MIMO ($n_t > 1, n_r > 1$)		
fading	$P_e \propto \gamma^{-n_t n_r}$	$C = \min(n_t, n_r)\log_2(1+\gamma)$

1.5 Increasing spectral efficiency: quadrature amplitude modulation (QAM) vs MIMO

The achievable link reliability (in terms of probability of error) and capacity in bps/Hz in SISO ($n_t = n_r = 1$), SIMO ($n_1 = 1, n_r > 1$), and MIMO ($n_t > 1, n_r > 1$) channels are summarized in Table 1.1. In non-fading SISO AWGN channels, the probability of error falls exponentially with increasing SNR, whereas the capacity grows only logarithmically with increasing SNR. With fading, the probability of error degrades to a linear fall with increase in SNR; this is a detrimental effect of fading in SISO channels. This performance degradation in fading can be alleviated by using more receive antennas, which offers receive diversity. That is, in SIMO fading channels, the probability of error falls with SNR as γ^{-n_r}. While this means better error performance in SIMO fading compared to SISO fading, the capacity of SIMO fading, like that of SISO fading, grows only logarithmically with increasing SNR. That is, in SISO and SIMO fading channels, significant power increase is needed to increase capacity. However, MIMO fading channels are attractive in terms of both achievable reliability as well as capacity. The probability of error in MIMO channels falls with SNR as $\gamma^{-n_t n_r}$, which approaches an exponential fall for large n_t, n_r. More importantly, the MIMO channel capacity increases linearly with the minimum of the number of transmit and receive antennas, which is much better than the logarithmic increase in capacity with increasing SNR.

Spectral efficiency in communication systems can be increased by increasing the size of the modulation alphabet (e.g., increasing M in M-QAM), or increasing the number of spatial dimensions for signaling (i.e., increasing n_t), or a combination of both. To achieve a given spectral efficiency, using a small modulation alphabet size and increasing number of antennas is more power efficient than using a small number of antennas and increasing the modulation alphabet size. This can be illustrated as follows. Consider the communication systems in Fig. 1.4. The system in Fig. 1.4(a) uses one transmit antenna and 64-QAM to achieve 6 bps/Hz spectral efficiency. The system in Fig. 1.4(b) achieves the same spectral efficiency using six transmit antennas and binary phase shift keying

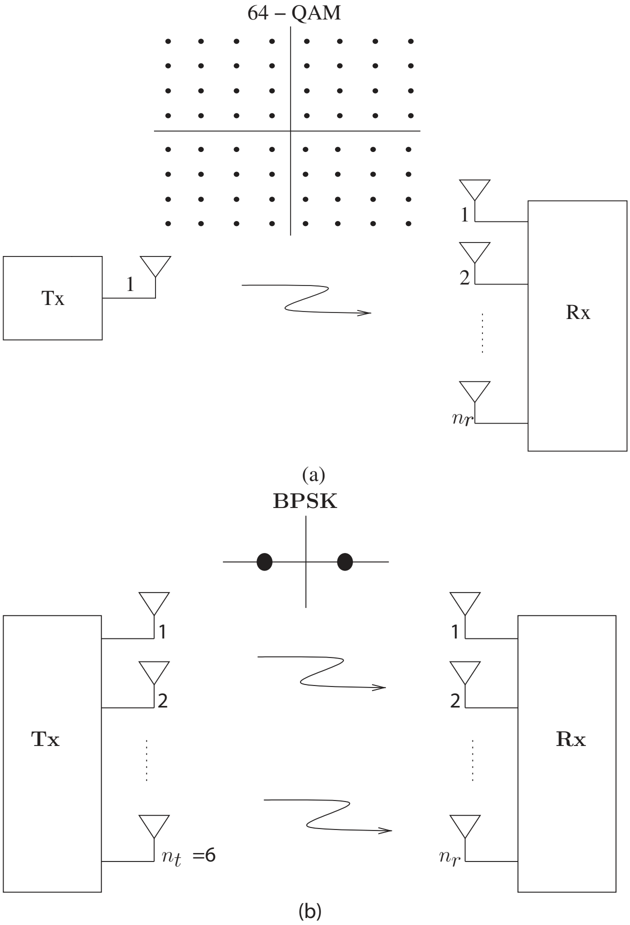

Figure 1.4 Communication systems with 6 bps/Hz spectral efficiency: (a) SISO/SIMO with 64-QAM. (b) MIMO with $n_t = 6$ and BPSK. (Rx: receiver; Tx: transmitter.)

(BPSK). The achieved bit error rates (BERs) (p_e) versus SNR (γ) performances of these systems are compared in Fig. 1.5. The performance of the 64-QAM with $n_t = n_r = 1$ is least power efficient. As mentioned earlier, in this SISO fading case p_e falls as γ^{-1}. This can be seen by noting that for a 10 dB increase in γ, the p_e falls by one order: e.g., p_e improves from 2×10^{-3} at $\gamma = 35$ dB to 2×10^{-4} at $\gamma = 45$ dB. By increasing the number of receive antennas from $n_r = 1$ to $n_r = 6$, keeping $n_t = 1$ and 64-QAM, the performance improves due to receive diversity. However, the MIMO system with $n_t = n_r = 6$ and BPSK significantly outperforms the SIMO and SISO systems with 64-QAM. At a p_e of 10^{-3}, the $n_t = n_r = 6$ MIMO system with BPSK is power efficient by more than 6 dB compared to the $n_t = 1, n_r = 6$ SIMO system with 64-QAM. This illustrates the better power efficiency of MIMO systems compared to SIMO and SISO systems with the same spectral efficiency. Therefore, increasing the num-