Diversity in Communication: From Source Coding to Wireless Networks*

Suhas Diggavi
School of Computer and Communication Sciences
Laboratory of Information and Communication Systems (LICOS)
École Polytechnique Fédérale de Lausanne (EPFL)
EPFL-IC-ISC-LICOS, Building INR, Room 112, Station 14,
Lausanne, Switzerland, CH 1015.
Email: suhas.diggavi@epfl.ch

Abstract

Randomness is an inherent part of network communications. We broadly define diversity as creating multiple independent instantiations (conduits) of randomness for conveying information. In the past few years a trend is emerging in several areas of communications, where diversity is utilized for reliable transmission and efficiency. In this chapter, we give examples from three topics where diversity is beginning to play an important role.

1 Introduction

One of the main characteristics of network communication is the uncertainty (randomness): randomness in users' wireless transmission channels, randomness in users' geographical locations in a wireless network, and randomness in route failures and packet losses in networks. The randomness we study in this chapter can have time scales of variation that are comparable to the communication transmission times. This can result in complete failures in communication and therefore affect reliability. Such "non-ergodic" losses can be combated if we somehow create independent instantiations of the randomness. We broadly define *diversity* as the method of conveying information through such multiple independent instantiations. The overarching theme of this chapter is how to create diversity and how we can use it as a tool to enhance performance. We study this idea through diversity in *multiple antennas*, *multiple users* and *multiple routes*.

*To appear as a book chapter in "Brain and Systems: New Directions in Statistical Signal Processing", edited by Simon Haykin, Jose Principe, Terry Sejnowski, and John McWhirter, MIT Press, to appear 2005.

The functional modularities and abstractions of the network protocol known as stack layering (Keshav, 1997) contributed significantly to the success of the wired Internet infrastructure. The layering achieves a form of information hiding, providing only interface information to higher layers, and not the details of the implementation. The physical layer is dedicated to signal transmission, while the data-link layer implements functionalities of data framing, arbitrating access to transmission medium and some error control. The network layer abstracts the physical and data-link layers from the upper layers by providing an interface for end-to-end links. Hence, the task of routing and framing details of the link layer are hidden from the higher layers (transport and application layers). However, as we will see, the use of diversity necessarily causes cross-layer interactions. These cross-layer interactions form a subtext to the theme of this chapter.

Wireless communication hinges on transmitting information riding on radio (electromagnetic) waves, and hence the information undergoes attenuation effects (fading) of radio waves (see Section 2 for more details). Such multipath fading is a source of randomness. Here diversity arises by utilizing independent realizations of fading in several domains: time (mobility), frequency (delay spread), and space (multiple-antennas). Over the past decade research results have shown that multiple-antenna spatial diversity (space-time) communication can not only provide robustness, but also dramatically improve reliable data rates. These ideas are having a huge impact on the design of physical layer transmission techniques in next-generation wireless systems. *Multiple-antenna diversity* is the focus of Section 3.

The wireless communication medium is naturally shared by several users using the same resources. Since the users' locations (and therefore their transmission conditions) are roughly independent, they experience independent randomness in local channel and interference conditions. Diversity in this case arises by utilizing the independent transmission conditions of the different users as conduits for transmitting information *i.e., multi-user diversity.* This can be utilized in two ways. One by allowing users access to resources when it is most advantageous to the overall network. This is a form of opportunistic scheduling and is examined in Section 4.1. The other by using the users themselves as relays to transmit information from source to destination. This is a form of opportunistic relaying, and is studied in Section 4.2. These multi-user diversity methods are the focus of Section 4.

In transmission over networks, random route failures and packet losses degrade performance. Diversity here would be achieved by creating conduits with independent probability of route failures. For example, this can be done by transmission over multiple routes with no overlapping links. A fundamental question that arises is how we can best utilize the presence of such *route diversity.* In order to utilize these conduits, multiple description source coding generates multiple codeword streams to describe a source (such as images, voice, video, etc.). The design goal is to have a graceful degradation in performance (in terms of distortion) when only subsets of the transmitted streams are received. In Section

5 we study fundamental bounds and design ideas for multiple description source coding.

Therefore, diversity not only plays a role in robustness, it can also result in remarkable gains in achievable performance over several disparate applications. The details of how diversity enhances performance are discussed in the sequel.

2 Transmission models

Since a considerable part of this chapter is about wireless communication, it is essential to understand some of the rudiments of wireless channel characteristics. In this section, we focus on models for point-to-point wireless channels and also introduce some of the basic characteristics of transmission over (wireless) networks.

Wireless communication transmits information by riding (modulation) on electromagnetic (radio) waves with a carrier frequency varying from a few hundred megahertz to several gigahertz. Therefore, the behavior of the wireless channel is a function of the radio propagation effects of the environment.

A typical outdoor wireless propagation environment is illustrated in Figure 1, where the mobile wireless node is communicating with a wireless access point (base station). The signal transmitted from the mobile may reach the access point directly (line-of-sight) or through multiple reflections on local scatterers (buildings, mountains, etc.). As a result, the received signal is affected by multiple random attenuations and delays. Moreover, the mobility of either the nodes or the scattering environment may cause these random fluctuations to vary with time. Time-variation results in the random waxing and waning of the transmitted signal strength over time. Finally, a shared wireless environment may incur interference (due to concurrent transmissions from other mobile nodes) to the transmitted signal.

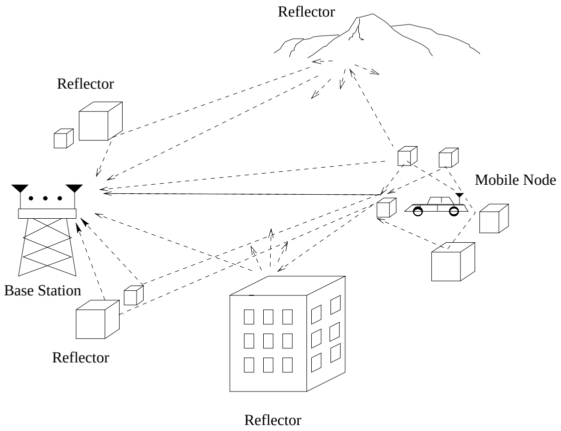

Figure 1: Radio Propagation Environment

The attenuation incurred by wireless propagation can be decomposed in three main factors: a signal attenuation due to the distance between communicating nodes (*path loss*), attenuation effects due to absorption in local structures such as buildings (*shadowing loss*), and rapid signal fluctuations due to constructive and destructive interference of multiple reflected radio wave paths (*fading loss*). Typically the path loss attenuation behaves as $\frac{1}{d^\alpha}$ as a function of distance d, with $\alpha \in [2, 6]$. More detailed models of wireless channels can be found in (Jakes, 1974; Rappaport, 1996).

2.1 Single user model

For the purposes of this chapter we start with the following model

$$y_c(t) = \int h_c(t;\tau)s(t-\tau)d\tau + z(t)\ , \tag{1}$$

where the transmitted signal $s(t) = g(t) * x(t)$ is the convolution of the information-bearing signal $x(t)$ with $g(t)$, the transmission shaping filter, $y_c(t)$ is the continuous time received signal, $h_c(t;\tau)$ is the response at time t of the time-varying channel if an impulse is sent at time $t - \tau$, and $z(t)$ is the additive Gaussian noise. The channel impulse response (CIR) depends on the combination of all three propagation effects and in addition contains the delay induced by the reflections.

To collect discrete-time sufficient statistics[1] of the information signal $x(t)$ we need to sample (1) faster than the Nyquist rate[2]. Therefore we focus on the following discrete-time model:

$$y(k) = y_c(kT_s) = \sum_{l=0}^{\nu} h(k;l)x(k-l) + z(k)\ , \tag{2}$$

where $y(k)$, $x(k)$, and $z(k)$ are the output, input, and noise samples at sampling instant k, respectively, and $h(k;l)$ represents the sampled time-varying channel impulse response of finite length ν. Modeling the channel as having a finite duration can be made arbitrarily accurate by appropriately choosing the channel memory ν.

Though the channel response $\{h(k;l)\}$ depends on all three radio propagation attenuation factors, in the time scales of interest the main variations come from the small-scale fading which is well modeled as a complex Gaussian random process.

Since we are interested in studying multiple-antenna diversity, we need to extend the model given in (2) to the multiple transmit (M_t) and receive (M_r) antenna case. The multi-input multi-output (MIMO) model is given by

$$\mathbf{y}(k) = \sum_{l=0}^{\nu} \mathbf{H}(k;l)\mathbf{x}(k-l) + \mathbf{z}(k)\ , \tag{3}$$

[1] The term *sufficient statistics* refers to a function (perhaps many-to-one) which does not cause loss of information about the random quantity of interest.

[2] To be precise, we need to sample (1) at a rate larger than $2(W_I + W_s)$, where W_I is the input bandwidth and W_s is the bandwidth of the channel time variation (Kailath, 1961).

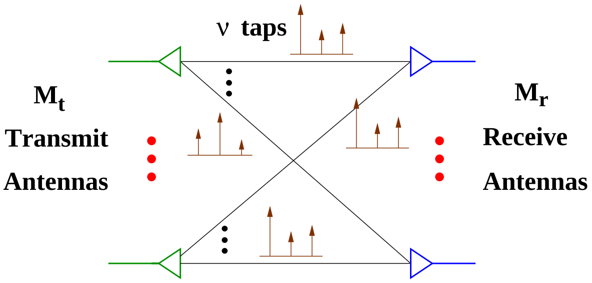

Figure 2: MIMO Channel Model

where the $M_r \times M_t$ complex[3] matrix $\mathbf{H}(k;l)$ represents the l^{th} tap of the channel matrix response with $\mathbf{x} \in \mathbb{C}^{M_t}$ as the input and $\mathbf{y} \in \mathbb{C}^{M_r}$ as the output (see Figure 2). The variations of the channel response between antennas arises due to variations in arrival directions of the reflected radio waves (Raleigh et al., 1994). The input vector may have independent entries to achieve high throughput (*e.g.,* through spatial multiplexing) or correlated entries through coding or filtering to achieve high reliability (better distance properties, higher diversity, spectral shaping, or desirable spatial profile; see Section 3). Throughout this chapter, the input is assumed to be zero mean and to satisfy an average power constraint, *i.e.*, $\mathbb{E}[||\mathbf{x}(k)||^2] \leq P$. The vector $\mathbf{z} \in \mathbb{C}^{M_r}$ models the effects of noise and is assumed to be independent of the input and is modeled as a complex additive circularly symmetric Gaussian vector with $\mathbf{z} \sim \mathbb{C}\mathcal{N}(0, \mathbf{R}_z)$, *i.e.,* a complex Gaussian vector with mean $\mathbf{0}$ and covariance $\mathbf{R}_z$. In many cases we assume white noise, *i.e.,* $\mathbf{R}_z = \sigma^2 \mathbf{I}$.

Finally the basic point-to-point model given in (3) can be modified for an important special case. Many of the insights can be gained for the *flat fading* channel where we have $\nu = 0$ in (3). Unless otherwise mentioned, we will use this special case for illustration throughout this chapter. Also we examine the case where we transmit a block or frame of information. Here we encounter another important modeling assumption. If the transmission block is small enough so that the channel time-variation within a transmission block can be neglected, we have a *block time-invariant* model. Such models are quite realistic for transmission blocks of lengths less than a millisecond and typical channel variation bandwidths. However, this does *not* imply that the channel remains constant during the entire transmission. Transmission blocks sent at various periods of time can experience different (independent) channel instantiations (see Figure 3). This can be utilized by coding across these different channel instantiations as will be seen in Section 3. Therefore, if the transmission block is of length T, for the *flat fading* case, the specialization of (3) yields,

$$\mathbf{Y}^{(b)} = \mathbf{H}^{(b)}\mathbf{X}^{(b)} + \mathbf{Z}^{(b)}, \tag{4}$$

[3] In passband communication, a complex signal arises due to in-phase and quadrature phase modulation of the carrier signal see (Proakis, 1995).

where $\mathbf{Y}^{(b)} = [\mathbf{y}^{(b)}(0), \ldots, \mathbf{y}^{(b)}(T-1)] \in \mathbb{C}^{M_r \times T}$ is the received sequence, $\mathbf{H}^{(b)} \in \mathbb{C}^{M_r \times M_t}$ is the block time-invariant channel fading matrix for transmission block b, $\mathbf{X}^{(b)} = [\mathbf{x}^{(b)}(0), \ldots, \mathbf{x}^{(b)}(T-1)] \in \mathbb{C}^{M_t \times T}$ is the "space-time" information transmission sequence and $\mathbf{Z}^{(b)} = [\mathbf{z}^{(b)}(0), \ldots, \mathbf{z}^{(b)}(T-1)] \in \mathbb{C}^{M_r \times T}$.

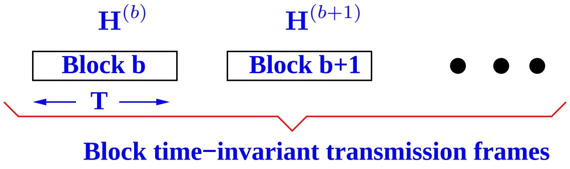

Figure 3: Block time-invariant model.

2.2 Network model

The wireless medium is inherently shared, and this directly motivates a study of multi-user communication techniques. Moreover, since we are also interested in multi-user diversity, we need to extend our model from the point-to-point scenario (2) to the network case. The general communication network (illustrated in Figure 4) consists of n nodes trying to communicate with each other. In the scalar flat-fading wireless channel, the received symbol $Y_i(t)$ at the i^{th} node is given by

$$Y_i(t) = \sum_{\substack{j=1 \\ j \neq i}}^{n} h_{i,j} X_j(t) + Z_i(t), \tag{5}$$

where $h_{i,j}$ is determined by the channel attenuation between nodes i and j. Given this general model, one way of abstracting the multi-user communication problem is through embedding it in an underlying *communication graph* $\mathcal{G}_C$ where the n nodes are vertices of the graph and edges of the graph represent a channel connecting the two nodes along with the interference from other nodes. The graph could be directed with constraints and channel transition probability depending on the directed graph. A general multi-user network is therefore a fully connected graph with the received symbol at each node described as a conditional distribution dependent on the messages transmitted by all other nodes. Such a graph is illustrated in Figure 5. We examine different communication topologies in Section 4 and study the role of diversity in networks.

3 Multiple-antenna diversity

The first form of diversity that we examine in some detail is that of multiple-antenna diversity. A major development over the past decade has been the emergence of space-time (multiple-antenna) techniques that enable high-rate, reliable communication over fading

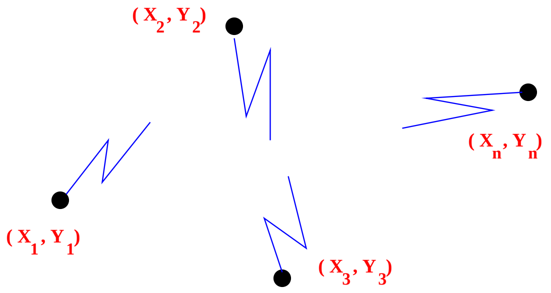

Figure 4: General multiuser wireless communication network.

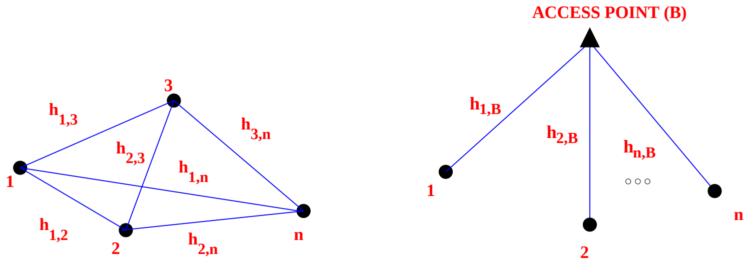

Figure 5: Graph Representation of Communication Topologies. On the left is a general topology and on the right is a hierarchical topology.

wireless channels. In this section we highlight some of the theoretical underpinnings of this topic. More details about practical code constructions can be found in (Tarokh et al., 1998; Diggavi et al., 2004b) and references therein.

Reliable information transmission over fading channels has a long and rich history, see (Ozarow et al., 1994) and references therein. The importance of multiple-antenna diversity was recognized early; see for example (Brennan, 1959). However, most of the focus until the mid 1990s was on receive diversity, where multiple "looks" of the transmitted signal were obtained using many receive antennas (see (3) using $M_t = 1$). The use of multiple transmit antennas was restricted to sending the same signal over each antenna, which is a form of repetition coding (Wornell and Trott, 1997).

During the mid 1990s several researchers started to investigate the idea of coding across transmit antennas to obtain higher rate and reliability (Foschini, 1996; Tarokh et al., 1998; Telatar, 1995). One focus was on maximizing the reliable transmission rate, *i.e.,* channel capacity without requiring a bound on the rate at which error probability diminishes (Foschini, 1996; Telatar, 1995). However, another point of view was explored where non-degenerate correlation was introduced between the information streams across the multiple transmit antennas in order to guarantee a certain bound on the rate at which the error probability diminishes (Tarokh et al., 1998). These approaches have led to the broad area of space-time codes, which is still an active research topic.

In Section 3.1 we first start with an understanding of reliable transmission rate over multiple-antenna channels. In particular we examine the rate advantages of multiple transmit and receive antennas. Then in Section 3.2 we introduce the notion of diversity order, which captures transmission reliability (error probability) in the high signal-to-noise ratio (SNR) regime. This allows us to develop criteria for space-time codes which guarantee a given reliability. Section 3.3 examines the fundamental trade-off between maximizing rate and reliability.

3.1 Capacity of Multiple-Antenna Channels

The concept of capacity was introduced in (Shannon, 1948), where it was shown that even in noisy channels, one can transmit information at positive rates with the error probability going to zero asymptotically in the coding block size. The seminal result was that for a noisy channel whose input at time k is $\{X_k\}$ and output is $\{Y_k\}$, there exists a number C such that

$$C = \lim_{T\to\infty}\left[\frac{1}{T}\sup_{p(x^T)} I(X^T;Y^T)\right], \tag{6}$$

where the mutual information is given by $I(X^T;Y^T) = \mathbb{E}_{X^T,Y^T}[\log(\frac{p(x^T,y^T)}{p(x^T)p(y^T)})]$, $p(\cdot)$ is the probability density function, and for convenience we have denoted $X^T = \{X_1,\dots,X_T\}$ and similarly for Y^T (Cover and Thomas, 1991). In (Shannon, 1948) it was shown that asymptotically in block length T, there exist codes which can transmit information at all rates below C with arbitrarily small probability of error over the noisy channel. Perhaps the most famous illustration of this idea was the formula derived in (Shannon, 1948) for the capacity C of the additive white Gaussian noise channel with noise variance σ^2 and input power constraint P:

$$C = \frac{1}{2}\log(1+\frac{P}{\sigma^2}). \tag{7}$$

In this section we will focus mostly on the *flat-fading* channels where, in (3) we have $\nu = 0$. The generalizations of these ideas for *frequency selective* channels (*i.e.,* $\nu > 0$) can be easily carried out; see (Biglieri et al., 1998; Diggavi et al., 2004b), and references therein. We begin with the case where we are allowed to develop transmit schemes which code across multiple (B) realizations of the channel matrix $\{\mathbf{H}^{(b)}\}_{b=1}^{B}$ (see Figure 3). In such a case, we can again define a notion of reliable transmission rate, where the error probability decays to zero when we develop codes across an asymptotically large number of transmit blocks (*i.e.,* $B \to \infty$). We examine this for a coherent receiver, where the receiver uses perfect channel state information $\{\mathbf{H}^{(b)}\}$ for each transmission block. But the transmitter is assumed not to have access to the channel realizations. To gain some intuition, consider first the case when each transmission block is large *i.e.,* $T \to \infty$. If we have one transmit antenna ($M_t = 1$), the channel vector response is a vector $\mathbf{h}^{(b)} \in \mathbb{C}^{M_r}$ (see (4) in Section 2).

Therefore the reliable transmission rate for any particular block can be generalized[4] $\{\mathbf{h}(k)\}$ from (7) as $\log(1+\frac{||\mathbf{h}^{(b)}||^2 P}{\sigma^2})$. Note that when we are dealing with complex channels (as is usual in communication with in-phase and quadrature-phase transmissions), the factor of $\frac{1}{2}$ disappears (Neeser and Massey, 1993) when we adapt the expression from (7). Now, if one codes across a large number of transmission blocks ($B \to \infty$), for a stationary and ergodic sequence of $\{\mathbf{h}^{(b)}\}$ we would expect to get a reliable transmission rate that is the average of this quantity. This intuition has been made precise in (Ozarow et al., 1994) and references therein for flat-fading channels ($\nu = 0$), even when we do not have $T \to \infty$, but we have $B \to \infty$. Therefore when we have only receive diversity, *i.e.,* $M_t = 1$, for a given M_r, it is shown (Ozarow et al., 1994) that the capacity is given by

$$C = \mathbb{E}\left[\log(1+\frac{||\mathbf{h}||^2 P}{\sigma^2})\right], \tag{8}$$

where the expectation is taken over the fading channel $\{\mathbf{h}^{(b)}\}$ and the channel sequence is assumed to be stationary and ergodic. This is called the *ergodic channel capacity* (Ozarow et al., 1994). This is the rate at which information can be transmitted if there is *no feedback* of the channel state ($\{\mathbf{h}^{(b)}\}$) from the receiver to the transmitter. If there is feedback available about the channel state, one can do slightly better through optimizing the allocation of transmitted power by "waterfilling" over the fading channel states. The problem of studying the capacity of channels with causal transmitter side-information was introduced in (Shannon, 1958a), where a coding theorem for this problem was proved. Using ideas from there and *perfect* transmitter channel-state information, capacity expressions that generalize (8) have been developed (Goldsmith and Varaiya, 1997). However, for fast time-varying channels the instantaneous feedback could be difficult, resulting in an outdated estimate of the channel being sent back (Viswanathan, 1999; Caire and Shamai, 1999). However, the basic question of impact of feedback on capacity of time-varying channels is still not completely understood, and for developing the basic ideas in this chapter, we will deal with the case where the transmitter does not have access to the channel state information. We refer the interested reader to (Biglieri et al., 1998) for a more complete overview of such topics.

Now let us focus our attention on the multiple transmit and receive antenna channel where again as before we consider the coherent case, *i.e.,* the receiver has perfect channel state information (CSI) $\mathbf{H}^{(b)}$. In the flat-fading case where $\nu = 0$, when we code across B transmission blocks, the mutual information for this case is

$$R^{(B)} = \frac{1}{BT} I(\{\mathbf{X}^{(b)}\}_{b=1}^{B}; \{\mathbf{Y}^{(b)}\}_{b=1}^{B}, \{\mathbf{H}^{(b)}\}_{b=1}^{B}),$$

[4] This can be seen by noticing that for $M_t = 1$, a sufficient statistics is an equivalent scalar channel, $\tilde{y}^{(b)} = \mathbf{h}^{(b)*}\mathbf{y}^{(b)} = ||\mathbf{h}^{(b)}||^2\mathbf{x}^{(b)} + \mathbf{h}^{(b)*}\mathbf{z}^{(b)}$. In this chapter, $|h|^2 = \bar{h}h$, where $\bar{h}$ denotes complex conjugation, and for a vector $\mathbf{h}$ we denote its 2-norm by $||\mathbf{h}||^2 = \mathbf{h}^*\mathbf{h}$, where $\mathbf{h}^*$ denotes the Hermitian transpose and $\mathbf{h}^t$ denotes ordinary transpose.

since we assume that the receiver has access to CSI. Using the chain rule of mutual information (Cover and Thomas, 1991), this can be written as

$$R^{(B)} = \frac{1}{BT}\left[I(\{\mathbf{X}^{(b)}\}_{b=1}^{B}; \{\mathbf{H}^{(b)}\}_{b=1}^{B}) + I(\{\mathbf{X}^{(b)}\}_{b=1}^{B}; \{\mathbf{Y}^{(b)}\}_{b=1}^{B} | \{\mathbf{H}^{(b)}\}_{b=1}^{B})\right]. \tag{9}$$

Using the the assumption that the input $\{\mathbf{x}(k)\}$ is independent of the fading process (as the transmitter does not have CSI), (9) is equal to

$$R^{(B)} = \frac{1}{BT}\mathbb{E}_{\mathcal{H}}\left[I\left(\{\mathbf{X}^{(b)}\}_{b=1}^{B}; \{\mathbf{Y}^{(b)}\}_{b=1}^{B} | \mathcal{H}^{(B)} = \{\mathbf{H}^{(b)}\}_{b=1}^{B}\right)\right]. \tag{10}$$

Now, if we use the memoryless property of the vector Gaussian channel obtained by conditioning on $\mathbf{H}^{(b)}$ and also due to the assumption[5] that $\{\mathbf{H}^{(b)}\}$ is i.i.d. over b, for when $B \to \infty$ we get that

$$\lim_{B\to\infty} \frac{1}{BT} I(\{\mathbf{X}^{(b)}\}_{b=1}^{B}; \{\mathbf{Y}^{(b)}\}_{b=1}^{B}, \{\mathbf{H}^{(b)}\}_{b=1}^{B}) = \mathbb{E}_{\mathcal{H}}[\log(\frac{|\mathbf{R}_z + \mathbf{H}\mathbf{R}_x\mathbf{H}^*|}{|\mathbf{R}_z|})], \tag{11}$$

where[6] the expectation is taken over the random channel realizations $\{\mathbf{H}^{(b)}\}$. An *operational* meaning to this expression can be given by showing that there exist codes which can transmit information at this rate with arbitrarily small probability of error (Telatar, 1995).

In general, it is difficult to evaluate (11) except for some special cases. If the random matrix $\mathbf{H}^{(b)}$ consists of zero-mean i.i.d. Gaussian elements (Telatar, 1995) showed that

$$C = \mathbb{E}_{\mathcal{H}}[\log(|\mathbf{I} + \frac{P}{M_t\sigma^2}\mathbf{H}\mathbf{H}^*|)] \tag{12}$$

is the capacity of the fading matrix channel[7]. Therefore in this case, to achieve capacity the optimal codebook is generated from an i.i.d. Gaussian input $\{\mathbf{x}^{(b)}\}$ with $\mathbf{R}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^*] = \frac{P}{M_t}\mathbf{I}$.

The expression in (12) shows that the capacity is dependent on the eigenvalue distribution of the random matrix $\mathbf{H}$ with Gaussian i.i.d. components. This important connection between capacity of multiple-antenna channels and the mathematics related to eigenvalues of random matrices (Edelman, 1989) was noticed in (Telatar, 1995) where it was shown that the capacity could be numerically computed using Laguerre polynomials (Telatar, 1995; Muirhead, 1982; Edelman, 1989).

Theorem 3.1 *(Telatar, 1995) The capacity C of the channel with M_t transmitters and M_r receivers and average power constraint P is given by*

$$C = \int_0^\infty \log(1 + \frac{P\lambda}{\sigma^2 M_t}) \sum_{k=0}^{T_{min}-1} \lambda^{T_{max}-T_{min}} [L_k^{T_{max}-T_{min}}(\lambda)]^2 \frac{k!}{k + T_{max} - T_{min}} e^{-\lambda} d\lambda \,,$$

[5] The assumption that $\{\mathbf{H}^{(b)}\}$ is i.i.d. is not crucial. This result is (asymptotically) correct even when the sequence $\{\mathbf{H}^{(b)}\}$ is a mean ergodic sequence (Ozarow et al., 1994). We use the notation $\mathbf{H}$ to denote the channel matrix $\mathbf{H}^{(b)}$ for a generic block b.

[6] For a matrix $\mathbf{A}$, we denote its determinant as $det(\mathbf{A})$ and $|\mathbf{A}|$, interchangeably.

[7] In (Foschini, 1996), a similar expression was derived without illustrating the converse to establish that the expression was indeed the capacity.

where $T_{max} = \max(M_t, M_r)$*,* $T_{min} = \min(M_t, M_r)$*, and* $L_k^m(\cdot)$ *is the generalized Laguerre polynomial of order* k *with parameter* m *(Gradshteyn and Ryzhik, 1994).*

■

In (Foschini, 1996), it was observed that when $M_t = M_r = M$ the capacity C grows linearly in M as $M \to \infty$.

Theorem 3.2 *(Foschini, 1996) For* $M_t = M_r = M$ *the capacity* C *given by (12) grows asymptotically linearly in* M, i.e.,

$$\lim_{M \to \infty} \frac{C}{M} = c^*(\mathrm{SNR}) \, , \tag{13}$$

where $c^*(\mathrm{SNR})$ *is a constant depending on SNR.*

■

This quantifies the advantage of using multiple transmit and receive antennas and shows the promise of such architectures for high rate reliable wireless communication.

To achieve the capacity given in (12), we require joint optimal (maximum-likelihood) decoding of all the receiver elements which could have large computational complexity. The channel model in (3) resembles a multi-user channel (Verdu, 1998) with user cooperation. A natural question to ask is whether the simpler decoding schemes proposed in multi-user detection would yield good performance on this channel. A motivation for this is seen by observing that for i.i.d. elements of the channel response matrix (flat-fading) the normalized cross-correlation matrix decouples (*i.e.,* $\lim_{M_r \to \infty} \frac{1}{M_r}\mathbf{H}^*\mathbf{H} \to \mathbf{I}_{M_t}$). Therefore, since nature provides some decoupling, a simple "matched filter" receiver (Verdu, 1998) might perform quite well. In this context a matched filter for the flat-fading channel in (3) is given by $\tilde{\mathbf{y}}(k) = \mathbf{H}^*(k)\mathbf{y}(k)$. Therefore, component-wise this means that

$$\tilde{\mathbf{y}}_i(k) = ||\mathbf{h}_i(k)||^2\mathbf{x}_i(k) + \sum_{\substack{j=1 \\ j \neq i}}^{M_t} \mathbf{h}_i^*(k)\mathbf{h}_j(k)\mathbf{x}_j(k) + \tilde{\mathbf{z}}_i(k), \quad i = 1, \dots, M_t. \tag{14}$$

By ignoring the cross-coupling between the channels we decode $\hat{\mathbf{x}}_i$ by including the "interference" from $\{\mathbf{x}_j\}_{j \neq i}$ as part of the noise. However, a tension arises between the decoupling of the channels and the added "interference" $\sum_{\substack{j=1 \\ j \neq i}}^{M_t} \mathbf{h}_i^*(k)\mathbf{h}_j(k)\mathbf{x}_j(k)$ from the other antennas, which clearly grows with the number of antennas. It is shown in (Diggavi, 2001) that the two effects exactly cancel each other.

Proposition 3.1 *If* $\mathbf{H}(k) = [\mathbf{h}_1(k), \dots, \mathbf{h}_{M_t}(k)] \in \mathbb{C}^{M_r \times M_t}$ *and* $\mathbf{h}_l(k) \sim \mathbb{C}\mathcal{N}(0, \mathbf{I}_{M_r})$*,* $l = 1, \dots, M_t$*, are i.i.d., then*

$$\lim_{\substack{M_r \to \infty \\ M_t = \lfloor \alpha M_r \rfloor}} \sum_{\substack{j=1 \\ j \neq i}}^{M_t} |\frac{\mathbf{h}_i^*(k)\mathbf{h}_j(k)}{M_r}|^2 = \alpha \ \text{ almost surely.}$$

Therefore, using this result it can be shown that the simple detector still retains the linear growth rate of the optimal decoding scheme (Diggavi, 2001). However, in the rate R_I achievable for this simple decoding scheme, we do pay a price in terms of rate growth with SNR.

Theorem 3.3 *If* $\mathbf{H}_{i,j} \sim \mathbb{C}\mathcal{N}(0,1)$*, with i.i.d. elements then*

$$\lim_{\substack{M_t \to \infty \\ M_t = \lfloor \alpha M_r \rfloor}} \frac{1}{M_t} I(\mathbf{Y}, \mathbf{H}; \mathbf{X}) \geq \lim_{\substack{M_t \to \infty \\ M_t = \lfloor \alpha M_r \rfloor}} R_I / M_t = \log(1 + \frac{\frac{P}{\sigma^2 \alpha}}{1 + \frac{P}{\sigma^2}}).$$

■

Multi-user detection (Verdu, 1998) is a good analogy to understand receiver structures in MIMO systems. The main difference is that unlike multiple access channels, the space-time encoder allows for cooperation between "users". Therefore, the encoder could introduce correlations that can simplify the job of the decoder. Such encoding structures using space-time block codes are discussed further in (Diggavi et al., 2004b) and references therein. An example of using the multi-user detection approach is the result in Theorem 3.3 where a simple matched filter receiver is applied. Using more sophisticated linear detectors, such as the decorrelating receiver and the MMSE receiver (Verdu, 1998), one can improve performance while still maintaining the linear growth rate. The decision feedback structures also known as successive interference cancellation, or onion peeling (Cover, 1975; Wyner, 1974; Patel and Holtzman, 1994) can be shown to be optimal, *i.e.,* to achieve the capacity, when an MMSE multi-user interference suppression is employed and the layers are peeled off (Cioffi et al., 1995; Varanasi and Guess, 1997). However, decision feedback structures inherently suffer from error propagation (which is not taken into account in the theoretical results) and could therefore have poor performance in practice, especially at low SNR. Thus, examining non-decision feedback structures is important in practice.

All of the above results illustrate that significant gains in information rate (capacity) are possible using multiple transmit and receive antennas. The intuition for the gains with multiple transmit and receive antennas is that there are a larger number of communication modes over which the information can be transmitted. This is formalized by the observation (Zheng and Tse, 2002; Diggavi, 2001) that the capacity as a function of SNR, $C(SNR)$, grows linearly in $\min(M_r, M_t)$, even for a finite number of antennas, asymptotically in the SNR.

Theorem 3.4

$$\lim_{SNR \to \infty} \frac{C(SNR)}{\log(SNR)} = \min(M_r, M_t). \tag{15}$$

■

In the results above, the fundamental assumption was that the receiver had access to *perfect* channel state information, obtained through training or other methods. When the channel is slowly varying, the estimation error could be small since we can track the channel variations and one can quantify the effect of such estimation errors. As a rule of thumb, it is shown in (Lapidoth and Shamai, 2002) that if the estimation error is small compared to $\frac{1}{SNR}$, these results would hold. Another line of work assumes that the receiver does not have *any* channel state information. The question of the information rate that can be reliably transmitted over the multiple-antenna channel without channel state information was introduced in (Hochwald and Marzetta, 1999) and has also been examined in (Zheng and Tse, 2002). The main result from this line of work shows that the capacity growth is again (almost) linear in the number of transmit and receive antennas, as stated formally next.

Theorem 3.5 *If the channel is block fading with block length T and we denote $K = \min(M_t, M_r)$, then for $T > K + M_t$, as $SNR \to \infty$, the capacity is*[8]

$$C(SNR) = K\left(1 - \frac{K}{T}\right)\log(SNR) + c + o(1) ,$$

where c is a constant depending only on M_r, M_t, T.

■

In fact (Zheng and Tse, 2002) go on to show that the rate achievable by using a training-based technique is only a constant factor away from the optimal, *i.e.*, it attains the same capacity-SNR slope as in Theorem 3.5. Further results on this topic can be found in (Hassibi and Marzetta, 2002). Therefore, even in the non-coherent block-fading case, there are significant advantages in using multiple antennas.

Most of the discussion above was for the flat-fading case where $\nu = 0$ in (3). However, these ideas can be easily extended for the block time-invariant frequency selective channels where again the advantages of multiple-antenna channels can be established (Diggavi, 2001). However, when the channels are not block time-invariant, the characterization of the capacity of frequency selective channels is an open question.

Outage: In all of the above results, the error probability goes to zero asymptotically in the number of coding blocks *i.e.,* $B \to \infty$. Therefore, coding is assumed to take place *across* fading blocks, and hence it inherently uses the *ergodicity* of the channel variations. This approach would clearly entail large delays, and therefore (Ozarow et al., 1994) introduced a notion of outage, where the coding is done (in the extreme case) just across one fading block, *i.e.,* $B = 1$. Here the transmitter sees only one block of channel coefficients, and therefore the channel is *non-ergodic*, and the strict Shannon-sense capacity is zero. However, one can

[8] Here the notation $o(1)$ indicates a term that goes to zero when $SNR \to \infty$.

define an outage probability that is the probability with which a certain rate R is possible. Therefore, for a block time-invariant channel with a single channel realization $\mathbf{H}^{(b)} = \mathbf{H}$ the outage probability can be defined as follows.

Definition 3.1 *The outage probability for a transmission rate of R and a given transmission strategy $p(\mathbf{X})$ is defined as*

$$P_{outage}(R, p(\mathbf{X})) = \mathbb{P}\left\{\mathbf{H} : I(\mathbf{X}; \mathbf{Y}|\mathbf{H}^{(b)} = \mathbf{H}) < R\right\}. \tag{16}$$

Therefore, if one uses a white Gaussian codebook ($\mathbf{R}_x = \frac{P}{M_t}\mathbf{I}$) then (abusing notation by dropping the dependence on $p(\mathbf{X})$) we can write the outage probability at rate R as

$$P_{outage}(R) = \mathbb{P}\left\{\log(|\mathbf{I} + \frac{P}{M_t\sigma^2}\mathbf{H}\mathbf{H}^*|) < R\right\}. \tag{17}$$

It has been shown in (Zheng and Tse, 2003) that at high SNR the outage probability is the same as the frame-error probability in terms of the SNR exponent. Therefore, to evaluate the optimality of practical coding techniques, one can compare, for a given rate, how far the performance of the technique is from that predicted through an outage analysis. Moreover, the frame-error rates and outage capacity comparisons in (Tarokh et al., 1998) can also be formally justified through this argument.

3.2 Diversity Order

In Section 3.1 the focus was on achievable transmission rate. A more practical performance criterion is probability of error. This is particularly important when we are coding over a small number of blocks (low delay) where the Shannon capacity is zero (Ozarow et al., 1994) and we are in the outage regime as was seen above. By characterizing the error probability, we can also formulate design criteria for space-time codes.

Since we are allowed to transmit a coded sequence, we are interested in the probability that an erroneous codeword[9] $\mathbf{e}$ is mistaken for the transmitted codeword $\mathbf{x}$. This is called the *pairwise error probability* (PEP) and is used to bound the error probability. This analysis relies on the condition that the receiver has perfect channel state information. However, a similar analysis can be done when the receiver does not know the channel state information, but has statistical knowledge of the channel (Hochwald and Marzetta, 2000).

For simplicity, we shall again focus on a flat-fading channel (where $\nu = 0$) and when the channel matrix contains i.i.d. zero-mean Gaussian elements, *i.e.*, $\mathbf{H}_{i,j} \sim \mathbb{C}\mathcal{N}(0, 1)$. Many of these results can be easily generalized for $\nu > 0$ as well as for correlated fading and other fading distributions. Consider a codeword sequence $\mathbf{X} = [\mathbf{x}^t(0), \ldots, \mathbf{x}^t(T-1)]^t$, where $\mathbf{x}(k) = [\mathbf{x}_1(k), \ldots, \mathbf{x}_{M_t}(k)]^t$ is defined in (4). In the case when the receiver has perfect

[9] For an information rate of R bits per transmission and a block length of T, we define the codebook as the set of 2^{TR} codeword sequences of length T.

channel state information, we can bound the PEP between two codeword sequences $\mathbf{x}$ and $\mathbf{e}$ (denoted by $P(\mathbf{x} \rightarrow \mathbf{e})$) as follows (Tarokh et al., 1998; Guey et al., 1999).

$$P(\mathbf{x} \rightarrow \mathbf{e}) \leq \left[\frac{1}{\prod_{n=1}^{M_t} (1 + \frac{E_s}{4N_0} \lambda_n)} \right]^{M_r} . \tag{18}$$

$E_s = \frac{P}{M_t}$ is the power per transmitted symbol, λ_n are the eigenvalues of the matrix $\mathbf{A}(\mathbf{x}, \mathbf{e}) = \mathbf{B}^*(\mathbf{x}, \mathbf{e})\mathbf{B}(\mathbf{x}, \mathbf{e})$ and

$$\mathbf{B}(\mathbf{x}, \mathbf{e}) = \begin{pmatrix} \mathbf{x}_1(0) - \mathbf{e}_1(0) & \dots & \mathbf{x}_{M_t}(0) - \mathbf{e}_{M_t}(0) \\ \vdots & \vdots & \vdots \\ \mathbf{x}_1(N-1) - \mathbf{e}_1(N-1) & \dots & \mathbf{x}_{M_t}(N-1) - \mathbf{e}_{M_t}(N-1) \end{pmatrix} . \tag{19}$$

If q denotes the rank of $\mathbf{A}(\mathbf{x}, \mathbf{e})$, (*i.e.*, the number of non-zero eigenvalues) then we can bound (18) as

$$P(\mathbf{x} \rightarrow \mathbf{e}) \leq \left[\prod_{n=1}^{q} \lambda_n \right]^{-M_r} \left(\frac{E_s}{4N_0} \right)^{-qM_r} . \tag{20}$$

We define the notion of diversity order as follows.

Definition 3.2 *A coding scheme which has an average error probability $\bar{P}_e(SNR)$ that behaves as*

$$\lim_{SNR \rightarrow \infty} \frac{\log(\bar{P}_e(SNR))}{\log(SNR)} = -d \tag{21}$$

as a function of SNR is said to have a diversity order of d.

In words, a scheme with diversity order d has an error probability at high SNR behaving as $\bar{P}_e(SNR) \approx SNR^{-d}$ (see Figure 6). One reason to focus on such a behavior for the error probability can be seen from the following intuitive argument for a simple scalar fading channel ($M_t = 1 = M_r$). It is well known that for particular frame b, the error probability for binary transmission, conditioned on the channel realization $h^{(b)}$, is given by $P_e(h^{(b)}) = Q\left(\sqrt{2SNR}\,|h^{(b)}|\right)$ (Proakis, 1995). Hence if $|h^{(b)}|\sqrt{2SNR} \gg 1$, then $P_e(h^{(b)}) \approx 0$, and if $|h^{(b)}|\sqrt{2SNR} \ll 1$, then $P_e(h^{(b)}) \approx \frac{1}{2}$. Therefore a frame is in error with high probability when the channel gain $|h^{(b)}|^2 \ll \frac{1}{SNR}$, *i.e.,* when the channel is in a "deep fade". Therefore the average error probability is well approximated by the probability that $|h^{(b)}|^2 \ll \frac{1}{SNR}$. For high SNR we can show that, for $h \sim \mathbb{C}\mathcal{N}(0, 1)$, $\mathbb{P}\left\{|h|^2 < \frac{1}{SNR}\right\} \approx \frac{1}{SNR}$, and this explains the behavior of the average error probability. Although this is a crude analysis, it brings out the most important difference between the additive white Gaussian noise (AWGN) and the fading channel. The typical way in which an error occurs in a fading channel is due to channel failure, *i.e.,* when the channel gain $|h|$ is very small, less than $\frac{1}{SNR}$. On the other hand, in an AWGN channel errors occur when the noise is large, and since the noise is Gaussian, it has an exponential tail causing this to be very unlikely at high SNR.

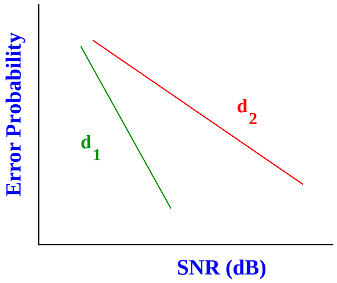

Figure 6: Relationship between error probability and diversity order.

Given the definition 3.2 of diversity order, we see that the diversity order in (20) is at most qM_r. Moreover, in (20) we notice that we also obtain a coding gain of $(\prod_{n=1}^{q} \lambda_n)^{1/q}$.

Note that in order to obtain the average error probability, one can calculate a naive union bound using the pairwise error probability given in (20) but this may not be tight. A more careful upper bound for the error probability can be derived (Zheng and Tse, 2003). However, if we ensure that *every* pair of codewords satisfies the diversity order in (20), then clearly the average error probability satisfies it as well. This is true when the transmission rate is held constant with respect to SNR, *i.e.,* a *fixed rate* code. Therefore, in the case of fixed-rate code design the simple pairwise error probability given in (20) is sufficient to obtain the correct diversity order.

In order to design practical codes that achieve a performance target, we need to glean insights from the analysis to state design criteria. For example, in the flat-fading case of (20) we can state the following rank and determinant design criteria (Tarokh et al., 1998).

Design criteria for space-time codes

To design practical space-time codes, design guidelines to achieve particular performance was given in (Tarokh et al., 1998). For the flat-fading channel, the following rank and determinant criteria were developed (Tarokh et al., 1998).

- *Rank criterion:* In order to achieve maximum diversity M_tM_r, the matrix $\mathbf{B}(\mathbf{x}, \mathbf{e})$ from (19) has to be full rank for any codewords $\mathbf{x}, \mathbf{e}$. If the minimum rank of $\mathbf{B}(\mathbf{x}, \mathbf{e})$ over all pairs of distinct codewords is q, then a diversity order of qM_r is achieved.

- *Determinant criterion:* For a given diversity order target of q, maximize $(\prod_{n=1}^{q} \lambda_n)^{1/q}$ over all pairs of distinct codewords.

Over the past few years, there have been significant developments in designing codes which can guarantee a given reliability (error probability). An exhaustive listing of all these developments is beyond the scope of this chapter, but we give a glimpse of the recent developments. The interested reader is referred to (Diggavi et al., 2004b) and references therein.

Pioneering work on trellis codes for Gaussian channels was done in (Ungerboeck, 1982). In (Tarokh et al., 1998), the first space-time trellis code constructions were presented. In this seminal work, trellis codes were carefully designed to meet the design criteria for minimizing error probability. In parallel a very simple coding idea for $M_t = 2$ was developed in (Alamouti, 1998). This code achieved maximal diversity order of $2M_r$ and had a very simple decoder associated with it. The elegance and simplicity of the Alamouti code has made it a candidate for next generation of wireless systems which are slated to utilize space-time codes. The basic idea of the Alamouti code was extended to orthogonal designs in (Tarokh et al., 1999). The publication of (Tarokh et al., 1998; Alamouti, 1998) created a significant community of researchers working on space-time code constructions. Over the past few years, there has been significant progress in the construction of space-time codes for coherent channels. The design of codes that are linear in the complex field was proposed in (Hassibi and Hochwald, 2002) and efficient decoders for such codes were given in (Damen et al., 2000). Codes based on algebraic rotations and number-theoretic tools are developed in (El-Gamal and Damen, 2003; Sethuraman et al., 2003). A common assumption in all these designs was that the receiver had perfect knowledge of the channel. Techniques based on channel estimation and the evaluation of the degradation in performance for space-time trellis codes was examined in (Naguib et al., 1998). In another line of work, non-coherent space-time codes were proposed in (Hochwald and Marzetta, 2000). This also led to the design and analysis of differential space-time codes for flat-fading channels (Hochwald and Sweldens, 2000; Hughes, 2000; Tarokh and Jafarkhani, 2000). This was also examined for frequency selective channels in (Diggavi et al., 2002a).

As can be seen, the topic of space-time codes is still evolving and we just have a snapshot of the recent developments.

3.3 Rate-Diversity Trade-off

A natural question that arises is how many codewords can we have which allow us to attain a certain diversity order. For a flat Rayleigh fading channel, this has been examined in (Tarokh et al., 1998; Lu and Kumar, 2003) and the following result was obtained[10].

Theorem 3.6 *If we use a transmit signal with constellation of size $|\mathcal{S}|$ and the diversity order of the system is qM_r, then the rate R that can be achieved is bounded as*

$$R \leq (M_t - q + 1) \log_2 |\mathcal{S}| \tag{22}$$

in bits per transmission.

[10] A constellation size refers to the alphabet size of each transmitted symbol. For example, a QPSK modulated transmission has constellation size of 4.

■

One consequence of this result is that for maximum (M_tM_r) diversity order we can transmit at most $\log_2 |\mathcal{S}|$ bits/sec/Hz. Note that the trade-off in Theorem 3.6 is established with a constraint on the alphabet size of the transmit signal, which may not be fundamental from an information-theoretic point of view. An alternate viewpoint of the rate-diversity trade-off has been explored in (Zheng and Tse, 2003) from a Shannon-theoretic point of view. In that work the authors are interested in the multiplexing rate of a transmission scheme.

Definition 3.3 *A coding scheme which has a transmission rate of $R(SNR)$ as a function of SNR is said to have a multiplexing rate r if*

$$\lim_{SNR\to\infty} \frac{R(SNR)}{\log(SNR)} = r. \tag{23}$$

Therefore, the system has a rate of $r\log(SNR)$ at high SNR. One way to contrast this with the statement in Theorem 3.6, is to note that the constellation size is also allowed to become larger with SNR. The naive union bound of the pairwise error probability (18) has to be used with care if the constellation size is also increasing with SNR. There is a trade-off between the achievable diversity and the multiplexing gain, and $d^*(r)$ is defined as the supremum of the diversity gain achievable by *any* scheme with multiplexing gain r. The main result in (Zheng and Tse, 2003) states the following.

Theorem 3.7 *For $T > M_t + M_r - 1$, and $K = \min(M_t, M_r)$, the optimal trade-off curve $d^*(r)$ is given by the piecewise linear function connecting points in $(k, d^*(k)), k = 0, \dots, K$ where*

$$d^*(k) = (M_r - k)(M_t - k). \tag{24}$$

■

If $r = k$ is an integer, the result can be notionally interpreted as using $M_r - k$ receive antennas and $M_t - k$ transmit antennas to provide diversity while using k antennas to provide the multiplexing gain. However, this interpretation is not physical but really an intuitive explanation of the result in Theorem 3.7. Clearly this result means that one can get large rates which grow with SNR if we reduce the diversity order from the maximum achievable. This diversity–multiplexing trade-off implies that a high multiplexing gain comes at the price of decreased diversity gain and is a manifestation of a corresponding trade-off between error probability and rate. This trade-off is depicted in Figure 7. Therefore, as illustrated in Theorems 3.6 and 3.7, the trade-off between diversity and rate is an important consideration both in terms of coding techniques (Theorem 3.6) and in terms of Shannon theory (Theorem 3.7).

The rank and determinant design criteria given in Section 3.2 are suitable for transmission when we have a fixed input alphabet. Since the rate-diversity trade-off can also be

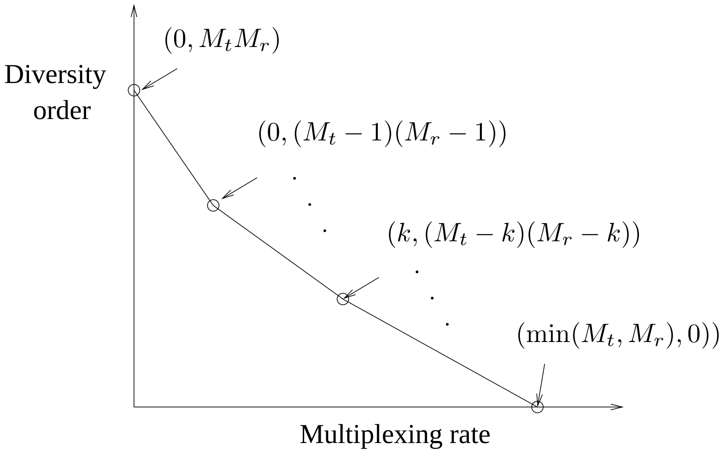

Figure 7: Rate-Diversity Trade-off Curve

explored in the context of the multiplexing rate, a natural question to ask is whether the same the code-design criteria apply in this context. For the diversity-multiplexing guarantees, it is not clear that the rank and determinant criterion is the correct one to use. In fact, in (El-Gamal et al., 2004b), it is shown that for designing codes with the multiplexing rate in mind, the determinant criterion is not necessary for *specific* fading distributions. However, it has been shown that the determinant criterion again arises as a sufficient condition for designing codes for the diversity-multiplexing rate trade-off for specific constructions (see (Yao and Wornell, 2003) and references therein). For these constructions, it is shown that the determinant of the code-word difference matrix plays a crucial role in the diversity-multiplexing optimality of the codes.

A different question was proposed in (Diggavi et al., 2004a, 2003), where it was asked whether there exists a strategy that combines high-rate communications with high reliability (diversity). Clearly the overall code will still be governed by the rate-diversity trade-off, but the idea is to ensure the reliability (diversity) of at least part of the total information. This allows a form of communication where the high-rate code opportunistically takes advantage of good channel realizations whereas the embedded high-diversity code ensures that at least part of the information is received reliably. In this case, the interest was not in a single pair of multiplexing rate and diversity order (r, d), but in a tuple (r_a, d_a, r_b, d_b) where rate r_a and diversity order d_a was ensured for part of the information with rate-diversity pair (r_b, d_b) guaranteed for the other part. A class of space-time codes with such desired characteristics have been constructed in (Diggavi et al., 2003, 2004a).

From an information-theoretic point of view (Diggavi and Tse, 2004) focused on the case when there is one degree of freedom (*i.e.,* $\min(M_t, M_r) = 1$). In that case if we consider $d_a \geq d_b$ without loss of generality, the following result was established in (Diggavi and Tse, 2005).

Theorem 3.8 *When* $\min(M_t, M_r) = 1$*, then the diversity-multiplexing trade-off curve is successively refinable, i.e., for any multiplexing gains* r_a *and* r_b *such that* $r_a + r_b \leq 1$*, the diversity orders* $d_a \geq d_b$*,*

$$d_a = d^*(r_a), \;\; d_b = d^*(r_a + r_b), \tag{25}$$

are achievable, where $d^*(r)$ *is the optimal diversity order given in Theorem 3.7.*

■

Since the overall code has to still be governed by the rate-diversity trade-off given in Theorem 3.7, it is clear that the trivial outer bound to the problem is that $d_a \leq d^*(r_a)$ and $d_b \leq d^*(r_a + r_b)$. Hence Theorem 3.8 shows that the best possible performance can be achieved. This means that for $\min(M_t, M_r) = 1$, we can design ideal *opportunistic* codes. This new direction of enquiry is being currently explored (Diggavi et al., 2005a; Diggavi and Tse, 2006).

4 Multi-user diversity

In Section 3, we explored the importance of using many fading realizations through multiple-antennas for reliable, high-rate, single-user wireless communication. In this section we explore another form of diversity where we can view different users as a form of *multi-user diversity*. This is because each user potentially has independent channel conditions and local interference environment. This implies that in Figure 5, the fading links between users are random and independent of each other. Therefore, this diversity in channel and interference conditions can be exploited by treating the independent links from *different* users as conduits for information transfer.

In order to explore this idea further we first digress to discuss communication topologies. As seen in Section 2 (see Figure 5), we can view the n-user communication network through the underlying graph $\mathcal{G}_C$. One topology which is very commonly seen in practice is obtained by giving special status to one of the nodes as the base station or access point. The other nodes can *only* communicate to the base station. We call such a topology the *hierarchical communication topology* (see Figure 5). An alternate topology that has emerged more recently is when the nodes organize themselves without a centralized base station. Such a topology is called an *ad hoc communication topology*, where the nodes relay information from source to destination, typically through multiple "nearest neighbor" communication hops (see also Figure 8). In both these topologies there is potential to utilize multi-user diversity, but the methods to do so are distinct. Therefore we explore them separately in Sections 4.1 and 4.2.

4.1 Opportunistic Scheduling

In the hierarchical topology, we distinguish between two types of problems; the first is the *uplink* channel where the nodes communicate to the access point (many-to-one communication or the *multiple access channel*), and the second is the *downlink* channel where the access point communicates to the nodes (one-to-many communication or the *broadcast channel*).

The idea of multi-user diversity can be further motivated by looking at the scalar fading multiple access channel. If the users are distributed across geographical areas, their channel responses will be different depending on their local environments. This is modeled by choosing the users' channels to vary according to channel distributions that are chosen to be independent and identical across users. The rate region for the uplink channel for this case was characterized in (Knopp and Humblet, 1995) where it was shown that in order to maximize the total information capacity (the sum rate), it is optimal to transmit *only* to the user with the best channel. For the scalar channel, the channel gain determines the best channel. The result in (Knopp and Humblet, 1995) when translated to rapidly fading channels results in a form of time-division multiple access (TDMA), where the users are not preassigned time slots, but are scheduled according to their respective channel conditions. Even if a particular user at the current time might be in a deep fade, there could be another user who has good channel conditions. Hence this strategy is a form of *multi-user diversity* where the diversity is viewed across users. Here the multi-user diversity (which arises through independent channel realizations across users) can be harnessed using an appropriate scheduling strategy. If the channels vary rapidly in time, the idea is to schedule users when their channel state is close to the peak rate that it can support. A similar result also holds for the scalar fading broadcast channel (Tse, 1997; Li and Goldsmith, 2001). Note that this requires feedback from the users to the base station about the channel conditions. The feedback could be just the received SNR. These results are proved on the basis of two assumptions. One is that all the users have identically distributed (*i.e.,* symmetric) channels and the other is that we are interested in long-term rates. We focus on the first assumption, and later briefly return to the question about delay.

In wireless networks, the users' channel is almost never symmetric. Nodes that are closer to the base station experience much better channels on the average than nodes that are further away (due to path loss, see Section 2). Therefore, using a TDMA technique that allows exclusive use of the channel to the best user would be inherently unfair to users who are further away. Suppose the long-term average rate $\{T_k\}$ is to be provided to the users. The criterion used in the result in (Knopp and Humblet, 1995) was the sum throughput of all the users, *i.e.,* $\max \sum_k T_k$. This criterion can be maximized by only scheduling the nodes with strong channels, and this could be an unfair allocation of resources across users. In order to translate the intuition about multi-user diversity into practice, one would need to ensure fairness among users. The idea in (Bender et al., 2000; Jalali et al., 2000;

Chaponniere et al., 2002), is to use a *proportionally fair* criterion for scheduling which maximizes $\sum_{k=1}^{K} \log(T_k)$. This idea is inherently used in the downlink scheduling algorithm used in IS-856 (Bender et al., 2000; Jalali et al., 2000; Chaponniere et al., 2002) (also known as the High Data Rate - HDR 1xEV-DO system).

The scheduling algorithm implemented in the 1xEV-DO system keeps track of the average throughput $T_k(t)$ of user k in a past window of length t_c. Let the rate that can be supported to user k at time t be denoted by $R_k(t)$. At time t, the scheduling algorithm transmits to the user with the largest $\frac{R_k(t)}{T_k(t)}$ among the active users. The average throughputs are then updated given the current allocation. Since this idea ensures fairness while utilizing multi-user diversity, it is an instantiation of an *opportunistic scheduler*.

This scheduling algorithm described above relies on the rates supported by the users to vary rapidly in time. But this assumption can be violated when the channels are constant or are very slowly time-varying. In order to artificially induce time-variations, (Viswanath et al., 2002, 2004) propose to use multiple transmit antennas and introduce random phase rotations between the antennas to simulate fast fading. This idea of phase-sweeping for multiple-antennas has been also proposed in (Weerackody, 1993; Hiroike et al., 1992) in the context of creating time diversity in single-user systems. With such artificially induced fast channel variations, the same scheduling algorithm used in IS-856 (outlined above) inherently captures the multi-user spatial diversity of the network. In (Viswanath et al., 2002), this technique is shown to achieve the maximal diversity order (see Section 3.2) for each user, asymptotically in number of (uniformly-distributed) users.

In a heavily loaded system (large number of users) and where there is a uniform distribution of users, the technique proposed in (Viswanath et al., 2002) is attractive. However, for lightly loaded systems, *or* when delay is an important QoS criterion, its desirability is less clear. Given that the technique proposed in (Viswanath et al., 2002) is based on a rate-based QoS criterion, it cannot provide delay guarantees for the jobs of different users. This motivates the discussion of scheduling algorithms for job-based QoS criteria.

In job-based criteria, the requests are assumed to come in at certain arrival times a_i, and we have information about the size s_i (say in bytes). *Response time* is defined to be $c_i - a_i$ where c_i is the time when a request was fully serviced and a_i is the arrival time of the request. This is a standard QoS criterion for a request. *Relative response* is defined as $\frac{c_i - a_i}{s_i}$ (Bender et al., 1998). Relative response was proposed in the context of heterogeneous workloads, such as the web, *i.e.,* requests for data of different sizes (thus, different s_i). The above criteria relate to guarantees per request; we could also give guarantees only over all requests. For example, the overall performance criterion for a set of jobs could be the l_∞ norm, namely, $\max_i(c_i - a_i)$ (*i.e.,* max response time) or $\max_i \frac{c_i - a_i}{s_i}$ (*i.e.,* max relative response). Other criteria based on average instead of maximum are also studied.

The new generation of wireless networks can support multiple transmission rates de-

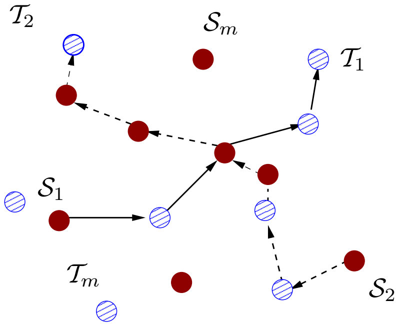

Figure 8: Routes from sources $\{\mathcal{S}_i\}$ denoted by filled circles to destinations $\{\mathcal{T}_i\}$ denoted by shaded circles.

pending on the channel conditions. Assuming an accurate communication-theoretic model for the physical layer achievable rates (as described in Section 3), job-scheduling algorithms are proposed and analyzed for various QoS criteria in (Becchetti et al., 2002). These algorithms utilize diverse job requirements of the users to provide provable guarantees in terms of the job-scheduling criteria.

These discussions just illustrate how multi-user diversity can be utilized in hierarchical networks. This form of opportunistic scheduling is an important part of the new generation of wireless data networks.

4.2 Mobile Ad Hoc Networks

In an ad hoc communication topology (network), one need not transmit information directly from source to destination, but instead can use other users which act as relays to help communication of information to its ultimate destination. Such multihop wireless networks have rich history (see, for example (Hou and Li, 1986) and references therein).

In an important step toward systematically understanding the capacity of wireless networks, (Gupta and Kumar, 2000) explored the behavior of wireless networks asymptotically in the number of users. In their setup, n nodes were placed independently and randomly at locations $\{S_i\}$ in a finite geographical area (a scaled unit disk). Also $m = \Theta(n)$ source and destination (S-D) pairs $\{(\mathcal{S}_i, \mathcal{T}_i)\}$ are randomly chosen as shown in Figure 8[11]. The model assumes that each source $\mathcal{S}_i$ has an infinite stream of (information) packets to send to its respective destination $\mathcal{T}_i$. The nodes are allowed to use any scheduling and relaying strategy through other nodes to send the packets from the sources to the destinations (see Figure 8). The goal is to analyze the best possible long term throughput per S-D pair asymptotically in the number of nodes n.

[11] We use the notation $f(n) = \Theta(g(n))$ to denote $f(n) = O(g(n))$ as well as $g(n) = O(f(n))$. Here $f(n) = O(g(n))$ means $\limsup_{n\to\infty} |\frac{f(n)}{g(n)}| < \infty$.

In (Gupta and Kumar, 2000), a single-user communication model was used where each node transmitted information to its intended receiver (relay or destination node), and the receiver considered the interference from other nodes as part of the noise. Therefore in the communication model, a successful transmission of rate R occurred when the signal-to-interference-plus-noise ratio (SINR) was above a certain threshold β. Clearly, such a communication model can be improved by attempting to decode the "interference" from other nodes using sophisticated multi-user decoding (Verdu, 1998). But such a decoding strategy was not considered by (Gupta and Kumar, 2000) and therefore this need not be an information-theoretically optimal strategy. In order to represent wireless signal transmission, the signal strength variation was modeled only through path loss (see Section 2) with exponent α. Therefore, if $\{P_i\}$ are the powers at which the various nodes transmitted, then the SINR from node i to node j is defined as

$$SINR = \frac{\frac{P_i}{|S_i - S_j|^\alpha}}{\sigma^2 + \sum_{\substack{k \in \mathcal{I} \\ k \neq i}} \frac{P_k}{|S_k - S_j|^\alpha}}, \tag{26}$$

where $\mathcal{I}$ is the subset of users simultaneously transmitting at some time instant. Next, we need to define the notion of throughput per S-D pair more precisely.

Definition 4.1 *For a scheduling and relay policy π, let $M_i^\pi(t)$ be the number of packets from source node $\mathcal{S}_i$ to its destination node $\mathcal{T}_i$ successfully delivered at time t. A long-term throughput $\tilde{\lambda}(n)$ is feasible if there exists a policy π such that for* every *source-destination pair*

$$\lim \inf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} M_i^\pi(t) \geq \tilde{\lambda}(n) \ . \tag{27}$$

We define the throughput $\lambda(n)$ as the highest achievable $\tilde{\lambda}(n)$.

■

Note that $\lambda(n)$ is a random quantity which depends on the node locations of the users. Our interest is in the *scaling law* governing $\lambda(n)$, *i.e.,* the behavior of $\lambda(n)$ asymptotically in n. One of the main results of (Gupta and Kumar, 2000) was the following.

Theorem 4.1 *There exist constants c_1 and c_2 such that*

$$\lim_{n \to \infty} \mathbb{P}\left\{\lambda(n) = \frac{c_1 R}{\sqrt{n \log n}} \text{ is feasible}\right\} = 1, \ \lim_{n \to \infty} \mathbb{P}\left\{\lambda(n) = \frac{c_2 R}{\sqrt{n}} \text{is feasible}\right\} = 0 \ .$$

■

Therefore, the long-term per-user throughput decays as $O(\frac{1}{\sqrt{n}})$, showing that high per-user throughput may be difficult to attain in large-scale (fixed) wireless networks. This result has been recently strengthened: it was shown by (Franceschetti et al., 2004), that $\lambda(n) = \Theta(\frac{1}{\sqrt{n}})$.

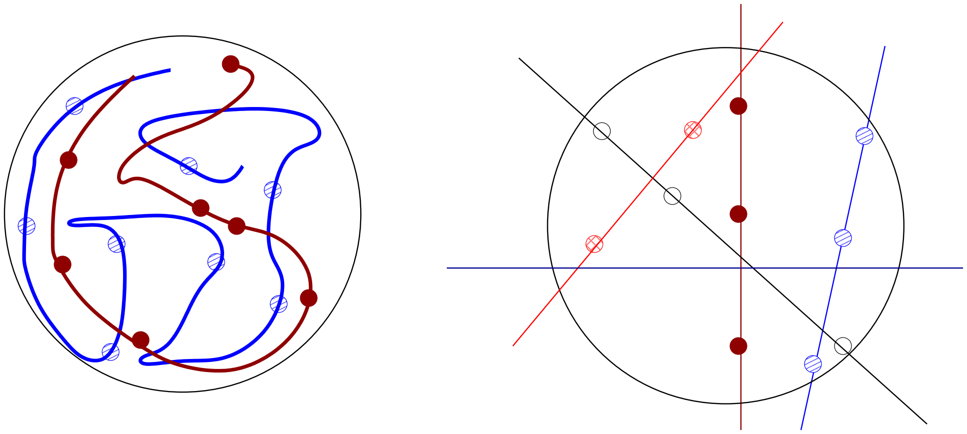

Figure 9: Mobility in ad hoc networks. The figure on the left shows a space-filling mobility model where the nodes uniformly cover the region. The figure on the right shows a limited one-dimensional mobility model where nodes move along *fixed* line segments.

One way to interpret this result is the following. If n nodes are randomly placed in a unit disk, nearest neighbors (with high probability) are at a distance $O(\frac{1}{\sqrt{n}})$ apart. It is shown in (Gupta and Kumar, 2000) that it is important to schedule a large number of simultaneous short transmissions, *i.e.,* between nearest neighbors. If randomly chosen source-destination pairs are $O(1)$ distance apart and we can only schedule nearest-neighbor transmissions, information has to travel $O(\sqrt{n})$ hops to reach its destination. Since there can be at most $O(n)$ simultaneous transmissions at a given time instant, this imposes a $O(\frac{1}{\sqrt{n}})$ upper bound on such a strategy. This is an intuitive argument, and a rigorous proof of Theorem 4.1 is given in (Gupta and Kumar, 2000) among other interesting results.

Note that the coding strategy in Theorem 4.1 was simple and the interference was treated as part of the noise. An open question concerns the throughput when we use sophisticated multi-user codes and decoding is used. Therefore, for such an information-theoretic characterization, understanding the rate region of the relay channel is an important component (Cover and Thomas, 1991). The relay channel was introduced in (van der Meulen, 1977) and the rate region for special cases was presented in (Cover and El-Gamal, 1979). Recently (Xie and Kumar, 2004; Leveque and Telatar, 2005; Gupta and Kumar, 2003), have established that even with network information-theoretic coding strategies, the per S-D pair throughput scaling law decays with the number of users n.

A natural question that arises is whether there is any mechanism by which one can improve the scaling law for throughput in wireless networks. Mobility was one such mechanism examined in (Grossglauser and Tse, 2002). In the model studied, random node mobility was allowed and the locations $\{S_i(t)\}$ vary in a uniform, stationary and ergodic manner over the entire disk (see Figure 9).

In the presence of such symmetric (among users) and "space-filling" mobility patterns, the following surprising result was established (Grossglauser and Tse, 2002).

Theorem 4.2 *There exists a scheduling and relaying policy π and a constant $c > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left\{\lambda(n) = cR \textit{ is feasible}\right\} = 1 \,. \tag{28}$$

■

Therefore, node mobility allows us to achieve a per-user throughput of $\Theta(1)$. The main reason this was attainable was that packets are relayed only through a finite number of hops by utilizing node mobility. Thus, a node carries packets over $O(1)$ distance before relaying it, and therefore (Grossglauser and Tse, 2002) show that, with high probability, if the mobility patterns are space-filling, the number of hops needed from source to destination is bounded instead of growing as $O(\sqrt{n})$ in the case of fixed (non-mobile) wireless networks (Gupta and Kumar, 2000). However, the above mobility model is a generous one since, (i) it is homogeneous, *i.e.,* every node has the same mobility process, and (ii) the sample path of each node "fills the space over time". This means that there is a non-zero probability that the node visits every part of the geographical region or area. A natural question is whether the throughput result in (Grossglauser and Tse, 2002) strongly depends on these two features of the mobility model.

In (Diggavi et al., 2005b), a different mobility model is introduced which embodies two salient features that many real mobility processes seem to possess (*e.g.,* cars traveling on roads, people walking in buildings or cities, trains, satellites circling earth), which are not captured by the model in (Grossglauser and Tse, 2002). First, an individual node typically visits only a small portion of the entire space, and rarely leaves this preferred region. Second, the nodes do move frequently within their preferred regions, and an individual region often covers a large distance. As an extreme abstraction of such mobility processes, (Diggavi et al., 2005b) studied mobility patterns where nodes move along a given set of one-dimensional paths (see Figure 9). In particular, the mobility patterns were restricted to random line segments and once chosen, the configuration of line segments are fixed for all time. Therefore, given the configuration, the only randomness arose through user mobility along these line segments. In order to isolate the effects of one-dimensional mobility from edge effects, (Diggavi et al., 2005b) studied a model in which the nodes are on a unit sphere but each node is constrained to move on a single-dimensional great circle. Therefore, a configuration in this case was a set of line segments (great circles) which were fixed throughout the communication period, and the nodes moved in randomly only on these one-dimensional paths. Thus, the homogeneity assumption in (Grossglauser and Tse, 2002) is now relaxed. In particular, there can be pairs of nodes that are far more likely to be in close proximity to each other than other pairs. For example, if two one-dimensional paths nearly overlap, the probability of close encounter between the nodes is significantly larger than for two paths that are "far apart". This lack of homogeneity implies, as shown in (Diggavi et al., 2005b), that there are configurations where constant throughput is unattainable even with mobility.

Since the capacity of such a mobile ad hoc network then depends on the constellation of one-dimensional paths, the question becomes one of scaling laws for a *random* configuration. Therefore, the configurations themselves are chosen randomly with each one-dimensional path (great circle) chosen independently and with an identical uniform distribution. Given such a random configuration, the question then becomes whether "bad" configurations (where the per S-D pair throughput is not $\Theta(1)$) occur often. One of the key ideas in (Diggavi et al., 2005b) was the identification and proof of *typical* ("good") configurations, on which the average long-term throughput per node is $\Theta(1)$. Intuitively the typical configurations defined in (Diggavi et al., 2005b) are those where the fraction of one-dimensional paths intersecting any given area is *uniformly* close to its expected number. That is, the empirical probability counts are *uniformly* close to the underlying probability of a random one-dimensional path intersecting that area. Therefore, even for a particular deterministically chosen configuration which satisfies the typicality condition, the per S-D pair throughput is $\Theta(1)$. One of the main results in (Diggavi et al., 2005b) is that if the one-dimensional paths are chosen (uniformly) randomly and independently, then for almost all constellations of such paths, the throughput per S-D pair is $\Theta(1)$. Therefore, for random configurations the probability of an *atypical* configuration is shown to go to zero asymptotically in network size n. Thus, although each node is restricted to move in a one-dimensional space, the same asymptotic performance is achieved as in the case when they can move in the entire two-dimensional region.

Theorem 4.3 *Given a configuration $\mathcal{C}$, there exists a scheduling and relaying policy π and a constant $c > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left\{\lambda(n) = cR \textit{ is feasible } |\mathcal{C}\right\} = 1 \tag{29}$$

for almost all configurations $\mathcal{C}$ as $n \to \infty$, i.e.*, the probability of the set of configurations for which the policy achieves a throughput of λ goes to 1 as $n \to \infty$.*

■

Next we give a flavor of the proof techniques used to prove Theorem 4.3. First, we examine a relaying strategy where at each time, every node carries *source packets*, which originate from that node, and *relay packets*, which originated from other nodes and are to be forwarded to their final destinations. In phase I, each sender attempts to transmit a source packet to its nearest receiver, who will serve as a relay for that packet. In phase II, each sender identifies its nearest receiver and attempts to transmit a relay packet destined

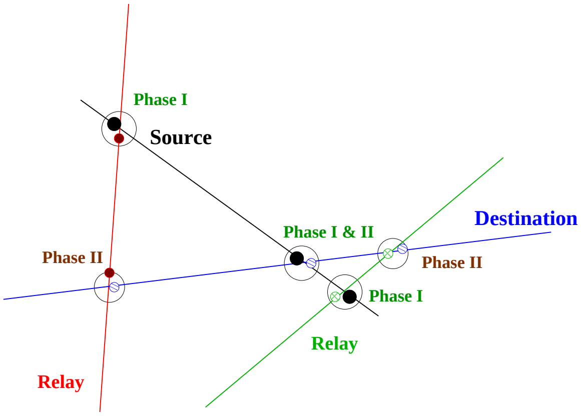

Figure 10: The relaying strategy for mobile nodes. In Phase I, the source attempts to transfer packets to relays. During Phase II, the relays attempt to transfer packets to the destination.

for it, if the sender has one (see Figure 10). As in (26), a successful transmission of rate R occurs when the signal-to-interference-plus-noise ratio (SINR) is above a certain threshold β. Note that it can be shown that if the source nodes attempt to "wait" till it encounters its destination, the per S-D pair throughput cannot be $\Theta(1)$. Therefore every source spreads its traffic to random intermediate nodes depending on the mobility. Moreover, each packet is forwarded successfully to *only one* relay, *i.e.,* there is no duplication. Mobility allows source-destination pairs to be able to relay information through several independent relay paths, since nodes have changing nearest neighbors due to mobility. This method of relaying information through independent attenuation links which vary over time is also a form of *multi-user diversity*. One can see this by observing that the transmission occurs over several realizations of the communication graph $\mathcal{G}_C$. The relaying strategy which utilizes mobility schedules transmissions over appropriate realizations of the graph. Conceptually, this use of independent relays to transmit information from source to destination is illustrated in Figure 11, where the strategy of Theorems 4.2 and 4.3 is used.

Intuitively, if the source is able to *uniformly* spread its traffic through each of its relays (see Figure 11) then we can expect to obtain $\Theta(1)$ throughput per S-D pair. In order for this to occur, we need to show two properties:

Property I: Every node spends the same order of time as the nearest neighbor to $\Theta(n)$ other nodes. This ensures that each source can spread its packets uniformly across $\Theta(n)$ other nodes, all acting as relays, and these packets can in turn be merged back into their respective final destinations.

Property II: When communicating with the nearest neighbor receiver, the capture probability is not vanishingly small even in a large system, even though there are $\Theta(n)$ interfering nodes transmitting simultaneously.

However, with one-dimensional mobility, it is shown in (Diggavi et al., 2005b) that there exists configurations where these properties cannot be satisfied. This is where the identification of *typical configurations* becomes important. For typical configurations through a detailed technical argument it is shown in (Diggavi et al., 2005b) that these properties hold. Moreover, for randomly chosen configurations, it is shown that such typical configurations occur with probability going to 1 asymptotically in n. Therefore, using these components, the proof of Theorem 4.3 is completed.

Throughput-delay trade-off: There is a dramatic gain in the per S-D pair throughput in Theorems 4.2 and 4.3 over Theorem 4.1 from $O(\frac{1}{\sqrt{n}})$ to $\Theta(1)$. A natural question to ask is whether there is a cost to this improvement. The results in Theorems 4.2 and 4.3 utilized node mobility to deliver the information from source to destination. Therefore, the time scale over which this is effective is dependent on the velocity of the nodes, which determines the rate of change of the topology. Hence we can expect there to be significantly larger packet delays for this scheme as compared to the fixed network. In some sense, the Gupta-Kumar result in Theorem 4.1 has a smaller throughput, but also has a smaller packet delay, since the delays depend on successful packet transmissions over the route and not the change in node topology. Hence a natural question to ask is whether there exists a fundamental trade-off between delay and throughput in ad hoc networks. This question was recently studied in (El-Gamal et al., 2004a) where the authors quantified this trade-off.

In order to quantify the trade-off there needs to be a formal definition of delay. In (El-Gamal et al., 2004a) delay $D(n)$ is defined as the sum of the times spent in every relay node. This definition does not include the queueing delay at the nodes, just the delay incurred in successful transmission of the packet on each single hop of the route. Given this definition of delay (El-Gamal et al., 2004a) established that for a fixed random network of n nodes, the delay-throughput trade-off for $\lambda(n) = O(\frac{1}{\sqrt{n\log(n)}})$, is $D(n) = \Theta(n\lambda(n))$. For a mobile ad hoc network as well, a weaker trade-off in throughput and delay was established.

The theoretical developments in Sections 4.1 and 4.2 indicate the strong interactions between the physical layer coding schemes and channel conditions with the networking issues of resource allocation and application design. This is an important insight we can draw for the design of wireless networks. Therefore, several problems which are traditionally considered as networking issues and are typically designed independent of the transmission techniques need to be reexamined in the context of wireless networks. As illustrated, diversity needs to be taken into account while solving these problems. Such an integrated approach is a major lesson learned from the theoretical considerations, and we develop another aspect of this through the study of source coding using route diversity in Section 5.

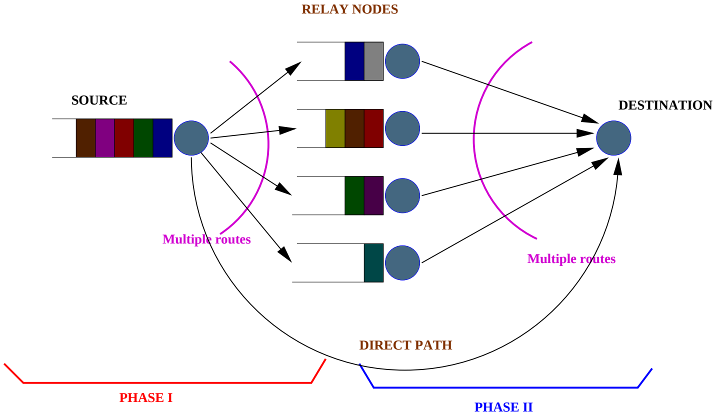

Figure 11: Multiuser diversity through relays.

5 Route diversity

The interest in Section 4.2, was the characterization of *long-term* throughput from source to destination. However, in applications, such as sensor networks (see, for example, (Pottie and Kaiser, 2000; Pradhan et al., 2002), and references therein), there could be node failures which lead to routes being disconnected through a transmission period. This might become particularly crucial when there are strong delay constraints, such as those in real-time data delivery. Such route failures can also occur in ad hoc networks (discussed in Section 4.2) as well as in wired networks. In multihop relay strategies, we could utilize the existence of multiple routes from source to destination in order to increase the probability of successfully receiving the information at the destination within delay constraints despite route (path) failures. This is a form of *route diversity* (see Figure 12) and was first suggested by (Maxemchuk, 1975) in the context of wired networks. Note that in a broad sense, the multiuser diversity studied in mobile ad hoc networks in Section 4.2 also utilizes the presence of multiple routes from source to destination. However, in that case the multiple routes were utilized to increase the long-term per S-D pair throughput. In the topic of this section we will utilize the multiple routes for low-delay applications.

We will examine this problem in the context of delivering a real-time source (like speech, images, video, etc.) with tight delay constraints. If the same information about the source is transmitted over both routes, then this is a form of repetition coding. However, when both routes are successful, there is no performance advantage. Perhaps a more sophisticated technique would be to send correlated descriptions of the source in the two routes such that each description is individually good, but they are different from one another so that if both routes are successful one gets a better approximation of the source. This is the basic idea behind *multiple description* (MD) source coding (El-Gamal and Cover, 1982). This notion can be extended to more than two descriptions as well, but in this section we will focus on

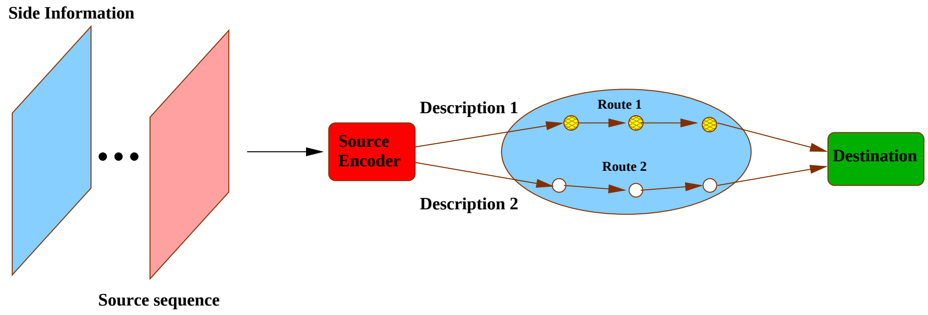

Figure 12: Route diversity.

the two-description case for simplicity. The idea is that the source is coded through several descriptions, where we require that performance (distortion) guarantees can be given to any subset of the descriptions and the descriptions mutually refine each other. This is the topic discussed in Sections 5.1 and 5.2.

In a packet-based network such as the Internet, packet losses are inevitable due to congestion or transmission errors. If the data does not have stringent delay constraints, error recovery methods typically ensure reliability either through a repeat request protocol or through forward error correction (Keshav, 1997). Another technique is through scalable (or layered) coding techniques which send a lower-rate base layer or coarser description of the source and send refinement layers to enhance the description. Such a technique is again dependent on reliable delivery of the base layer, and if the base layer is lost, the enhancement layers are of no use to the receiver. Therefore, such layered techniques are again inherently susceptible to route failures. These arguments reemphasize the need to develop multiple description (MD) source coding schemes. Note that the layered coding schemes form a special case of such MD coding scheme, where guarantees of performance are not given for individual layers, but the layers refine the coarser description of the source.

An important application for future wireless networks could be real-time video. There has been significant research into robust video coding in the presence of packet errors (Reibman and Sun, 2000). The main problem that arises in video is that the compression schemes typically have motion compensation, which introduces memory into the coded stream. Therefore, decoding the current video frame requires the availability of previous video frames. If previous frames are corrupted or lost, the decoder is required to develop methods to conceal such errors. This is an active research topic especially in the context of wireless channels (Girod and Farber, 2000). However, an appealing approach to this problem might be through route diversity and MD coding, and this is briefly discussed in Section 5.1.

5.1 Multiple Description (MD) Source Coding

In order to formalize the requirement of the MD source coder, we study the setup shown in Figure 13. As mentioned earlier, we will illustrate the ideas using only the two-description MD problem. Given a source sequence $\{X(k)\}$, we want to design an encoder that sends two descriptions at rate R_1 and R_2 over the two routes such that we get *guaranteed approximations* of the source when either route fails, or when both succeed. In Section 5.2 we develop techniques that achieve such an objective. In order to understand the fundamental bounds on the performance of such techniques, we need to examine the problem from an information-theoretic point of view. The main tool to do this is given in *rate-distortion* theory (Cover and Thomas, 1991). This theory describes fundamental limits of the trade-off between the rate of the representation of a source and the quality of the approximation. Not surprisingly, the origins of this theory are in (Shannon, 1948, 1958b). In order to give some of the basic ideas, we first make a short digression on the rudiments of this theory.

Rate-distortion function: Given a source sequence $X^T = \{X(1), \ldots, X(T)\}$ from a given alphabet $\mathcal{X}$, the *source encoder* needs to describe it using R bits per source sample (*i.e.,* with a total of RT bits for the sequence). Equivalently we map the source to the index set $\mathcal{J} = \{1, \ldots, 2^{RT}\}$. The goal is that given this description a decoder is able to *approximately* reconstruct the source sequence by the sequence $\hat{X}^T = \{\hat{X}(1), \ldots, \hat{X}(T)\}$. This is accomplished by constructing a function $f : \mathcal{J} \to \hat{\mathcal{X}}^T$, and $\hat{\mathcal{X}}$ is the alphabet over which the reconstruction is done. Common examples for the alphabet are $\mathcal{X} = \mathbb{R} = \hat{\mathcal{X}}$, or the binary field. The *distortion measure* $\tilde{d}(X^T, \hat{X}^T)$ quantifies the quality of the approximation between the reconstructed and original source sequence. Typically, the distortion measure is a single-letter function constructed as

$$\tilde{d}(X^T, \hat{X}^T) = \frac{1}{T} \sum_{i=1}^{T} d(X(i), \hat{X}(i)), \tag{30}$$

where $d(X, \hat{X})$ denotes the quality of the approximation for each sample. Common examples are $d(X, \hat{X}) = |X - \hat{X}|^2$ and Hamming distance (Cover and Thomas, 1991).

The simplest framework to give performance bounds is to analyze the performance of a source encoder for an independent and identically distributed random source sequence. Typically, the interest is in the *average distortion* over the set of input sequences, for the given probability distribution associated with the source sequence. Therefore, the average distortion is $\mathbb{E}[\tilde{d}(X^T, \hat{X}^T)]$, and the problem becomes one of quantifying the smallest rate R that be used to describe the source with average fidelity D, asymptotically in the block length T. This is called the rate-distortion function $R(D)$ and can be given an *operational* meaning by proving that there exist source codes that can achieve this fundamental bound (Cover and Thomas, 1991). The central result in single source rate-distortion theory is that

$R(D)$ is characterized as

$$R(D) = \min_{p(\hat{x}|x):\mathbb{E}[d(x,\hat{x})]\leq D} I(X;\hat{X}), \tag{31}$$

where, as before, $I(X;\hat{X})$ represents the mutual information between X and $\hat{X}$ (Cover and Thomas, 1991). A simple instantiation of this result is the special case where we want $D = 0$, *i.e.,* the lossless case. In this case, one can see that $R(0) = H(X)$, where $H(X)$ is the entropy of the source. Another important special case is when the source sequence comes from a Gaussian distribution, $X \sim \mathcal{N}(0, \sigma_x^2)$, and we are interested in the squared error distortion metric, *i.e.,* $d(X,\hat{X}) = |X - \hat{X}|^2$. In this case, (31) evaluates to $R(D) = \frac{1}{2}\log\frac{\sigma_x^2}{D}$, for $D \leq \sigma_x^2$ and zero otherwise. Another way of writing this is in terms of the distortion-rate function $D(R)$, which characterizes the smallest distortion achievable for a given rate. In the Gaussian case we see that $D(R) = \sigma_x^2 2^{-2R}$. We will interchangeably consider these two quantities.

The result in (31) guarantees only that the average distortion does not exceed D. However, under some regularity conditions, the rate-distortion function remains the same even when we require that the probability of the distortion $\tilde{d}(X^T, \hat{X}^T)$ exceeding D to go to zero (Cover and Thomas, 1991; Berger, 1977). The characterization of the rate-distortion function given in (31) has also been extended in many other ways including sources with memory (Cover and Thomas, 1991).

Armed with this background, we can now formulate the question on the fundamental rate-distortion bounds on multiple description (MD) source coding. The *multiple description* source encoder needs to produce two descriptions of the source using R_1, R_2 bits per source sample respectively. We can formally describe the problem by requiring that the reconstructions $\{\hat{X}_1(k)\}, \{\hat{X}_2(k)\}, \{\hat{X}_{12}(k)\}$ use these descriptions to approximately reconstruct the source (see Figure 13). As in the "single-description" case, we accomplish this by constructing functions

$$f_1 : \mathcal{J}_1 \rightarrow \hat{\mathcal{X}}^T, \;\; f_2 : \mathcal{J}_2 \rightarrow \hat{\mathcal{X}}^T, \;\; f_{12} : \mathcal{J}_1 \times \mathcal{J}_2 \longrightarrow \hat{\mathcal{X}}^T, \tag{32}$$

where $\mathcal{J}_i = \{1, \ldots, 2^{R_i T}\}, i = 1, 2$, and $\hat{\mathcal{X}}$ is the alphabet over which the reconstruction is done. We want the approximations to give average fidelity guarantees of

$$\mathbb{E}[\tilde{d}(X^T, \hat{X}_1^T)] \leq D_1, \;\; \mathbb{E}[\tilde{d}(X^T, \hat{X}_2^T)] \leq D_2, \;\; \mathbb{E}[\tilde{d}(X^T, \hat{X}_{12}^T)] \leq D_{12}. \tag{33}$$

The rate-distortion question in this context is to characterize the bounds on the tuple $(R_1, D_1, R_2, D_2, D_{12})$. Therefore, we are interested in characterizing the achievable *rate-distortion region* described by the tuple $(R_1, D_1, R_2, D_2, D_{12})$. As can be seen, this seems like a much more difficult question than the single-description problem for which there is a complete characterization. As a matter of fact, the complete characterization of the MD rate region is still an open question.

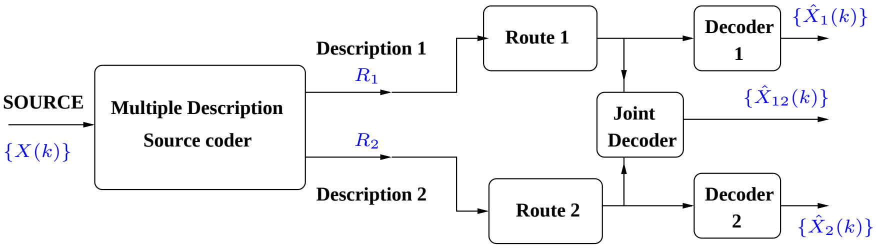

Figure 13: Multiple description (MD) source coding.

This problem was formalized in 1979, and in (El-Gamal and Cover, 1982) a theorem was proved which demonstrated a region of the tuple $(R_1, D_1, R_2, D_2, D_{12})$ for which MD source codes exist.

Theorem 5.1 *(El-Gamal and Cover, 1982) Let $X(1), X(2), \ldots$ be a sequence of i.i.d. finite alphabet random variables drawn according to a probability mass function $p(x)$. If the distortion measures are $d_m(x, \hat{x}_m), m = 1, 2, 12$ then an achievable rate region for tuples $(R_1, R_2, D_1, D_2, D_{12})$, is given by the convex hull of the following.*

$$\begin{aligned} R_1 &\geq I(X; \hat{X}_1), \quad R_2 \geq I(X; \hat{X}_2), \\ R_1 + R_2 &\geq I(X; \hat{X}_{12}, \hat{X}_1, \hat{X}_2) + I(\hat{X}_1; \hat{X}_2) \end{aligned} \tag{34}$$

for some probability mass function $p(x, \hat{x}_1, \hat{x}_2, \hat{x}_{12})$ such that, $\mathbb{E}[d_t(X, \hat{X}_t)] \leq D_t$, $t = 1, 2, 12$.

■

This region was further improved in (Zhang and Berger, 1987) to a larger region for which MD source codes exist. However, what is unknown is whether these characterizations completely exhaust the set of tuples that can be achieved, *i.e.,* a converse for the MD rate-distortion region. There are some special cases for which there are further results (Ahlswede, 1985; Fu and Yeung, 2002), and references therein. There has also been recent work on achievable rate-regions for more than two descriptions (Venkataramani et al., 2003; Pradhan et al., 2004). However, in these cases as well the complete characterization is unknown.

The *only* case for which the MD region is completely characterized is that for memoryless Gaussian sources with squared error distortion measures and specifically for two descriptions[12]. In (Ozarow, 1980) it was shown that the two-description MD region given in (El-Gamal and Cover, 1982) was also applicable to the Gaussian case with squared error distortion where the alphabet is not finite. Moreover it was shown that the region in

[12] Interestingly, this result is specifically for two descriptions and does not immediately extend to the general case.

Theorem 5.1 was in fact the complete characterization by proving a converse (outer bound) to the rate region. In this context the source was modeled as a sequence of i.i.d. Gaussian random variables $X \sim \mathcal{N}(0, \sigma_x^2)$ and the squared error distortion measure was chosen, *i.e.,* $d_m(x, \hat{x}_m) = |x - \hat{x}_m|^2, m = 1, 2, 12$. Therefore, specializing the result in Theorem 5.1 to the Gaussian case yields the following complete characterization of the set of all achievable tuples $(R_1, R_2, D_1, D_2, D_{12})$ (El-Gamal and Cover, 1982; Ozarow, 1980).

$$\begin{aligned} D_1 &\geq \sigma_x^2 e^{-2R_1}, \quad D_2 \geq \sigma_x^2 e^{-2R_2}, \\ D_{12} &\geq \frac{\sigma_x^2 e^{-2(R_1+R_2)}}{1 - \left[\sqrt{\left(1 - \frac{D_1}{\sigma_x^2}\right)\left(1 - \frac{D_2}{\sigma_x^2}\right)} - \sqrt{\left(\frac{D_1}{\sigma_x^2}\right)\left(\frac{D_2}{\sigma_x^2}\right) - e^{-2(R_1+R_2)}}\,\right]^2}. \end{aligned} \tag{35}$$

In order to interpret this result, consider the following. As seen before, for a single-description Gaussian problem, the minimum distortion for a given rate is $D(R) = \sigma_x^2 2^{-2R}$. Therefore, the distortions D_1, D_2 clearly need to be governed by the single description bound, and this explains the first two inequalities in (35). However, in the MD problem we also need to bound the distortion D_{12} when both descriptions are available. From the single-description bound it is clear that we would have $D_{12} \geq D(R_1 + R_2) = \sigma_x^2 2^{-2(R_1+R_2)}$. Therefore, a natural question is whether this bound on D_{12} can be achieved with equality. However, the result in Theorem 5.1 shows that this is not possible unless $D_1 = \sigma_x^2$ or $D_2 = \sigma_x^2$. Here is where the tension between the two descriptions manifests itself. We examine the tension in the symmetric case, when we have $D_1 = D_2 = D$, $R_1 = R_2 = R$ and specialize it for the unit variance source $\sigma_x^2 = 1$. If we want the individual descriptions to be as efficient as possible (*i.e.,* $D = e^{-2R}$), then we see that $D_{12} \geq \frac{D}{2-D}$ which is far larger than $D(R_1 + R_2) = e^{-2(R_1+R_2)} = D^2$. For small D, we see that D_{12} is approximately $\frac{D}{2}$ which is much larger than D^2. Therefore, if we ask that the individual descriptions be close to optimal themselves, then they do not mutually refine each other very well. This reveals the tension between getting small the distortions D_1, D_2 of individual descriptions and a small D_{12}. We need to make the individual descriptions coarser in order to get more mutual refinement in D_{12}.

One important real-time application is that of video coding. This can be viewed as a sequence of individual frames which are correlated to each other. The traditional way of encoding video is by describing the "current" frame differentially with respect to the previous frame. This is done through a block-matching technique where the "closest" (in terms of squared distance) blocks from the previous frame are matched to blocks in the current frame, and then only the differences are transmitted. The rationale behind this idea is that blocks are only relatively displaced due to motion of objects in the video, and hence this mechanism is called motion compensation in the literature (Reibman and Sun, 2000). Note that in this scheme, the encoder *explicitly* uses the knowledge of the previous frame. Clearly, when there are packet/route errors and the previous frame is not received at the destination, the reconstruction is difficult since the previous reference frame is not available. Therefore, several fixes to this problem have been developed over the past two decades see (Girod and Farber, 2000) and references therein.

In a more abstract framework, we can think of the video as a sequence of correlated random variables which we are trying to describe efficiently. In (Witsenhausen and Wyner, 1980) an alternate approach was taken by considering the video coding problem as a source coding problem with *side-information*. In this setting, after encoding and transmitting the "previous" frame, the "current" frame develops an encoder which does *not* explicitly depend on the knowledge of the previous frame. The basic idea of this scheme arises from encoding schemes and decoding described in (Slepian and Wolf, 1973; Wyner and Ziv, 1976). Since the encoder does not explicitly use the side-information (previous frame) it can be designed such that the computational complexity is shifted from the encoder to the decoder. Such an architecture is attractive for applications where the encoder needs to be simple but the decoder can be more complex. This idea has been developed comprehensively in (Puri and Ramchandran, 2003), where practical coding techniques are developed with such applications in mind.

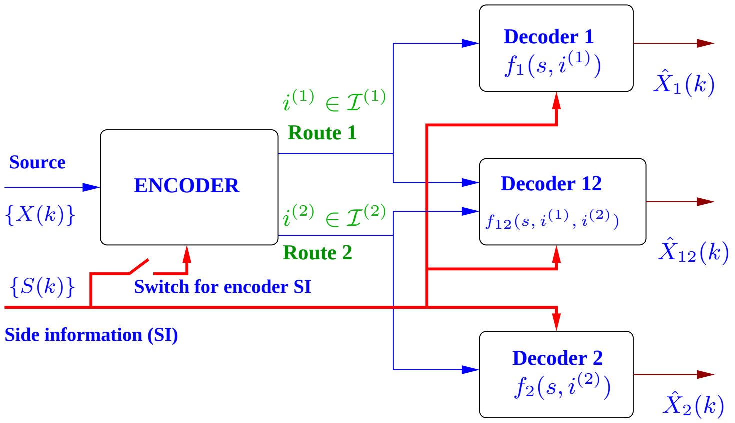

Figure 14: Multiple description source coding with side information.

However, even with this idea the robustness to route failures which is inherent to MD coding is not captured. Motivated by this, (Diggavi and Vaishampayan, 2004) considered the MD problem with side information (see Figure 14). In this abstract setting, we want to encode a source $\{X(k)\}$ when the decoder has knowledge of a correlated process $\{S(k)\}$ as side-information. For example, in the setting of (Witsenhausen and Wyner, 1980; Puri and Ramchandran, 2003), the side-information could be the previous frame. In order to describe the source in the presence of route diversity, we can pose a MD problem, but now with side information as shown in Figure 14. Clearly this is a generalization of the MD problem, and an achievable rate region was established for this problem in (Diggavi and Vaishampayan, 2004).

Theorem 5.2 *Let $(X(1), S(1)), (X(2), S(2)) \dots$ be drawn i.i.d. $\sim Q(x, s)$. If* only *the decoder has access to the side information $\{S(k)\}$, then $(R_1, R_2, D_1, D_2, D_{12})$ is achievable if there exist random variables (W_1, W_2, W_{12}) with probability mass function $p(x, s, w_1, w_2, w_{12}) = Q(x, s)p(w_1, w_2, w_{12}|x)$, that is, $S \leftrightarrow X \leftrightarrow (W_1, W_2, W_{12})$ forms a Markov chain, such that*

$$\begin{aligned} R_1 &> I(X; W_1|S), \;\; R_2 > I(X; W_2|S) \\ R_1 + R_2 &> I(X; W_{12}, W_1, W_2|S) + I(W_1; W_2|S) \end{aligned} \tag{36}$$

and there exist reconstruction functions f_1, f_2, f_{12} which satisfy

$$\begin{aligned} D_1 &\geq \mathbb{E}[d_1(X, f_1(S, W_1))], \;\; D_2 \geq \mathbb{E}[d_2(X, f_2(S, W_2))] \\ D_{12} &\geq \mathbb{E}[d_{12}(X, f_{12}(S, W_{12}, W_1, W_2))]. \end{aligned} \tag{37}$$

■

This result gives an achievable rate region, but the complete characterization for this problem is open. A slightly improved region to Theorem 5.2 is also found in (Diggavi and Vaishampayan, 2004). However, it is unknown whether this region exhausts the achievable rate region. But for the case when both the source and the side information are jointly Gaussian, and we are interested in the squared error distortion, a complete characterization of the rate-distortion region was obtained in (Diggavi and Vaishampayan, 2004).

In more detail the result was the following. Let $(X(1), S(1)), (X(2), S(2)) \dots$ be a sequence of i.i.d. jointly Gaussian random variables. With no loss of generality this can be represented by

$$S(k) = \alpha \left[X(k) + U(k) \right], \tag{38}$$

where $\alpha > 0$ and $\{X(k)\}, \{U(k)\}$ are independent Gaussian random variables with $\mathbb{E}[X] = 0 = \mathbb{E}[U]$, $\mathbb{E}[X^2] = \sigma_X^2$, $\mathbb{E}[U^2] = \sigma_U^2$. As considered in Theorem 5.2, only the decoder has access to the side information $\{S(k)\}$. If the distortion measures are $d_m(x, \hat{x}_m) = ||x - \hat{x}_m||^2, m = 1, 2, 12$ then it is shown in (Diggavi and Vaishampayan, 2004) that the set of all achievable tuples $(R_1, R_2, D_1, D_2, D_{12})$ are given by

$$D_1 > \sigma_{\mathcal{F}}^2 e^{-2R_1}, \quad D_2 > \sigma_{\mathcal{F}}^2 e^{-2R_2}, \quad D_{12} > \frac{\sigma_{\mathcal{F}}^2 e^{-2(R_1+R_2)}}{1 - (\sqrt{\tilde{\Pi}} - \sqrt{\tilde{\Delta}})^2}, \tag{39}$$

where $\sigma_{\mathcal{F}}^2 = \frac{\sigma_X^2 \sigma_U^2}{\sigma_X^2 + \sigma_U^2}$ and $\tilde{\Pi}, \tilde{\Delta}$ are given by

$$\tilde{\Pi} = \left(1 - \frac{D_1}{\sigma_{\mathcal{F}}^2}\right)\left(1 - \frac{D_2}{\sigma_{\mathcal{F}}^2}\right), \quad \tilde{\Delta} = \left(\frac{D_1}{\sigma_{\mathcal{F}}^2}\right)\left(\frac{D_2}{\sigma_{\mathcal{F}}^2}\right) - e^{-2(R_1+R_2)}. \tag{40}$$

The result in (39) also shows that the rate-distortion region in this case is the same as that achieved when *both* encoder and decoder have access to the side information. That is, in the Gaussian case, the rates that can be achieved are the same whether the switch in Figure

14 is open or closed. In (Wyner and Ziv, 1976) it was shown that in the single-description Gaussian case, the decoder-only side-information rate-distortion function coincided with that when both encoder and decoder were informed of the side-information. The result in (39) establishes that this is also true in the Gaussian two-description problem with decoder side-information. However, the encoding and decoding techniques to achieve these rate tuples are very different when the encoder has access to the side information than when it does not. This shows that there might be efficient mechanisms to construct MD video coders which are robust to route failures. Some of the code constructions that bring this idea to fruition are discussed in Section 5.2.

5.2 Quantizers for Route Diversity

The results given in Section 5.1 show the existence of codes that can achieve the rate tuples given in Theorems 5.1 and 5.2, but there are no *explicit* constructions. In this section we explore explicit coding schemes which utilize the presence of route diversity.

As seen in Section 5.1, the single-description rate-distortion function quantifies the fundamental limits of the trade-off between the rate of the representation of a source and its average fidelity. The result in (30) showed the existence of such codes. Explicit constructions of these codes are called *quantizers* (Gray and Neuhoff, 1998; Gersho and Gray, 1992). More formally, quantizers map a sequence $\{X(1), \ldots, X(T)\}$ of source samples into a "representative" reconstruction $\{\hat{X}(1), \ldots, \hat{X}(T)\}$ through an *explicit* mapping which is typically computationally efficient. *Scalar quantizers* operate on a single source sample $X(k)$ at a time. Most current systems use scalar quantizers (Jayant and Noll, 1984). However, rate-distortion theory tells us that using sequences is important, and hence *vector quantizers* use sequences of source samples, *i.e.,* $T > 1$ for quantization. Quantization techniques for single description have been quite well studied and understood (Jayant and Noll, 1984; Gersho and Gray, 1992; Gray and Neuhoff, 1998).

The rudiments of the MD coding ideas arose in the 1970s at Bell Laboratories. In (Jayant, 1981) a very simple idea of channel splitting was proposed and analyzed. The basic idea was to oversample a speech signal and send the odd samples through one channel and the even ones through another. However, this technique is not very efficient in terms of rate. Many such simple coding techniques were being considered at Bell laboratories, but the ideas were not archived. These questions actually motivated the information-theoretic formulation of the MD problem described in Section 5.1. The systematic study of coding for multiple descriptions was initiated in (Vaishampayan, 1993). Its publication resulted in a spurt of recent activity on the topic (see, for example (Vaishampayan et al., 2001; Goyal and Kovacevic, 2001; Diggavi et al., 2002b; Ingle and Vaishampayan, 1995), and references therein). More recently the utility of MD coding in conjunction with route diversity has also created interest in the networking community (see (Apostolopoulos and Trott, 2004),

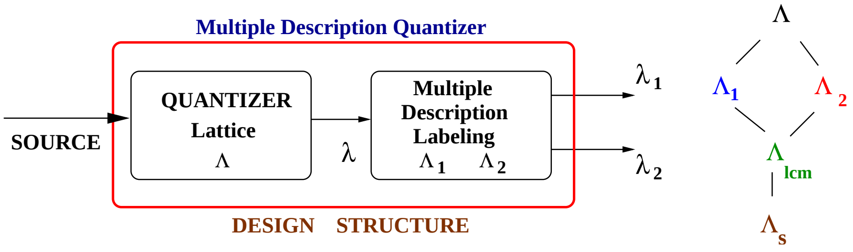

Figure 15: Structure of Multiple Description quantizer.

and references therein).

The basic idea introduced in (Vaishampayan, 1993) constructed scalar quantizers for the MD problem. This was done specifically for the symmetric case, where $D_1 = D_2$ and $R_1 = R_2$. This symmetric construction was extended to structured (lattice) vector quantizers in (Vaishampayan et al., 2001). The symmetric case has been further explored by several other researchers (Goyal and Kovacevic, 2001; Ingle and Vaishampayan, 1995). The importance of structured quantizers is in the computational complexity of the source encoder. For example, just as in channel coding, trellis-based structures are also important in source coding. Such structures have also been proposed for the symmetric MD problem (Buzi, 1994; Jafarkhani and Tarokh, 1999). In general, unstructured quantizers based on training on some source samples can also be constructed, but the computational complexity of such techniques is much higher than structured (lattice) quantizers and therefore they are less attractive in practice. Such unstructured quantizers have been considered in the literature (Fleming et al., 2004). Our focus in this chapter will be on structured quantizers for which we have computationally efficient encoders as well as techniques to analyze their performance.

In general we would like to design MD quantizers that can attain an arbitrary rate-distortion tuple, and not just the symmetric case. This is motivated by applications where the multiple routes have disparate capacities (and therefore rate requirements) as well as different probabilities of route failures. In these cases, we need to design *asymmetric* MD quantizers which give graceful degradation in performance with route failures. Such a structure was studied in (Diggavi et al., 2002b) and is depicted in Figure 15.

We illustrate the ideas of MD quantizer design from (Diggavi et al., 2002b), using a scalar example. In Figure 16, the first line represents a uniform scalar quantizer. If we take a single source sample $X(k) \in \mathbb{R}$, then the uniform quantizer maps this sample to the closest "representative" point $\hat{X}$ on the one-dimensional (scaled integer) lattice Λ. Loosely, a T-dimensional lattice is a set of regularly spaced points in $\mathbb{R}^T$ for which any point can be chosen as the origin and the set of points would be the same. A more precise notion is based on the set of points forming an additive group (Conway and Sloane, 1999).

Each of the representation points is given a unique label λ and this label is transmitted to the receiver. The transmission rate depends on the number of labels. Typically a finite set of points $2M$ is used to represent the labels. In a straightforward manner, this translates to a rate of $\log(2M)$ bits per source sample. If the source either has finite extent or a finite second-order moment, such a quantizer would have a bounded squared error distortion. If the representative points are separated by a distance of Δ, then the worst-case squared error distortion between a source sample and the representative is $\frac{\Delta^2}{4}$ for source samples $X(k) \in [-\frac{(M+1)\Delta}{2}, \frac{(M+1)\Delta}{2}]$. For a uniform distribution of the source in the region $X(k) \in [-\frac{(M+1)\Delta}{2}, \frac{(M+1)\Delta}{2}]$, the average distortion is $\frac{\Delta^2}{12}$ (Gersho and Gray, 1992).

The mapping described above is a single-description uniform scalar quantizer. The MD scalar quantizer needs to map every source sample to an *ordered pair* of representation points $(\hat{X}_1, \hat{X}_2)$. The labels (λ_1, λ_2) of this pair are used to send information over the two routes. For example, we could send the label λ_1 over the first route and label λ_2 over the second route. Now, in Figure 16 we have illustrated this by choosing coarser scalar quantizers in the second and third lines for the representations $\hat{X}_1$ and $\hat{X}_2$ respectively. These quantizers are also one-dimensional lattices Λ_1 and Λ_2 respectively. These representations $\hat{X}_1, \hat{X}_2$ in themselves give coarser information about the source sample, *i.e.,* have a larger distortion than the "finer" quantizer Λ shown in the first line. Now, we need to represent the source sample $X(k)$ by a pair of representation points from Λ_1 and Λ_2. We want to choose this pair in such a way that if either of the labels is lost due to route failure, then we are still guaranteed a certain distortion. However, if both labels are received, *i.e.,* both routes are successful, then we need to get a smaller distortion. This means that the label pair have to mutually refine each other's representations.

One such labeling technique is illustrated in Figure 16. Each point in the coarser lattices $\hat{X}_1$ in Λ_1 and $\hat{X}_2$ in Λ_1 is given a label λ_1 and λ_2 respectively. The idea is then to give a pair of labels (λ_1, λ_2) to each of the points on the fine lattice Λ. Every lattice point in Λ gets a *unique* label pair (λ_1, λ_2). Once this *labeling function* is constructed, then we can form the multiple description (MD) scalar quantizer by doing the following two steps. First, reduce the source sample $X(k) \in \mathbb{R}$ to its closest representative in Λ, $\hat{X}$ with label λ, *i.e.,* apply a uniform scalar quantizer to $X(k)$. Given this $\hat{X}$, and the labeling function, we know the pair (λ_1, λ_2) that represents $\hat{X}$. The second step is to associate $\hat{X}_1$ with the reconstruction given by the label λ_1 in the first coarse quantizer Λ_1, and similarly for $\hat{X}_2$ in Λ_2. These operations are what the structure in Figure 15 represents. Therefore, in this design, the main task is to construct the labeling function for each point in Λ. Given the label pair (λ_1, λ_2), the encoder sends the index associated with λ_1 on route 1 and the index for λ_2 on route 2.

Before describing the labeling function, we examine the decoder structure in the MD scalar quantizer described above. First recall that the labeling function is designed so that

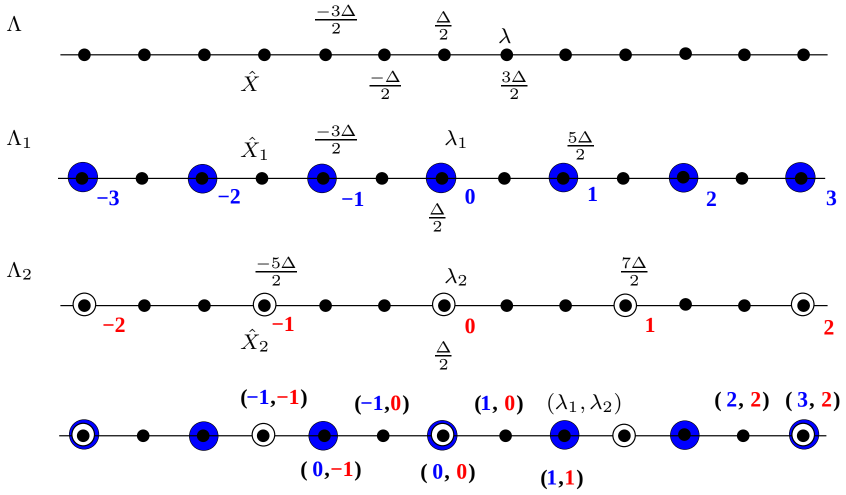

Figure 16: Scalar quantizer labeling example. The top line is a uniform scalar quantizer that maps source points $X(k)$ to a set of discrete representatives $\hat{X}$. The second and third line show coarser uniform scalar quantizers. The last line puts together the combination of the coarser quantizers to give an ordered pair (λ_1, λ_2) as a label to every lattice point λ in the fine quantizer.

any particular pair (λ_1, λ_2) is *uniquely* associated with a particular λ. Therefore, if *both* routes succeed, then the receiver is able to reconstruct λ and as a consequence $\hat{X}$. This means that the distortion in this case is that associated with the fine quantizer Λ, *i.e.*, the average fidelity is $\frac{\Delta^2}{12}$. Now suppose route 1 succeeds and route 2 fails, then the receiver has only λ_1 and does not know λ_2. For example, suppose in Figure 16, the label pair $(-1, 0)$ was chosen at the source encoder, *i.e.*, $\lambda_1 = -1, \lambda_2 = 0$. Now, the receiver knows that the encoder was trying to send one of the two points $(-1, 0)$ or $(-1, -1)$ and since route two failed it does not know which. More generally, in this situation, the receiver knows that λ belongs to the set of points in Λ which have the *same* first label λ_1 but have *different* second label λ_2. Now, assume that the decoder uses the reconstruction $\hat{X}_1$ associated with label $\lambda_1 = -1$ in Λ_1 (see second line in Figure 16). Therefore, for this particular example, the worst-case error due to this choice is $\frac{9\Delta^2}{4}$. This example also shows that the labeling function directly affects the decoder distortion. The design of the labeling function is the central part of the MD quantizer. The reconstruction $\hat{X}_1$ can use the mean of the set of all points in Λ associated with the *same* first label λ_1 which may improve the distortion. Note that in general this might not coincide with the reconstruction associated with λ_1. For design simplicity this reconstruction need not be taken into account in designing the labeling function, but rahter can be used only at the decoder to improve the final distortion.

In general, we would need to construct a labeling function for all the points in Λ. However, we describe a particular design which solves a smaller problem and then expands its solution to Λ (Vaishampayan et al., 2001; Diggavi et al., 2002b). We will illustrate this idea using the example shown in Figure 16.

In the last line of Figure 16, we have depicted the overlay of the two coarse one-dimensional lattices Λ_1, Λ_2 along with Λ. We see that there is a repetitive pattern after every six points in Λ. This is not a coincidence, because Λ_1 was formed by taking every second point in Λ and Λ_2 by taking every third. The least common multiple is six and therefore we would expect the pattern to repeat. The basic idea is to just form a labeling function for these six points and then "shift" these labels to tile the entire lattice Λ. For example, in Figure 16, consider the point which we have labeled as $(2, 2)$ on the last line. This was done in the following manner. Notice that the repeating pattern of six points can be anchored by the points where both the Λ_1 and Λ_2 points coincide. In Figure 16, these are the points which have overlapped circles on the last line. We can think of all points in Λ with respect to these anchor points. For example, the point labeled $(2, 2)$ is one point to the left of such an overlap point and is "equivalent" to the point labeled $(-1, 0)$. More precisely, it is in the same *coset* as the other point with respect to the intersection lattice Λ_s, which is formed by the anchor points. Therefore, we get the label by shifting the label of $(-1, 0)$ with respect to its cosets. In this case, note that $\lambda_1 = -1$ in Λ_1 is two points to the left of the anchor point $(0, 0)$. Therefore, the corresponding point with respect to the anchor point $(3, 2)$ is $\lambda_1 = 2$ and hence the first label for the point of interest is $\lambda_1 = 2$. Next, the corresponding point of the label $\lambda_2 = 0$ in Λ_2 with respect to the anchor point $(3, 2)$ is $\lambda_2 = 2$. This gives us the label $(2, 2)$ which is shown in the Figure 16. In a similar manner, given the labeling for the six points, we can construct the labeling for all points in Λ by the shifting technique described above. Actually, the six points correspond to the discrete Voronoi region of the point $(0, 0)$ of the intersection lattice of the anchor points. Therefore, we can focus on constructing labels for the points in the Voronoi region of the intersection lattice. Note that in the example of Figure 16, the intersection lattice had an index of six which is exactly the least common multiple of the indices of lattices Λ_1, Λ_2 in Λ. This is also true when the indices of Λ_1, Λ_2 in Λ are not co-prime (Diggavi et al., 2002b).

Let $V_{\Lambda_s:\Lambda}(0)$ be defined as the Voronoi region of the intersection lattice. Our problem is to develop the labeling function for the points in $V_{\Lambda_s:\Lambda}(0)$ in order to satisfy the individual distortion constraints D_1, D_2. This is accomplished by using a Lagrangian formulation in (Diggavi et al., 2002b). This formulation reduces to finding the labeling scheme $\alpha(\lambda) = (\alpha_1(\lambda), \alpha_2(\lambda))$ so as to minimize,

$$\sum_{\lambda \in V_{\Lambda_s:\Lambda}(0)} \left[\gamma_1 \|\lambda - \alpha_1(\lambda)\|^2 + \gamma_2 \|\lambda - \alpha_2(\lambda)\|^2 \right]. \tag{41}$$

For this minimization problem we need to choose the appropriate labels $(\alpha_1(\lambda), \alpha_2(\lambda)) =$

(λ_1, λ_2). This is done by observing the following identity.

$$\gamma_1\|\lambda - \lambda_1\|^2 + \gamma_2\|\lambda - \lambda_2\|^2 = \frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}\|\lambda_2 - \lambda_1\|^2 + (\gamma_1 + \gamma_2)\|\lambda - \frac{\gamma_1\lambda_1 + \gamma_2\lambda_2}{\gamma_1 + \gamma_2}\|^2.$$

This results in the following design guideline. The labeling problem is split into two parts: (i) Choose $|V_{\Lambda_s:\Lambda}(0)|$ "shortest" pairs (λ_1, λ_2) (*not* all pairs of (λ_1, λ_2) are used). (ii) Assign these pairs to lattice points $\lambda \in V_{\Lambda_s:\Lambda}(0)$. The second design can be solved very efficiently using linear programming methods. The solution of this labeling problem illustrates an important feature of the MD quantizer design that is quite distinct from the single-description case. It can happen that particular labels of each description can be non-contiguous, *i.e.,* not all points λ which get the same label, say λ_1, need to occur contiguously. This is quite different from the single description case, where the labels are assigned to contiguous intervals. Also, the labels generated in this systematic manner are non trivial and difficult to hand craft.

The labeling scheme described for the scalar quantizer actually illustrates a more general principle which is applicable to MD vector quantizers (Diggavi et al., 2002b). We use a chain of lattices as illustrated in Figure 15, *i.e.,* we use a fine lattice Λ and two coarser sublattices Λ_1, Λ_2. These lattices have an intersection lattice Λ_{lcm} one of whose Voronoi regions is what we label. The idea of using sublattice shifts as done above to generate the labels using only the labels of this Voronoi region can also be generalized (Diggavi et al., 2002b). One such example of the labels of the Voronoi region for a two-dimensional lattice is shown in Figure 17. Therefore, the vector quantizer proceeds as follows. We first reduce point $X^T \in \mathbb{R}^T$ using a fine lattice Λ, and then using the labeling function we find (λ_1, λ_2). Then as before λ_1 is sent over the first route and λ_2 is sent over the second route. The decoder also proceeds in a manner similar to the scalar quantizer described above.

As seen above, the crux of the MD quantizer design problem is to construct the appropriate labeling function. In (Diggavi et al., 2002b) it is shown that an appropriate labeling function, along the lines described for the scalar quantizer, can be constructed very efficiently using a linear program. In fact (Diggavi et al., 2002b) shows that such a labeling scheme is very close to being optimal in terms of the rate distortion result given in Theorem 5.1 in the high-rate regime.

5.3 Network Protocols for Route Diversity

In order to utilize route diversity in a network, one of the most important components is clearly the design of MD source coding techniques studied in Section 5.2. However, an equally important question is the design of routing techniques that can enable the use of MD source coding. In this section we briefly examine these issues from a networking point of view.

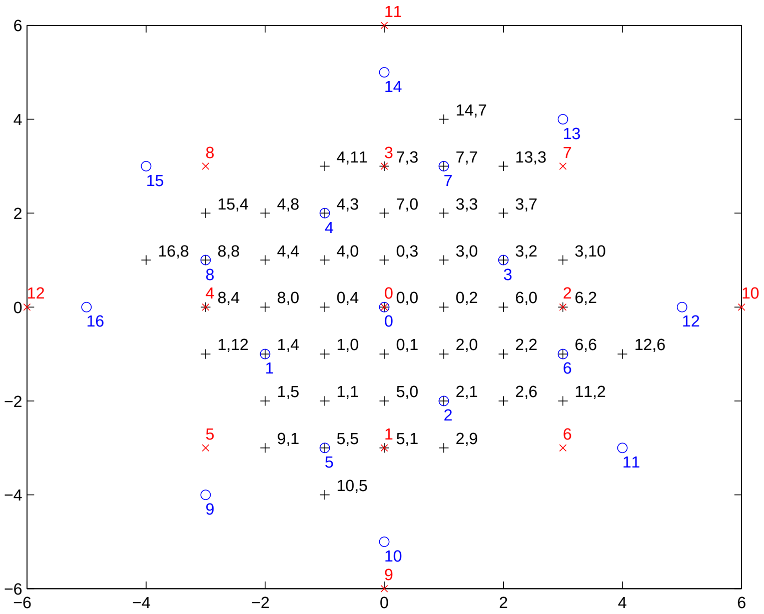

Figure 17: Labels for a two-dimensional integer lattice example.

In order to create route diversity, we need to have multiple routes which are disjoint, in that they do not share common links. This can be done through IP *source routing* (Keshav, 1997). Source routing is a technique whereby the sender of a packet can specify the route that a packet should take through the network. In the typical IP routing protocol, each router will choose the next hop to forward the packet by examining the destination IP address. However, in source routing, the "source" (*i.e.,* the sender) makes some or all of these decisions. In strict source routing (which is virtually never used), the sender specifies the exact route the packet must take. The more common form is loose source record route (LSRR), in which the sender gives one or more hops that the packet must go through. Therefore, the sender can take a MD code and send each of the description using different routes by explicitly specifying them in the IP source routing protocol. An alternate technique might be to use an *overlay* network where there is an application that collects the different descriptions and sends them through different relay nodes in order to create route diversity. This discussion shows that creating route diversity is architecturally not difficult even using the provisions within the IP protocol (Keshav, 1997).

This discussion from a networking point of view also exposes the inherent interactions required between the routing and application layers of the networking protocol stack. Such "inter layer" interactions become particularly important in wireless networks, where route failures could occur more frequently than in wired networks. Therefore, in this case diversity, albeit at a much higher layer in the IP protocol stack, again becomes quite important.

6 Discussion

In this chapter we studied the emerging role of diversity with respect to three disparate areas. The idea of using multiple instantiations of randomness attempts to turn the presence of randomness to an advantage. For example, in multiple-antenna diversity, the degrees of freedom provided by the space diversity is utilized for increased rate or reliability. In mobile ad hoc networks, the random mobility is utilized to route information from source to destination.

To realize the benefits promised by the use of diversity, we need to have interactions across networking layers. For example, in opportunistic scheduling studied in Section 4.1 the transmission rates that can be supported by the physical layer interact with the resource allocation (scheduling), which is normally only a functionality of the data-link layer. In the multi-user diversity studied in mobile ad hoc networks (see Section 4.2) the routing of the packets interacted with the physical layer transmission. Finally, the MD source coding studied in Section 5 necessitated an interaction between source coding (application-layer functionality) and routing.

These examples of cross-layer protocols are increasingly becoming important in reliable network communication. Diversity is the common thread among several of these cross-layer protocols. The advantages of using diversity in these contexts are just beginning to be realized in practice. There might be many more areas where the ideas of using diversity could have an impact, and this is a topic of on going research.

References

R. Ahlswede. The rate distortion region for multiple descriptions without excess rate. *IEEE Transactions on Information Theory*, 31(6):721–726, November 1985.

S.M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451–1458, October 1998.

J G. Apostolopoulos and M D. Trott. Path diversity for enhanced media streaming. *IEEE Communications Magazine*, 42(8):80–87, August 2004.

L. Becchetti, S. Diggavi, S. Leonardi, A. Marchetti-Spaccamela, S. Muthukrishnan, T. Nandagopal, and A. Vitaletti. Parallel scheduling problems in next generation wireless networks. In *ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 238–247, 2002. See also longer version by same authors in *Networks*, 45(1):9-22, January, 2005.

M. Bender, S. Chakrabarti, and S. Muthukrishnan. Flow and stretch metrics for scheduling continuous job streams. In *Proceedings of the Annual Symposium on Discrete Algorithms (SODA '98)*, pages 270–279, 1998.

P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi. CDMA/HDR; A bandwith-efficient high-speed wireless data service for nomadic users. *IEEE Communications Magazine*, 38(7):70–77, July 2000.

T. Berger. Multiterminal source coding. In G. Longo, editor, *The information theory approach to communications*, pages 172–231. Springer-Verlag, New York, 1977.

E. Biglieri, J. Proakis, and S. Shamai. Fading channels: Information-Theoretic and Communications aspects. *IEEE Transactions on Information Theory*, 44(6):2619–2692, October 1998.

D. Brennan. Linear diversity combining techniques. *Proceedings IEEE*, 47:1075–1102, June 1959.

L. Z. Buzi. Design of structured vector quantizers for diversity communication systems, May 1994. M.S. thesis, Department of Electrical Engineering, Texas A&M University.

G. Caire and Shlomo Shamai. On the capacity of some channels with channel state information. *IEEE Transactions on Information Theory*, 45(6):2007–2019, September 1999.

E F. Chaponniere, P J. Black, J M. Holtzman, and D N C. Tse. Transmitter directed, multiple receiver system using path diversity to equitably maximize throughput, 2002. United States patent, # 6,449,490.

J M. Cioffi, G P. Dudevoir, M V. Eyuboglu, and G D. Forney. MMSE decision-feedback equalizers and coding: Parts I & II. *IEEE Transactions on Communications*, 43(10):2582–2604, Oct. 1995.

J.H. Conway and N.J.A. Sloane. *Sphere packings, lattices, and groups.* Springer, New York, 3rd edition, 1999.

T M. Cover. Some Advances in Broadcast Channels. In A. Viterbi, editor, *Advances in Communication Theory.* Academic Press, San Francisco, 1975. Volume 4 of Theory and Applications.

T M. Cover and A. El-Gamal. Capacity theorems for the relay channel. *IEEE Transactions on Information Theory*, 25(5):572–584, September 1979.

T M. Cover and J A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

M O. Damen, A. Chkeif, and J.-C. Belfiore. Lattice codes decoder for space-time codes. *IEEE Communication Letters*, 4:161–163, May 2000.

S. N. Diggavi. On achievable performance of spatial diversity fading channels. *IEEE Transactions on Information Theory*, 47(1):308–325, January 2001. Also see Proceedings of the IEEE International Symposium on Information Theory, pp 396, 1998.

S. N. Diggavi and D N C. Tse. On successive refinement of diversity. In *Allerton Conference on Communication, Control, and Computing, Allerton, Illnois*, 2004.

S. N. Diggavi and D N C. Tse. Fundamental limits of diversity-embedded codes over fading channels. In *IEEE International Symposium on Information Theory (ISIT)*, pages 510–514, 2005.

S. N. Diggavi and D N C. Tse. On opportunistic codes and broadcast codes with degraded message sets. In *IEEE Information Theory Workshop*, 2006.

S N. Diggavi and V. A. Vaishampayan. On multiple description source coding with decoder side information. In *IEEE Information Theory Workshop, San Antonio, Texas*, October 2004.

S N. Diggavi, N. Al-Dhahir, A. Stamoulis, and A R. Calderbank. Differential space-time coding for frequency-selective channels. *IEEE Communications Letters*, pages 253–255, June 2002a.

S N. Diggavi, N.J.A. Sloane, and V. A. Vaishampayan. Asymmetric Multiple Description Lattice Vector Quantizers. *IEEE Transactions on Information Theory*, 48(1):174–191, January 2002b.

S. N. Diggavi, N. Al-Dhahir, and A. R. Calderbank. Diversity embedded space-time codes. In *IEEE Global Communications Conference (GLOBECOM)*, pages 1909–1914, 2003.

S. N. Diggavi, N. Al-Dhahir, and A. R. Calderbank. Diversity embedding in multiple antenna communications. In P. Gupta, G. Kramer, and A. J. van Wijngaarden, editors, *Network Information Theory*, pages 285–302. AMS volume 66, Series on Discrete Mathematics and Theoretical Computer Science, 2004a. Appeared as a part of DIMACS workshop on Network Information Theory, March 2003.

S N. Diggavi, N. Al-Dhahir, A. Stamoulis, and A R. Calderbank. Great Expectations: The value of spatial diversity to wireless networks. *Proceedings of the IEEE*, 92(2):217–270, February 2004b.

S. N. Diggavi, S. Dusad, A R. Calderbank, and N. Al-Dhahir. On embedded diversity codes. In *Allerton Conference on Communication, Control, and Computing*, 2005a.

S. N. Diggavi, M. Grossglauser, and D N C. Tse. Even one-dimensional mobility increases the capacity of wireless networks. *IEEE Transactions on Information Theory*, 51(11):3947–3954, 2005b. Also Proceedings IEEE International Symposium on Information Theory, 2002, pp 388.

A. Edelman. *Eigenvalues and condition numbers of random matrices.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.

A. El-Gamal, J. Mammen, B. Prabhakar, and D. Shah. Throughput-delay trade-off in wireless networks. In *Proceedings of the IEEE INFOCOM*, pages 464–475, 2004a.

A. A. El-Gamal and T. M. Cover. Achievable rates for multiple descriptions. *IEEE Transactions on Information Theory*, 28:851–857, November 1982.

H. El-Gamal and M.O Damen. Universal space-time coding. *IEEE Transactions on Information Theory*, 49(5):1097–1119, May 2003.

H. El-Gamal, G. Caire, and O. Damen. Lattice coding and decoding achieve the optimal diversity-multiplexing of MIMO channels. *IEEE Transactions on Information Theory*, 50(6):968–985, June 2004b.

M. Fleming, Q. Zhao, and M. Effros. Network vector quantization. *IEEE Transactions on Information Theory*, 50(8):1584–1604, August 2004.

G.J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal*, 1(2):41–59, September 1996.

M. Franceschetti, O. Dousse, D. Tse, and P. Thiran. Closing the gap in the capacity of random wireless networks. In *IEEE International Symposium on Information Theory*, 2004. See also preprint ”On the throughput capacity of random wireless networks” at http://fleece.ucsd.edu/∼massimo/.

F-W. Fu and R. Yeung. On the rate-distortion region for multiple descriptions. *IEEE Transactions on Information Theory*, 48(7):2012–2021, July 2002.

A. Gersho and R M. Gray. *Vector quantization and signal compression*. Kluwer, Boston, 1992.

B. Girod and N. Farber. Wireless video. In A. Reibman and M.-T. Sun, editors, *Compressed Video over Networks*. Marcel Dekker, 2000.

A. Goldsmith and P. Varaiya. Capacity of fading channels with channel side information. *IEEE Transactions on Information Theory*, 43(6):1986–1992, November 1997.

V K. Goyal and J. Kovacevic. Generalized multiple description coding with correlating transforms. *IEEE Transactions on Information Theory*, 47(6):2199–2224, September 2001.

I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, San Diego, 1994.

R M. Gray and D L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44:2325–2383, October 1998.

M. Grossglauser and D. N. C. Tse. Mobility Increases the Capacity of Ad-hoc Wireless Networks. *IEEE/ACM Transactions on Networking*, 10(4):477–486, August 2002.

J-C. Guey, M P. Fitz, M R. Bell, and W-Y Kuo. Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels. *IEEE Transactions on Communications*, 47(4):527–537, April 1999.

P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, March 2000.

P. Gupta and P. R. Kumar. Towards an information theory of large networks: an achievable rate region. *IEEE Transactions on Information Theory*, 49(8):1877–1894, August 2003.

B. Hassibi and B. Hochwald. High-Rate Codes That Are Linear in Space and Time. *IEEE Transactions on Information Theory*, pages 1804–1824, July 2002.

B. Hassibi and T.L. Marzetta. Multiple-antennas and isotropically random unitary inputs: the received signal density in closed form. *IEEE Transactions on Information Theory*, 48(6):1473–1484, June 2002.

A. Hiroike, F. Adachi, and N. Nakajima. Combined effects of phase sweeping transmitter diversity and channel coding. *IEEE Transactions on Vehicular Technology*, 41(5):170–176, May 1992.

B. Hochwald and W. Sweldens. Differential unitary space-time modulation. *IEEE Transactions on Communications*, pages 2041–2052, December 2000.

B M. Hochwald and T L. Marzetta. Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading. *IEEE Transactions on Information Theory*, 46(2):543–564, March 2000.

B M. Hochwald and T L. Marzetta. Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Transactions on Information Theory*, 45(1):139–157, January 1999.

T.-C Hou and V. O. K. Li. Transmission range control in multihop packet radio networks. *IEEE Transactions on Communications*, 34(1):38–44, January 1986.

B. L. Hughes. Differential Space-Time Modulation. *IEEE Transactions on Information Theory*, 46 (7):2567–2578, November 2000.

A. Ingle and V. A. Vaishampayan. DPCM system design for diversity systems with applications to packetized speech. *IEEE Transactions on Speech and Audio Processing*, 1:48–58, January 1995.

H. Jafarkhani and V. Tarokh. Multiple description trellis coded quantizers. *IEEE Transactions on Communications*, 47:799–803, June 1999.

W.C. Jakes. *Microwave Mobile Communications*. IEEE Press, 1974.

A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficiency high data rate personal communication wireless system. In *Proceedings Vehicular Technology Conference VTC 2000-Spring*, pages 1854–1858, 2000.

N D. Jayant and P. Noll. *Digital coding of waveforms*. Prentice Hall, Englewood Cliffs, 1984.

N S. Jayant. Sub-sampling of a DPCM speech channel to provide two "self-contained" half rate channels. *Bell Systems Technical Journal*, 60(4):501–509, April 1981.

T. Kailath. Channel characterization: Time-variant dispersive channels. In E. J. Baghdady, editor, *Lectures on Communication System Theory*, pages 95–123. McGraw-Hill, New York, 1961.

S. Keshav. *An Engineering Approach to Computer Networking*. Addison Wesley, 1997.

R. Knopp and P. Humblet. Information capacity and power control in single-cell multiuser communications . In *IEEE International Conference on Communications (ICC)*, pages 331 –335, June 1995.

A. Lapidoth and S. Shamai. Fading channels: How perfect need "perfect side information" be? *IEEE Transactions on Information Theory*, 48(5):1118–1134, May 2002.

O. Leveque and E. Telatar. Information theoretic upper bounds on the capacity of large extended ad hoc wireless networks. *IEEE Transactions on Information Theory*, March 2005.

L. Li and A. Goldsmith. Optimal Resource Allocation for Fading Broadcast Channels- Part I: Ergodic Capacity. *IEEE Transactions on Information Theory*, 47(3):1083–1102, March 2001.

H F. Lu and P V. Kumar. Rate-diversity trade-off if space-time codes with fixed alphabet and optimal constructions for PSK modulation. *IEEE Transactions on Information Theory*, 49(10): 2747–2752, October 2003.

N. Maxemchuk. *Dispersity Routing in Store and Forward Networks.* PhD thesis, University of Pennsylvania, Philadelphia, 1975.

R J. Muirhead. *Aspects of multivariate statistical theory.* Wiley, New York, 1982.

A. Naguib, V. Tarokh, N. Seshadri, and A.R. Calderbank. A space-time coding modem for high-data-rate wireless communications. *IEEE Journal on Selected Areas in Communications*, pages 1459–1477, October 1998.

F.D. Neeser and James L. Massey. Proper complex random processes with applications to information theory. *IEEE Transactions on Information Theory*, 39:1293–1302, July 1993.

L. Ozarow. On a source coding problem with two channels and three receivers. *Bell Syst. Tech. J.*, 59:1909–1921, December 1980.

L H. Ozarow, S Shamai, and A. D. Wyner. Information theoretic considerations for cellular mobile radio. *IEEE Transactions on Vehicular Technology*, 43(2):359–378, May 1994.

P. Patel and J. Holtzman. Analysis of a simple successive interference cancellation scheme in a DS/CDMA system. *IEEE Journal Selected Areas in Communications*, 12(5):796 –807, June 1994.

G.J. Pottie and W.J. Kaiser. Wireless Integrated Network Sensors. *Communications of the ACM*, 43(2):51–58, May 2000. See also cens.ucla.edu/.

S S. Pradhan, J. Kusuma, and K. Ramchandran. Distributed compression in a dense microsensor network. *IEEE Signal Processing Magazine*, 2:51–60, March 2002.

S S. Pradhan, R. Puri, and K. Ramchandran. n-channel symmetric multiple descriptions – part i:(n, k)source-channel erasure codes. *IEEE Transactions on Information Theory*, 50(1):47–61, January 2004.

J G. Proakis. *Digital Communications.* McGraw Hill, New York, 3 edition, 1995.

R. Puri and K. Ramchandran. PRISM: an uplink-friendly multimedia coding paradigm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 856–859, 2003.

G. Raleigh, S N. Diggavi, A F. Naguib, and A. Paulraj. Characterization of fast fading vector channels for multi-antenna communication systems. In *Proceedings 28th IEEE Asilomar Conference on Signals, Systems and Computers*, pages 853–857, 1994.

T. Rappaport. *Wireless Communications.* IEEE Press, 1996.

A. Reibman and M.-T. Sun. *Compressed Video over Networks.* Marcel Dekker, 2000.

B. A. Sethuraman, B. S. Rajan, and V. Shashidhar. Full-diversity, high-rate space-time block codes from division algebras. *IEEE Transactions on Information Theory*, 49(10):2596–2616, 2003.

C E. Shannon. A mathematical theory of communications. *Bell Systems Technical Journal*, 27: 623–656, 1948.

C E. Shannon. Channels with side-information at the transmitter. *IBM Journal of Research and Development*, 2:289–293, October 1958a.

C E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, pages 142–163, 1958b.

D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, July 1973.

V. Tarokh and H. Jafarkhani. A differential detection scheme for transmit diversity. *IEEE Journal on Selected Areas in Communications*, 18(7):1169 –1174, July 2000.

V. Tarokh, N. Seshadri, and A.R. Calderbank. Space-time codes for high data rate wireless communications: Performance criterion and code construction. *IEEE Transactions on Information Theory*, 44(2):744–765, March 1998.

V. Tarokh, H. Jafarkhani, and A.R. Calderbank. Space-Time Block Codes from Orthogonal Designs. *IEEE Transactions on Information Theory*, pages 1456–1467, July 1999.

I Emre. Telatar. Capacity of Multiple Antenna Gaussian Channels. *AT&T Technical Memorandum*, 1995. Also published in European Transactions on Telecommunications, 10(6), 585-595, November-December, 1999.

D N C. Tse. Optimal power allocation over parallel gaussian broadcast channels. In *Proc. IEEE International Symposium on Information Theory (ISIT), Ulm, Germany*, page 27, 1997.

G. Ungerboeck. Channel Coding with Multilevel/Phase Signals. *IEEE Transactions on Information Theory*, 28(1):55–67, January 1982.

V. A. Vaishampayan. Design of multiple description scalar quantizers. *IEEE Trans. Information Theory*, 39:821–834, May 1993.

V A. Vaishampayan, N. J. A. Sloane, and S Servetto. Multiple Description Vector Quantization with Lattice Codebooks: Design and Analysis. *IEEE Transactions on Information Theory*, 47(5): 1718–1734, July 2001.

E C. van der Meulen. A survey of multiway channels in information theory. *IEEE Transactions on Information Theory*, 23(1):1–37, January 1977.

M. Varanasi and T. Guess. Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel. In *Asilomar Conference on Signals, Systems and Computers*, pages 1405–1409, 1997.

R. Venkataramani, G. Kramer, and V. K. Goyal. Multiple description coding with many channels. *IEEE Transactions on Information Theory*, 49(9):2106–2114, September 2003.

S. Verdu. *Multiuser Detection.* Cambridge University Press, Cambridge, 1998.

P. Viswanath, D N C. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48(6):1277–1294, June 2002.

P. Viswanath, R. Laroia, and D N C. Tse. Methods and apparatus for transmitting information between a basestation and multiple mobile stations, 2004. United States patent, number 6,694,147.

H. Viswanathan. Capacity of Markov channels with receiver CSI and delayed feedback. *IEEE Transactions on Information Theory*, 45(2):761–770, March 1999.

V. Weerackody. Characteristics of a simulated fast fading indoor radio channel. In *IEEE Vehicular Technology Conference*, pages 231–235, 1993.

H. Witsenhausen and A D. Wyner. Interframe coder for video signals, 1980. United States patent, # 4,191,970.

G. Wornell and M. Trott. Efficient Signal Processing techniques for exploiting transmit antenna diversity on fading channels. *IEEE Transactions on Signal Processing*, 45(1):191–205, Jan 1997.

A. D. Wyner. Recent Results in the Shannon Theory. *IEEE Trans. on Information Theory*, 20(1): 2–10, January 1974.

A D. Wyner and J. Ziv. The rate distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22:1–10, January 1976.

L.-L. Xie and P R. Kumar. A network information theory for wireless communication: scaling laws and optimal operation. *IEEE Transactions on Information Theory*, 50(5):748–767, May 2004.

H. Yao and G. Wornell. Achieving the full MIMO diversity-multiplexing frontier with rotation based space-time codes. In *Allerton Conference on Communication, Control, and Computing*, 2003.

Z. Zhang and T. Berger. New results in binary multiple descriptions. *IEEE Transactions on Information Theory*, 33:502–521, July 1987.

L. Zheng and D N C. Tse. Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel. *IEEE Transactions on Information Theory*, 48(2): 359–383, February 2002.

L. Zheng and D N C. Tse. Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels. *IEEE Transactions on Information Theory*, 49(5):1073–1096, May 2003.